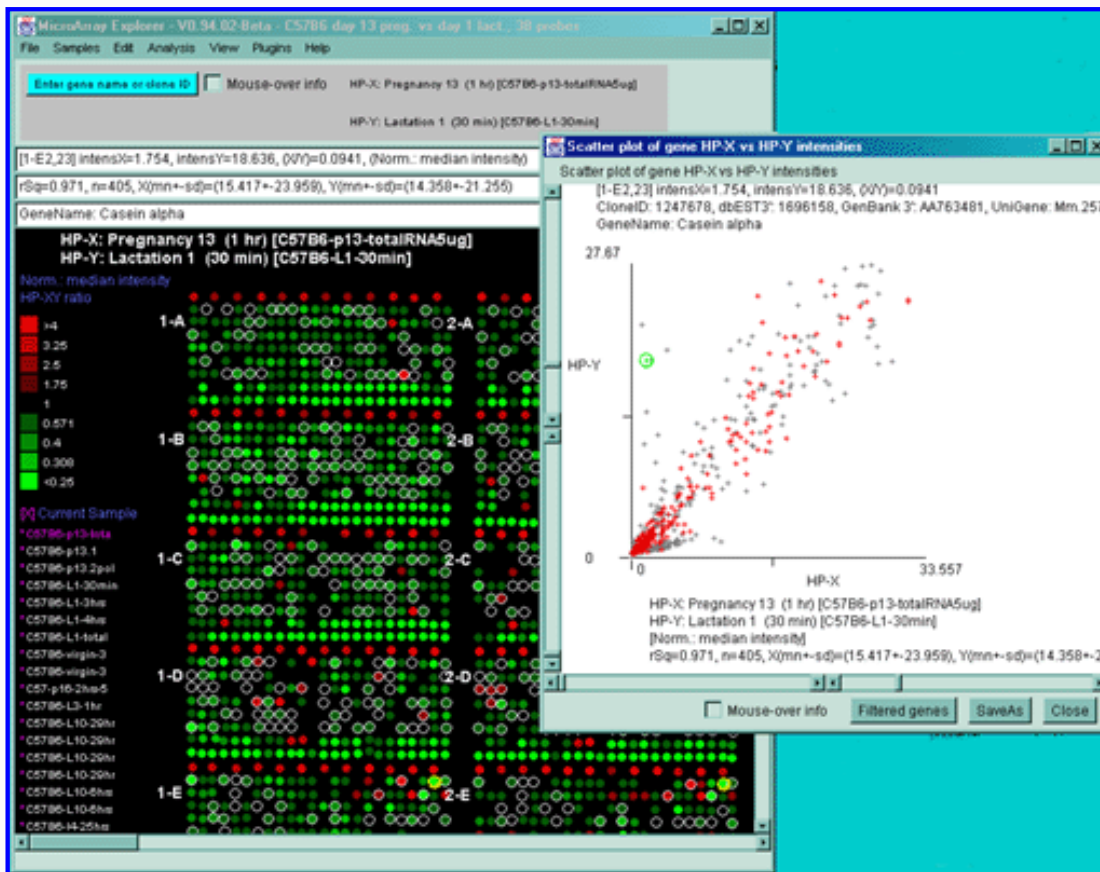# Reference Manual++ - MAExplorer Microarray Exploratory Data Analysis

  **Legend

(Click on figures to show higher resolution versions)

*** DRAFT ***
$Date: 2002/08/19 16:09:41 $ MicroArray Explorer-V0.96.11 "Millenium" edition

Peter F. Lemkin
Laboratory of Experimental and Computational Biology
Center for Cancer Research, National Cancer Institute - Frederick, Frederick, MD 21702

MAExplorer home | Table of Contents | Overview | Introduction | Menu summary | PDF documents |
Newsletters |
Plugins | Quick start | Short tutorial | Advanced Tutorial | Glossary | Figures | Tables | Index | Help desk

SourceForge	http://maexplorer.sourceforge.net/
or	
LECB/NCI	http://www.lecb.ncifcrf.gov/MAExplorer

++ **Note:** This hypertext manual is divided into chapters and appendices Web pages. These may be printed individually from your Web browser by (1) clicking in the text window to be printed, and (2) using the "Print Frame" in Netscape or "Print" in Internet Explorer. Some of the chapters (eg. 2) have many images. The entire manual may be downloaded at one time with low resolution figures and is suitable for printing in the Web browser. You may also download a an Adobe acrobate PDF file version of the entire manual with the lower resolution figures (~5Mb). The Unix script for creating the full reference manual from the individual HTML pages is CreateMaeFullRefManual.do.

The MAExplorer is a Java-based bioinformatics exploratory data-analysis and data-mining facility for analyzing sets of quantitative spotted cDNA or oligonucleotide microarray data (Lemkin et al., 2000) - (see (Schulze, 2001) for a review of microarray technology).

Prior to its release on SourceForge, MAExplorer was developed by Dr. Peter Lemkin (LECB/NCI-Frederick) with help from Gregory Thornwall (SAIC) and Jai Evans (DECA/CIT, NIH). It was initially developed for analyzing ^{33}P labeled membrane array data from the mouse mammary tissue from Mammary Genome Anatomy Project (MGAP) http://mammary.nih.gov/ with the help of many researchers in the Laboratory of Genetics and Physiology, NIDDK under Dr. Lothar Hennighausen. Since the early work with MGAP it was extended to work with other types of arrays and nucleotide labeling methods. These include spotted Cy3/Cy5 glass slides, spotted membranes, non-geometric chip data, and other chip supports with different geometries and numbers of duplicate spots/gene, clones as well as oligo chip data such as Affymetrix. A wizard tool called Cvt2Mae was developed to make it easier for other researchers to convert their data to the format required by MAExplorer. Cvt2Mae was developed by Peter Lemkin, Greg Thornwall and Bob Stephens (ABCC/SAIC).

This document describes the MAExplorer's functionality, provides tutorials and contains documentation for using it with other arrays.

With this program, you may: 1) analyze expression of individual genes; 2) analyze expression of gene families and clusters; 3) compare expression patterns for multiple hybridized samples.

MAExplorer is written in Java and runs as a stand-alone application you may download to your computer. Using it with your data requires that you download the program. The stand-alone version is able to extend analysis methods using open-source Java Plugin Extensions to MAExplorer.

Although MAExplorer began out as a Java applet for use with with Web browsers for the MGAP Web database (http://www.lecb.ncifcrf.gov/mae), we have depricated its use because of many problems with running large Java applets in some Web browsers. Instead, we recommend downloading the MGAP array data and run it with MAExplorer after you have installed it on your computer. The public MGAP data set is included as a demonstration database when you download MAExplorer (from LECB/NCI)- so you may access the MGAP data either way (with the initial download or downloaded later).

> **Notation:** MAExplorer uses the notation that the sample *probe* total mRNA is labeled and then hybridized against the known cDNA *targets* tethered to the microarray. Because of this notation, we refer to a hybridized sample as a HP. An alternative notation that reverses these terms is also commonly used (see "Chipping Forecast", *Nature Genetics* supplement, Jan, 1999, pg 1). Also, because arrays may be constructed from either spotted clones or oligonucleotides, we refer to hybridized chip DNA from any of these sources genericlly as "genes".

Throughout this document we use the abbreviations HP for hybridized sample, GC for gene class. These and other terms are explained in the Glossary and Index . There are a number of figures and tables illustrating various features of MAExplorer throughout this manual. Figures are presented at low-resolution. By clicking on the lower-resolution figure, the high-resolution versions can be viewed.

> *NOTES: because MAExplorer is under development, there may be occasional problems with some of its functionality. There may also be some problems (mostly bad HTML links) with migrating from LECB/NCI to the SourceForge Web site. Some operations that are under development are labeled with "[Future]" in this manual. We welcome your suggestions for improvements as well as letting us know about problems that you encounter. Occasionally the manual or the figures in the manual may not be quite in phase with the software. Please notify us of problems or suggestions by E-mail so we can try to fix or implement them. If you are a bioinformatics developer and would be interested on working with the MAExplorer project, consider joining the MAExplorer development team.*

**Icon Legend

Data from a 38 sample subset of hybridized samples from the [MGAP](#) mouse microarray database. This screen illustrates a synthetic pseudoarray image showing the ratios of duplicated grids of genes comparing day 13 pregnancy in C57B6 mouse (sample HP-X 'set') with Lactation day 1 (sample HP-Y 'set'). The color scale of the spots is indicated on the left as is the current data normalization mode (median). Genes with white circles are named genes and were selected by the data filter. A scatter plot of this data is shown on the right with genes passing the data filter indicated as red + and those not passing the filter (i.e. ESTs, calibration DNA, user's genes) shown as gray + symbols. A single gene was selected by clicking on it in the array image and has a yellow circle (grid 1-D) and a corresponding green circle in the scatter plot. Information on that gene is indicated above the array and at the top of the scatter plot. MAExplorer can also be used to view *mean data* from sets of samples e.g. Day 13 pregnancies from C57B6 (3 HP-X samples) vs. Day 1 Lactation (4 HP-Y samples) at [low](#) or [high](#) resolution.

Table of Contents

R.8 Software design of the MAExplorer data mining tool (PDF) or (PPT)

MAExplorer - Overview

MAExplorer is a bioinformatics microarray data mining Java application that may help in the discovery of genes regulated in cancer and other diseases. MAExplorer is generally run as as a [stand-alone application](#) on a local computer. By running as a local application, it is able to access your local disk to save the state of your data mining session as well as plots and reports. Using the previously saved data mining state, you can continue a data-mining session at a later date after exiting the program.

- MAExplorer helps perform computer data-mining of multiple samples hybridized with microarrays. Data mining is the process of attempting to find relevant patterns of information from large sets of data. MAExplorer enables the investigator to:

  1. organize the hybridized sample data by experimental condition, (including: disease state, dose response, developmental stage, strain, time course, knock-in/-out, shock treatment, etc.) so an investigator can design data mining experiments relevant to those conditions or a subset of conditions from a particular database.
  2. compare gene expression patterns between *sets* of different hybridized samples (denoted HP-X and HP-Y 'sets') for comparing mean changes between replicate sets of samples. An ordered expression profile *list* HP-E of samples is used for finding similar expression patterns across genes for a sequence of samples such as from the cell cycle, developmental stage, or conditions.
  3. use data mining techniques of graphical direct-manipulation (which requires the real-time response of local computation), statistical, clustering and spreadsheet techniques, and connectivity to other Internet genomic databases to get additional information on individual genes. The latter leverages the maintenance resources of other groups to allow transparent access to that data.
  4. explore, compare, and record analyses between researchers in their own group and for sharing with other investigators (i.e. groupware).

- Spots in microarray images are [quantified](#) into tab-delimited data files using programs such as generated by Axon's GenePix$^{(TM)}$, Scanalyze, Molecular Dynamic's ImageQuant-NT$^{(TM)}$, Research Genetics' Pathways$^{(TM)}$, and other systems. This data is transformed using the [Cvt2Mae wizard](#) data conversion tool to quantification data files, a print-file ([Gene In Plate Order table or GIPO](#)), a list of DB samples file, and a MAExplorer configuration file. These may be copied to your local computer file system (in stand-alone operation) or a MAExplorer-compatible Web database server where they are loaded on demand by MAExplorer. The file formats schema used by MAExplorer is discussed in [Appendix C](#)). The data conversion tool [Cvt2Mae (Appendix C.6)](#) helps convert user's data sets to MAExplorer format.

- Upon starting, MAExplorer uses a [".mae" startup file](#) to specify the subset of samples to be used from the database, and initial parameters to use. It then loads a configuration file which describes additional files including the gene-in-plate-order table which maps spot position to clone ID and other genomic information, and a samples database file containing a list of the names of the quantified spot data files for the hybridized samples being analyzed. Later, you may request additional sample files or data from the local file system or Web database server when requested by the user. The .mae file format is discussed in [Appendix D](#). Users may save data mining sessions to create new .mae startup files. These may be used at a later for continuing their sessions in the future.

- The investigator interacts directly with the system by selecting entries from menus ([Section 2](#)), selecting data by clicking on spots in the microarray image, selecting points in graphic plots or cells in spreadsheets, manipulating threshold sliders, or typing in gene names, clone IDs, GenBank IDs, UniGene IDs, LocusLink IDs, etc. Data reports may be exported to Excel spreadsheets or used dynamically to access other genomic Web databases. With the stand-alone version, you may save the full resolution plots in GIF files and report tables in tab-delimited text files.

Recommended Hardware

Because data mining is a computationally intensive activity, a reasonable level of computation resources are required for adequate response. The same Java program runs on a variety of operating systems including Windows 95/98/Me/NT/2000, Macintosh OS8/9/X, Solaris, Linux, etc. so the choice of computer is not that critical. We recommend the following hardware:

- A computer with at least 500Mhz CPU speed (Intel). For other CPUs such as the PowerPC (Macintosh), Sparc (Sun), etc. it should have a corresponding capability (for more powerful CPU chips, a lower CPU speed may be fine). For large data sets

(order of 100 or more) having a large number of samples with many spots, a much faster system with much more memory is desirable.

- At least 128Mbytes of memory. Although it will work with less memory, we don't recommend it as it is underpowered. For large data sets, more memory (eg. 256Mb or more) is desirable.
- Adequate disk space for the data sets required. The MAExplorer distribution itself, including the MGAP demonstration database and a Java Virtual Machine, is on the order of 24 Mbytes. This Reference Manual with both low and high resolution figures is on the order of 11 Mbytes.

Addition of user defined analysis methods using Java Plugins

We have provided the ability for users to add their own Java Plugin Extensions to MAExplorer. These extend the capabilities of the core MAExplorer program to other more sophisticated analysis methods created by users and allow interaction with specialized genomic servers. This is described in Appendix E, Section 2.6, and in the MAExplorer Plugins Web page.

---

1. Introduction

This hyperlinked manual provides a detailed description of the MAExplorer conventions (Section 1) and operation (Section 2). The latter contains many figures of computer screens showing the operations described in the Section. Section 3 discusses typical scenarios in using MAExplorer for data-mining microarrays and contains a brief introduction to the process of data mining. Section 4 lists currently known bugs and the revision history. Appendix A is a short tutorial. Appendix B is a more advanced tutorial. Appendix C describes the files required by MAExplorer and how they may be created for using MAExplorer with other array data. It also describes the data conversion tool Cvt2Mae. The Appendix D covers downloading, installing and running MAExplorer as a stand-alone Java application on a local computer. Appendix E discusses design issues for the MAExplorer Java program and supporting Web servers. Users may create new analytic methods and add them as MAExplorer Plugins as Java extensions. There is a glossary of terms used in MAExplorer. There is also a List of Figures, a List of Tables, and an Index to help find material of interest.

MAExplorer is normally used as stand-alone program

Figure 1 gives an overview of the system. Note that MAExplorer does *not* perform spot quantification from raw scanned images - it is used for the *subsequent* data mining analysis of quantified spot data. Figures 1.1.1 through 1.1.4 describe this in more detail.
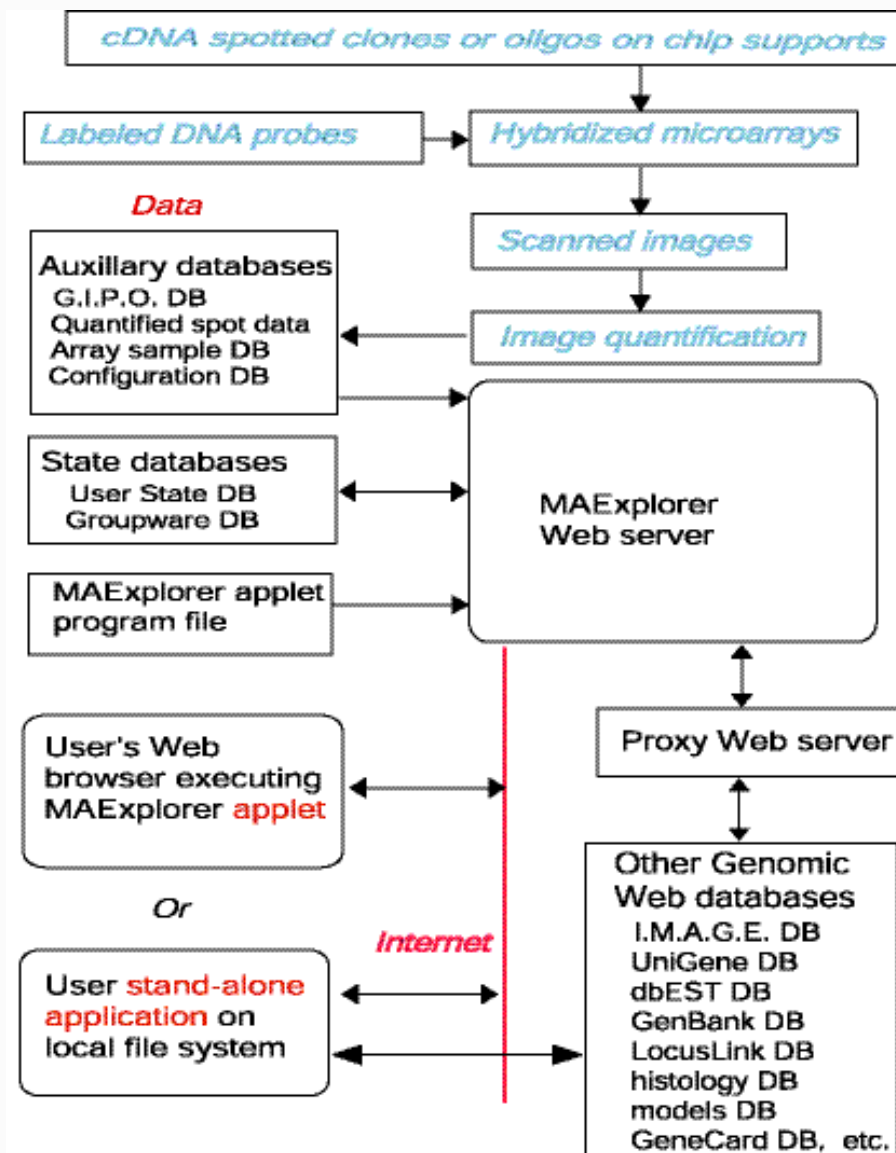
Figure 1. Overview of MAExplorer exploratory data analysis system. Initial data preparation steps are performed *prior* to analysis by MAExplorer and are indicated by cyan italics at the top of the figure. The primary data consists of quantified microarray image data as well as corresponding qualitative clone ID, gene-in-plate-order (GIPO), gene name, hypertext base references and related information. After the microarrays are hybridized, they are scanned and spots quantified using image spot quantification programs. These lists are then saved for each array in a tab-delimited file. Microarray image quantification may be performed by various software such as Axon's GenePix$^{(TM)}$, Scanalyze, Molecular Dynamics ImageQuant$^{(TM)}$, Research Genetics' Pathways$^{(TM)}$, etc. When used as an applet, this auxiliary databases and the MAExplorer Jar files are copied to the Web server or local file system (in the case of the stand-alone version) where they are then available to be downloaded by users. When a user invokes a Web page containing the Java applet, it first downloads the applet that then downloads auxiliary databases including a configuration file that describes the array data. It then downloads the subset of quantified microarray spot data files requested for the set of hybridized samples being investigated. Additional samples may be downloaded at any time. When the user selects an operation that requires access to Web databases not residing on the MAExplorer Web server, implicit Java security restrictions prevent the applet from going directly to these other Web servers. Instead, it requests the MAExplorer proxy server request the data from the foreign Web server, and then returns it back to the user's Web browser. When used as a stand-alone application, data may be saved on the local computer for local off-line use, and direct access to other Internet genomic databases may be made without using a proxy server.
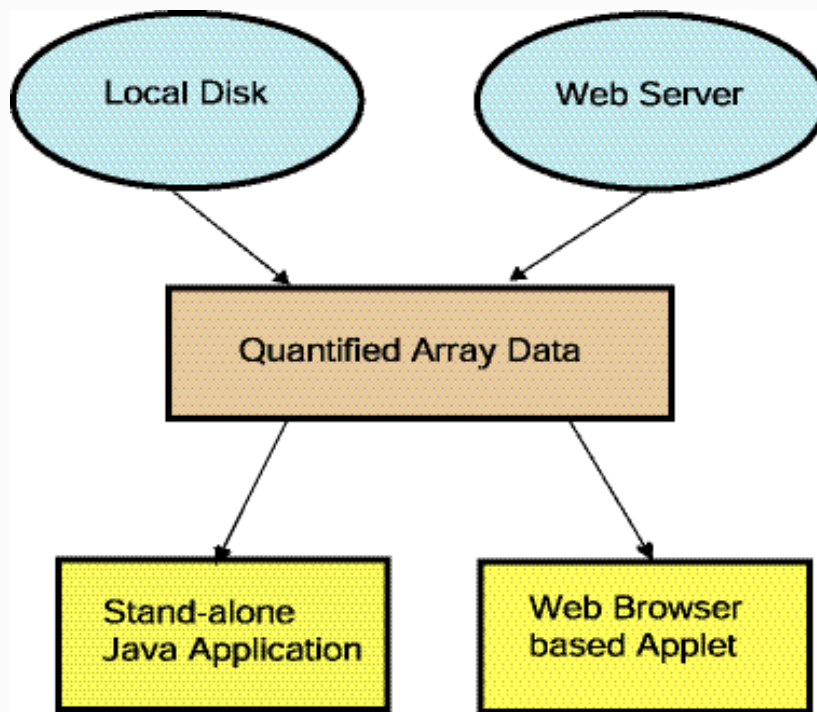
Figure 1.1.1 Overview of MAExplorer exploratory data analysis system. MAExplorer may be used either as a stand-alone application or as a Web browser applet. In the case of the applet, it may only access quantified array data from the Web server that launched the applet.

Overview of data preparation for quantified spot data used by MAExplorer

Figure 1.1.2 Overview of data preparation for quantified spot data used by MAExplorer. MAExplorer handles quantified spot data as shown in this figure. Arrays are hybridized against labeled samples are scanned and spots are quantified into spot data files. Quantified spot data is represented as tab-delimited data with data for one spot/row. Each spot is identified in this file by its grid coordinates (grid, grid row, grid column) with image (X,Y) coordinates being optional. Quantified spot data includes the raw spot intensity for each channel (in the case of multiple channels such as Cy3, Cy5, etc.). If the original data has background spot intensity values, then that may be included as well - otherwise no background data will be available for background correction. The spot data is discussed in more detail in Section 1.1 and Appendix C.1, and Appendix C.3.
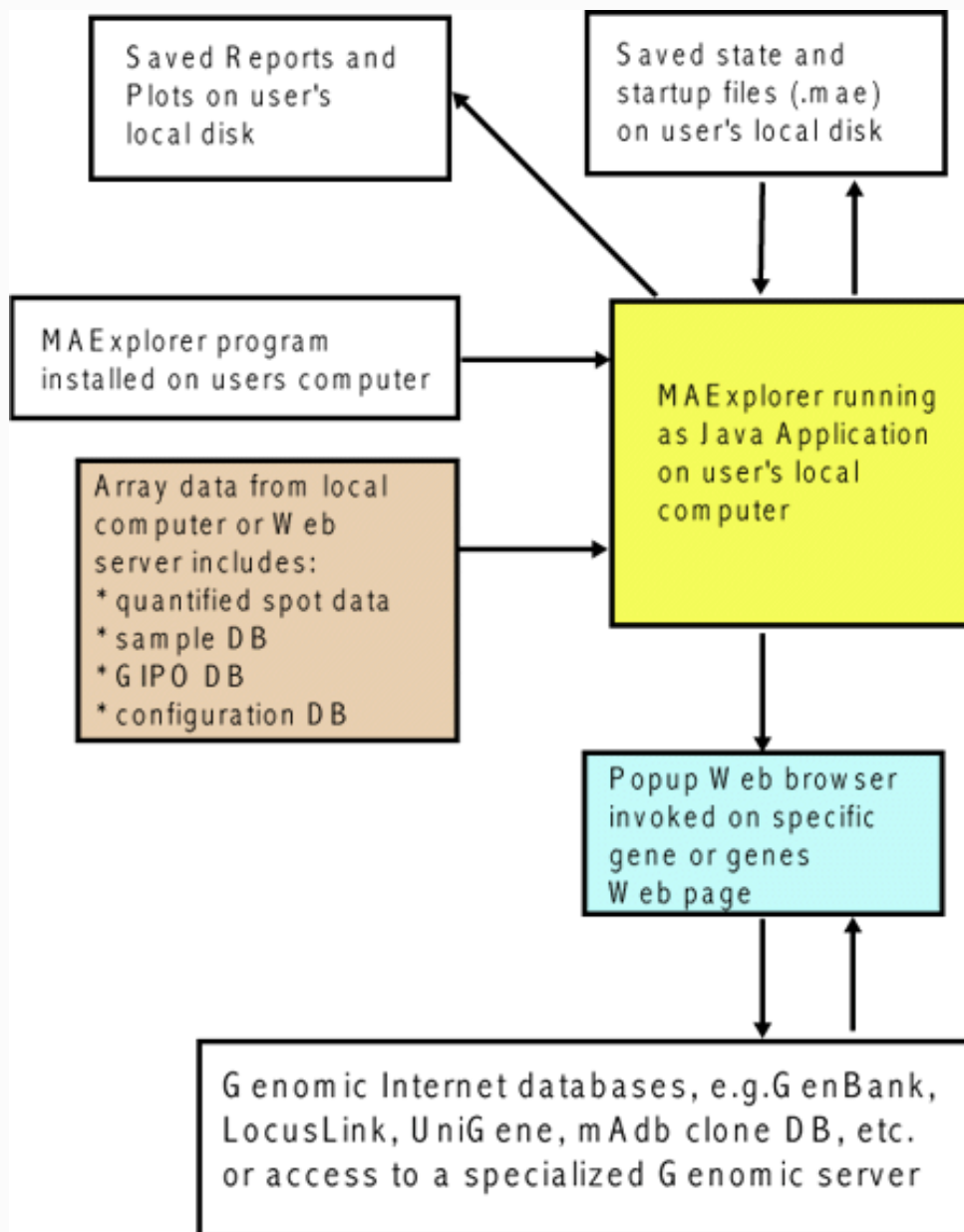
Figure 1.1.3 Overview of running MAExplorer as a stand-alone application. The preferred way of running MAExplorer is as a stand-alone application. There are distinct advantages in running MAExplorer as an application in that data and the exploration state may be saved on the users local computer, direct access to genomic servers is easier (no proxy server required - see Figure 1.4). MAExplorer plugin extensions (MAEPlugins) may only be used with the stand-alone version. Since MAExplorer is packaged for download for a variety of operating systems, using this method is not difficult to set up and the MAEPlugins should run on a variety of operating systems.
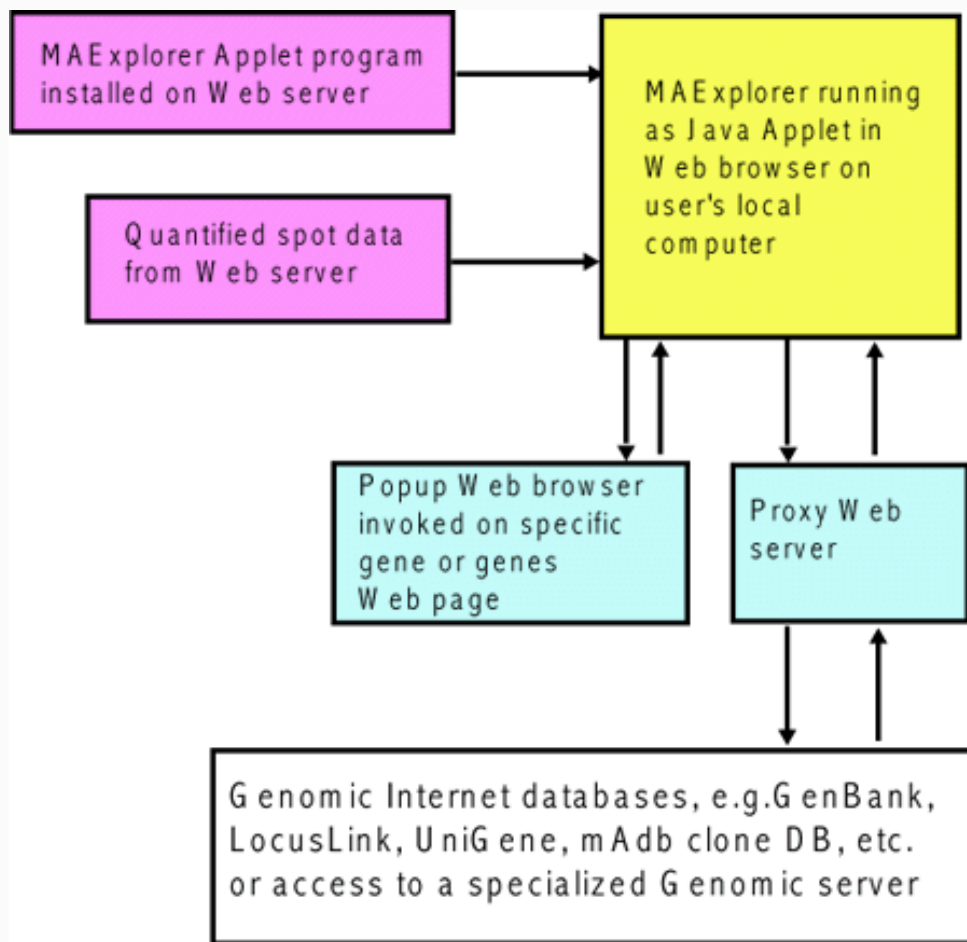
Figure 1.1.4 Overview of running MAExplorer as a Web browser applet. An alternative way of running MAExplorer on existing databases is as a Web-browser applet. There advantage of this method is that no software installation is required on the user's computer. However, the user may not save data and the exploration state on their local computer. Furthermore, direct access to genomic servers requires a proxy server. MAExplorer plugin extensions (MAEPlugins) may not be used with the the applet version. The Mammary Genome Anatomy Program (MGAP) originally used the MAExplorer applet.

Example of a MAExplorer database - http://www.lecb.ncifcrf.gov/mae - the public MGAP DB

The Mammary Genome Anatomy Program (MGAP) microarrays of cDNA clones from mouse mammary tissue (collaboration with Research Genetics) were hybridized with ^{33}P radio-labeled samples. These were then used to charge fluorescing plates. See the MGAP site for more documentation on the database and preparation procedures. The hybridized arrays are scanned on a phospho-imager scanner at high resolution. Spot data was quantified from these images using the Research Genetics' "Pathways 2.01" program which generated tab-delimited data files. This data also includes the microarray grid point locations (field, grid, grid row, grid col) from the associated microarray description data files (grid-in-plate-order data). These files are then automatically loaded over the Internet from the Web server by MAExplorer into the user's Web browser as required.

1.1 Microarrays and notation used with MAExplorer

In general, microarrays are hybridized using cDNA samples derived from mRNA labeled with either radio-label, biotin, fluorescent dyes, or other methods (see Schulze, 2001) for review of the technology). MAExplorer may be used to construct databases using ^{33}P radio-labeled and fluorescent double labeled (i.e. Cy3/Cy5) ratio data arrays with different GIPO geometries.

Definition of "intensity" for single-labeled samples

MAExplorer uses the term "intensity" in slightly different ways dependent on whether you are using the single radio-labeled or fluorescent double labeled data. For single radio-labeled data, "intensity" is the raw quantified data value as measured by the image scanner. Raw data must be normalized between samples in order to compare it between samples. Therefore, to compare N samples, first normalize the data and then compare them.

Definition of "intensity" for fluorescent double-labeled samples

For fluorescent double-labeled data, the Cy3 and Cy5 dye-labeled (for example) measurements are the raw quantified data values as measured by the image scanner. In this case, "intensity" is defined as the ratio of Cy3 to Cy5 (i.e. Cy3/Cy5). If you wish to look at the ratio as Cy5/Cy3, you may flip the two channels on a per-sample basis (see Section 2.2.2 for more details).

Issues of experimental design of microarray experiments

Some of the issues involved in experimental design (setting up experiments) based on the types of arrays are discussed in Section 3.1.1 for (Cy3/Cy5)-labeled as well as ^{33}P-labeled samples. Poorly designed experiments will not yield significant statistical results, so attention should be paid to developing an adequate and robust design for your data given costs of doing experiments as well as statistical constraints on analyzing the data.

Actual and "Pseudoarray" image geometry

The main MAExplorer windows contains a pseudoarray image for visualization purposes. It may or may not correspond the spot positions on the actual array. This array geometry is defined by the number of replicate Fields (normally 1) each of which contains a number of grids (also called "blocks") containing a number of rows/grid and columns/grid of spots. If there is no explicit array geometry or spot (X,Y) coordinate data available but simply gene identifiers and intensity data, then an arbitrary pseudoarray geometry is generated. If there is an explicit array geometry, then it waill draw the pseudoarray using this geometry. The database configuration determines which method will be used and is discussed in Appendix C.5. If there is no explicit grid geometry, the number of spot Locations (e.g., IncyteID, Affymetrix probe_set) may be used to synthesize a set of grids of a size that is reasonable for viewing with MAExplorer. This is done in the Cvt2Mae array data conversion program when the array geometry (#grids, #rows/grid, #columns/grid) is not known. This conversion is not done in MAExplorer itself.

Gene coordinate numbering on the microarray

Gene coordinate numbering corresponds to the numbering for particular array geometries. In the case where there is no array coordinates, a pseudoarray geometry is constructed to serve as a basis to display the microarray pseudoimage. These are described by *grids* (or blocks), each consisting of *grid rows* by *grid columns* of spots. The grids may be repeated on the array and constitute duplicate *fields*. In MAExplorer we refer to grids by letter names (A,B,C,...) and fields by F1 and F2. If you are using Cy3/Cy5 ratio data and the Cy3 and Cy5 data is available as independent channels for each HP sample, then operations that use F1 and F2 will use the Cy3 and Cy5 data for various operations such as scatter plots (Cy3 vs Cy5), etc. If there is only one field in an array (i.e. no duplicate grids), then when MAExplorer is run, operations and menus describing F1 and F2 operations will not be available.

For duplicate (F1 and F2) spots allows us to get an estimate of the hybridization variance within an array and is used to compute the (F1,F2) gene coefficient of variation (CV) used in the gene data Filter to remove noisy data before looking for additional differences.

Setting the "current gene" to a specific gene by "Master gene ID"

The MAExplorer uses the concept of the "current gene" to indicate a particular gene to be analyzed. You may interrogate the microarray database or Internet databases for data on the current gene or to use it in one of the operations. For example, you might cluster genes by expression profiles to find other genes with profiles *similar* to the current gene.

Various gene identifiers may be present in the GIPO data file associated with the array. One of these is selected to as a unique identifier to represent genes in the MAExplorer database. Normally, the **Master gene ID** is defined as the Clone ID. However if the Clone ID is not present, but the GenBank ID is, it will use the latter as the identifier. If neither GenBank nor Clone ID is present, it will use GenBank5' then GenBank3' if present. If that is not present, it will use the UniGene ID if is present. If that is not present, it will use dbEST5' then dbEST3' if present. If that is not present, it will use LocusLink LocusID if present. Finally, if none of those identifiers are present, you can specify a 'Generic ID' that is related to some other database gene identifier such as a 'Location' identifier.

The current gene may be specified by clicking on a spot in the microarray image or on a point in the popup scatter plot, or a gene ID cell in a report.

Setting the "current gene" by Gene Name Guesser

In addition, the user may type a specific gene name or clone ID into a popup Gene Name Guesser dialog text window. This is invoked by clicking on the blue button "Enter gene name or clone ID" at the top right in the control panel. When the "guesser" window pops up, start typing the gene name or clone ID in the blue text entry field. You select either the Gene Names, Clone ID, UniGene ID, GenBank, GenBank 3' or GenBank 5',dbEST 3', dbEST 5', or LocusID identifier. Then you may start typing letters and it will match all names or identifiers which are prefixed with the sub-string you have typed so far. As you type more characters, it will limit the list of possible completions of what you are typing. After selecting the gene you want, you then press the "Done" button to use this entry to set the current gene and remove the guesser popup window. You may press the "Clear" button to clear what you have typed and the "Cancel" button to cancel the current gene selection process.

Setting the "Edited Gene List" subset of genes using wildcard names

You may also define a set of genes from the guesser window using wildcard names where the character '*' matches zero or more characters. First you specify a sub-string common to gene names. Then press the "Set E.G.L." (set 'Edited Gene List') button. For example (see Figure 2.3.1), you could find all oncogenes and proto-oncogenes by typing "*ONCO*" in the guesser. It automatically enables the View 'Edited Gene List' in the array that shows genes in the E.G.L. enclosed in magenta boxes.

The current gene cluster

Some operations involving clustering will automatically assign the gene cluster to the E.G.L. This includes clustering of genes

similar to a selected (i.e. current) gene and K-means clustering. In the case of K-means clustering, the cluster you select by picking a gene belonging to that cluster will cause it to be defined as the current cluster and also assigned to the E.G.L. This will be discussed in more detail in the section on clustering.

Saving full resolution plots as GIF files in stand-alone mode

The various plots may be saved as full resolution GIF files when running MAExplorer in stand-alone mode. The various plots have "SaveAs" buttons which appear in stand-alone mode. Saving your intermediate results may be useful for documenting your data mining session or for subsequent publication. (Here is an example of a full resolution clustergram of 38 MGAP hybridized samples for 1076 named and EST genes).

Saving Text windows as .txt files in stand-alone mode

The various text windows may be saved as .txt files when running MAExpplorer in stand-alone mode. The various text windows have "SaveAs" buttons which appear in stand-alone mode. Saving your intermediate results may be useful for documenting your data mining session or for subsequent publication.

1.2 Microarray image quantification

Quantification data for all genes in a hybridized sample (x and y coordinates, intensity, background density) is obtained by reading data from a *quantification* file for that hybridized sample. The quantification file for each hybridized sample resides on the local file system (for stand-alone) or MAExplorer Web server (for applet use) and is derived from image quantification programs such as Axon's GenePix(TM) program, Scanalyze, Molecular Dynamics' ImageQuant(TM) program, Research Genetics' Pathways(TM) program, etc.   These programs are independent of MAExplorer and are not part of our downloadable software distribution. Normalization between hybridized samples must be performed to allow comparison between different hybridized array samples. File formats are discussed in Appendix C).

1.2.1 Ratio and Zscore comparison of data from different hybridized samples

Because of variation between hybridized samples, data is normalized. Methods that are pure scaling transformations (such as Median, Scale to 65K, By Calibration DNA, By Use Gene Set, etc.) allow you to compare data using the ratio between two normalized sets of data. We define the ratio for two samples as follows:

```
    ratio(x,y,c) = I_xc / I_yc
where:
    samples x,y have values I_xc and I_yc for the same
    gene c in samples HP-X and HP-Y
```

The Zscore method transforms the data such that it can not be used with the ratio comparison. Instead we use the Zdiff(x,y) method for comparing Zscore developed by Mark Vawter (Vawter, 2000). Zscores typically cover the range of -3.0 to +3.0 (standard deviations) with a transformed mean of 0.0. Therefore the Zdiff will typically cover the range of -6.0 to +6.0.

```
Let
    Zscore(p,c) = (I_pc - mean_p)/stdDev_p
where:
    I_pc is the intensity of gene c for sample p. Sample p has mean_p
    and stdDev_p

Then,
    Zdiff(x,y,c) = Zscore(x,c) - Zscore(y,c),
where:
```

```
    samples x,y have Zscore(x,c) and Zscore(y,c) normalized values for the
    same gene c in samples HP-X and HP-Y, or HP-X 'sets' and HP-Y 'sets'.
```

Table 1.2.1 Displays affected by the normalization mode. When comparing two hybridized samples or sets of hybridized samples, the metric used is either ratio or Zdiff depending on whether the Zscore normalization was selected in the Normalization menu. This will affect a variety of data displays and data Filter methods listed here. In addition, all of the other graphics (EP plots, intensity histogram plot, cluster plots including clustergrams and dendrograms) are also affected by the normalizations.

1. Pseudocolor X/Y ratio, X-Y Z-diff, and other pseudoarray images
2. the 3-line gene data displayed in the main MAExplorer window
3. the gene data display at the top of the scatter plot when clicking on a point in the scatter plot
4. report tables of genes with the highest/lowest X/Y (Cy3/Cy5) (F1/F2) ratio or X-Y (F1-F2) Zdiff
5. Ratio histogram plot of X/Y (Cy3/Cy5) (F1/F2) ratios or X-Y (Cy3-Cy5) (F1-F2) Zdiff data
6. data Filter: Spot Intensity [SI1:SI2] range or Zdiff [Z1:Z2] range for HP data
7. data Filter: Intensity [I1:I2] range or Zdiff [Z1:Z2] range for (HP-X/HP-Y) or (HP-X - HP-Y) data
8. data Filter: Ratio [R1:R2] range or Zdiff [Z1:Z2] range for (HP-X/HP-Y) or (HP-X - HP-Y) data
9. data Filter: Ratio [CR1:CR2] range or Zdiff [CZ1:CZ2] range for (Cy3/Cy5) or (Cy3-Cy5) data of a single sample
10. Range and scale of data in EP plots, cluster plots, clustergrams, dendrograms, etc.

1.3 Microarray image and plot display

The MAExplorer displays one microarray **pseudoarray image** of the hybridized samples. This is either for a single sample, the ratio of two samples, the average of replicate samples or the ratio of two sets of replicate samples, the ratio Cy3/Cy5 or Cy5/Cy3, or other mappings. Section 2.4.4.1 Show microarray pseudoarray images menu describes these options and shows some examples.

The **Filter** menu is used to select a set of data filters that determines which genes are selected. These are highlighted in the array image in different ways - with a red (white) circle in the intensity (ratio) pseudoarray image each spot meeting the range threshold criteria. How these are highlighted depends on which **Plot** menu **Show Microarray** method and **View** menu modes were selected. If the **Show 'Edited Gene List'** (EGL) option is set in the **View** menu, genes in the EGL will appear as magenta squares. The "Filter mode" is always present and shows genes meeting various Filter criteria (to be discussed). The user may interactively define a list of genes by clicking on them when the **Click to add gene to edited gene list** option is set in the **Edit** menu. Alternatively, you can click on a gene with the Control key pressed to add a gene to the EGL or with the Shift key pressed to delete a gene from the EGL.

Types of pseudoarray image displays

There are several differnt types of pseudoarray images that may be displayed. The current type is set in the **Show Microarray** submenu in the **Plot menu** selections including **Pseudograyscale intensity** that approximates the intensity of a single sample or average of samples. The **Pseudocolor Red(X)-Yellow-Green(Y) HP-X/HP-Y ratio or Zdiff** and **Pseudocolor Red(Cy5)-Yellow-Green(Cy3) Cy3/Cy5 (or F1/F2) ratio or Zdiff** add the two samples or channels together as separate Red+Green channels to give a color spectrum. The **Pseudocolor HP-X/HP-Y ratio or Zdiff Pseudocolor Cy3/Cy5 (or F1/F2) ratio or Zdiff** gives a color spectrum from a low ratio (zdiff) value (Green) to a high value (Red) with a value of 1.0 (0.0) of Black. The **Pseudocolor (HP-X,HP-Y) 'sets' p-value** shows the p-Value between two X and Y sets in a color spectrum.. If the **Original image** is set and the image file is in the database, it will pop up a separate Web browser window to display it. The Pseudograyscale display is a grayscale image, with higher concentration genes appearing darker, on a light blue background. The pseudocolor HP-X/HP-Y ratio of spots image is constructed using a color scale going from bright green (<1) to black (=0) to bright red (>1) on a black background. For the pseudocolor Zdiff of (X-Y), the color scale goes from bright green (<0) to black (=0) to bright red (>0). If the **dichromasy** switch is set in the View menu, that a different set of colors is selected that may be easier for some people to differentiate. If the **Use dual HP-X & HP-Y 'sets' else single samples** toggle in the Samples menu is set, it displays the mean HP-X data in the left and HP-Y in the right for doing a side by side comparison.

In all of the pseudoarray images, the grids in the image are labeled *field#-GridLetter* (e.g. 1-C, 2-B, etc). This allows them to be clearly identified as the user scrolls over the image that is larger than the visible computer window.

Popup windows

MAExplorer starts with the main pseudoarray image windows. This window contains the pull-down menus where you may issue commands. As you perform various operations, new windows may popup for some of these commands. For most of these windows, you may click on the "Close" button or click on the close window icon associated with your operating system (generally one of the buttons at the top of the popup window). However, some windows were designed to not close when you do this. In particular the "State sliders" are *not* able to be closed unless the associated data filtering or clustering operation is closed. When you close the associated operation will automatically close the state slicer window.

The current sample sample, HP-X, and HP-Y

In MAExplorer, a hybridized array sample is abbreviated HP. The underlying data comparison model assumes, as a minimum, the comparison of two different experimental conditions represented by samples HP-X and HP-Y. A good way to think about this is that these variables are the two axes of a scatter plot (one of the displays you may generate). The HP-X and HP-Y may be thought of as containing data from either single hybridized samples or containing mean data from multiple replicate sets of sample. The HP-X and HP-Y are assigned using the **Set current HP-X** and **Set current HP-Y** in the **Samples** menu (hybridized sample is abbreviated HP in MAExplorer. The sets are most easily changed using **Choose HP-X, HP-Y and HP-E** to select the currently active samples. The contents of the of multiple sample HP-X and HP-Y 'sets' may alternatively be changed using the **Edit HP-X & HP-Y 'sets' of samples by source** submenu, and the HP-E list of samples using the **Edit HP-E list of samples by source**. Assigning single samples to either HP-X or HP-Y may be done from the Samples menu. However, it is easier to do it by clicking on the pseudoarray image. First click on the magenta "[X]" or "[Y]" Current Sample box at the top of the list of switch between HP-Y and HP-Y. Whichever is visible ([X] or [Y]) is the one that will be the HP sample assigned. Then simply click on the magenta "*" to the left of the sample name for the sample you wish to assign.

Hybridized samples are selected from a list of all of the sample samples in the database. To make it easier to select a HP, they may be selected from submenus by their developmental stage (if supported by your particular database) or from a list of all samples in the database located on the left side of the pseudoarray image. If a sample has never been loaded during a session, it will be loaded when you request it.

The last sample selected is called the *current sample* or current HP. That is the sample that is displayed in the pseudoarray image in the primary MAExplorer window when using display modes requiring a single sample.

Using 'sets' of HP-X and sets of HP-Y

Multiple samples may be assigned to the to the HP-X or HP-Y *sets*. These are assigned using the **Edit HP-X and HP-Y 'sets' of microarrys** in the **Samples** menu. The multiple sets are enabled by setting the **Use HP-X and HP-Y 'sets' else single samples** checkbox in the **Samples** menu. Then, when statistical calculations are performed on that data, it will use the means, std-deviations, etc. from each of these sets rather than individual samples.

The HP-E sample list for computing expression profiles

You may cluster sets of genes with similar expression profiles across a set of hybridized samples. The set of HP samples used in doing these profiles is specified by **Edit expression profile 'list (HP-E)** in the **Samples** menu. The **Choose HP-X, HP-Y, and HP-E** command may also be used for defining the members and order of the samples in the HP-E 'list'. Then, gene intensity expression profiles may be created in a popup window for hybridized samples in the HP-E set by using the Expression profile plot commands in the Plot menu. Several of these plots may be created on the screen at the same time. Clicking on a vertical data line in the plot will show the name of the HP, its intensity and coefficient of variation (CV) of the (F1,F2) data for this gene. Note that you can order the hybridized samples in the HP-E set by the order in which they are added.

Data 'Filters' - the intersection of one or more data tests

A set of genes may be computed by taking the intersection selected gene sets. These sets are determined by various logical, data

range and statistical tests. Genes passing each test are assigned to a gene subset which in turn are used in the gene intersection computation. The final gene subset is used in array, plots, and reports, and subsequent data filtering. Changing any test parameters causes the data filter to be re-computed.
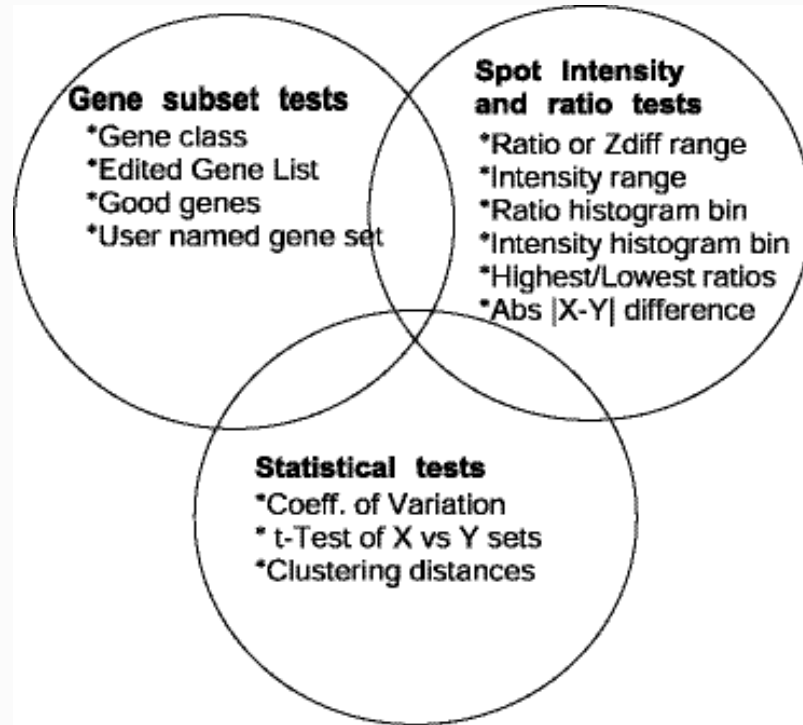
Figure 1.3 Data Filter Venn diagram. This illustrates some of the logical, data range and statistical tests criteria available using the MAExplorer data Filter paradigm. Note that multiple criteria may be selected from each of these categories. The extreme case, probably never used, could use all tests.

1.4 Exploratory data analysis - overview

MAExplorer may be used to perform various data explorations by looking for patterns correlated with different sets of hybridized samples or with expression profiles of genes. This is discussed in more detail throughout this manual and later in Section 3 on Exploratory Data Analysis. Detailed descriptions of all commands are given in Section 2 Menus.

A first-approximation approach to data-mining might be to sequentially constrain the data of interest to find some changes and then to report on those changes. We have arranged these commonly performed first-pass operations as submenu entries in the Analysis Menu. The submenus are:

1. GeneClass Menu
2. Normalization Menu
3. Filter Menu
4. Plot Menu
5. Cluster Menu
6. Report Menu
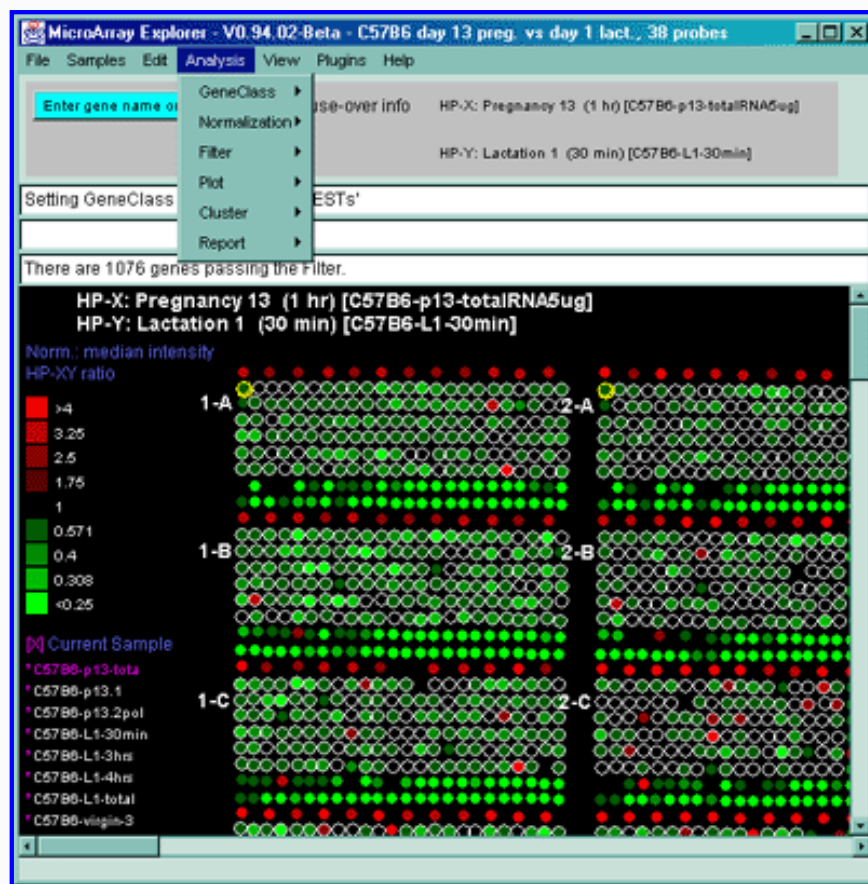
Figure 1.4 Screen view of MAExplorer main window with Analysis Menu. The menu structure of MAExplorer was designed to allow users to quickly perform commonly used data-mining operations. Other menus are used for modifying the data (File, Samples, Edit, and View menus) or accessing on-line Help menu information in a separate Web browser popup window. MAExplorer menus are similar to most Windows PC applications where pull-down menu selections are used to invoke operations. The current hybridized array sample is displayed as a pseudocolor ratio image of median normalized spot intensities. Clicking on a spot assigns it as the current gene with data being reported in the top most message area. The names of the current HP-X and HP-Y samples are listed above that area. In general, clicking on spots, points in plots or cells in spreadsheet reports will assign the it as the current gene and access Web genomic databases if enabled.

In addition to displaying the hybridized sample pseudoarray images, derived data may be viewed in various types of plots. These include scatter plots, histograms, ratio-histograms, expression profiles, gene clustering, etc. Data may be presented as table reports presented as either active spreadsheets that can access genomic databases by clicking on cells or as tab-delimited Excel-compatible tables that may be cut (if your windowing system supports this) and pasted into an Excel spreadsheet.

The selected HP-X and HP-Y samples are used when generating scatter plots, ratio histograms and other graphics. Scatter plots and ratio histograms may also be performed on the left and right sides of the currently displayed HP array (fields F1 and F2 respectively if array data has duplicate spots for the same genes).

A MAExplorer database contains a table identifying genes, so data is accessible by gene name as well or by sub-strings identifying a set of genes (e.g. "onco" that could be used to find any oncogene or proto-onco gene in the database).

When the program starts, it displays the microarray image of the first hybridized sample in the HP-X set of samples initially specified. If you specify a new HP-X or HP-Y sample, then it changes the pseudoarray image to correspond to that array. You may change the current HP-X or HP-Y sample from either the Samples pull-down menu or by clicking on a sample in the **Active Sample** list in the left of the pseudoarray image. If you click the mouse on or near a spot, it will *latch* onto that spot and define it as the *current gene*.

Note: In Figure 1.4, genes that pass the MAExplorer data Filters are indicated by red (white) circles around spots in the

pseudograyscale (pseudocolor) intensity (ratio) image. The pseudoarray image shows the gene data as replicate grids of spots if there are two fields Field 1 (left set of grided spots) and Field 2 (right set of grided spots). If there is no duplicate spot data, then only Field 1 is shown.

If background correction is enabled in the Normalization menu, then intensity is reported in the message displays as *intensity'* otherwise as *intensity*. Normalization should also be used between hybridized samples - whether the data is ratio data (i.e. Cy3/Cy5) or single sample intensity arrays.

1.4.1 Saving the state of a data-mining session in stand-alone mode

If you are running MAExplorer in stand-alone mode, you may save the state of your session for later use using the "Save DB" or "SaveAs DB" commands. Then, the checkpointed database could be accessed using the "Open file DB command". It currently saves: the gene sets, condition (HP) lists, current HP-X, HP-Y and HP-E lists, data Filter options and slider value settings, display options, clustering options, normalization options, etc. We recommend using the "SaveAs ... DB" so you can save the state under a different name rather than overriding the original state. This way you could backup to the original state if you wanted to. The "SaveAs DB" and "Open file DB" commands are described in the File menu.

1.4.2 Logging messages and command history

Often a user would like to review measurements of particular genes and to review the list of commands they issued (also called the command history). Various data measurements as well as many other types of information in the three text lines in the status area of the main window may optionally be recorded in a popup message log (Section 2.5.1) and the command history may also be reviewed in a separate popup message log (Section 2.5.2). If you are running the stand-alone version, the logs may be saved. Otherwise, you could cut and paste the log data into other word processing applications.

1.5 Quick start - demonstration of MAExplorer

MAExplorer is normally used as a stand-alone application. You may download the stand-alone application (see Appendix D). The SourceForge download does not include any data. The NCI/LECB download also include a demo data set of 50 hybridized samples from the MGAP database. In any case, you can explicitly download the data at any time at http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database.zip>.

Setting up MAExplorer to work with user-specific data is discussed later in this manual in Appendix C.
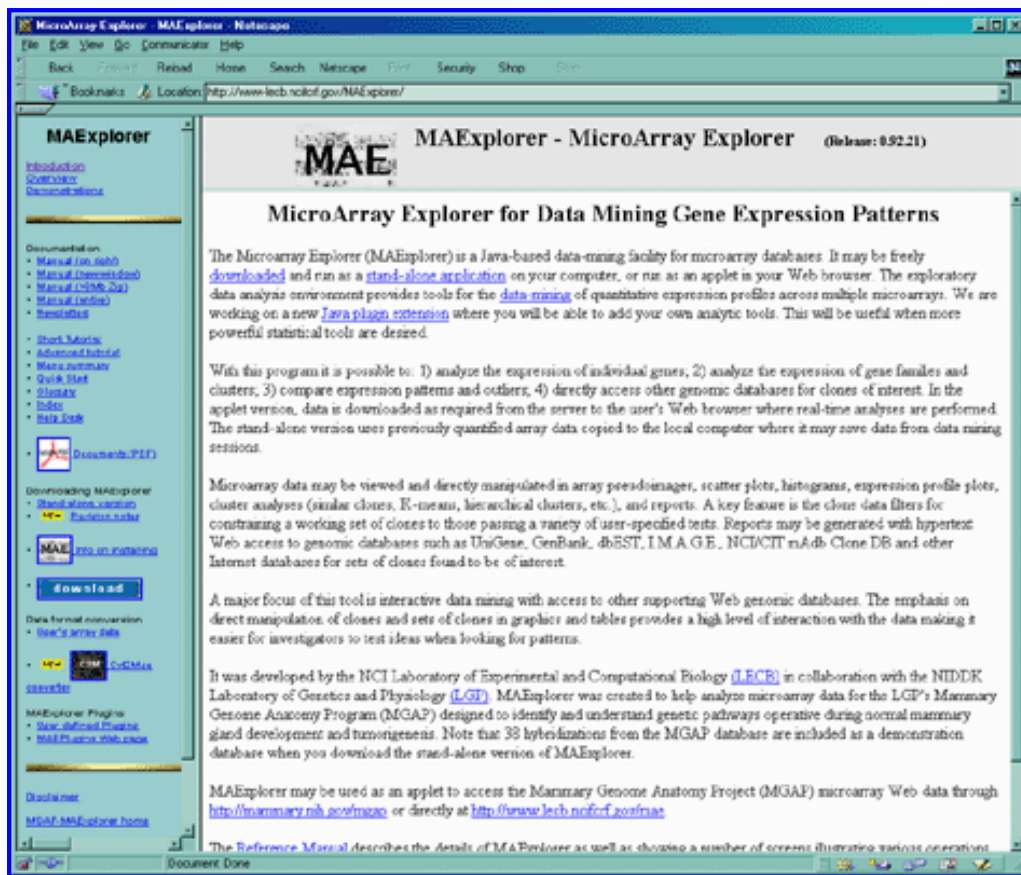
Figure 1.5.1 The MicroArray Explorer home page at http://maexplorer.sourceforge.net/. The table of contents in the left panel lists an introduction and short tutorial, several demonstration databases. Below that are links to documentation including this reference manual, glossary and index. The Export version discusses running MAExplorer with other arrays and as a stand-alone version. The Download application is a Web page for downloading and installing the stand-alone Java application on your computer.

You may start MAExplorer in your Web browser from the MGAP Startup DB. This offers several preset public databases consisting of sets of hybridized samples as well as the empty database. After you have clicked on a particular startup database, it will begin loading MAExplorer - indicated by a red box with a "Loading..." message in the top window of your browser. After MAExplorer starts, this message changes to a white box with "Reading DB" while it downloads the data files required. Finally, when it is ready for your interaction, it displays a white box with a green "Ready".

NOTE: for Web browser invocation, the MAExplorer applet works with Netscape 4.7, Internet Explorer 5.0, and HotJava on a Windows (95/98/NT/2000/XP) system or a Solaris Unix system. Macintosh and SGI systems seem to hang at times because of Web browser problems. However, it works on all other systems as a stand-alone Java application that you may download and install on your computer. You might want to review these Web browser restrictions.

After the MAExplorer is started and the menus become active, you may switch the preset hybridized samples to other samples using the **Samples** pull-down menu. The last hybridized sample loaded becomes the "current hybridized sample" and its image is the one displayed.

Exiting MAExplorer

If you are in MAExplorer and want close the program and exit, you may use the **Quit** command in the **File** menu or click on the "close application" button (found in the upper right hand corners of MAExplorer windows put there by your operating system).

1.6 Tutorials for using MAExplorer

There are a number of things you may do in this data mining facility.

1. Analyze expression of individual genes
2. Analyze expression of gene families and clusters
3. Compare expression patterns in multiple hybridizations

We wrote two tutorials to help you understand its capabilities. We recommend you first try the short tutorial before attempting the advanced tutorial. The latter demonstrates some of the more advanced capabilities.

---

2. MAExplorer menus

A MAExplorer analysis is performed using various interactive controls. These commands are selected from pull-down menus located on the "menu bar" at the top of the main MAExplorer window. The primary menus are:

- **2.1 File** - database, file access operations
- **2.2 Samples** - select lists of hybridized sample conditions
- **2.3 Edit** - edit gene and condition subsets, E.G.L. and preferences
- **2.4 Analysis** - primary analysis menus
    - **2.4.1 GeneClass** - select gene subset for gene class data Filter
    - **2.4.2 Normalization** - select gene intensity normalization mode
    - **2.4.3 Filter** - select data filters to compute gene subset of interest
    - **2.4.4 Plot** - pseudoarray image, scatter, histograms, expression profile popup plots
    - **2.4.5 Cluster** - perform cluster analysis on data filtered genes
    - **2.4.6 Report** - generate popup spreadsheet reports of genes and samples
- **2.5 View** - setup genomic gene data views preferences
- **2.6 Plugins** - add and execute new MAEPlugin methods
- **2.7 Help** - popup documentation on MAExplorer and specific database

Menu notation

In the following menus, selections that are *sub-menus* are indicated by a '▶'. Selections prefaced with a '☑' and indicate '☐' indicate that the command is a checkbox that is enabled and disabled respectively. Selections prefaced with a '◉' and indicate '◯' indicate that the command is a multiple choice "radio button" that is enabled and disabled respectively, and that only one member of the group is allowed to be on at a time. The default values set for an initial database are shown in the menus. Selections prefaced with a '#' indicate that the commands are available only when MAExplorer is run in the stand-alone mode. Selections prefaced with a '*' commands requires access to the backend Web server [Future]. Selections that are not currently available will be grayed out in the menus of the running program.

The following Sections 2.1 through 2.7 describe the pull-down menus in detail.

2.1 File menu

The **File** menu operations includes options and submenus providing access to database data from disk and Web servers, state saving, and groupware to share states between collaborators.

In stand-alone mode, the user may select the database subset to be loaded from either a Web server or a local file system. When used as an applet, this is pre-determined by the Web page where MAExplorer is started. Opening a disk DB, 'Open disk DB', also restores any user defined gene sets and other parts of the exploratory state that were present when the 'Save ... disk DB' was invoked.

In the following menus, selections that are *sub-menus* are indicated by a '▷'. Selections prefaced with a '☑' and indicate '☐' indicate that the command is a checkbox that is enabled and disabled respectively. Selections prefaced with a '◉' and indicate '◯' indicate that the command is a multiple choice "radio button" that is enabled and disabled respectively, and that only one member of the group is allowed to be on at a time. Selections prefaced with a '#' indicate that the commands are available only when MAExplorer is run in the stand-alone mode. Selections prefaced with a '*' commands requires access to the backend Web server [Future]. Selections that are not currently available will be grayed out in the menus of the running program.

When used as an applet connected to a Web database server, databases may be divided into public and collaborator projects. Users accessing protected collaborator projects will be required to log-in to the server and a popup login request will appear.

- **Databases**▷ - open and save databases of hybridized samples
- **Exploratory 'state'**▷ - save and restore the user's data-mining explorations
- **Groupware**▷ - share some exploration states with collaborators [Future]
- **Quit** - exit MAExplorer

2.1.1 Databases menu

The **Databases** submenu is currently available *only in stand-alone mode* and contains the following selections for opening and saving databases.

[In the future], each user will be able to save the state of their exploration into a password protected directory of named states on a Web server (e.g. doing a 'Save ... Web DB' command. Later, they could restore that state from the Web server by doing an 'Open Web DB' command). Users would be required to register with that server to set up a unique state-saving area. Once this facility was setup, users may selectively allow other user's to view selected data implementing a groupware environment for improving collaboration.

- # **New project** - assign new project `.mae` database file. This assigns a project to either a local disk directory (or to a Web DB [Future]).
- # **Set project directory** - change the project directory and then do an "Open disk DB" command (or "Open Web DB"). Note: used with "New project" [Future].
- # **Open disk DB** - open an existing disk `.mae` database file
- # **Open Web DB** - open MAExplorer database with Web server that supports MAExplorer. [Future]
- # **Save disk DB** - save current database as a `.mae` startup disk file
- # **Save as disk DB** - save current database as a new `.mae` startup disk file
- # **Save Web DB** - save current database as a `.mae` startup Web DB file [Future]
- # **Save as Web DB** - save current database as a new `.mae` startup Web DB file [Future]
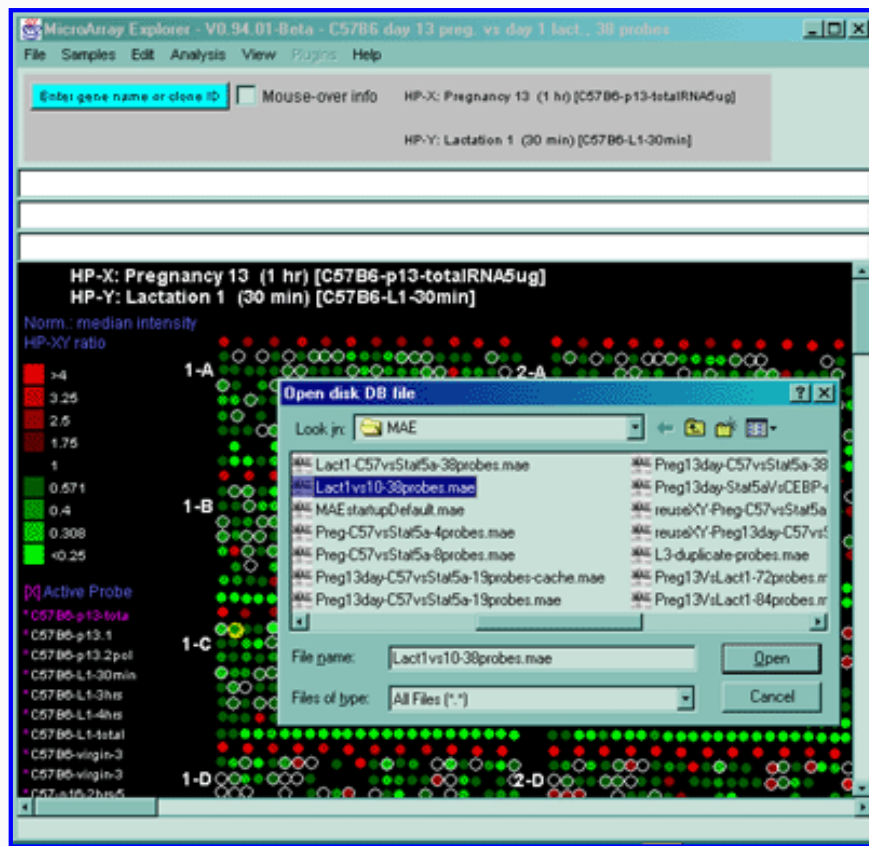
Figure 2.1.1 Example of the "Open file DB" command. The file browser is opened in the current project directory with the name of the currently opened file. You may select another .mae startup database file to load in the current project. You may also "cruise" the file system and load an .mae file from a different project directory. The "Set project" command makes this easier since it gives you a list of available projects that you may change directly. The projects must have been setup on your computer previously. The "New project" command can be used for setting up new projects or projects.
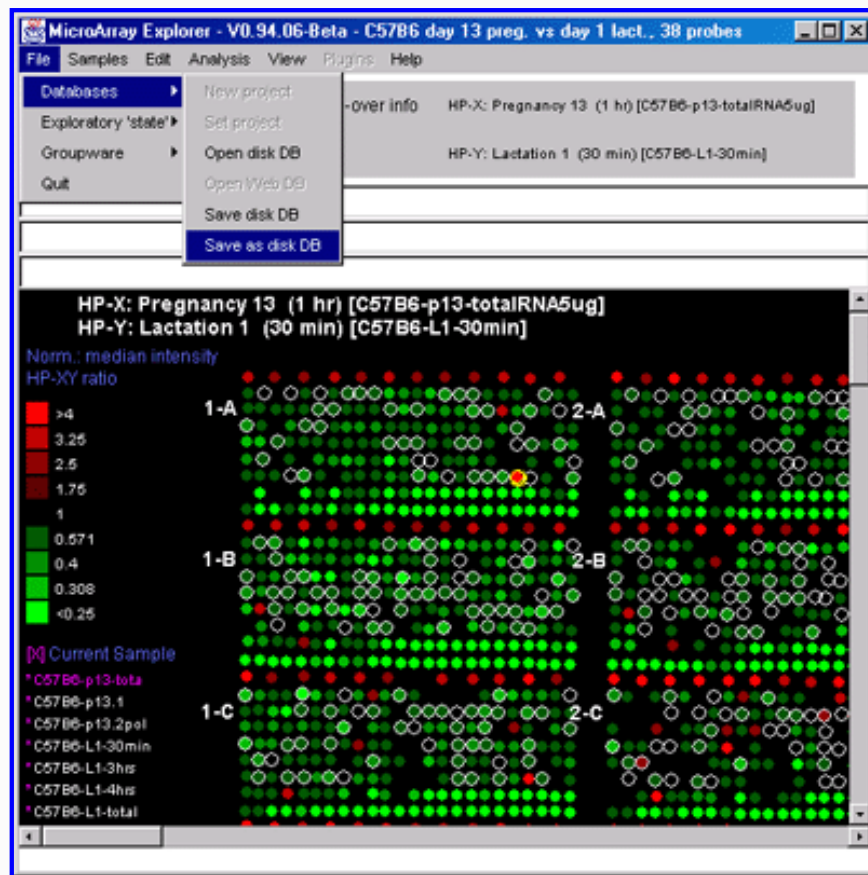
Figure 2.1.2 Example of saving a user session in a new startup file using the "SaveAs DB" command. The file browser is opened in the current project directory with the name of the currently opened file. You may enter another .mae file name to save your current session. Then when you restart MAExplorer using this new file, it will restore the data mining state to where you left off (except that no popup windows are opened).

2.1.2 Exploratory state menu

The **Exploratory 'state'** submenu contains the following selections for saving and later using the user's state of the exploration. If MAExplorer is being run on a local computer, no login is required.

- *** Register** - register your login name and request a password so you may login in the future [Future]
- *** Login** - log on to the MAExplorer web server. This is used for accessing protected projects and for saving accessing your "state" data [currently only used for MGAP DB]
- *** Open user's state** - restore a user's previously saved exploratory state file [Future]
- *** Save user's state** - save the user's current exploratory state in a file on the server [Future]
- *** Directory of user's states** - list the user's named exploratory states [Future]
- *** Delete user's state** - delete a particular user state file [Future]

2.1.3 Groupware facility for sharing user states menu [Future]

The groupware facility allows users to share their state data with other users. However, If MAExplorer is being run on a local computer, then groupware may not be available since it depends on using specific Web servers with MAExplorer-specific groupware services available. [We first developed these concepts in the context of 2D protein gels. They include: WebGel (Lemkin et al., 1999b) for Internet exploration of 2DE databases, Xconf (Lemkin et al., 1993) an early X-windows based image conferencing similar to CU-See-Me or ALO Instant messenger for sharing images in a conference over the net, Flicker (Lemkin, 1997) an image comparison over the internet using a a Java applet, and GELLAB-II (Lipkin and Lemkin, 1981) a system for data mining - see GELLAB-II Poster on the Web) that embody many of the concepts used in MAExplorer.] The **Groupware** submenu

contains the following selections:

- **Open another user's state** - read another user's exploratory state file if they granted you access [Future]
- **Share user state** - allow another user to access a particular exploratory state [Future]
- **'Unshare' user state** - disallow another user access a particular exploratory state [Future]

Saving the state and databases on the local file system

The current state of an exploratory data analysis session may be saved on the local file system. The state consists of: gene sets; HP-X and HP-Y hybridized sample sets; HP-E hybridized sample lists; thresholds and switch setting preferences; etc. The user may save their current state and restore a previous state at any time. Restoring the state will overide the current database.

Groupware sharing of intermediate exploratory results [FUTURE]

Each registered user will be able to save the current state of their exploration of the data in named *User State* files on the back-end Web server using the **Save user's state** command. The user may keep multiple named states on the back-end secure server where they be accessed to restore the state at a future time using the **Open user's state** command. The user can request a list of their states with the **Directory of user's states** command. They may remove a particular state with **Delete user's state**.

A registered user may allow another registered user to access their state or states (using the **Open another user's state** command) if the user owning the data had granted them permission. The **Share user state** and **Unshare user state** commands control these permissions. There are two special share-users defined: **public** to allow unlimited read-only access to the state they specify, and **private** to disallow all access to a user state.

2.2 Samples menu

Each experimental condition sample is represented by a hybridized sample (abbreviated HP in MAExplorer). The **Samples** menu operations include operations to select the current hybridized sample or samples. The simplest model of a MAExplorer analysis assumes (at least) two microarray hybridized samples variables HP-X, HP-Y whose data may be plotted against one another or compared. The default is a single HP-X sample and a single HP-Y sample.

The first menu command, "Choose HP-X, HP-Y and HP-E", entries lets you change the *current* HP-X 'set', HP-Y 'set', and HP-E 'list' hybridized samples. The second menu command, "Set Samples from lists", lets you change the *current* HP-X and HP-Y, HP-Y samples as well as the HP-X 'set', HP-Y 'set', and HP-E 'list' samples. You may change the current HP-X or HP-Y sample by clicking on the sample name directly in the list of sample names on the left side of the pseudoarray image (see Figure 2.2.3 legend).

The third menu entry, "Edit use (Cy5/Cy3) else (Cy3/Cy5) for each HP", lets you swap data channels for Cy3/Cy5 data for individual samples.

Other menu commands list the status of the current HP-X 'set', HP-Y 'set', or HP-E 'list', and define condition class names that are associated with the HP-X 'set' and HP-Y 'set'. The last menu entry, "Use HP-X & HP-Y 'sets' else single samples", lets you switch between using HP-X and HP-Y as single samples of sets of multiple samples. For example, if you are using a scatter plot of X and Y, it will switch the data being plotted from a comparison of single samples to a comparison of means of sets of samples depending on the status of the switch. Sets of samples are used extensively in data explorations.

- **Choose HP-X, HP-Y and HP-E** - change the currently active samples.
- **Set Samples from lists** ▷ - define HP-X, HP-Y, HP-X 'set', HP-Y 'set', or HP-E 'list' from lists of samples.
- ---------------------
- **Edit use (Cy5/Cy3) else (Cy3/Cy5) for each HP** for use with ratio data. This selectively swaps (Cy3,Cy5) data entries so

may use (carefully!) dye-swap data for replicates. Only available for ratio data.
- ----------------------
- **List HP-X & HP-Y sample 'sets'** - list the samples in the HP-X and HP-Y 'sets'.
- **List HP-E sample 'list'** - list the samples in the ordered HP-E 'list'.
- **Define HP-X class name** - for set of HP-X samples
- **Define HP-Y class name** - for set of HP-Y samples
- ☐ **Use HP-X & HP-Y 'sets' else single samples** - toggle between using HP-X and HP-Y 'sets' of multiple samples or single HP-X and HP-Y samples.

Figure 2.2.1 shows setting the HP-X, HP-Y, HP-E lists of samples using the "Chooser" - the preferred method. Figure 2.2.2 shows setting the HP-X sample from the menus. Figure 2.2.3 shows changing the current HP-X sample by clicking on a sample name in the microarray image.
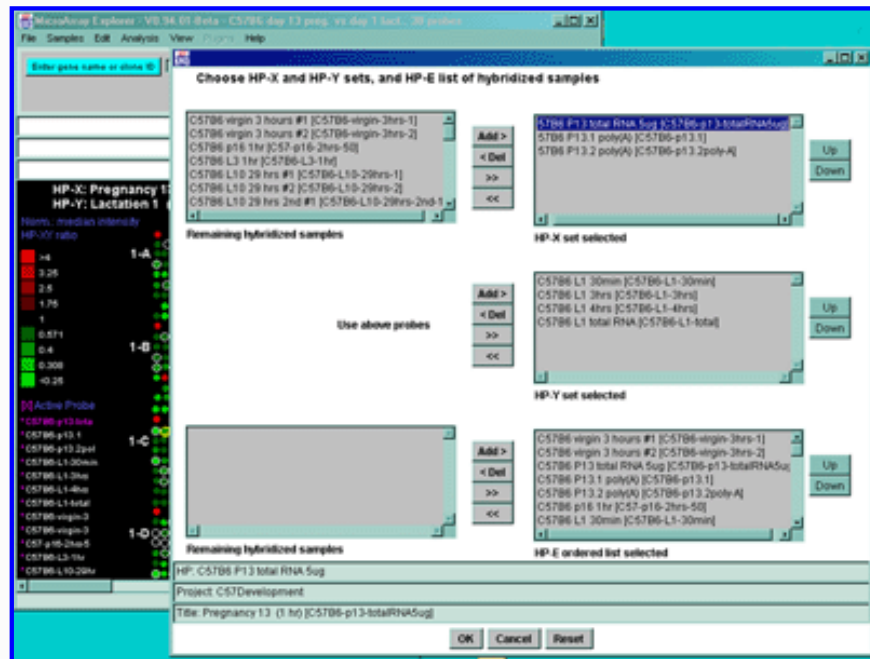
Figure 2.2.1 Samples menu - selecting lists of samples by using the "chooser". The hybridized samples assigned to the current HP-X, current HP-Y, set of HP-X, set of HP-Y and expression profile list HP-E may be changed from the **Samples** pull down menu using the **Choose HP-X, HP-Y and HP-E** option lets you graphically change the currently active sample HP-X, HP-Y sets and E-list.

Selecting HP X, Y, or E samples using the pull-down menus

Figure 2.2.2 Samples menu - selecting samples by source characteristics. The hybridized samples assigned to the current HP-X, current HP-Y, set of HP-X, set of HP-Y and expression profile list HP-E may be changed from the **Samples** pull down menu. The specific "By Source" menus shown here are from the MGAP database. This figure shows the user changing the current X sample from the developmental stages submenu that is part of the "By Source" submenu. Alternatively, samples containing a keyword or part of a keyword can be found using a "guesser" popup window that allows the use of wild cards. This is invoked using the "From list of all H.P.s" submenu. For example, you could specify "*pregnancy*" to find all samples of containing that word.

Figure 2.2.3 Changing the current sample to either the HP-X or HP-Y sample by clicking on a sample name at the left edge in the microarray pseudoarray image. The current sample is indicated in magenta. Click on the magenta "*" adjacent to the new name you want to select and it will change the HP-X sample. To switch between setting HP-X and HP-Y, click on the **[X] Current Sample** box to change the sample to HP-Y. You can click on **[Y] Current Sample** box to change it back to HP-X. Then clicking on a sample name will set it to the current HP-X or HP-Y that was selected. This figure shows that the user had selected [Y] and C57B6-L10-29hrs for the new HP-Y sample.

2.2.1 Selecting hybridized samples with Chooser or pull-down menu sample lists

The **Set Samples from lists** submenu lets you define HP-X, HP-Y, HP-X 'set', HP-Y 'set', or HP-E 'list' from lists of samples. It contains four submenus:

- **Set current HP-X sample** ▷ - i.e. single HP-X sample from the list of H.P.s
- **Set current HP-Y sample** ▷ - i.e. single HP-Y sample from the list of H.P.s
- **Edit HP-X & HP-Y 'sets' of samples by source** ▷ - edit sets of X and Y samples by source for advanced statistics comparisons.
- **Edit HP-E expr. profile 'list' by source** ▷ - edit samples by source in the HP-E ordered list of samples for expression profile statistics.

The **Set current HP-X sample** and **Set current HP-Y sample** commands offer another way to set the single current X and Y sample (see Figure 2.2.3 above for the preferred way using the "Chooser").

The **Edit HP-X & HP-Y 'sets' of samples by source** menu allows the user to define HP-X and HP-Y as *sets* having *multiple* hybridized samples. Then, the mean values of the genes are used when comparing HP-X with HP-Y.

- **Add sample to -X** ▷ - add the selected sample to HP-X set.

- **Add sample to -Y** ▷ - add the selected sample to HP-Y set.
- **Rmv sample from -X** ▷ - remove the selected sample from HP-X set.
- **Rmv sample from -Y** ▷ - remove the selected sample from HP-Y set.

The **Edit HP-E expr. profile 'list' by source** menu allows the user to define an ordered list of samples for use in expression profile statistics. Then, an expression vector of normalized quantification values (one for each sample in the HP-E list) is computed for each gene. Note: to place the samples in a particular order, start with an empty HP-E set and then add them in the order you desire.

- **Add sample to HP-E** ▷ - add the selected H.P. to HP-E list.
- **Rmv sample from HP-E** ▷ - remove the selected sample from HP-E list.

The Convention for pull-down menu sample selection lists

We use a common sample selection scheme when selecting a sample from a pull-down menu list. This sub-sample "By Source" option is only available if your database was set up to allow sub-sample source names in the Samples database.

- **By Source** ▷ - selects the sample from one of additional submenus. These database-specific menu entries might be categories such as developmental stage, tumor models, time series, etc and are set up for a specific database in its configuration.
- **From list of all samples** - select a sample from a list of all samples in the database accessible to the user.

For example, the **By Source** database-specific entries for the MGAP database includes the following submenus.

- Mammary Development ▷
- Developmental models ▷
- Tumor models ▷
- Knock-out mice ▷
- Transgenic mice ▷
- Natural mutants ▷
- Human specimens ▷

The **From list of all samples** selection pops up a hybridized sample guesser dialog window. As with the gene name guesser, you can start typing in the name of a sample and it will give you a list of HPs that match that initial string. You then click on the sample you want and then press the **Done** button.

2.2.2 Swapping selected samples' (Cy3,Cy5) channels in ratio data dye-swap experiments

The **Edit use (Cy5/Cy3) else (Cy3/Cy5) for each HP** command may be used to selectively swap (Cy3,Cy5) data entries so the user may use the samples (carefully, since gene labeling efficiency is not always symmetric!) dye-swap data for replicates. This is only available for ratio data. It swaps the data contained in MAExplorer (memory only) so that the Cy3 data is swapped for the Cy5 data. For example, consider the case where there are two materials A and B hybridized in two experiments and labeled as follows: E1 (A=Cy3,B=Cy5) and E2 (A=Cy5,B=Cy3). Then, assuming uniform symmetric labeling (which is generally *not* the case - although it might be true for a subset of genes), then one might average data from E1 and E2 if the data from E1 (or E2) were swapped. This is shown in the following figure.
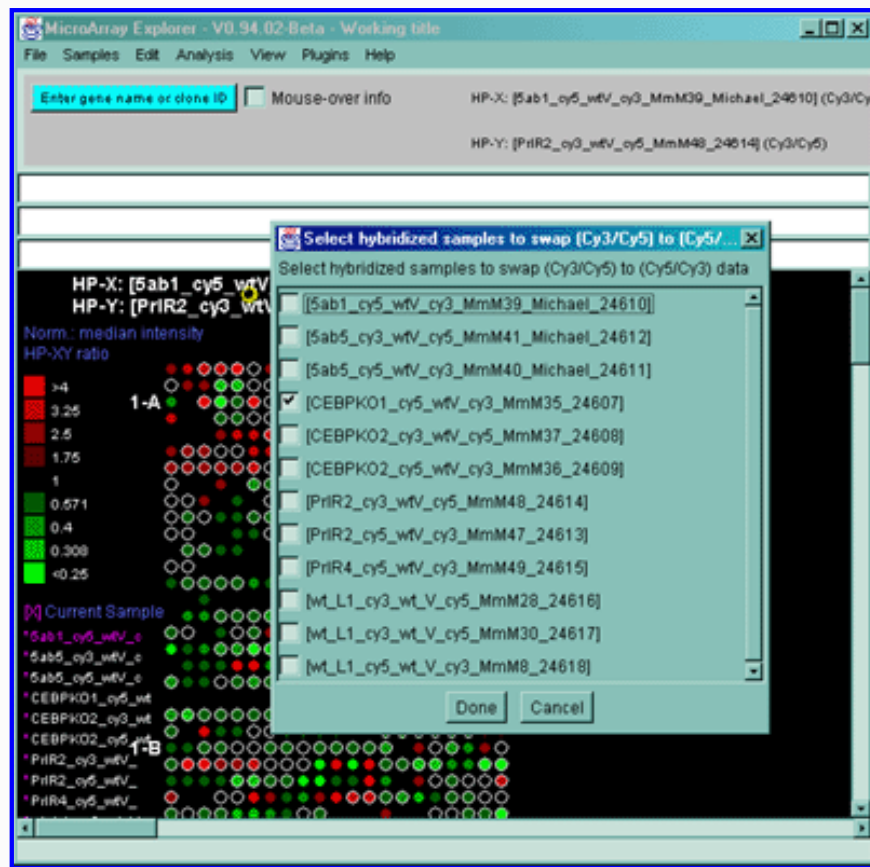
Figure 2.2.4 Samples menu - selectively swapping (Cy3,Cy5) data channels for particular samples. This is only operative if your database contains Cy3/Cy5 ratio labeling data. This is useful in databases containing subsets of dye-swap experiments mixed in with other samples that are not dye-swapped.

2.2.3 Viewing sample HP-X, HP-Y, and HP-E partitions

You setup sets of HP samples for the HP-X and HP-Y sample sets and HP-E expression list of samples using the Chooser (above). The current contents of these lists may be viewed using the **List HP-X & HP-Y sample 'sets'** to list the samples in the HP-X and HP-Y 'sets'. The **List HP-E sample 'list'** may be used to list the samples in the ordered HP-E 'list'.

2.2.4 Defining sample condition 'class' names

When using sets of conditions, the HP-X and HP-Y 'sets', you will probably want to assign meaningful names to these sets. The commands **Define HP-X class name** and
• **Define HP-Y class name** assign a free text line name for the sets.

2.2.5 Toggling between single HP-X (-Y) samples and HP-X (-Y) sets

When MAExplorer first starts up, it assumes that you wish to treat the data as single samples so that HP-X and HP-Y are assigned to single samples. However, if you want to work with sets of multiple samples then you must toggle the state using the **Use HP-X & HP-Y 'sets' else single samples** check box command. This toggles the state between treating the data as multiple samples (HP-X and HP-Y 'sets') or as single HP-X and HP-Y sample samples.

2.3 Edit menu

The **Edit** menu operations include operations to modify the 'edited gene list' that is set from a variety of Filters as well as manually this menu. The user may perform set operations (union, intersection, and difference) on named sets of gene and sets of sample experiments (conditions). User preferences are also set in this menu.

Sets of genes or HP condition lists are very useful for tracking complex data-mining sequences of analysis. For example, derived named gene sets may be used in successive data filters and for reports. For example, one could do the following experiment given four different types of HPs for (e.g. virgin, pregnancy, lactation, and involution)

> *First compare two HPs using a statistical test such as a t-test. Then save the resulting set of genes under the name "virgin vs. pregnancy". Then compare the next two HPs and save the resulting genes under the name "lactation vs. involution". Finally, compute the difference of genes found in "virgin vs. pregnancy" that are not found in "lactation vs. involution". This resulting gene set could then be saved (e.g. with a name "Genes found in virgin vs. pregnancy, but not in lactation vs. involution"). Similarly, taking the intersection of these two named sets shows genes that are common between the two sets. Taking the union shows genes found in either of the two named sets.*

The Edit menu contains the following main selections. All of these entities and preferences are saved as part of the startup state when you do a (File | Databases | SaveAs ... DB).

- **User edited gene list** - edit the user defined 'Edited Gene List' or E.G.L.
- **Sets of genes** - operations on named sets of genes.
- **Sets of Conditions** - operations on lists of hybridized samples (i.e. experimental conditions).
- **Preferences** - set various statistical limits and other parameters.

2.3.1 User edited gene list - the 'Edited Gene List' menu

You may define and edit arbitrary sets of genes using the **User edited gene list** submenu to modify the 'Edited Gene List' (EGL). This has sub-modes of operation for adding or removing genes from the image by clicking on spots. If the **Show 'Edited Gene List'** mode is set, you may see exactly which genes you have defined by the magenta squares drawn around each gene in the EGL. Many of the clustering operations will leave the current cluster in the EGL. The commands include:

- ☐ **Show 'Edited Gene List'** - toggle showing the EGL as magenta boxes in the pseudoarray image. If enabled, genes set by manual selection or as the result of some filtering operations
- ◉ **Don't edit** - clicking on a spot does nothing (i.e. disable the 'click to add (remove) genes to (from) the E.G.L.'.
- ☐ **Set 'Edited Gene List' to Filtered genes** - set the 'Edited Gene List' to the current genes defined by the Filter.
- **Click to add gene to E.G.L. (Ctrl/click)** - clicking on a spot adds the corresponding gene to the 'Edited Gene List'.
- ☐ **Click to remove gene from E.G.L. (Shift/click)** - clicking on a spot removes the corresponding gene from the 'Edited Gene List'.
- **Set 'Edited Gene List' to Filtered genes** capture the current Filtered genes into the E.G.L.
- **Clear 'Edited Gene List'**

This gives you the functionality of adding and deleting genes from a user defined list of genes to be analyzed. The EGL may be used with the gene-set operations discussed in Section 2.3.2. You may also define genes in the EGL using the "Gene Name Guesser" shown in Figure 2.3.1.
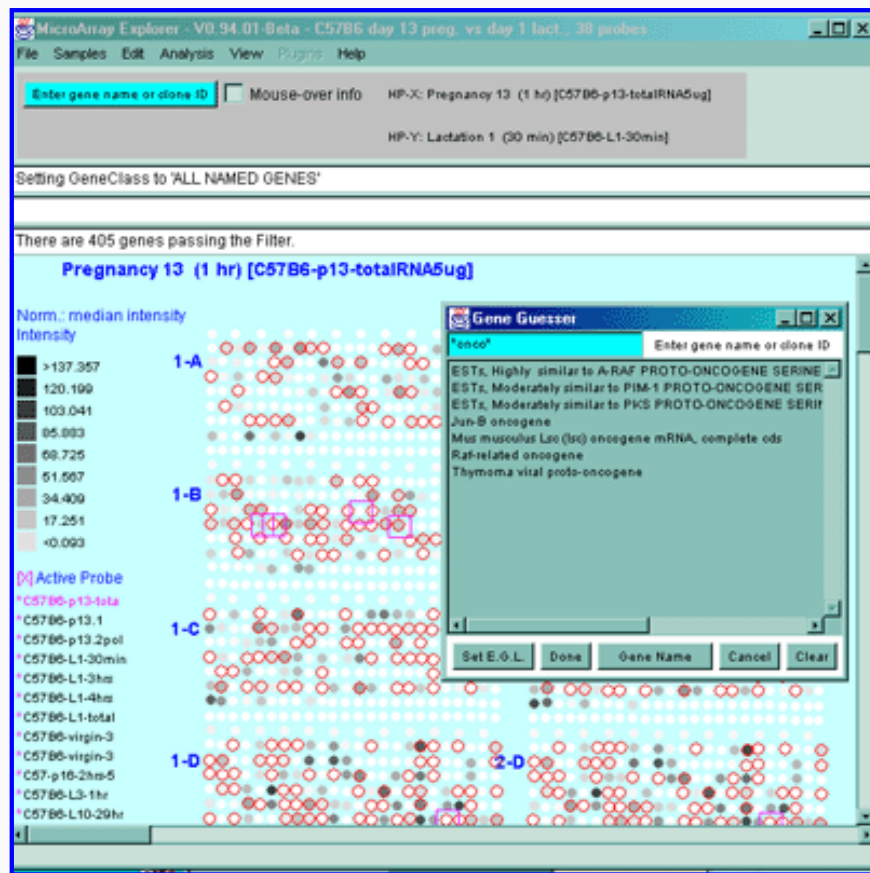
Figure 2.3.1 Edited Gene List defined from the Gene Name Guesser using wildcards. The Edited Gene List was defined as the set of genes containing the sub-string "onco" in it. The sub-string was specified to the popup guesser window as "*onco*" using '*' characters as wildcard symbols indicating that it should match any or no characters. The button **Gene Name** may be toggled through a set of other identifiers including Clone ID, UniGene ID, dbEST 3', dbEST 5', GenBank 3', and GenBank 5', LocusID, etc. depending on what identifiers are available in your database. The user then pressed the **Set E.G.L.** button on the guesser window that sets the E.G.L. to those genes. If you have enabled the View menu "Show 'edited gene list', then the genes in the EGL. are viewed as magenta squares seen in the pseudoarray image. You many to do additional editing to manually add or remove genes that you want to change in the set. If a 2D scatter plot was being used, EGL labeled genes would appear there as well. To select a particular gene as the current gene, click on the gene you want in the list, then press the **Done** button.

2.3.2 Sets of genes menu

These commands let you do comparisons of sets of genes generated under different criteria. In addition, you may compute derived gene sets from existing gene sets using set operations (**OR, AND, DIFFERENCE**). You may also normalize the data by a gene subset. The user may save the genes defined by: 1) by the Filter, or 2) the manually defined **'Edited Gene List'**. The gene set resulting from a binary gene set operation OR (union), AND (intersection), or DIFFERENCE are saved in a new named gene set. The set difference (A-B) is defined as the gets in set A that are not in set B. Genes in set B that are not in set A are ignored. The 'User Filter Gene Set' may be set to any gene set and may then be used as part of the gene Filter cascade. The 'User Normalization Gene Set' may be set to any gene set and may then be used to normalize gene intensity values across hybridized samples. (See normalization algorithm for more information on this method.)

If you are running MAExplorer in stand-alone mode, the current named gene sets are saved when you save the DB using the **Save disk DB** or **Save as disk DB** selections in the Databases submenu of the File menu. The gene sets are saved in a *State* sub directory as ".cbs" files and are used to restore the gene sets when restarting MAExplorer on a .mae startup file. The .mae startup file saves the names of the .cbs files that are shared among the various startup files for a given project. The implication then is that if you change and save a gene set in one startup database, it will change in other startup databases when they load that gene set. The advantage is that different startup databases may view a gene set produced by another database. The *Sets of genes* operations in the Edit menu include:

- **List saved gene sets** - popup a windows with the catalog of named gene subsets showing how many genes are in each subset. If the contents of any gene set is changed or a set is added or removed, the list is dynamically updated.
- **Save Filtered genes as gene set** - assign the Filtered genes to a named gene subset
- **Save 'Edited Gene List' as gene set** - assign the 'Edited Gene List' to a named gene subset
- **Assign 'User Filter Gene Set'** - for use as a Filter option
- **Assign 'User Normalization Gene Set'** - for use as a Normalization option
- **OR (Union) of 2 gene sets** - set the 'Edited Gene List' to the union of two named gene subsets (i.e. genes that are found in either set).
- **AND (Intersection) of 2 gene sets** - set the 'Edited Gene List' to the intersection of two named gene subsets (i.e. genes that belong to both sets).
- **Difference of 2 gene sets** - set the 'Edited Gene List' to the difference of two named gene subsets (i.e. the first set less genes in the second set found in the first set).
- **Rename gene set** - rename a saved gene set
- **Load gene set from disk file** - load a specific gene set from a user specified disk file (stand-alone mode only)
- **Remove gene set** - remove a saved gene set

The following is an example of **List saved gene sets** state listing the catalog of named gene subsets in some of the MGAP data. Note that sets #1 to #11 are fixed by the data in the GIPO file and may not be changed by the user. Sets #12 to #14 are assignable from other sets or in the case of the E.G.L, by various MAExplorer operations. Sets #1 through #14 may not be removed whereas #15 and higher may be removed.

```
User Gene Sets
Set# |#genes| title
======================
 #1 |1727| ALL GENES
 #2 |394|  ALL NAMED GENES
 #3 |246|  ESTs similar to genes
 #4 |456|  ESTs
 #5 |1096| All genes and ESTs
 #6 |1681| Good genes
 #7 |40|   Replicate genes
 #8 |0|    HousekeepingGenes
 #9 |96|   Calibration DNA
 #10 |77|  Your plates
 #11 |46|  Empty wells
--------- User Assignable ----------
 #12 |0|   User Filter Gene Set
 #13 |60|  Edited Gene List
 #14 |0|   Normalization Gene Set
--------- User definable-----------
 #15 |60|  The 60 genes closest to Carbonic Anhydrase-III
 #16 |30|  Named genes in the 60 genes closest to CA-III
 #17 |4|   Replicate genes in the 60 genes closes to CA-III
```

The following figure illustrates selecting sets by name for gene set operations.
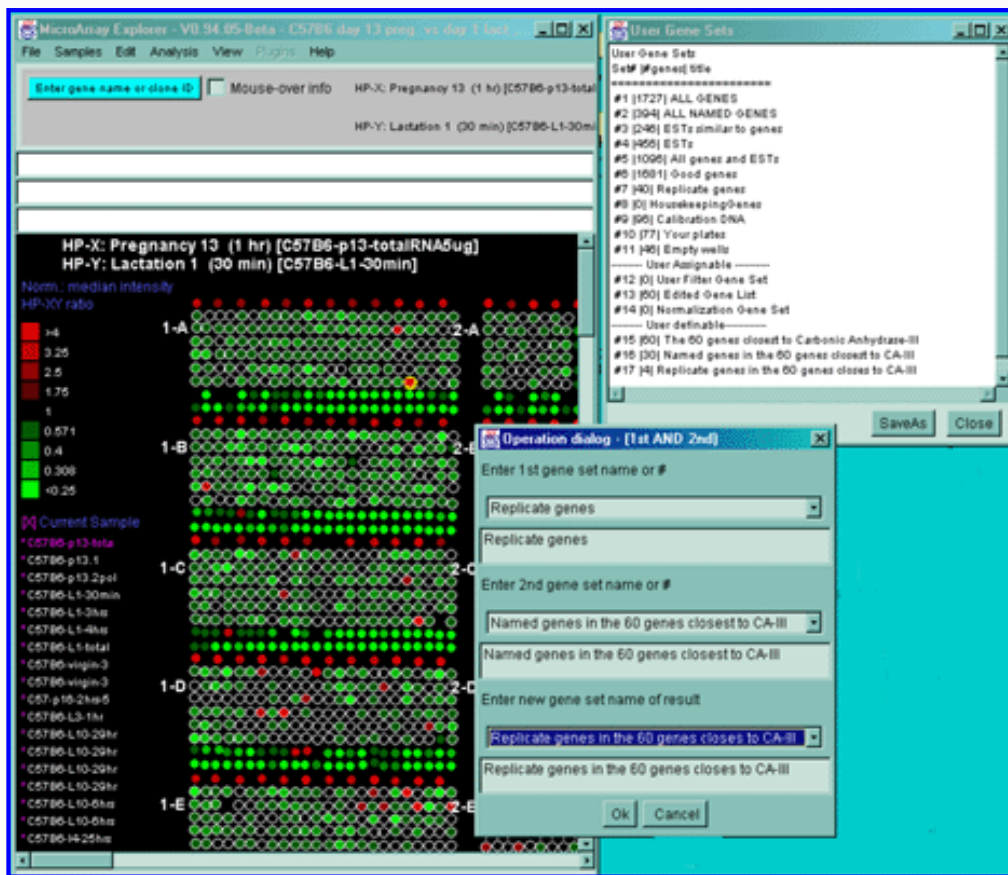
Figure 2.3.2 Selection of gene sets for binary gene set operations. This example computes the Boolean AND of two sets "ALL NAMED GENES" and "60 genes closest to CA-III from Named and Ests", and then the AND of the "Replicates" with the previous result. The first result is save in the set called "The 60 genes closest to Carbonic Anhydrase-III". The second result is saved in the called set "Named genes in the 60 genes closest to CA-III". Finally, the third result is saved in the set named "Replicate genes in the 60 genes closes to CA-III".

2.3.3 Sets of sample conditions menu

In addition, MAExplorer can operate on sets of hybridized samples. For example, a sample set might be replicate hybridized samples from the same biological experiment sample, or it could be repeated experiments of different but the same types of samples. (One must be careful in mixing data between the two cases because of the different expected sources of variance). This means you can treat multiple replicate samples as a distribution and compare the mean values for each gene in one set of samples with the mean values for another set of samples. We call these sets of hybridized samples *conditions lists* or *HP lists*. You may then put one or more HP samples into a condition set. These sets in turn can be used for computing statistics on clonal differences between different condition sets. Note each condition set may have multiple (i.e. different) samples. These condition sets are saved with the user state when doing a (File | Databases | SaveAs DB). As with sets of genes, there are a number of operations to manipulate HP condition set in the **Sets of Conditions** menu that includes:

- **List saved HP condition lists** - list the saved HP condition lists.
- **List contents of saved HP condition list** - for a particular condition.
- **Save HP-X as condition list** - save current HP-X 'set' to a named HP condition list.
- **Save HP-Y as condition list** - save current HP-Y 'set' to a named HP condition list.
- **Save HP-E as condition list** - save current HP-E 'list' to a named HP condition list.
- **Assign saved condition list to HP-X** - set the current HP-X 'list' to the saved condition list.
- **Assign saved condition list to HP-Y** - set the current HP-Y 'list' to the saved condition list.
- **Assign saved condition list to HP-E** - set the current HP-E 'list to the saved condition list.
- **OR (Union) of 2 condition lists** - make a new condition that is the union of two named condition lists (i.e. conditions that are found in either list).
- **AND (Intersection) of 2 condition lists** - make a new condition that is the intersection of two named condition lists (i.e. conditions that are found in both lists).

- **Difference of 2 condition lists** - make a new condition that is the difference of two named condition lists (i.e. the first list less conditions in the second list).
- **Rename HP list** - rename a saved HP condition list
- **Load HP condition list from disk file** - [Future]
- **Remove HP list** - remove a saved HP condition list

The following is an example of **List saved HP condition lists** state listing the catalog of named HP condition lists.

```
Condition Lists
===============
Condition[1] #HPs 2, [Initial HP-X: C57B6 pregnancy day 13]
Condition[2] #HPs 2, [Initial HP-Y: Stat5a (-,-) pregnancy day 13]
Condition[3] #HPs 4, [Initial HP-E expression list]
```

The following is an example of **List contents of saved HP condition list** state.

```
Condition List #1 [Initial HP-X: C57B6 pregnancy day 13]
=================================
HP[1] Pregnancy 13  (1 hr) [C57B6-p13-totalRNA5ug]
HP[2] Pregnancy 13  (1 hr) [C57B6-p13.2poly-A]
```

2.3.4 Setting user preferences menu

The **Preferences** submenu is used to set various data labels, statistical limits and other parameters. These include:

- ☐ **#Use Web DB** - if Web DB was defined, get data from the Web
- **#Define Web DB** - (re)define Web DB URL name for access when restart database
- ☑ **#Web DB data caching** - if Web DB was defined, cache data on local computer if getting data from the Web
- **Define HP-X class name** - for set of HP-X samples
- **Define HP-Y class name** - for set of HP-Y samples
- **Define DB name** - (re)define the local database name
- **Define DB title** - (re)define the local database title
- ----------------------
- **Adjust all Filter threshold scrollers** - popup the state scroller window with all of the thresholds. This is useful when you want to adjust thresholds **before** you enable data Filtering or clustering. If you are logging messages, then closing the window will print the current values of all of the threshold sliders in the log.
- **Set max # genes in highest/lowest report** - sets the number of genes N to report or Filter.
- ----------------------
- **Font Family** - set the font family (to improve readability)
- **Font Size** - set the font size (to improve readability)
- ----------------------
- ☐ **Cluster on Filtered genes, else all genes** - genes to use when clustering from current gene [Future]

The **Font Family** submenu is used to set the text font family. This may be useful if your computer is missing some fonts or some fonts are easier to read than others. Note: some fonts may not work well on your computer. If this is the case, try another font. When you save the data mining session with the "SaveAs file DB", it also saves the font you have set. For some plots or popup text-windows, you may have to regenerate the popup window to see the font changes.

- **Ariel**
- **Courier**
- **Helvetica**
- **MonoSpaced**
- **SanSerif** - the default font
- **TimesRoman**

Figure 2.3.4.1 Popup window allowing you to adjust all threshold slider values">. The **Adjust all Filter threshold scrollers** command allows you to pre-adjust all threshold slider values used in data filtering and in clustering. It may be easier to set the approximate range before invoking the clustering operation because changing a parameter will recluster your data.
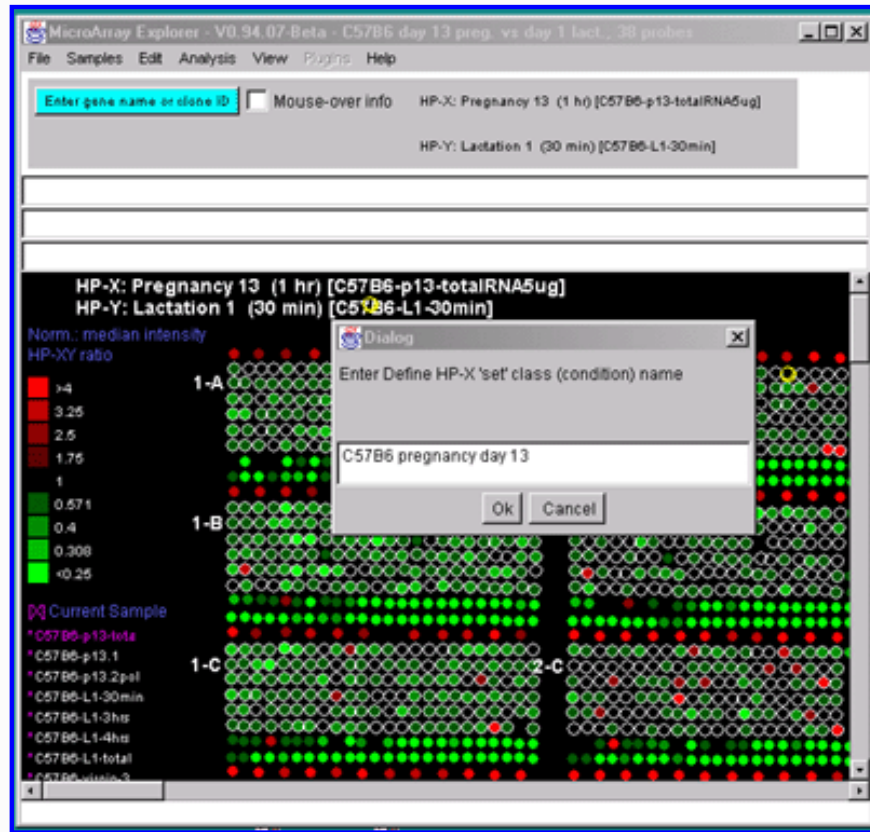
The **Define HP-X (HP-Y) class name** command may be used to change the names of the HP-X (HP-Y) experimental condition sets. These names are used in various labels in the main window, popup plots and reports, etc. The commands to change various names of database components are in the Preferences submenu in the Edit menu.

2.4 Analysis menu

The Analysis menu (see Figure 1.4) contains an ordered list of six primary menus that may be used, in that order, to perform an initial analysis. In more complex analyses, the sequence of operations will vary and include commands selected from other menus or will use these menus in different order. The Analysis submenus are as follows:
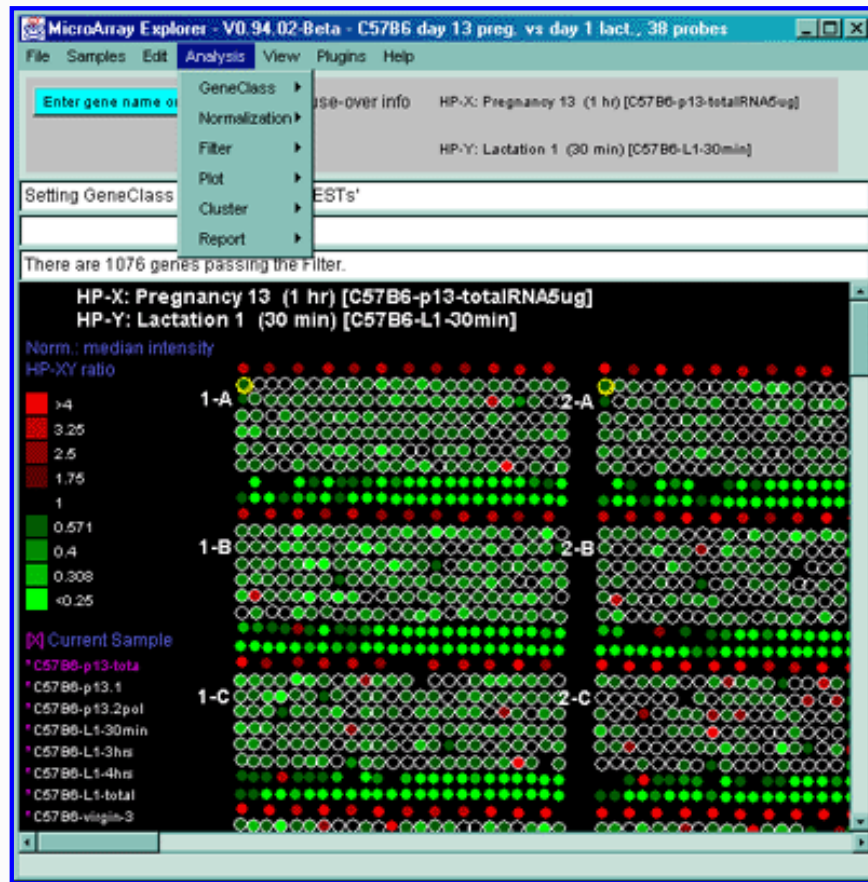
Figure 2.4 MAExplorer main window with Analysis Menu. The menu structure of MAExplorer was designed to allow users to quickly perform commonly used data-mining operations as a first approximation analysis.

2.4.1 GeneClass menu

A gene class (e.g. all named genes, ESTs, oncogenes, etc.) is a set of genes that belongs to the class of genes in the universe of genes in the particular microarray database. MAExplorer may restrict the set of genes by "Gene Class" membership (currently includes All Genes, All named Genes, ESTs similar to genes, Unknown ESTs, All genes and ESTs, Good genes, Replicate genes (i.e. with more than one copy of the gene in the array), Calibration DNA, genes from user's plates). The additional Gene Class list of names depends on its availability in a specific database.

Figure 2.4.1.1 Gene Class menu. The user may select a subset of genes that belong to one of the classes of genes. This shows the user selecting the set of "All named genes" that are indicated with red (white) circle over the spots in the array intensity (ratio) pseudoarray image.

Figure 2.4.1.2 Example of all replicated genes occurring more than once in the array. This was selected by using the GeneClass 'Replicate genes'. You may use the data Filter "Filter by genes with replicates" instead of the GeneClass. This has the advantage that you may use other GeneClasses (e.g. ESTs, or All named genes, etc.). Alternatively, you can find all of the replicates for a particular gene by 1) use the Gene Guesser to find the particular gene you want; 2) press "Set E.G.L." to save it as an Edited Gene List; 3) enable the Filter "Filter by E.G.L." at the same. This will show all occurrences of that gene.

The set of all genes constitutes a number of different gene classes. It is possible to restrict the subsequent analysis to a particular subset of these genes called a gene class. The **GeneClass** menu operations include operations to select the current set of genes to analyze from the set of all genes by their membership in a gene class.

- **All genes** - select the gene class of *all genes*
- **All named genes** - select the set of *all named genes* where the gene name is known
- **ESTs similar to genes** - select genes that are *ESTs similar to named genes*
- **ESTs** - select genes that are *unknown ESTs*
- **All genes and ESTs** - select genes that are either *named genes*, *ESTs* or *ESTs similar to named genes*
- **Good genes** - select the predetermined set of good genes marked in the GIPO file. If none are indicated, then it defaults to **All genes**.
- **Replicate genes** - select the genes there are at least 2 copies of the gene <u>replicated</u> on the array. Note: <u>duplicated</u> genes (i.e. F1, F2, etc) are not considered replicates for this purpose.
- **Housekeeping genes** - select the predetermined set of housekeeping genes (if present) [Future]
- **Calibration DNA** - select the predetermined set of Calibration DNA spots for microarray normalization (if present)
- **Your plates** - select genes from the investigator's plates (if present - see Table 2.4.1 below). The 'Your plate' menu name is specified in the database configuration setup file (see Appendix C.5.1-C).
- **Empty wells** - select spots that are empty wells on the array (if present - see Table 2.4.1 below). The 'Empty well' menu name is specified in the database configuration setup file (see Appendix C.5.1-C).
- **Set Gene Class subset** ▶ - select one of the gene class subsets [Future]

Some of the above gene classes are deduced from the gene name supplied with the Gene In Plate Order (GIPO) file for the array. We use the following automatic classification rules shown in Table 2.4.1.

Table 2.4.1 Rules for the automatic classification of gene names into the default Gene Class sets. The gene name is analyzed alphabetic-case independently.

Gene class	Rule for class membership
All genes	all genes on the array
All named	genes not starting with "EST"
ESTs similar to genes	genes starting with "EST,"
ESTs	genes with the name "EST"
Replicate genes	genes with multiple copies
Calibration DNA	genes using the configuration file name "calibDNAname" (optional - see Appendix Table C.4.1)
Your plates	clones using the configuration file name "yourPlates" (optional - see Appendix Table C.5.1-C))
Empty Wells	empty wells where no spot exists on the array indicated by keywords "empty", "empty well" or "EmptyWell" (optional - see Appendix Table C.5.1-C))
Good Genes	spots on the array where the GIPO QualCheck data was used and was valid. If it was not used, then it assumes all spots are good. (optional - see Appendix Table C.4.1)

2.4.1.1 GeneClass ontology subsets [Future]

If the **Set Gene Class subset** were activated, it <u>might</u> include categories such as the following. If the categories exist and the data is made available to MAExplorer, then it is possible to specify gene subsets by Gene Class name. *It is the responsibility of the database creator to define a mapping table supporting these named subsets of named genes.*

- *Apoptosis*
- *Bcl-2 family*
- *Cdk inhibition*
- *Cell adhesion*
- *Cell cycle*
- *Cell surface*
- *Chemokines*
- *Cyclins*
- *Cytokines*
- *Cytoskeleton*
- *DNA binding*
- *DNA recomb*
- *DNA repair*
- *DNA synthesis*
- *G-proteins*
- *Growth factors*
- *Heat shock*
- *Interferons*
- *Interleukins*
- *Ion channel*
- *Milk Genes*
- *Motility*
- *Oncogenes*
- *Proteases*
- *Protein turnover*
- *Receptors*
- *Receptors*
- *Signal transduction*
- *Stress response*
- *Transcription*
- *Transport*
- *Tumor suppressors*

2.4.1.2 Simulating Gene Class ontologies using Gene Set operations

You can effectively implement finding ontology subsets for Gene Class subsets using the following procedure. The trick is to repeatedly define an E.G.L. gene subset using the gene name guesser to find the genes of interest and save it as a named gene subset. Edit out genes you don't want. Then you would repeatedly do the OR of gene sets of interest, saving the result as a new named set. Then doing the OR of another gene set with the set you just created, etc.

Procedure

1. Use the Gene Name Guesser (see Figure 2.3.1) to collect genes belong to a particular ontology. Pressing "Set E.G.L" puts the genes from the guesser into the EGL set. You can then delete any genes that don't belong by clicking on the gene in the pseudoarray image with the SHIFT key pressed.
2. Save the EGL as a name gene set with a meaningful name representing the ontology construct using the (Edit | Save 'Edited Gene List' as named gene set) (see Section 2.3.2 for gene set editing commands).
3. Repeat steps [1] and [2] to gather additional sets.
4. Merge appropriate gene sets to get the complete ontology constructs using the (Edit | OR (Union) of 2 gene sets) as required.
5. Finally, set the (Edit | Assign 'User Filter Gene Set') to select a particular gene class to use, and then
6. Enable the data Filter by setting (Filter | Filter by 'User Filter Gene Set' membership) (Section 2.4.3).

2.4.2 Normalization menu

The **Normalization** menu operations include operations to normalize gene intensity data between hybridized samples. This is critical in being able to compare samples because of differences in amount of sample, labeling efficiency and variations in scanner operation including gain and baseline settings. There are several methods available including normalizing by Zscore, median, log mean, Zscore of logs, calibration DNA, housekeeping genes, etc. The specific microarray image quantification is determined the image analysis program being used to pre-process the arrays.

Note: although this set of normalization methods is limited, it is adequate for some analyses of the data. We are in the process of adding more normalization methods through MAEPlugin methods.

- **Zscore of intensity** - normalize by the (intensity-mean)/stdDev of raw intensities for all spots in each sample.
- **Median intensity** - normalize by the median of the raw intensities for all spots in each sample (the default normalization).
- **Log median intensity** - normalize by the log of median scaled raw intensities for all spots in each sample.
- **Zscore log intensity, stdDev** - normalize by the Zscore of the log intensity using $(\log(\text{intensity})-\text{mean}_{\log})/\text{stdDev}_{\log}$, standard deviation for all spots in each sample.
- **Zscore log intensity, mnAbsDev** - normalize by the Zscore of the log intensity using $(\log(\text{intensity})-\text{mean}_{\log})/\text{meanAbsDev}_{\log}$, mean absolute deviation for all spots in each sample.
- **By Calibration DNA set of genes** - normalize by the sum of the 'Calibration DNA' genes for each sample (if it exists in your database).
- **By 'User Normalization Gene Set'** - normalize by the sum of the genes in a user defined gene set in each sample. You assign this gene set using the (Edit menu | Gene sets | Assign 'User Normalization Gene Set') operation.
- **By housekeeping gene set** - normalize each HP data set by the sum of the intensity values for known housekeeping genes in each sample (if it exists in your database).
- **Scale intensity data to 65K** - scale the data for the microarray by 65535/maxIntensity for each sample.

- ☐ **Unnormalized** - do *not* scale data between samples. I.e. use the raw data.
- ----------------------
- ☐ **Use background intensity correction** - enable/disable background correction to gene intensity measurements.
- ☐ **Use ratio median intensity correction** - enable/disable ratio median correction to clone intensity measurements by multiplying the ratio (Cy3/Cy5) by medianCy5/medianCy3 intensies. If background correction is enabled, correct by (medianCy5-medianBkgdCy5)/(medianCy3-medianBkgdCy3).

2.4.2.1 Intensity background correction

The background intensity data from the spot quantification programs may be used to correct spot intensity. Background may be specified as either a global value or on a per-spot basis. If the array images have low background, then this may not be too much of a problem if no background values are available.

Some software quantification software (e.g. Research Genetics' Pathways 2.01) measures background globally as: BGLow (low background), BGAvg (Average background), BGRms (root mean square background). For MGAP, MAExplorer uses the BGLow value when you request background subtraction. These values are read from the MAExplorer Samples DB file (see Appendix Table C.2.1.1 For other quantification programs, background may be available on a per-spot basis in the quantification files. It the latter is available in your data, it will be used if background correction is enabled (see Appendix C.3).

The background corrected intensity I'_{ij} is computed from the raw intensity I_{ij} and background intensity $bkgrd_{HPi}$ for H.P. i and spot j as follows:

$$I'_{ij} = I_j - bkgrd_{HPi}$$

Ratio computation for Cy3 and Cy5 data

For most MAExplorer operations, the intensity of a gene is generally computed as the mean intensity of the spots (background corrected or not) which duplicate that gene on the microarray. When working with dual hybridized samples using Cye-3 and Cye-5-dUTP labeling that results in green and red fluorescence, this can be used in self-normalizing intensity for each hybridized clone array using the Cy3/Cy5 ratio. If local background is available, then the ratio can be computed for HP h and spot j as

$$(Cy3_{hj} - BkgrdCy3_{hj}) / (Cy5_{hj} - BkgrdCy5_{hj})$$

2.4.2.2 Normalization between microarrays to allow comparison

The normalization of quantitative data is crucial when comparing data between different microarray samples. There are a number of different schemes possible. One is to normalize by the sum of known calibration, housekeeping genes or other "constant expression" genes in the microarray. Another is to sum the background corrected integrated density for all spots in an array and to normalize individual gene measurements by that sum. These methods are now described in more detail. As the MAEPlugins facility becomes available, we will be adding a number of more sophisticated gene-specific normalization methods that take many of the problems specific to microarrays into account.

Normalizing by scaled Zscore of intensity

The "normalized Zscore of intensity" method normalizes each hybridized sample by the mean and standard deviation of the raw intensities for all of the spots in that sample. The mean intensity mnI_i and the standard deviation sdI_i are computed for the raw intensity of 'Good genes'. It is useful for standardizing the mean (to 0.0) and the range of data between hybridized samples to about -4.0 to +4.0. When using the Zscore, you compute Zdiff(erences) not ratios. The Zscore intensity $Zscore_{ij}$ for intensity I_{ij} for HP i and spot j is computed as

```
      Zscore_ij = (I_ij - mnI_i)/sdI_i,
```
and
```
      Zdiff_j(x,y) = Zscore_xj - Zscore_yj.
```

Normalizing by the median of intensity

The "Median intensity" method normalizes each hybridized sample by the median of the raw intensities of 'Good genes' for all of the spots in that sample. It is a useful normalization to use when you want to compute X/Y ratios between hybridized samples.

```
      Im_ij = (I_ij/ medianI_i)
```

Normalizing by the log of median of intensity

The "Log median intensity" method normalizes each hybridized sample by the log of median scaled raw intensities of 'Good genes' for all of the spots in that sample. The value 1.0 is added to the intensity value to avoid taking the log(0.0) when intensity has zero value. This is a useful normalization to use when you want to compute X/Y ratios between hybridized samples and compress the scale. Because we are computing a log, we report the difference between HP-X and HP-Y as (X-Y) instead of a ratio (X/Y).

```
      Im_ij = log(1.0 + (I_ij/ medianI_i))
```

Normalizing by scaled Zscore of log intensity, standard deviation

The "Normalize by Zscore of log intensity, stdDev" method normalizes each hybridized sample by the mean and standard deviation of the logs of the raw intensities for all of the spots in that sample. The mean log intensity $mnLI_i$ and the standard deviation log intensity $sdLI_i$ are computed for the log of raw intensity of 'Good genes'. Then the Zscore intensity $ZlogS_{ij}$ for HP i and spot j is

```
      ZlogS_ij = (log(I_ij) - mnLI_i)/sdLI_i
```

Normalizing by scaled Zscore mean absolute deviation of log intensity

The "Normalize by Zscore of log intensity, mean absolute deviation" method normalizes each hybridized sample by the mean and mean absolute deviation of the logs of the raw intensities for all of the spots in that sample. The mean log intensity $mnLI_i$ and the mean absolute deviation log intensity $madLI_i$ are computed for the log of raw intensity of 'Good genes'. Then the Zscore intensity $ZlogA_{ij}$ for HP i and spot j is

```
      ZlogA_ij = (log(I_ij) - mnLI_i)/madLI_i
```

By 'User Normalization Gene Set'

This method is useful a subset of genes have been determined to have relatively constant expression across the set of samples. It normalizes by the sum of intensities for a subset of genes defined by the user in the 'User Normalization Gene Set' (Section 2.3.2)using the gene set editing commands. Normalizing by the sum of genes uses the Igs_{ij} that is computed for microarray HPi with intensities I_{ij} for all genes j in the gene subset.

```
      Igs_i = Sum (I_ij)

                genes j
                i in HPi
```

Then, the normalized intensity I'_{ij} is computed as:

$$I'_{ij} = I_{ij}/Igs_i$$

By 'Calibration DNA' set

If a predefined set of calibration DNA genes are available on the array, they may be used to normalize density values between the samples. The calibration DNA genes are defined by special gene names that are declared in the Configuration file using the 'calibDNAname' parameter (see Appendix C Table C.5.1(C)). If there is no calibration DNA, this entry is not used. The algorithm is the same as "User Normalization Gene Set" (above), but the set is predefined as the genes flagged as calibration DNA. For example, in the MGAP database, these spots are the "mouse genomic DNA" spots so the Configuration file entry would be calibDNAname="m.g. DNA".

Scaling intensity data to 65K

Another method "Scale intensity data to 65K" scales the maximum intensity of each sample to 65K (the maximum intensity). Since the raw scanned data is often 16-bits, it can have a maximum value of 65535 (2^{16}-1) and so this does minimum scaling. This method may make it easier to view the data initially using the pseudoarray image. However, it may not properly scale the data between arrays and should probably not be used in quantitative comparisons.

No normalization

You may also want to look at the raw intensity (or Cy3 and Cy5 channel) data. Turning off normalization gives you the raw data read into MAExplorer.

2.4.2.3 Using different normalizations to 'see' different data views

Changing the normalization method will sometimes make differences between data sets more apparent. The following figure shows the same data in two different scatter plots but with two different normalizations.
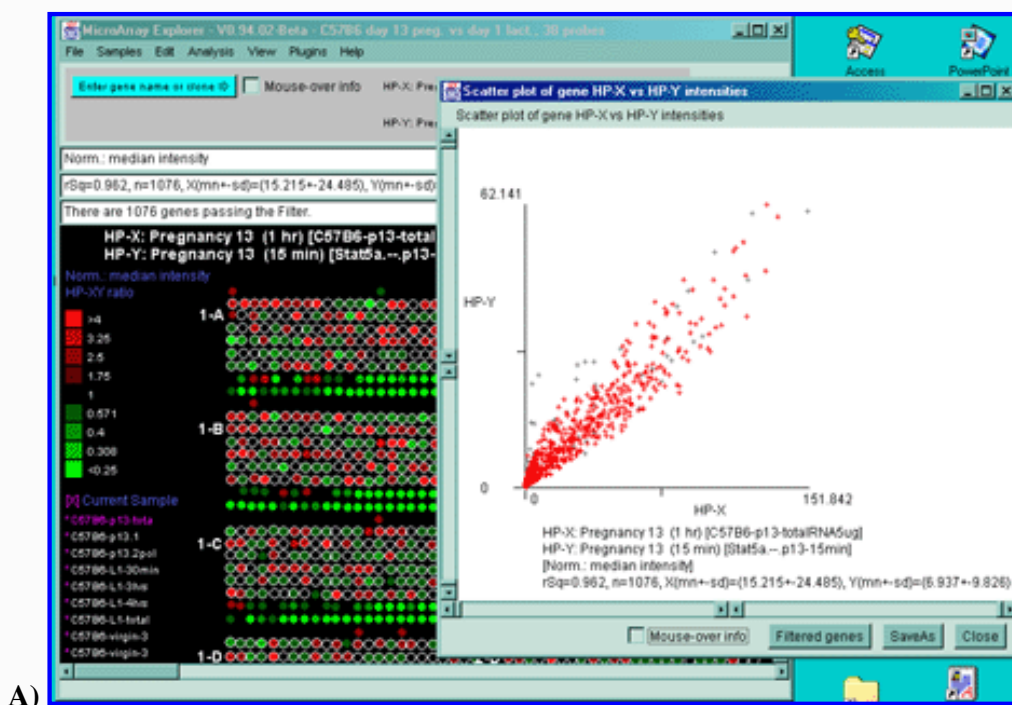
A)
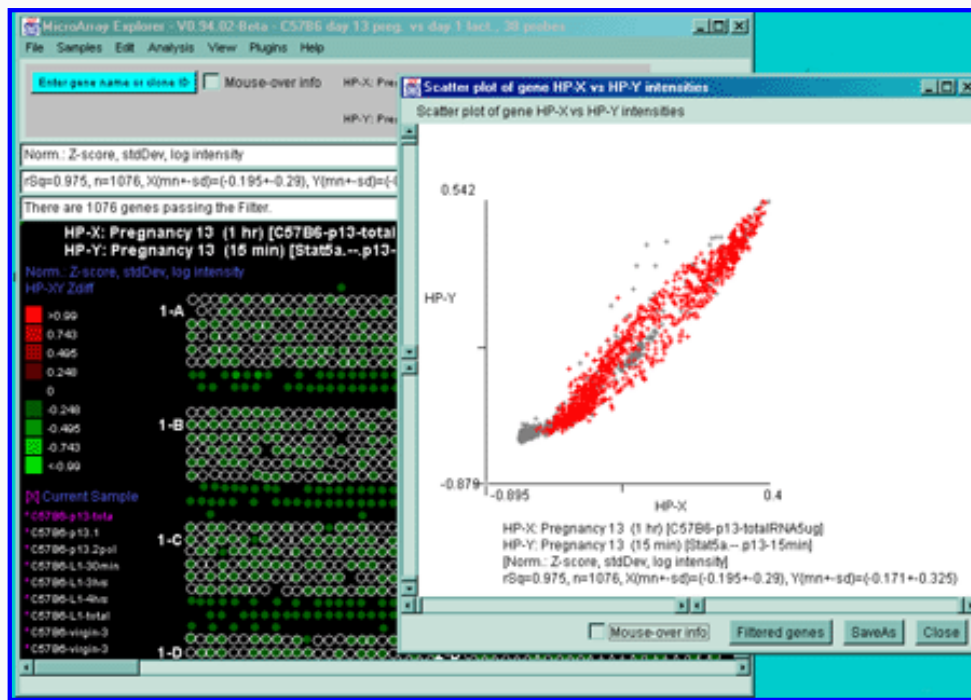
B)

Figure 2.4.2.3 Scatter plot of HP-X and HP-Y 'sets' data. HP-X is C57B6 pregnancy day 13 and HP-Y is Stat5a (-,-) pregnancy day 13 filtered by "All named genes and ESTs". **A)** A scatter plot using the Median normalization. **B)** A scatter plot using the Zscore of the logs normalization. Notice how the Casein alpha outlier is more apparent in the case of the Zscore log normalization. The skewed plot is characteristic of much microarray data. Some normalization methods (not currently included in MAExplorer) can compensate for these some of these artifacts (Dutoit, 2000) and are planned for future MAEPlugins.

2.4.3 Filter menu

The final set of genes presented for display, plotting, reports, etc. is determined by a cascade of gene "data filters" that generate a restricted gene set. The cascade is computed in real-time using the intersection of individual criteria and tests selected by the user. Examples of Filter criteria include: membership in a particular gene set, ratio (HP-X/HP-Y) within a range, passing statistical t-tests, etc.
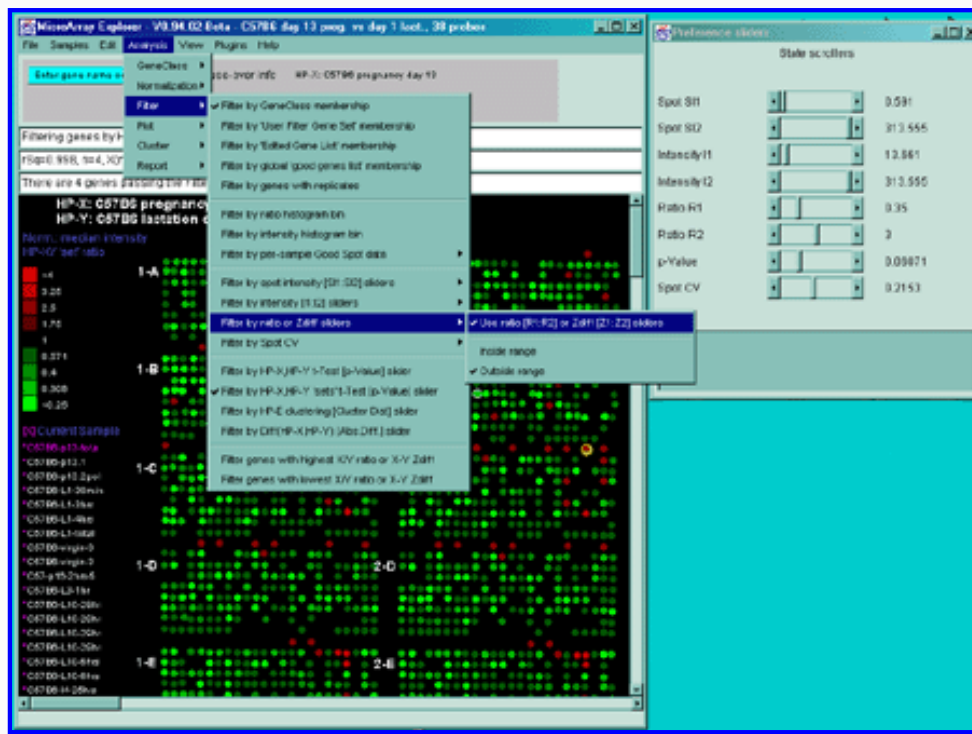
Figure 2.4.3 Filter menu. The Filter menu is a cascade of data filters that restrict the set of genes passing all filters that have been enabled and whatever the criteria was that was set for those filters. This figure shows the GeneClass filter set to "All genes and ESTs", the spot CV filter and Ratio (X/Y) range filters being set interactively by the scroll bars on the right. The genes that pass the filter are indicated with a red (white) circle in the array intensity (ratio) pseudoarray image.

The **Filter menu** options are used to restrict the set of genes by pre-filtering the data with a series of cascaded filter criteria and tests. The resulting subset of genes passing the filter are then used in the plots, reports and other data analysis methods. Some of the filters require additional parameters that are set by the *State scrollers*. The user will automatically be prompted for changes to these scollers (a threshold scrollers window will pop up) when the filter is activated or change. These values may also be set from the **Adjust all Filter threshold scrollers** entry in the **Preferences** submenu in the **Edit** menu. The filters are broken up into subgroups in the following menu with the grouping haveing more to do with the criteria (i.e. gene set membership, data range, or statistical tests).

- ☑ **Filter by GeneClass membership** - only include genes that are members of the current GeneClass.
- ☐ **Filter by 'User Filter Gene Set' membership** - only include genes that are members of the current 'User Filter Gene Set'.
- ☐ **Filter by 'Edited Genes List' membership** - only include genes that are members of the 'Edited Gene List'.
- ☐ **Filter by global 'Good Genes List' membership** - only include genes that are members of the list of good genes. [These genes are indentified by a QualCheck entry in the GIPO database file.]
- ☐ **Filter by 'Genes with replicates'** - only include genes that genes that have at least 2 copies of the gene replicated on the array. Note: duplicated genes (i.e. F1, F2, etc) are not considered replicates for this purpose.
- -------------------
- ☐ **Filter by ratio or Zdiff histogram bin** - only include genes that are in the range of the ratio or Zdiff histogram bin you have clicked on (should be set from histogram plot, but may be turned off here)
- ☐ **Filter by intensity or (Cy3/Cy5) histogram bin** - only include genes that are in the range of the intensity histogram bin you have clicked on (should be set from histogram plot, but may be turned off here)
- **Filter by positive intensity data** ▷ - filter by positive intensity data if the data may contain negative numbers. Otherwise it will use both positive and negative data. If the database has 2 channels (F1, F2) or (Cy3,Cy5) each channel is checked. If the background correction is enabled, the background corrected values are tested to see if any of them are negative.
- **Filter by per-sample Good Spot data** ▷ - filter out genes that have that do not have "Good Spot" values (defined by the optional QualCheck spot data on a per-sample (i.e. HP) basis. See the list of codes in Appendix C.4). If there is no such spot

quality data, then all spots are considered "good".

- -------------------
- **Filter by spot intensity [SI1:SI2] sliders** ▷ - filter by individual spot intensity (Cy3 and Cy5 channels if ratio data) within [SI1:SI2] threshold ranges
- **Filter by [I1:I2] sliders** ▷ - filter by gene expression (or Cy3/Cy5 if ratio data) within [I1:I2] threshold ranges
- **Filter by ratio or Zdiff sliders** ▷ - filter by gene ratios or Zdiff values within [R1:R2] or [Z1:Z2] threshold ranges (depending on the normalization method)
- **Filter by Cy3/Cy5 HP-X ratio or Zdiff sliders** ▷ - filter by gene ratios or Zdiff values within [CR1:CR2] or [CZ1:CZ2] threshold ranges (depending on the normalization method). This is useful for filtering data from a single sample.
- **Filter by spot CV** ▷ - filter out genes that do not meet minimum Coefficient of Variation (CV) values of spot replicates (F1 and F2 for the same HP, replicates in HP-X and HP-Y 'sets' of samples etc.).
- -------------------
- ☐ **Filter by HP-X,HP-Y t-Test [p-value] slider** - only include genes that meet the HP-X,HP-Y t-Test criteria if they have (F1,F2) duplicate spot (this is a weak form of the t-Test).
- ☐ **Filter by HP-X,HP-Y 'sets' t-Test [p-value] slider** - only include genes that meet the HP-X,HP-Y 'sets' t-Test criteria (only works if using HP-X and HP-Y 'sets' mode where there are replicate samples).
- ☐ **Filter by HP-E clustering [Cluster dist] slider** - only include genes that meet the clustering criteria (alternatively, see the Cluster menu commands).
- ☐ **Filter by Diff(HP-X,HP-Y) [Abs.Diff.] slider** - only include genes whose absolute difference between mean HP-X and HP-Y (single or 'sets') is < threshold.
- -------------------
- ☐ **Filter N genes with highest X/Y ratio or X-Y Zdiff** - look at highest ratios or Zdiff values. The value of N is set in the Edit menu preferences.
- ☐ **Filter N genes with lowest X/Y ratio or X-Y Zdiff** - look at lowest ratios. The value of N is set in preferences. N is set in the Edit menu preferences.

The **Filter by positive intensity data** submenu filter contains options that specify which spot intensity values are to be considered when excluding negative quantified spot data. Note: this filter only makes sense if your data might have negative values (e.g. Affymetrix chip "Avg Diff" data) or a background corrected value that is less than 0.0. The filter is enabled by setting the "Filter by spots with positive intensity" checkbox. Negative intensity values may occur with some types of arrays quantification programs. In the "Check spots for positive values mode" submenu, you may set the samples where the test may be applied to spots from the current HP, the single (HP-X,HP-Y) samples, (HP-X,HP-Y) 'sets' (replicated spots), or samples in the HP-E list selected to be used in the filter. If there are (F1,F2) or (Cy3/Cy5) data, then each spot must meet the threshold criteria.

- ⦿ **Current HP** - spots in current sample spots
- ☐ **HP-X & HP-Y** - spots in X and Y single samples
- ☐ **HP-X or HP-Y 'sets'** - spots in the HP-X set or HP-Y set
- ☐ **HP-X & HP-Y 'sets'** - spots in both the HP-X set and HP-Y set
- ☐ **HP-E** - spots in HPs in expression profile list

The **Filter by Good Spot data** submenu filter contains options that specify spots based on their quality. It filters out genes that have that do not have "Good Spot" values defined by the optional QualCheck spot data. (See the list of codes in [Appendix C.4](#)). If there is no such spot quality data, then all spots are considered "good". The filter is enabled by setting the "Filter by spots with Good Spot values" checkbox. All spots for the specified samples must meet the criteria. In the "Check spots for Good Spot mode" submenu, you may set the samples where the test may be applied to spots from the current HP, the single (HP-X,HP-Y) samples, (HP-X,HP-Y) 'sets' (replicated spots), or samples in the HP-E list selected to be used in the filter.

- ⦿ **Current HP** - spots in current sample spots
- ☐ **HP-X or HP-Y** - spots in X and Y single samples
- ☐ **HP-X 'set'** - spots in HP-X set
- ☐ **HP-Y 'set'** - spots in HP-Y set
- ☐ **HP-X or HP-Y 'sets'** - spots in HP-X set and HP-Y set
- ☐ **HP-E** - spots in HPs in expression profile list

The **Filter by spot intensity [SI1:SI2] sliders** submenu contains options that determines how individual spot intensity thresholding is to be applied in the Filter.

- ☐ **Use spot intensity [SI1:SI2] sliders** - use spot intensity thresholding
- ◉ **Inside** - test inside of [SI1:SI2] range
- ○ **Outside** - test outside of [SI1:SI2] range
- ------------------
- **Use data mode** ▶ - specify which samples are tested
- **Compare channels meeting range** ▶ - specify which additional constraints are used. This is useful for finding genes with high or low expression but that has some samples that have opposite expression.

The **Use data mode** submenu filter contains options that specify which spot intensity values are to be considered of the single sample (F1 and F2 replicated spot intensity data, or Cy3/Cy5 for ratio data), or the (HP-X,HP-Y) 'sets' of replicated samples is to be used in the filter. If there are single sample (F1,F2) or (Cy3/Cy5) data, then each spot must meet the threshold criteria.

- ◉ **Current HP** - spots in current sample spots
- ☐ **HP-X & HP-Y** - spots in X and Y single samples
- ☐ **HP-X & HP-Y 'sets'** - spots in HP-X set and HP-Y set
- ☐ **HP-E** - spots in HPs in expression profile list

The **Compare channels meeting range** submenu specifies which additional constraints are to be used. If required by the (**AT MOST channels**, **AT LEAST channels**, **PRODUCT OF channels**, **SUM OF channels**) commands, the Percent SI OK scroll bar will appear which covers the range of 0% to 100%.

- ◉ **ALL channels** - ALL channels must meet the range specification
- ☐ **ANY channels** - ANY channels may meet the range specification
- ☐ **AT MOST channels** - AT MOST Percent SI OK channels may meet the range specification
- ☐ **AT LEAST channels** - AT LEAST Percent SI OK channels may meet the range specification
- ☐ **PRODUCT of channels** - the PRODUCT of all channels must meet the range specification
- ☐ **SUM of channels** - the SUM of all channels must meet the range specification

The **Filter by [I1:I2] sliders** submenu contains options that determines how spot expression (intensity or (Cy3/Cy5) ratio value) thresholding is to be applied in the Filter:

- ☐ **Use intensity [I1:I2] sliders** - use spot intensity thresholds I1 (lower) and I2 (upper)
- ◉ **Inside** - test for intensity inside of [I1:I2]
- ○ **Outside** - test for intensity outside of [I1:I2]

The **Filter by ratio or Zdiff sliders** submenu contains options that determines how spot-ratio thresholding is to be applied in the Filter. The spot ratio is mean HP-X / mean HP-Y for sets of samples. The spot Zdiff is used if one of the Zscore normalization methods is active and is computed as (mean HP-X - mean HP-Y) for sets of samples.

- ☐ **Use ratio [R1:R2] or Zdiff [Z1:Z2] sliders** - use spot ratio [R1:R2] or Zdiff [Z1:Z2] range thresholds
- ○ **Inside** - test inside of [R1:R2] or [Z1:Z2] range
- ◉ **Outside** - test outside of [R1:R2] or [Z1:Z2] range

The **Filter by Cy3/Cy5 HP-X ratio or Zdiff sliders** submenu contains options that determines how spot Cy3/Cy5 HP-X ratio thresholding is to be applied in the Filter. The spot ratio is Cy3/Cy5 for normalized data unless one of the Zscore methods is used. In that case, the Zdiff is used and is computed as (Cy3 - Cy5) for sets of samples. If HP-X 'sets' is used, then it computes the mean Cy3 value and the mean Cy5 value and uses those values in the above computations.

- ☐ **Use ratio [R1:R2] or Zdiff [Z1:Z2] sliders** - use spot ratio [R1:R2] or Zdiff [Z1:Z2] range thresholds
- ○ **Inside** - test inside of [R1:R2] or [Z1:Z2] range
- ◉ **Outside** - test outside of [R1:R2] or [Z1:Z2] range

The **Filter by spot CV** submenu filter contains options that specify how the Coefficient Of Variation of the (F1,F2) or (HP-X,HP-Y) 'sets' (replicated spots) is to be used in the filter. The (F1,F2) CV is available only if there are duplicate spots on the HPs.

- ☐ **Use spot [CV] slider** - apply one of the spot CV filter modes as a Filter and popup a CV slider to set the threshold
- **CV spot filter mode** ▷ - select samples to be used in computing the CV
- ☐ **Use mean else max of CVs** - compute the CV as the maximum or the mean of the CVs of the samples selected

Filtering out genes with high replicate spot variation

The **Spot CV filter mode** submenu contains options to select how the spot CV filter is to be applied. It computes the maximum value of CV for all of the samples in the particular sample set specified. That maximum value is then used for the spot CV filter test. Genes may be filtered out having a large difference between spot quantification values of corresponding duplicate spots. You may compute the coefficient of variation CV_j for the two values ($f1_j$ and $f2_j$ for a particular gene j.

$$CV_j = 2|f1_j - f2_j|/(f1_j + f2_j)$$

If the database only has one field but replicate HPs, then you may use the **HP-X & HP-Y 'sets'** CV_j to filter the genes. Then CV_j values are tested against a CV threshold slider value to eliminate genes with a high coefficient of variation.

- ◉ **Current HP** - CV of (F1,F2) for each gene in current sample [if duplicate spots are available on each sample]
- ☐ **HP-X or HP-Y** - CV of (F1,F2) for HP-X and HP-Y single samples [if duplicate spots are available on each sample]
- ☐ **HP-X 'set'** - CV of spots in HP-X set
- ☐ **HP-Y 'set'** - CV of spots in HP-Y set
- ☐ **HP-X or HP-Y 'sets'** - CV of spots in the HP-X set or HP-Y set
- ☐ **HP-X and HP-Y 'sets'** - CV of spots in both the HP-X set and HP-Y set
- ☐ **HP-E** - CV of HPs in expression profile list

2.4.3.1 Data filtering using multiple gene data filters

Any or all of the data filters may be selected simultaneously. In particular, if you select filters that use parameter threshold scrollers, they will be added to a state scroller window (see Figure 2.3.4.1 for details to allow adjustment of ALL sliders simultaneously). You may change various thresholds and see the effect in real time. Note: some of the scrollers are more sensitive to low values. Therefore, we set them to respond non-linearly with a more precise vernier at the low end.
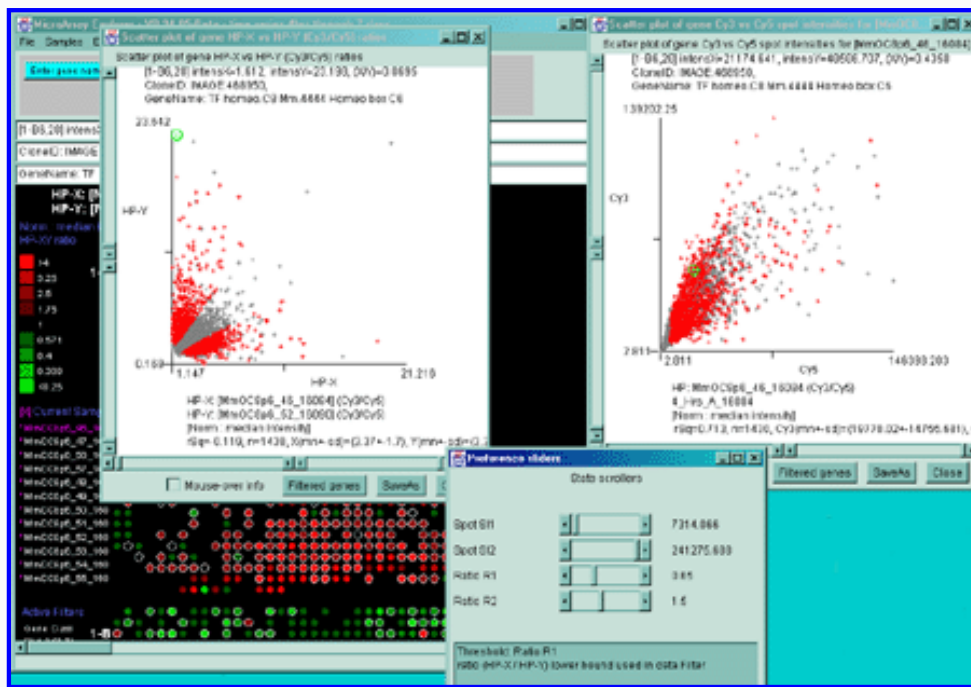
Figure 2.4.3.1 Filtering using multiple scrollers. This example is of Cy3/Cy5 time series data. It filters normalized spot intensity of the Cy3 and Cy5 channels independently ([SI1:SI2] inside range) where low intensity spots are eliminated. It then filters out genes outside of the [R1:R2] ratio range.

Filtering using positive
intensity data - ignoring
negative data

Figure 2.4.3.2 Using the Positive Intensity data Filter. This allows removing negative data if the data contains negative intensity values (e.g. Some Affymetrix data has negative Average Difference values which could be read as Intensity for MAExplorer).

2.4.4 Plot menu

The Plot menu lets you display a pseudoarray image, scatter plots, ratio and intensity histograms,and expression profile plots. The pseudoarray image is displayed in the main MAExplorer window. All of the other plots are displayed in popup windows. Depending on the particular plot, multiple instances may be allowed. The **Plot** submenus are:

- **Show Microarray** - display the pseudoarray image for the current HP sample
- **Scatter plots** - display various scatter plots of selected pairs of HP samples
- **Histograms** - display both ratio and intensity histograms of HP sample data
- **Expression profile plots** - display expression profile plots of genes or gene subsets

You may switch between different representations of the microarray spot pseudoarray image. It may be viewed as several different types of pseudo images including an intensity gray value and a pseudo-color Red/Black/Green image for ratio (HP-X/HP-Y) and Zscore (HP-X - HP-Y) data. The p-Value results of comparing a HP-X 'set' with a HP-Y 'set' of samples can be displayed as a color spectrum pseudoarray image. Depending on the origin of the array data, it may have the same verisimilitude as the original arrays.

Otherwise, it is displayed in a generic pseudoarray image containing grids that will fit the window - these are not the same as the original array image (see . However, the pseudoarrays are useful to getting a rough idea of the global changes in the data between arrays and how may genes pass the data filter.

When enabled using one of the commands in the Section 2.4.5 Clustering menu, cluster data appears as blue circles or squares drawn as overlays on the pseudoarray image. These options are discussed in the section on clustering. If you are doing clustering K-means clustering, the current cluster is displayed in the scatter plot if the latter is active.

Scatter plots, ratio and intensity histograms of the mean (HP-X/HP-Y) or (HP-X/HP-Y) 'set' data, or the F1/F2 or Cy3/Cy5 data. F1/F2 or Cy3/Cy5 plots are available if the data exists in your particular database. That might be the case with replicate spots or with Cy3/Cy5 data. If the normalization is set to a Zscore or log mean mode, it will compute Zscore scatter plots and histograms.

Clicking on spots in an array image or points in scatter plots sets the current gene and will bring up data on the gene or (optionally) access corresponding data from GenBank, UniGene, mAdb Clone, etc. databases in a popup Web browser. Clicking on a bin in a ratio or intensity histogram plot filters out all genes except for those in the range of that bin.

Expression profiles plots of selected genes or subsets of genes for all samples in the HP-E list. These are active plots with data reported when the user clicks in the plot.

Clicking on a spot (i.e. gene) in the microarray pseudo image or on a point (i.e. gene) in the scatter plot, it will define that gene as the "current gene" that is used in other operations. The current gene is indicated in both plots with a green circle around it. Similarly, you may modify the 'Edited Gene List' from either the pseudoarray image or the scatter plot. When viewing is enabled, it overlays those genes with magenta squares.
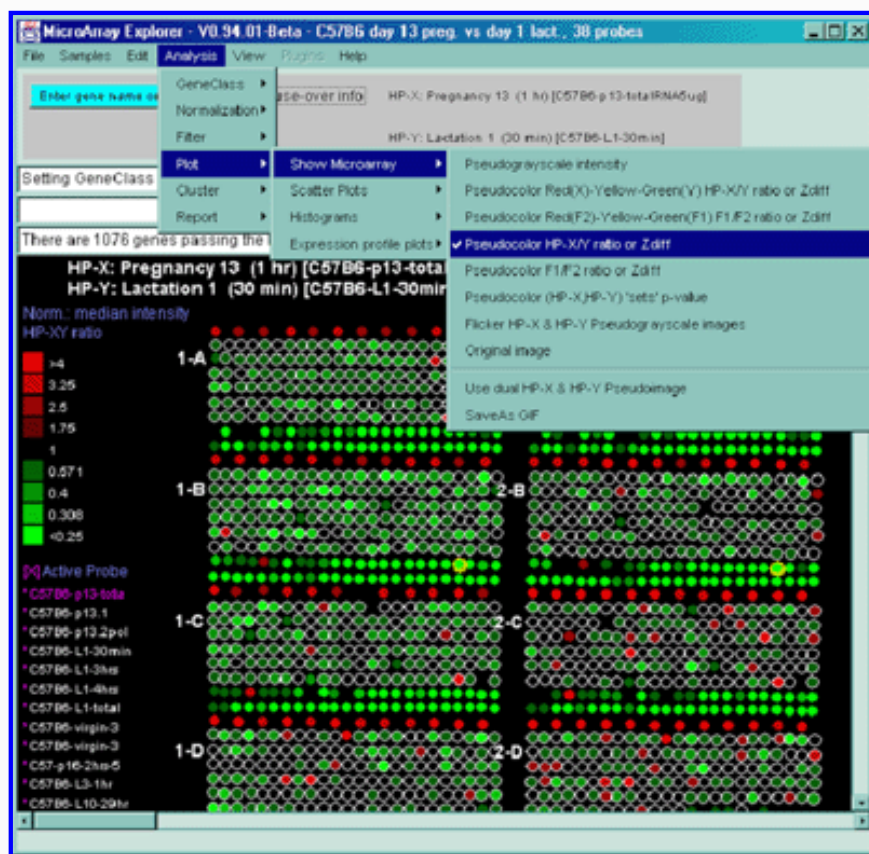
Figure 2.4.4 Plot menu - selecting Ratio Pseudoarray image. This displays a pseudocolor show in the scale on the left that indicates the ratio of the value of the HP-X sample / HP-Y sample (or 'sets' if the option to use HP-X and HP-Y 'sets' is enabled.) If The data is Cy3/Cy5 data, then this displays the ratio of the ratios using the current normalization. Various other pseudoarray image representations could be used.

2.4.4.1 Show microarray pseudoarray images menu

You may show the [pseudoarray image](#) of the current hybridized samples using several modalities. The grayscale pseudoarray image is generated from the quantified spot data. If the data contains the actual spot positions of the genes (as generated by the various array image quantification program), the spots may be drawn using a scaled version of those coordinates. Otherwise, a generic set of grids (and fields in there are multiple fields) is synthesized to represent the spot positions. Pseudoarray images may also be useful as an alternative modality for displaying X/Y ratio or X-Y Zdiff data. If the normalized intensities are the same, then the spot will appear as black with the overall spot intensity depending on the spot concentrations. High ratios and Zdiffs will be red and low values green as shown in [Table 2.4.4.1](#). The p-Value results of comparing a HP-X 'set' with a HP-Y 'set' of samples can be displayed as a color spectrum pseudoarray image.

If the database that was loaded contains only one sample, the pseudoarray image display defaults to the **pseudograyscale spot intensity** mode. If there is at least one HP-X and one HP-Y sample, then the **Pseudocolor HP-X/Y ratio or Zdiff** mode is the initial default display. If there are duplicate spots for each gene, you may generate a **Pseudocolor F1F2 ratio or Zdiff** mode image. If you are using Cy3/Cy5 ratio data and the data is available as independent channels for each HP, then you may plot Cy3 vs Cy5 for individual samples.

When available on the database server, the original image may be displayed in a separate popup Web browser.

- **Pseudograyscale intensity** - display a pseudograyscale microarray pseudoarray image for the current H.P. where higher intensity spots are mapped to black.
- **Pseudocolor Red(X)-Yellow-Green(Y) HP-XY ratio or Zdiff** - display the HP-X and HP-Y microarrays as a pseudocolor image as either a ratio (HP-X/HP-Y) or Zdiff (HP-X - HP-Y) if Zscore normalization is in effect. The HP-X (red) and HP-Y (green) components are *added together* so that high total intensity values are yellow and low values are black. High HP-X/HP-Y ratio values are more red and low values are more green.
- **Pseudocolor Red(Cy5)-Yellow-Green(Cy3) Cy3/Cy5 data** - display the Cy3 and Cy5 microarray channels as a pseudocolor array image. The the Cy5 (red) and Cy3 (green) components are *added together* so that high total intensity values are yellow and low values are black. High Cy3/Cy5 ratio values are more green and low values are more red. Note: this is only available with Cy3, Cy5 ratio data. Note: if the "Use ratio median corrections" is enabled, the Cy3 channel is scaled using the to the domain of the Cy5 channel by (median Cy5/median Cy3).
- **Pseudocolor HP-X/Y ratio or Zdiff** - display the HP-X and HP-Y data as a pseudocolor array image as either a ratio (HP-X/HP-Y) or Zdiff (HP-X - HP-Y) if Zscore normalization is in effect. The high ratios or Zscores are red and low values are green. Values in the middle are black.
- **Pseudocolor (HP-X,HP-Y) 'sets' p-value** - displays each spot as a pseudocolor proportional to the p-Value in a t-Test of the HP-X 'set' vs HP-Y 'set'. This can only be used when the "Use HP 'sets' ..." option is enabled and there are at least 2 samples in both the HP-X 'set' and the HP-Y 'set'. Note that unless proper normalization and filtering is used to remove poor quality data, some of the p-Values will not be significant and there will be a high false-positive rate. Use this display with that in mind.
- **Pseudocolor Cy3/Cy5 (or F1/F2) ratio or Zdiff** - display the Cy3 and Cy5 (or F1 and F2 for duplicate spotted grids) corresponding spots for the same genes as a pseudocolor array image as either a ratio (Cy3/Cy5) or Zdiff (Cy3-Cy5) if Zscore normalization is in effect. The high ratios or Zscores are red and low values are green. Values in the middle are black.
- **Flicker HP-X & HP-Y Pseudoimages** - toggle flickering the HP-X and HP-Y pseudograyscale intensity array images.
- **Original image** - display the original microarray 8-bit image (if available) for the current HP sample in a separate Web browser window.
- --
- **Use dual HP-X & HP-Y Pseudoimage** - If have replicate spots (F1,F2), toggle the display between F1 data in left grids and F2 data in the right grids to mean HP-X data in the left grids and HP-Y in the right grids for side by side comparisons.
- **"Scale pseudoarray image by 1/100 to zoom low-range values** - rescales intensity and (Cy3+Cy5) and (HP-X + HY-Y) (Red-Yellow-Green) plots to so mid values are easier to visualize.
- **SaveAs GIF** - save the current microarray pseudo image as a full resolution GIF file specified by the user in a popup file browser window (stand-alone mode only).

Table 2.4.4.1. Pseudocolors assigned to spots to represent data in the X/Y ratios or X-Y Zdiffs pseudocolor array images.
Each color represents the normalized X/Y ratio or X-Y Zdiff depending on Normalization mode. The 9 colors of the boxes represent the normalized expression ranges.

									Normalization mode - RBG
bright green			dark green	Black	dark red			bright red	
									Normalization mode - dichromasy
bright blue			dark blue	Black	dark orange			bright orange	
<0.250X	0.307X	0.400X	0.571X	1.000X	1.75X	2.50X	3.25X	>4.00X	Ratio data
<-3.0	-2.25	-1.50	-0.75	0.00	0.75	1.50	2.75	>3.0	Zscore data
<-0.99	-0.742	-0.495	-0.247	0.000	0.247	0.495	0.742	>0.99	Zscore Log data

Clicking on a particular gene will report its specific quantification and identification values (See Section 3.3 on gene quantification). If the **Enable display current gene in popup** *genomic DB* **Web Browser** option is set in the **View** menu, then it will also pop up a Web browser with the corresponding to the particular *genomic DB* data for that database if it exists.

The same data is shown in a variety of normalization and display formats.

2.4.4.1.1 Examples of microarray intensity data pseudoarray image

The relative intensity may be displayed for the current sample (last HP-X or HP-Y selected) or two samples (HP-X or HP-Y samples or HP-X or HP-Y 'sets' of samples). To show two samples side by side, enable the **Use dual HP-X & HP-Y Pseudoimage** in the Show Microarray submenu. To show averaged set data in the dual mode, enable the "Use HP-X & HP-Y 'sets' option in the Samples menu. The grayscale value reflects the current normalization mode.

Pseudoarray image show intensity plot for median normalized data for a single sample

Figure 2.4.4.1.1.1 Pseudoarray intensity image of median normalized intensities of the current HP sample (C57B6 virgin 10 weeks from MGAP database). The graylevel scale on the left edge of the pseudoarray image indicates the spot intensity. All pseudoarray images have scales that vary depending on the type of pseudoarray being displayed.
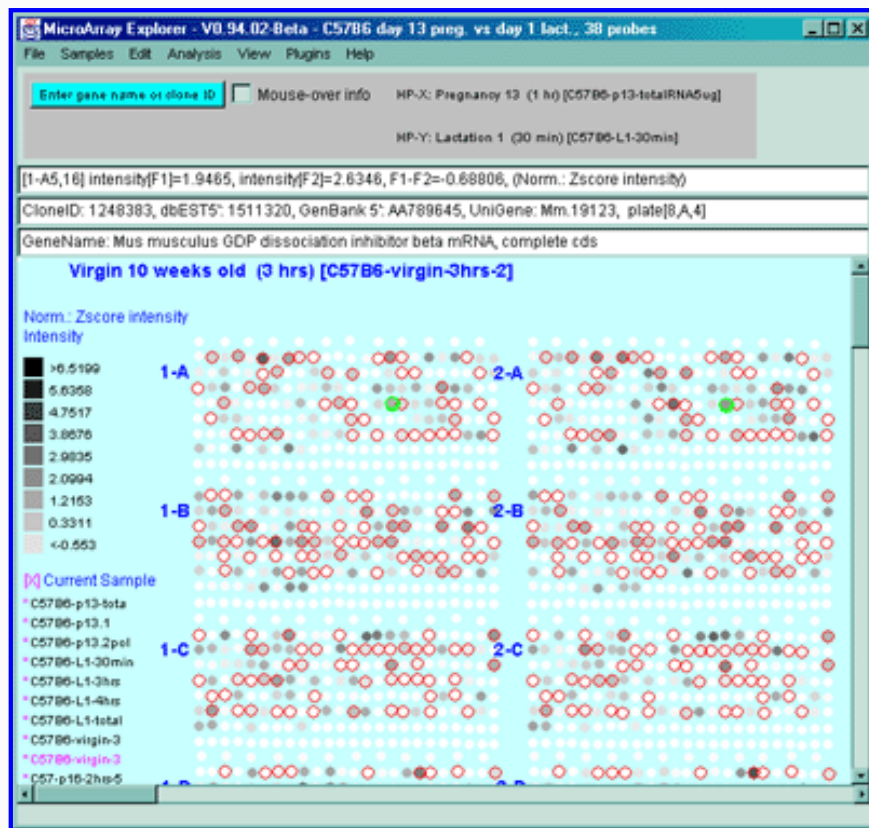
Figure 2.4.4.1.1.2 Pseudoarray intensity image of Zscore normalized intensities of the current HP (C57B6 virgin 10 weeks from MGAP database).
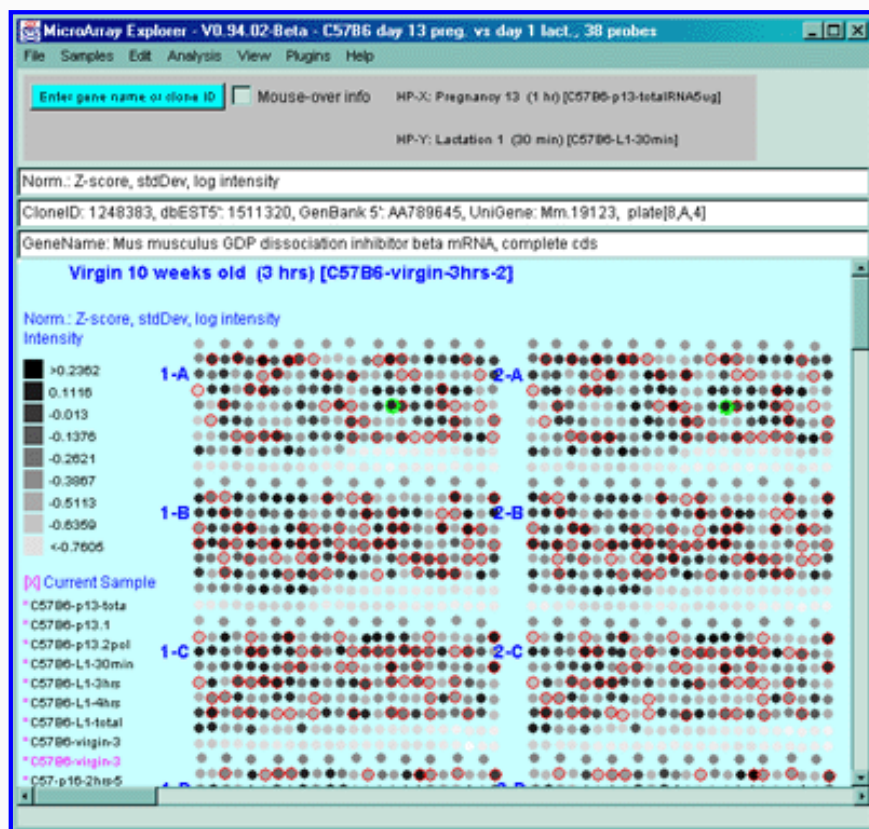
Figure 2.4.4.1.1.3 Pseudoarray intensity image of ZscoreLog normalized intensities of the current HP (C57B6 virgin 10 weeks from MGAP database).
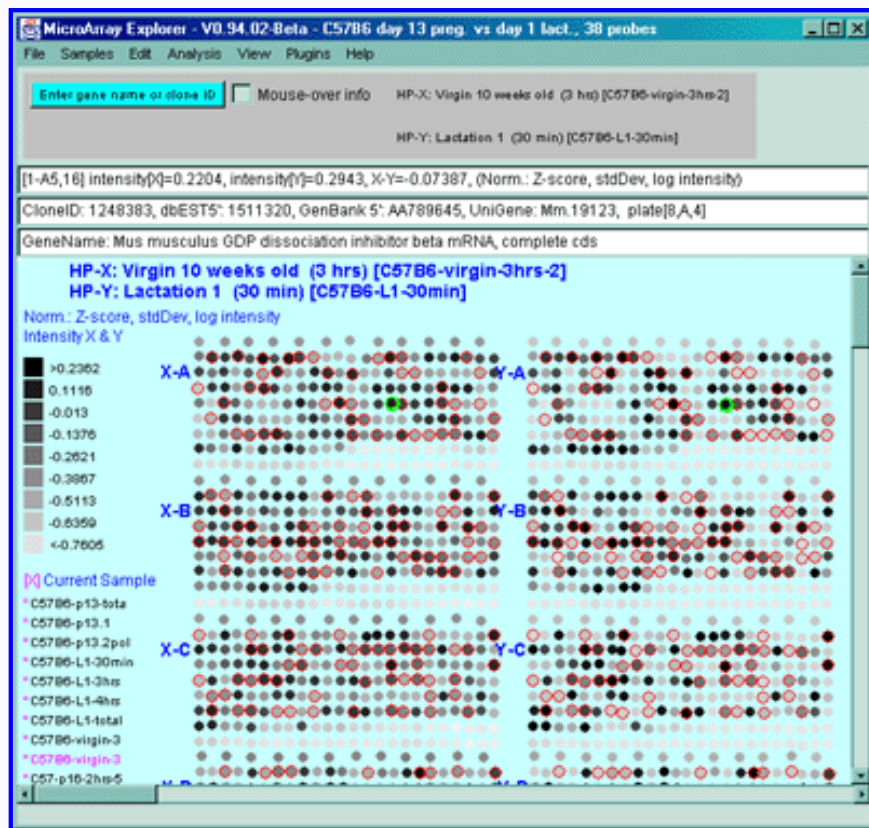
Figure 2.4.4.1.1.4 Pseudoarray intensity image of ZscoreLog normalized intensities of the dual HP-X and HY-Y individual samples. The Plot menu Show Microarray submenu toggle "Use dual HP-X & HP-Y samples" option is set. HP-X is a C57B6 pregnancy day 13 and HP-Y is a Stat5a (-,-) pregnancy day 13.
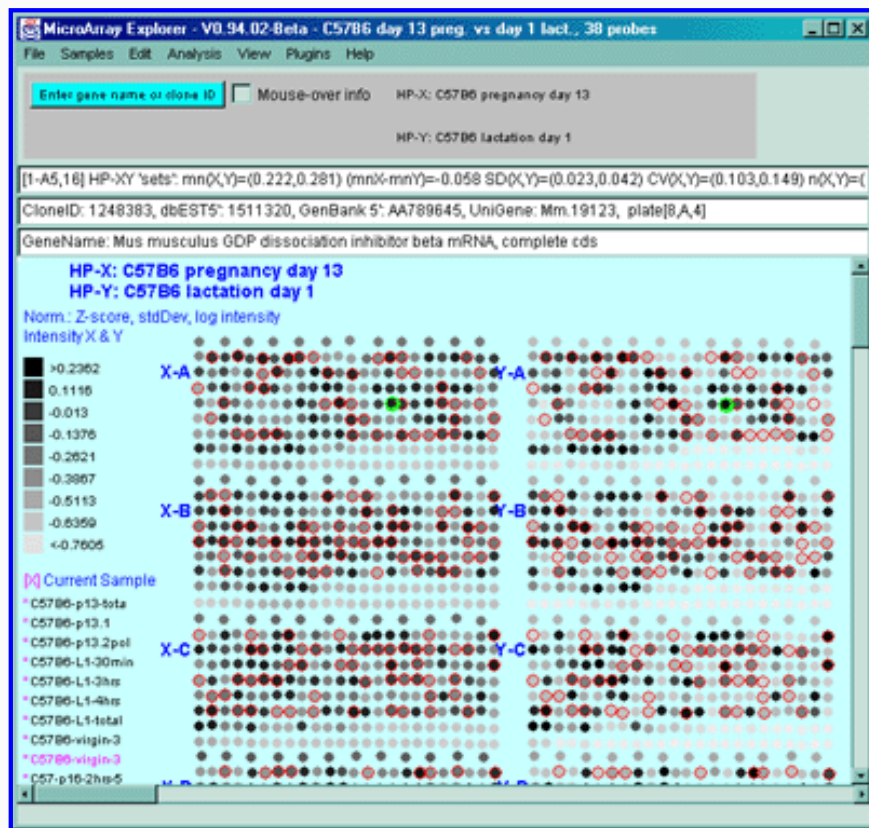
Figure 2.4.4.1.1.5 Pseudoarray intensity image of ZscoreLog normalized intensities of the dual HP-X and HY-Y sample 'sets'. The Plot menu Show Microarray submenu toggle "Use dual HP-X & HP-Y samples" option is set. The "Use HP-X & HP-Y 'sets' option in the Samples menu. HP-X is the mean of three 'C57B6 pregnancy day 13' and HP-Y is the mean of three 'Stat5a (-,-) pregnancy day 13'.

2.4.4.1.2 Example of microarray ratio or Zdiff data pseudocolor image

The ratio (HP-X/HP-Y) or Zdiff HP-X - HP-Y) normalized intensity data may be displayed as a pseudoarray image for HP-X and HP-Y, or HP-X and HP-Y 'sets' of samples. To show averaged set data in the dual mode, enable the "Use HP-X & HP-Y 'sets' option in the Samples menu. The colors of the 9 scale boxes represent the normalized expression ranges and is assigned according to the current normalization mode listed in the table.
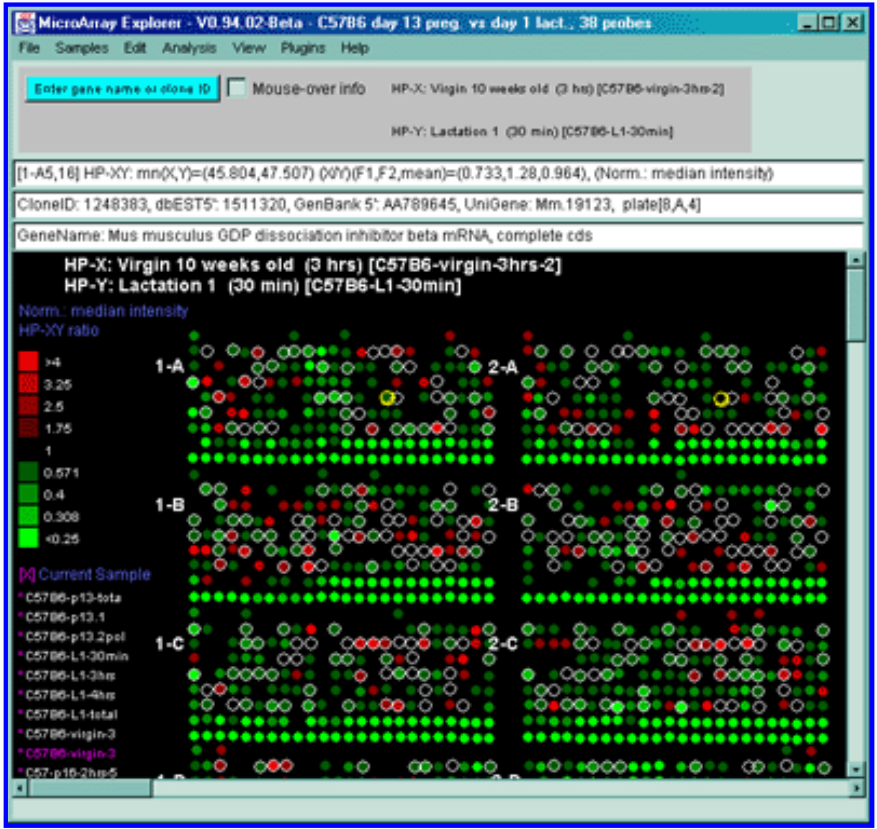
Figure 2.4.4.1.2.1 Pseudocolor array image of median normalized X/Y ratios. HP-X is C57B6 pregnancy day 13 and HP-Y is Stat5a (-,-) pregnancy day 13. Each spot's color represents the normalized X/Y ratio depending on Normalization mode. The color of the box is one of 9 colors representing the normalized expression ranges and assigned according to the table "Ratio mode".

Pseudoarray image show color ratio plot for median normalized data for HP-X and HP-Y 'set' samples

Figure 2.4.4.1.2.2 Pseudoarray color image of normalized X/Y 'set' mean value ratios. Mean of three HP-X C57B5 pregnancy day 13 samples and mean of three HP-Y Stat5a (-,-) pregnancy day 13 samples. Each spot's color represents the normalized X/Y 'set' ratios depending on Normalization mode. The color of the box is one of 9 colors representing the normalized expression ranges and assigned according to the table "Ratio mode".
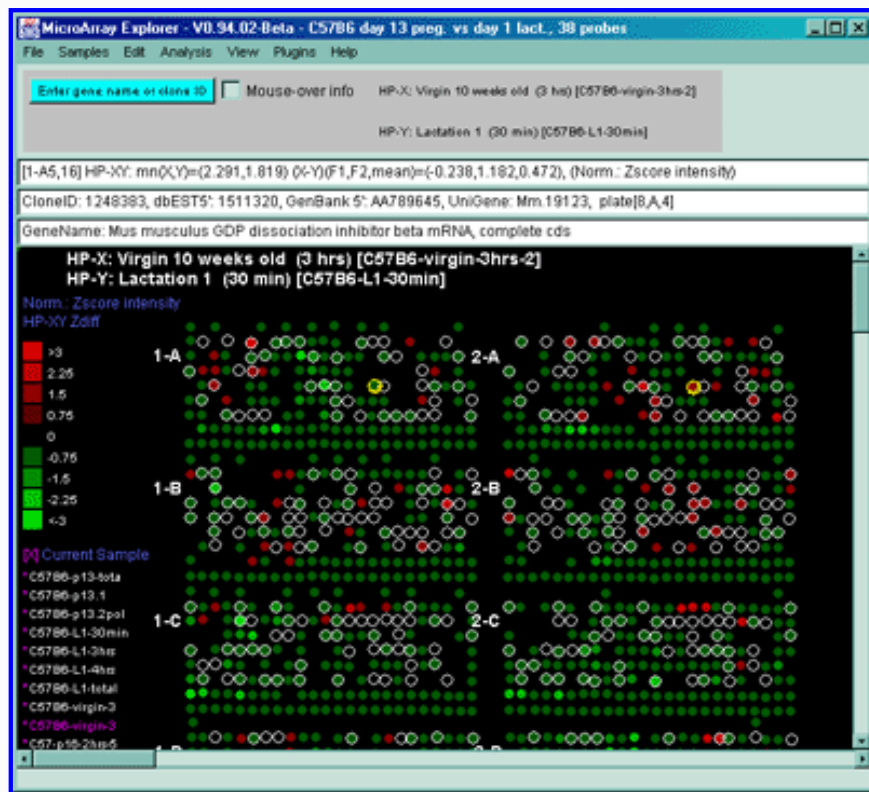
Figure 2.4.4.1.2.3 Pseudoarray color image of X-Y Zdiffs. HP-X is C57B6 pregnancy day 13 and HP-Y is Stat5a (-,-) pregnancy day 13. Each spot's color represents the normalized X-Y Zdiff depending using the Zdiff normalization mode. The color of the box is one of 9 colors representing the normalized expression ranges and assigned according to the table "Zdiff mode".
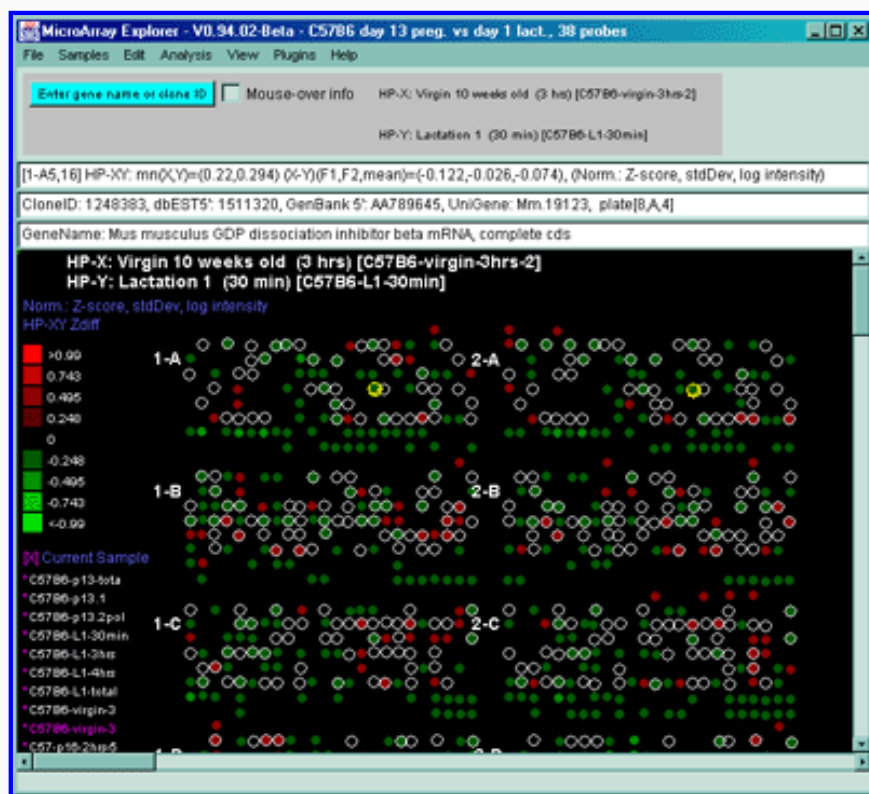
Figure 2.4.4.1.2.4 Pseudoarray color image of X-Y Zdiff of log data. HP-X C57B5 pregnancy day 13 sample and HP-Y Stat5a (-,-) pregnancy day 13 sample. Each spot's color represents the normalized X/Y ratio depending on ZdiffLog with StdDev normalization mode. The color of the box is one of 9 colors representing the normalized expression ranges and assigned according to the table "ZdiffLog mode".
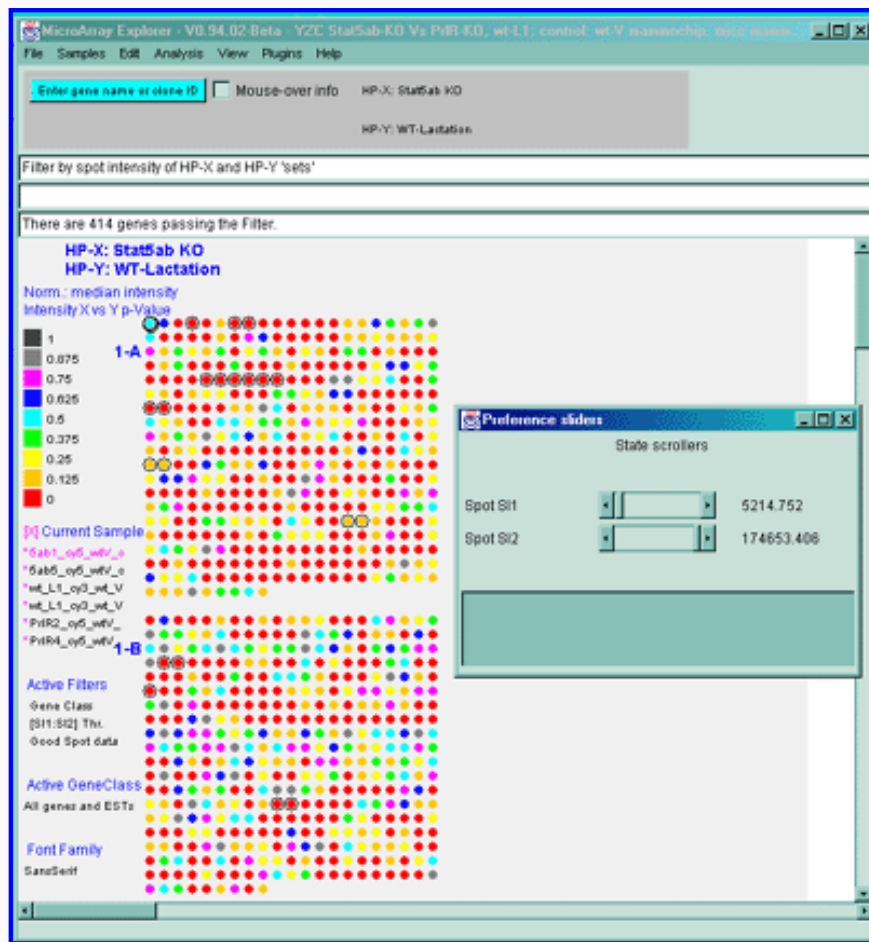
Figure 2.4.4.1.2.5 Pseudoarray image showing color-coded p-values for t-test comparison of HP-X and HP-Y 'set' samples. The HP-X and HP-Y sets both have 2 samples each (more is obviously much better). The data was normalized using the Median and a spot intensity [SI1:SI2] data filter was applied to eliminate some of the noisy data. Each spot's color represents a p-value in the range indicated in the scale in the left edge of the image. Note that although all spots are assigned a p-Value, many may not be very significant because adequate preprocessing of the data (such as normalization, and low intensity spot removal, etc.). So use this display with care.

2.4.4.2 Scatter plots menu

Scatter plots include **HP-X vs HP-Y intensity** for comparing data between HP-X and HP-Y samples (or sets if the HP-X and -Y sample 'sets' mode is enabled in the **Samples** menu - as is shown in Figure 2.2.1). You may zoom into any area of the scatter plot as is shown in Figure 2.4.4.4.2(C). If there are duplicate spots for each gene, you may plot **F1 vs F2 intensity** (or Cy3 vs Cy5 if using ratio data) for comparing replicate data (or Cy3 and Cy5 ratio data channels) within the same sample It will also compute the correlation coefficient for the data and display it in the plot and in the message panel. The data is the intensity values using the current normalization method. If you are analyzing ratio Cy3/Cy5 data, you may compare Cy3 or Cy5 of the HP-X sample against Cy3 or Cy5 of the HP-Y sample. If you are in stand-alone mode, a **SaveAs GIF** button will also be available. This saves the current plot as a full resolution GIF file specified by the user in a popup file browser window.

```
rSq=0.974, n=1728, X(mn+-sd)=(4.477+-7.845), Y(mn+-sd)=(12.379+-24.810)
```

The **Scatter plots** submenu includes:

- **HP-X vs HP-Y intensity** - display a scatter plot of HP-X vs HP-Y intensity data of H.P. selected (average of F1+F2 fields). If HP-X/-Y 'sets' are enabled, then the plot is of the mean set values.
- **HP F1 vs F2 intensity** - display a scatter plot of F1 vs F2 (or Cy3 vs Cy5) intensity data of the current selected sample
- --------------------
- **HP-X Cy3 vs HP-Y Cy3 intensity** - display a scatter plot of HP-X(Cy3) vs HP-Y(Cy3) intensity data of the currently selected HP-X and HP-Y samples (ratio data only).
- **HP-X Cy5 vs HP-Y Cy5 intensity** - display a scatter plot of HP-X(Cy5) vs HP-Y(Cy5) intensity data of the currently selected HP-X and HP-Y

samples (ratio data only).
- **HP-X Cy3 vs HP-Y Cy5 intensity** - display a scatter plot of HP-X(Cy3) vs HP-Y(Cy5) intensity data of the currently selected HP-X and HP-Y samples (ratio data only).
- **HP-X Cy5 vs HP-Y Cy3 intensity** - display a scatter plot of HP-X(Cy5) vs HP-Y(Cy3) intensity data of the currently selected HP-X and HP-Y samples (ratio data only).

If Cy3/Cy5 ratio data is being analyzed, then the "HP F1 vs F2 intensity" menu entry becomes

- **HP Cy3 vs Cy5 intensity** - display a scatter plot of Cy3 vs Cy5 intensity data of the current sample selected (i.e. being displayed).

The following figure illustrates some of the scatter plots and zoomed regions using the scroll bars on the horizontal and vertical axes.

A)

B)
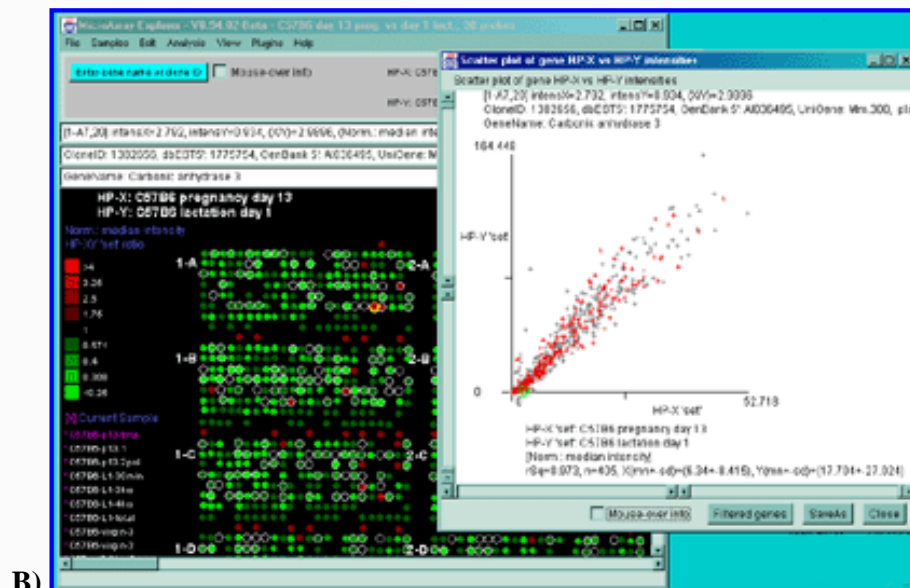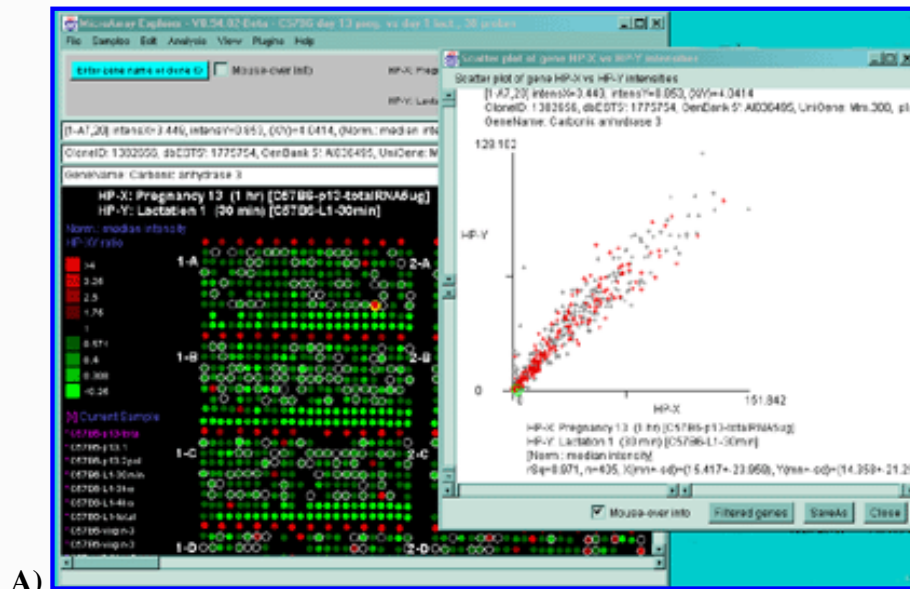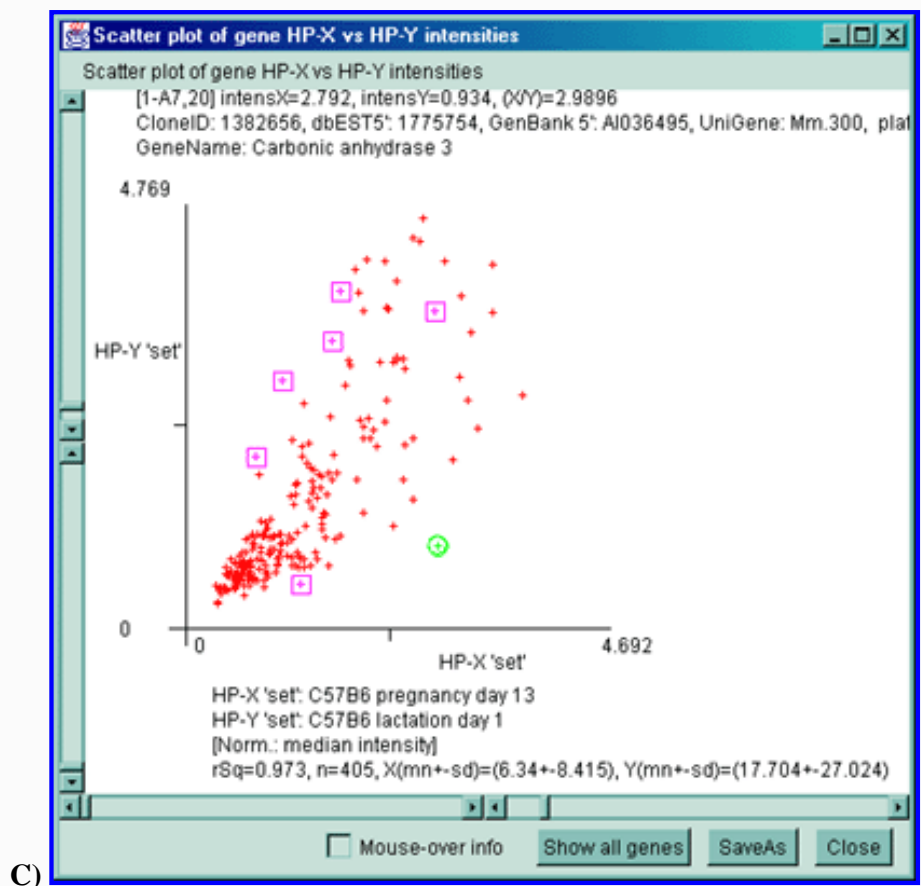
C)

Figure 2.4.4.2 Scatter plot of HP-X and HP-Y single sample data. HP-X is C57B6 pregnancy day 13 and HP-Y is pactation day 1. **A)** An active scatter plot may be generated for the current HP-X and HP-Y samples filtered by "All named genes". **B)** similar plot for HP-X and HP-Y 'sets' of replicate samples (3 pregnancy and 4 lactation samples in the sets respectively). Clicking on a point in the plot sets the current gene. **C)** Zoomed up region (of **B**) at the bottom of the plot showing more detail and filtered by just "All named genes". Zooming is performed by adjusting the X or Y axes limits scroll bars. Note the points enclosed in magenta boxes indicate genes in the E.G.L. gene list.

Scatter plots of data from multiple channels on the same sample

It is also possible to plot the separate channels within a single sample against each other. For example F1 vs F2 in samples with replicate data and Cy3 vs Cy5 in samples with separate ratio data channels.
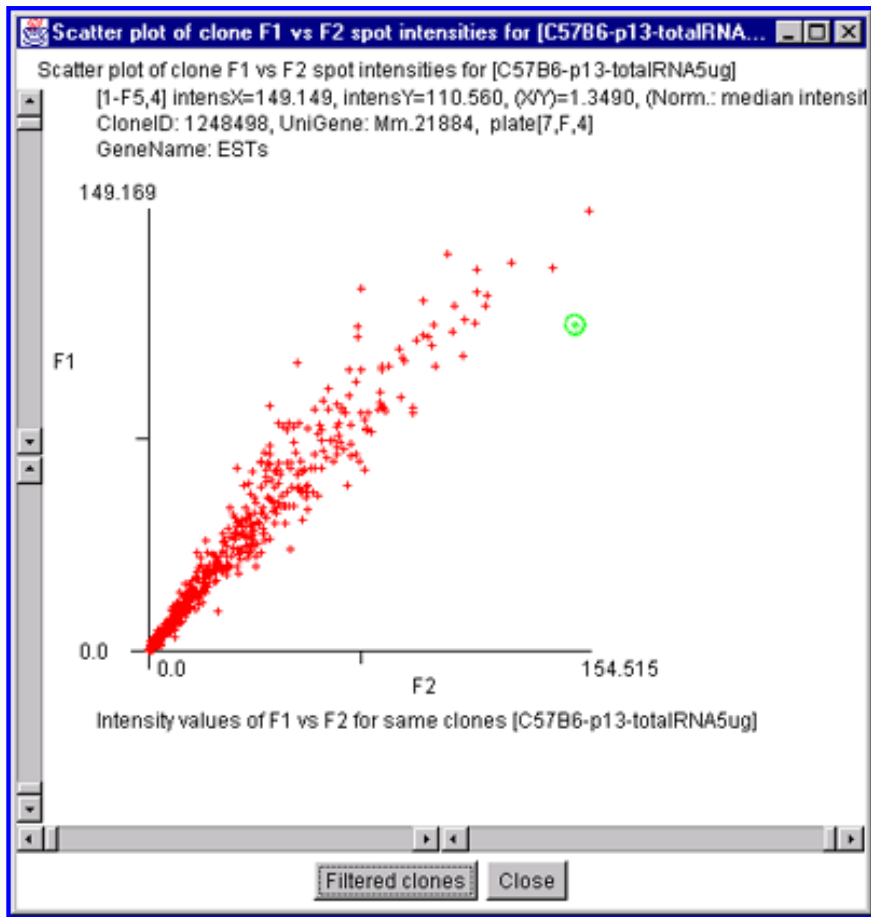
A)
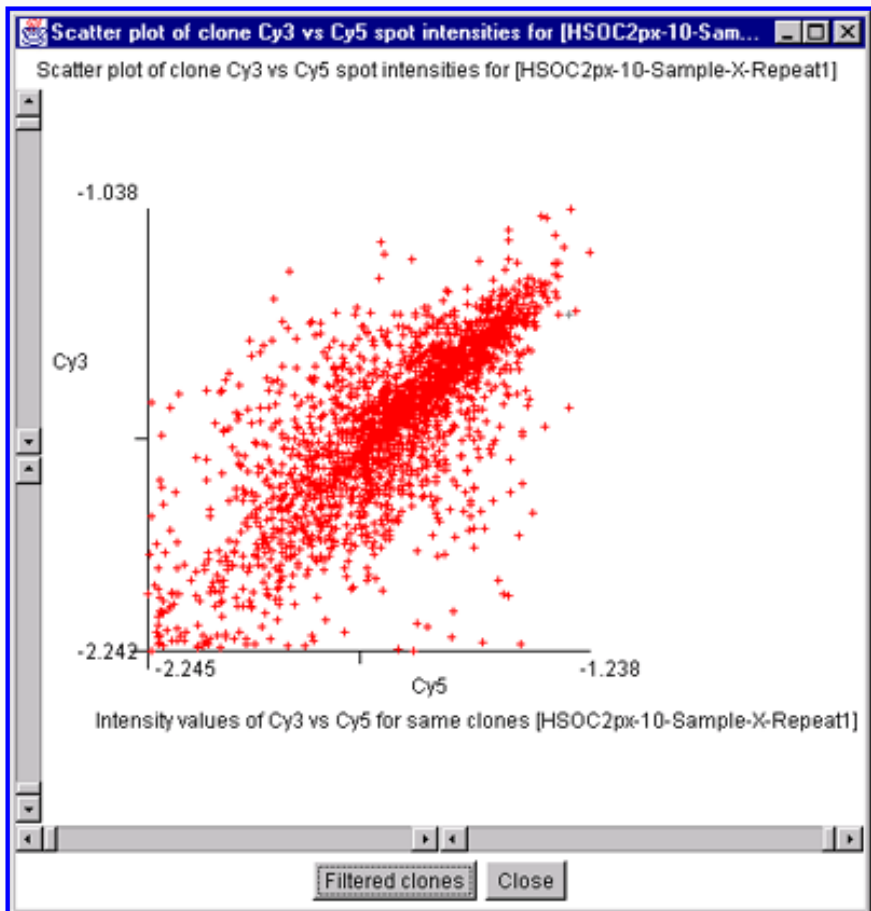
B)
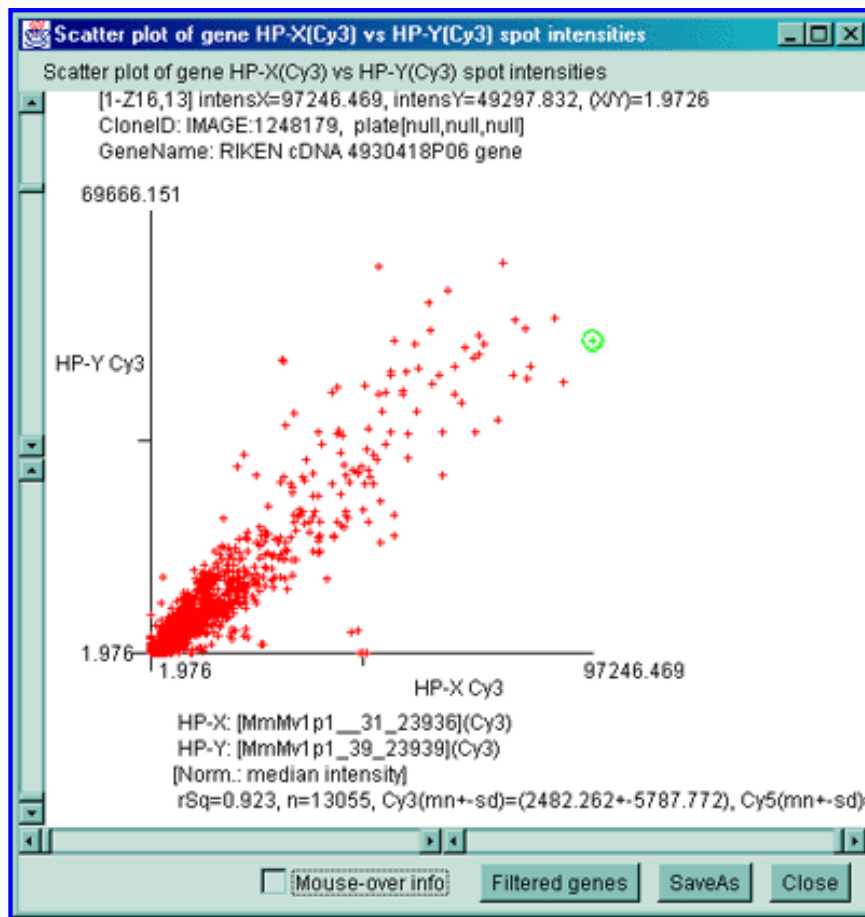
C)
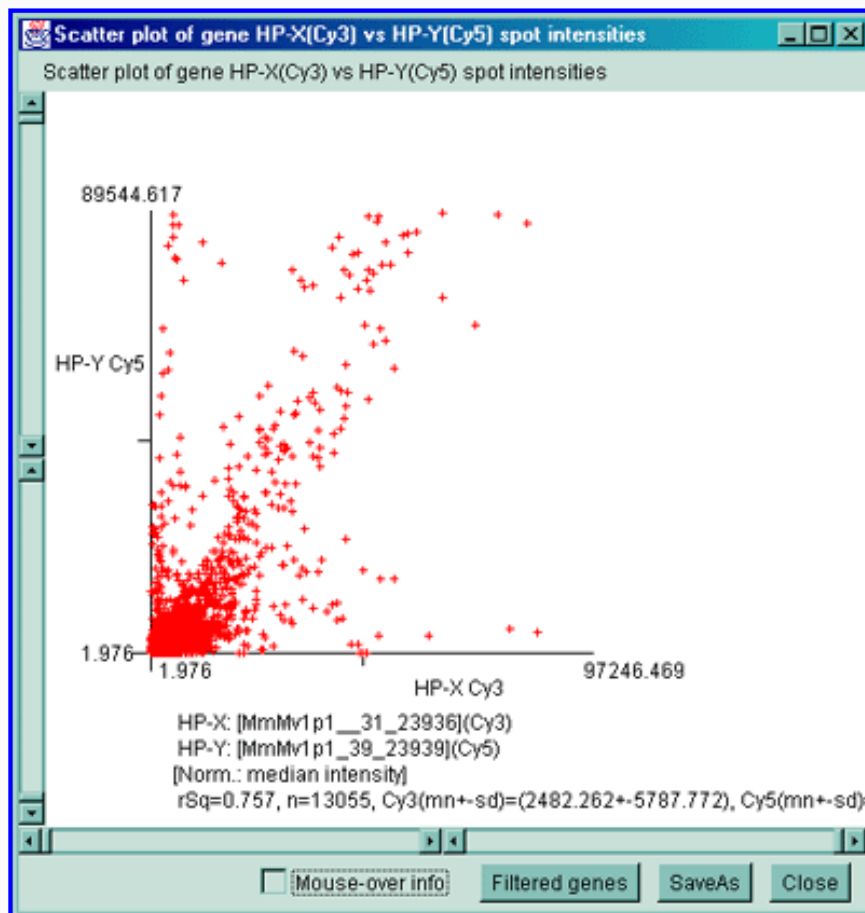
D)

Figure 2.4.4.2.1 Scatter plot of multiple channel data from a single sample. A) F1 Vs F2 data for a C57B6 pregnancy day 13 sample. **B)** Cy3 vs Cy5

data for a NCI mAdb mouse array sample. **C)** Scatter plot of individual Cy3 channels from two different ratio Cy3/Cy5 data hybridized samples. **C)** Scatter plot of individual Cy3 channel of HP-X compared with Cy3 channel of HP-Y for ratio Cy3/Cy5 data hybridized samples. **D)** Scatter plot of individual Cy3 channel of HP-X compared with Cy5 channel of HP-Y for ratio Cy3/Cy5 data hybridized samples.

2.4.4.3 Histogram plots menu

You may compare ratios or Zdiffs of data using the **HP-XY ratios or Zdiff** command to display a ratio histogram of Filtered intensity data from two samples selected from the Samples menu. The **HP-XY 'set' ratio or Zdiff** is used if there are multiple samples in the HP-X or HP-Y sets, then the mean values in each of the sets is used in the calculations. If there are duplicate spots for each gene, you may plot the **F1F2 ratio or Zdiff** histogram of the F1/F2 ratios or F1-F2 Zdiff values for normalized data for each spot in the currently displayed sample. If you are in stand-alone mode, a **SaveAs GIF** button will also be available to save the current plot as a full resolution GIF file specified by the user in a popup file browser window.

The **Intensity** selection plots a histogram of the gene intensity data values for each Filtered spot (gene) in the current hybridized sample.

The **Histograms** submenu includes:

- **HP-XY ratio or Zdiff** - display ratio or Zdiff histogram of data from selected HP-X and HP-Y samples (mean F1+F2 data if duplicate spots).
- **HP-XY 'sets' ratio or Zdiff** - display ratio or Zdiff histogram of mean data from HP-X and HP-Y 'set's of samples (mean data).
- **F1F2 ratio or Zdiff** - display ratio or Zdiff of histogram of F1 F2 duplicate spot data from the current sample.
- **HP Intensity** - (or **HP (Cy3/Cy5)** if ratio data) display a histogram of spot intensity (or Cy3/Cy3 ratio) data for the current sample.

If Cy3/Cy5 ratio data is being analyzed, then the F1F2 histogram menu entry becomes

- **Cy3Cy5 ratio or Zdiff** - display ratio or Zdiff of histogram of Cy3 Cy5data from the current sample.

The following figures illustrates the histograms. You may use the histogram to specify ranges of [I1:I2] or {R1:R2} for data filtering in the histogram by specifying the corresponding histogram bins. This is described in the figure legend.

A)

B)
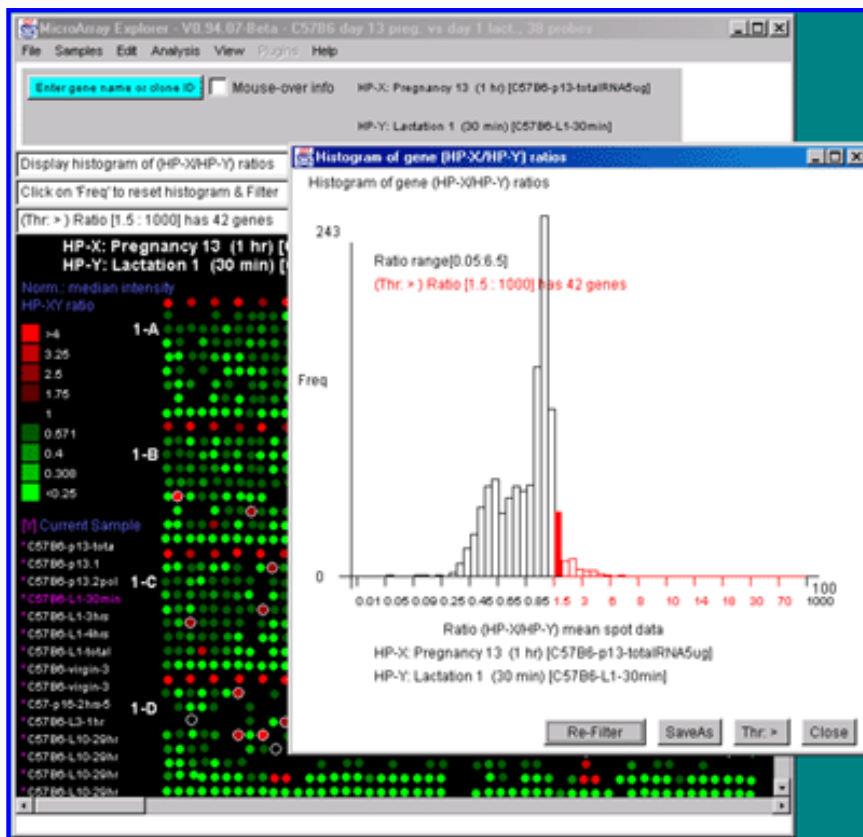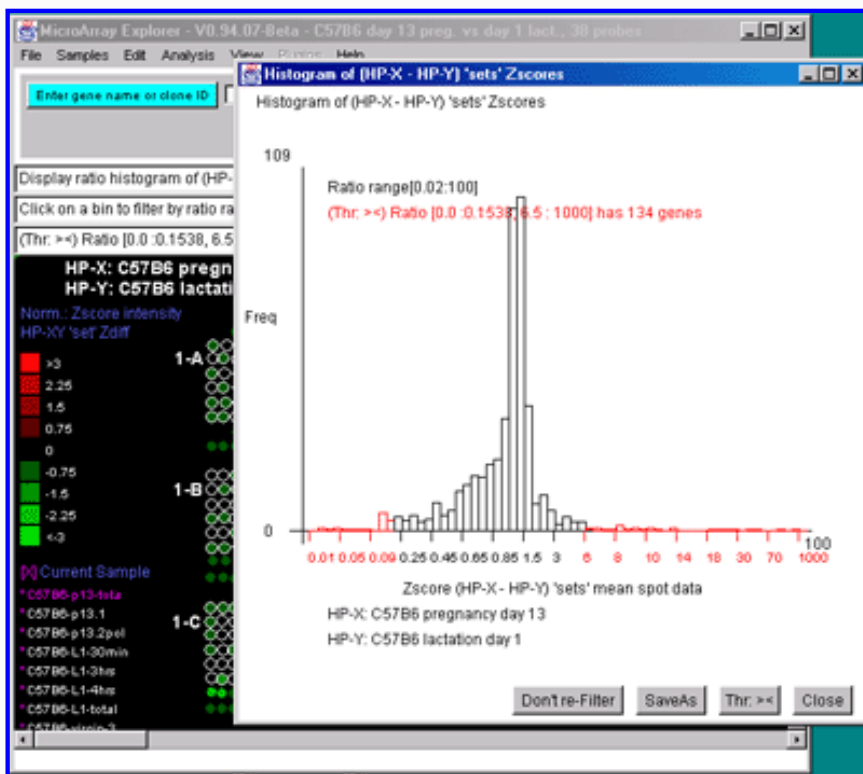
Intensity histogram of
current sample data, median
normalization

C)

Figure 2.4.4.3 Histogram plots. A) Ratio histogram of HP-X/HP-Y data with particular histogram bin selected with the constraint set to filter all genes
> that bin. HP-X is 13 day pregnancy C57B6 and HP-Y is day 1 lactatation. The selected bin thresholds are then used in the Filter with the resulting

Filtered genes shown in the array image. **B)** Zdiff histogram of HP-X - HP-Y 'sets' for same data as **(A)** but with the >< threshold constraint set to find genes outside of the symmetric histogram range. **C)** Intensity histogram of HP-X data filtered by [I1:I2] intensity range. As with ratio histograms, you can do additional filtering by selecting a particular histogram bin that is then used in the Filter. Filtering was disabled for the intensity histogram. To apply the filter, the "Don't re-Filter" button would be toggled to the "Re-Filter" state. The threshold constraints include: **=, >, <, >, <>,** and **><**. Note that each time you click on the "Thr:" button, it cycles to the next option in the threshold constraints list.

2.4.4.4 Expression profile plots menu

You may generate an individual expression profile plot (EP plot) or a scrollable list of EP plots. The order list of hybridized samples to plot are specified by the HP-E set. In the latter case, the genes are specified by the data Filter.

You many generate as many individual expression profile plots as you want using the **Display a gene's expr. profile for HP-E** command. However, only the last one will be active and will be updated with different genes as you click on them in the microarray image scatter plot. This could be used to compare the EP plots for several different genes. First view the EP plot for one gene, then create a new EP plot for the second gene, etc.

If you use the **Display Filtered genes expr. profiles for HP-E** command, it will generate a scrollable list of expression profile plots for all of the genes passing the Filter. If the number of genes is very large, it may take a while.

You may interrogate a line corresponding to a particular HP sample in a EP plot by moving the mouse over the line and then selecting the line. This will cause the name of the HP, its intensity and CV to appear in the plot. If the **Err** check box is set, then the mean of the intensity is indicated by a short horizontal bar and the +- CV by red vertical error bars above and below the mean. If the plot style **Line** button is pressed, then the plot style is cycled between Line (vertical lines for each point), Circle (small circles at each point), and Curve (circles are connected). Pressing the button repeatedly cycles through: **Line** (i.e. vertical vars), **Circle**, or **Curve** (i.e. continuous curve of all samples). In the case of mean expression profiles used in K-means clustering, the standard deviation is used in place of the CV value. The various clustering methods have **EP plots** buttons. When they are invoked, the scrollable list of EP plots is sorted by the clustering method ordered list of genes. This enables you to view the data in the same order as that produced by the cluster analysis. If the zoom **nnX** button is pressed, then all of the plots are magnified by nn-fold to make low intensity plots more visible. Pressing the button repeatedly cycles through: 1X, 2X, 5X, 10X and 20X. It does not change the data itself. The **Show HP names** button pops up a numbered list of all HP entries used in the expression profile. If you are in stand-alone mode, a **SaveAs GIF** button will also be available for the EP overlay mode (Figure 2.4.4.4.1) or individual EP plot. This saves the current plot as a full resolution GIF file specified by the user in a popup file browser window.

The **Expression profile plots** submenu contains:

- **Display a gene's expr. profile for HP-E** - popup a window and display the expression profile of a gene when click on a spot in the Image or a point in the scatter plot.
- **Display Filtered genes expr. profiles for HP-E** - popup a scrollable montage of expression profiles of the list of Filtered genes.
- --------------------
- ☐ **Use EP overlay else EP list** - display Filtered genes as an overlay plot of expression profiles, else as a scrollable list of EP plots.
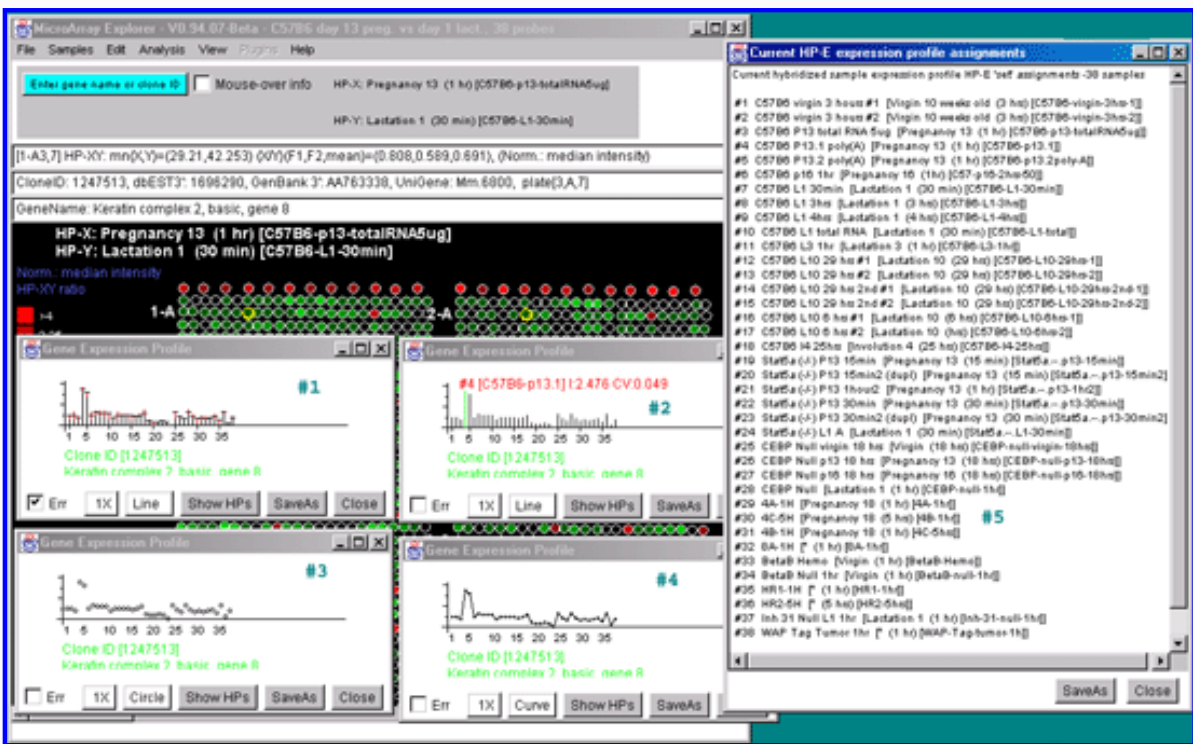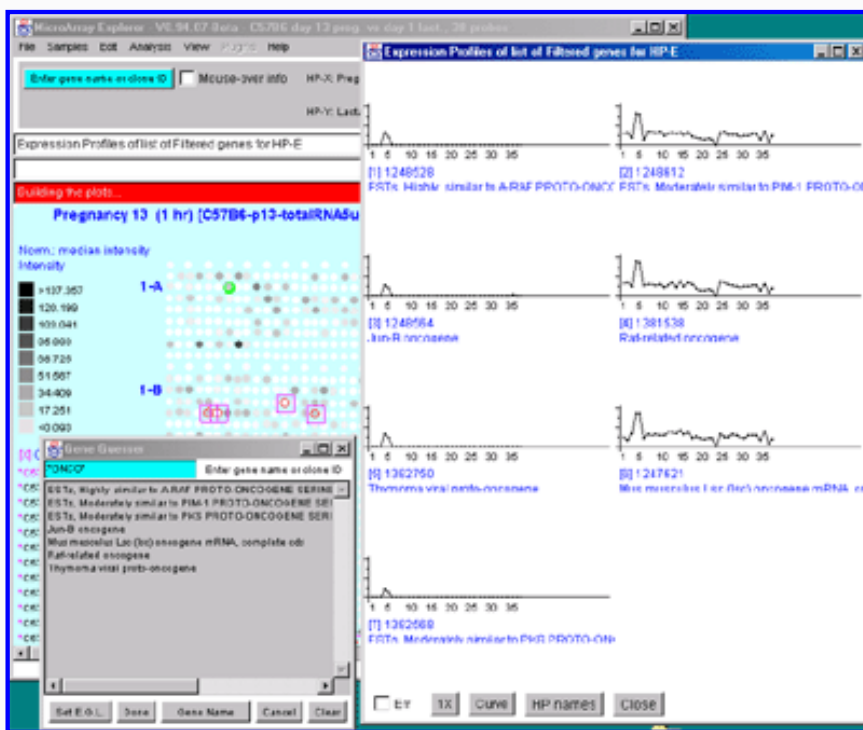
Figure 2.4.4.4 Expression profile plots. A) Individual expression profile plots may be created by clicking on any gene. Multiple instances may also be created. Here we show some of the presentation options for the 38 sample MGAP database. Error bars are computed for the standard error for that sample. There are three different plotting options: line, circle and curve. #1 is the default line plot with error bars. #2 is the line plot without the error bars but clicking on line 7 to find out which sample it is and what the intensity value is. #3 is the circle plot with error bars, and #4 is the curve plot without error bars. Window #5 shows the list of samples corresponding to the 38 points in the EP plots. **B)** List of EPplots of the oncogenes and proto-oncogenes in the database (set by the guesser with "onco" and "Set E.G.L." and the Edited Gene List Filter). The list would become scrollable if there were more than 10 profiles. Setting the current gene would scroll the list to the EPplot for the current gene.
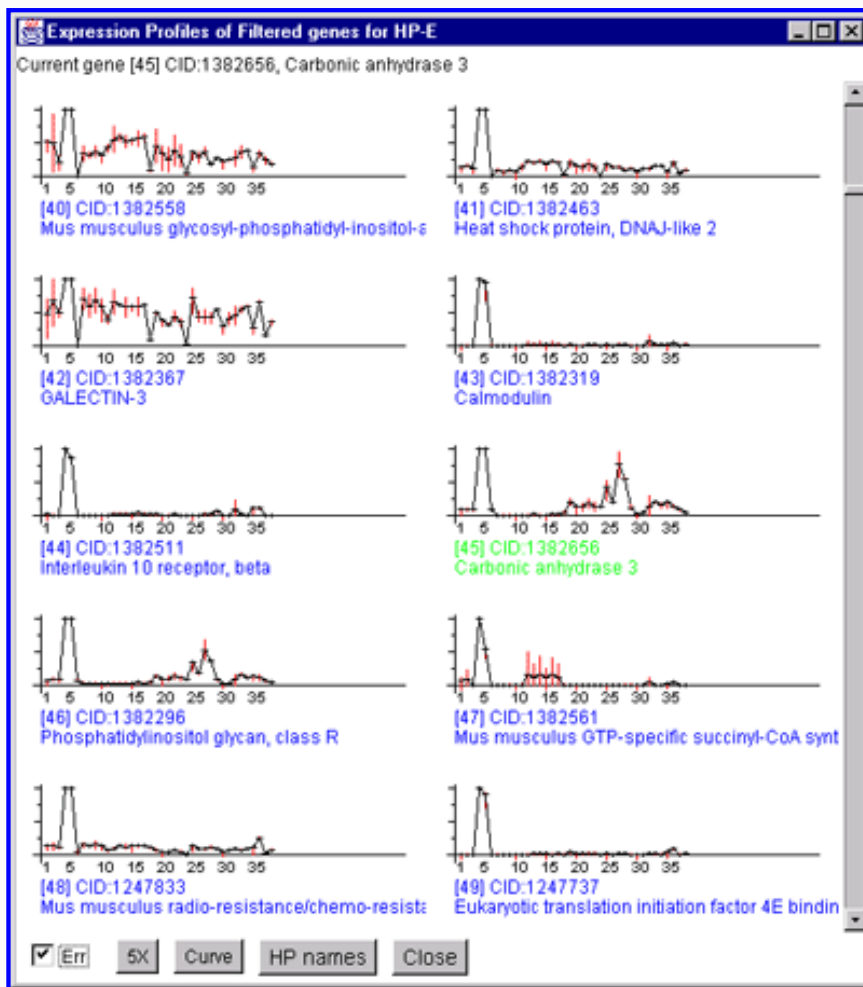
A)
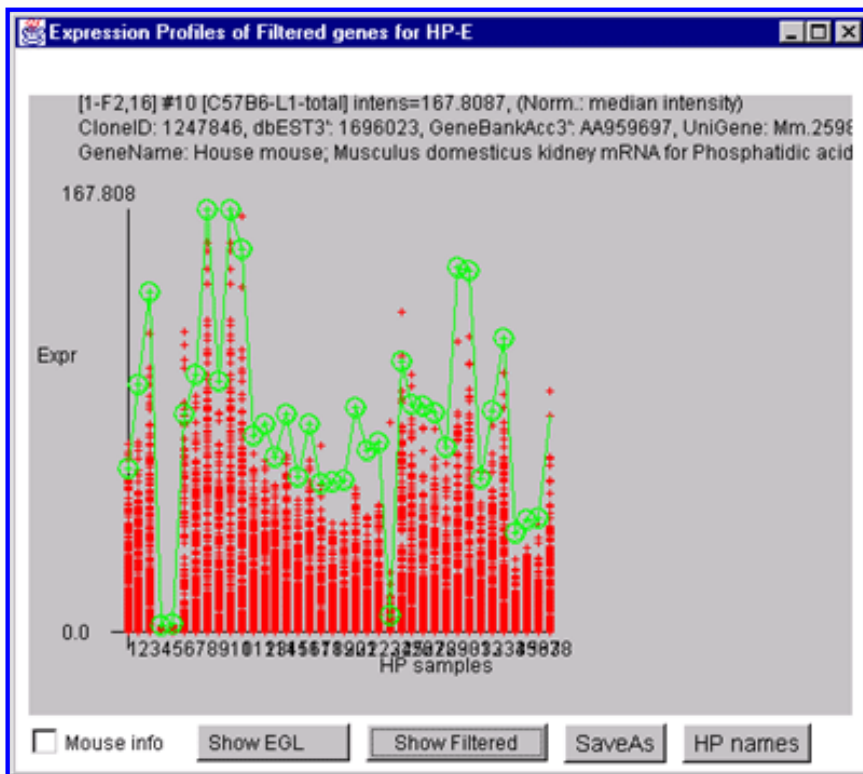
B)
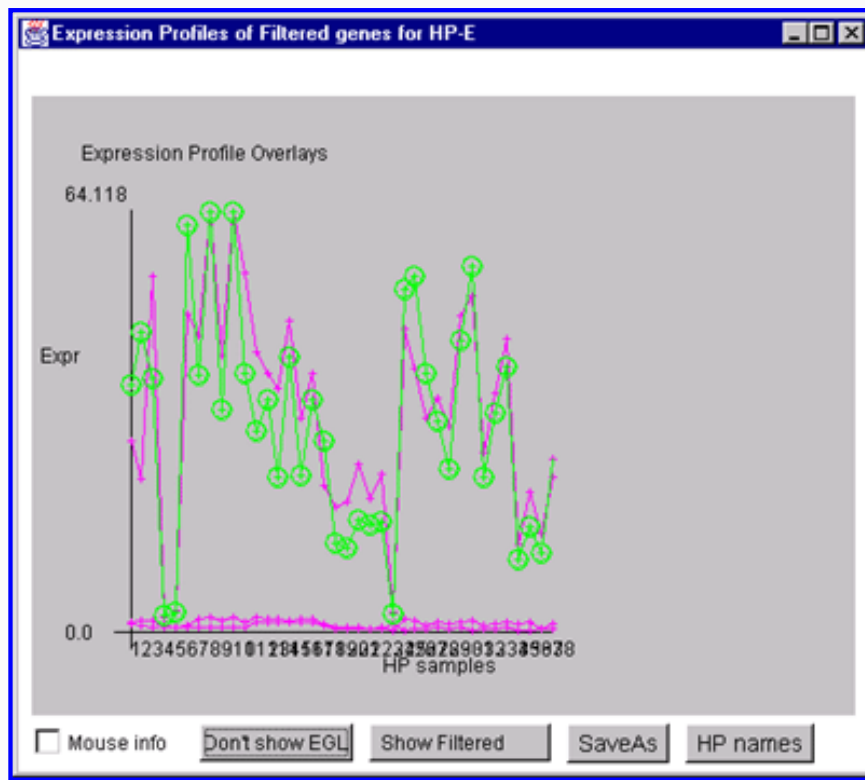
C)

Figure 2.4.4.4.1 Expression profile plots. A) Scrollable list of EP plots of Filtered named genes centered at Carbonic anhydrase III. **B)** Overlay plot of all named Filtered genes. **C)** Overlay plot of all ONCO or PROTO-ONCO genes with the draw EGL option active so the graphs are drawn for these genes.

2.4.5 Cluster menu

The Clustering menu lets perform various types of gene and condition clustering operations. When you invoke a clustering operation it will popup one or more windows and may modify the pseudoarray image. Some of the popup windows include clustergram and dendrogram analysis plots used with the hierarchical clustering.

When enabled, cluster data appears as blue circles or squares drawn as overlays on the pseudoarray image. These options are discussed in the section on clustering.

Cluster analysis plots include finding a subset of genes or subsets of samples based on cluster analysis of expression profile similarity measures. These show genes belonging to particular clusters, or genes that cluster well with specified genes. Cluster methods include: finding genes similar to the current selected gene within a "distance" threshold; K-means-like clustering where you specify a seed gene and the number of clusters; and hierarchical clustering with clustergram and dendrogram graphics.
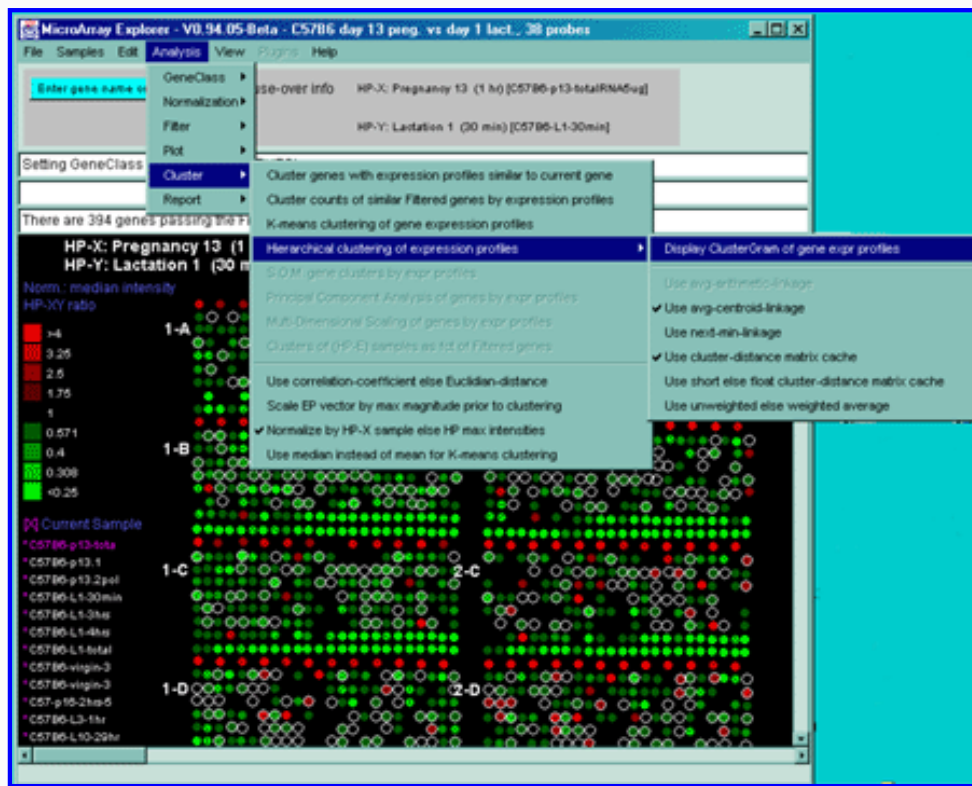
Figure 2.4.5 Cluster Menu options. The hierarchical clustering option is being selected.

Use of clustering to find patterns of similar gene expression

Clustering is a way of possibly finding co-expressed genes that exhibit similar expression changes in a set of samples. Genes may show similar co-expression, but that does not prove they are co-regulated at the same point in a pathway - merely that measurements of those genes in a particular set of experiments show similar expression. However, identifying genes with similar expression for which some information is already known about some of the genes may be useful as a starting point to help figure out gene function and possibly aspects of its pathways in cell function using additional experiments and analysis.

There are many methods for doing clustering - each with advantages and disadvantages. We present three methods in MAExplorer and plan on adding a variety of more powerful methods through the MAEPlugin facility under development.

These methods may find genes belonging to particular clusters or genes that cluster well with particular genes. Gene clusters are sets of genes whose expression profiles are found to be similar according to a particular metric. We now define what we mean by "similar". The order list of hybridized samples used in computing the expression profiles are those in the HP-E list. MAExplorer has two different dissimilarity measures for C_{ij}: Euclidean distance $LSQdist_{ij}$ and Pearson correlation coefficient r_{ij}. These are computed as follows and are tested against the cluster distance threshold (set by the slider in the preferences sliders). Let n= |HP-E|, the number of samples in the expression profile. We define similarity as (1.0 - normalized dissimilarity).

> *Hint: when working with very large data sets with many samples, it may be useful to pre-adjust the distance and/or number of clusters threshold sliders to an approximate range using the (Edit Menu | Preferences | Adjust all Filter threshold scrollers). This is because once the clustering starts, it does not (currently) let you abort the clustering to change the threshold value.*

```
LSQdist_ij = Sqrt( Sum ( D'_hj - D'_hi) **2 ) / n

            h in HP-E

            i,j in Filtered genes, i not j
```

```
Let,
    sum_ij = Sum( D'_hj * D'_hi ),
    mn_i = (1/n)Sum( D'_hi ),
    mn_j = (1/n)Sum( D'_hj ),
    sumSq_i = Sum( D'_hi *  D'_hi ),
    sumSq_j = Sum( D'_hj *  D'_hj ),

then,
          [sum_ij - n*(mn_i * mn_j)]
    r_ij = -------------------------------------------------------
          [Sqrt(sumSq_i - n*n*mn_i*mn_i) * Sqrt(sumSq_j - n*n*mn_j*mn_j)]

                  h in HP-E
                  i,j in Filtered genes, i not j
```

The **Cluster plots** submenu contains a number of clustering methods. Pressing the Escape key during a long cluster operation will abort the operation. If you are in stand-alone mode using the ClusterGram, a **SaveAs GIF** button will also be available for saving the current plot as a full resolution GIF file specified by the user in a popup file browser window.:

- ☐ **Cluster genes with expression profiles similar to current gene** - click on gene in image to find other genes with similar HP-E expression profiles whose cluster distance is less than the cluster distance threshold. The larger the blue box, the higher the similarity.
- ☐ **Cluster counts of similar Filtered genes by expression profiles** - draw blue circles around filtered genes indicating the number of other genes whose cluster similarity is less than the cluster distance threshold. The larger the circle, the more similar genes were found. Clicking on a gene switches to the above mode.
- ☐ **K-means clustering of gene expression profiles** - draw magenta circles around the N primary-node gene clusters representing the gene closest to representing the center of the cluster. Each of the nodes is a maximum distance from all other nodes in the recursive definition of nodes. N is determined by the State Scroller "# of Clusters". Changing N will recompute the clusters. It then pops up a scrollable text window with the clusters and indicates which genes belong to it. If you select the **EP plot** button, it will draw the expression profiles for the clustered genes. The **Mn-Cluster-Report** button will generate report for all genes sorted by K-means cluster. Summaries can be generated using the **Mean EP plot** and **Mn-Cluster-Report** buttons. The **SaveAs GeneSets** button saves all of the clusters as named Gene Sets ("Cluster #1", "Cluster #2", etc). If you change the filter or current gene, you should explicitly use the **Recompute Clusters** button to regenerate the new set of clustered genes. When you recompute the K-means clusters, it uses the current gene as the initial node.
- ☐ **Hierarchical clustering of expression profiles** ▷ - this computes the hierarchical clustering of the expression profiles (normalized by HP-X sample data for each gene) of Filtered genes. The hierarchical clusters are displayed in an ordered gene clustergram and optional dendrogram. Sub-regions of the clustergram may be explored in more detail using the **EP-subset plot** button, or a report of the ordered genes can be created using the **ClustGram Report** Note: you may add (remove) genes you select from the Clustergram to the E.G.L. by holding the Control(Shift) key while clicking on the gene name.
- ☐ **S.O.M. gene clusters by expr profiles** - [Future MAEPlugin]
- ☐ **Multi-Dimensional Scaling of genes by expr profiles** - [Future MAEPlugin]
- ☐ **Multi-Dimensional Scaling of genes by exprprofiles** - [Future MAEPlugin]
- ☐ **Clusters of (HP-E) samples as fct of Filtered genes** - [Future]
- --------------------
- ☐ **Use correlation-coefficient else Euclidian-distance** - use the (1.0 - correlation coefficient) as the distance metric instead of the default Euclidean distance.
- ☐ **Scale EP vector by max magnitude prior to clustering** - scale each sample in the EP by the max magnitude for all sample values in the EP.
- ☑ **Normalize by HP-X sample else HP max intensities** - normalize data by the corresponding HP-X sample data for each gene or the maximum raw intensity for each HP in the expression profile.
- ☐ **Use median instead of mean for K-means clustering** - use the clustering (see (Bickel, 2001)).

The **Hierarchical Cluster plots** submenu contains:

- ☐ **Display ClusterGram of gene expr profiles** - compute the hierarchical clustering of the expression profiles (normalized by HP-X sample data for each gene) of Filtered genes. Then display the hierarchical clusters in an ordered gene clustergram and optional dendrogram when the **dendrogram** checkbox is selected. Expression profile plots of the clustergram may be explored in more detail using the **EP plot** button that generates a scrollable list of all EP plots ordered by the same order as the clustergram. A full report of the ordered genes expression profiles may be created using the **ClustGram Report** button.
- --------------------

- ☐ **Use avg-arithmetic-linkage** - set the hierarchical clustering linkage method to the average arithmetic linkage of sub-clusters.[Future]
- ☑ **Use avg-centroid-linkage** - set the hierarchical clustering linkage method to average centroid linkage of sub-clusters (default).
- ☐ **Use next-min-linkage** - set the hierarchical clustering linkage method to the next minimum distance sub-cluster linkage in random order.
- ☑ **Use cluster-distance matrix cache** - if you do not have enough memory for clustering large gene sets, disable the cache. It will take MUCH longer without the cache. When clustering, if there is not enough memory available for the cache, it will warn you and suggest you either reduce the number of genes being clustered or use a computer with more memory.)
- ☐ **Use short else float cluster-distance matrix cache** - if there is not enough memory for the set of genes you wish to cluster and you still want to use the cache, you can use 16-bit (i.e. short) data instead of the 32-bit (i.e. float) data. The results will be less precise.
- ☐ **Use un-weighted else weighted average** - set the hierarchical clustering vector averaging to un-weighted (the default weights it by the number of genes in that sub-cluster). Otherwise using weighted gives equal (0.50) weighting to each sub-cluster.

Handling of hierarchical clustering of large numbers of genes - problem with slow response

The hierarchical clustering algorithm uses a gene-gene floating point (i.e. 32-bit) distance matrix of order N^2 (for N data filtered genes). This means that if you are experiencing a slow response, this may be due to several factors some of which you may not be able to control. You might:

1. If you determine that your computer is "paging", use a computer with more memory. The currently distributed stand-alone version will use up to 256 Mbytes of chip memory.
2. If that is not possible, reduce the number of genes being clustered. Even if you have enough memory, the computation is still high if N is large.
3. Set the **Use short else float cluster-distance matrix cache** option in the cluster plot menu. This reduces the memory requirements for the distance matrix by 1/2.

2.4.5.1 Cluster genes with expression profiles similar to current gene

The **Cluster genes with expression profiles similar to current gene** is used to find genes with similar HP-E expression profiles as measured by the least square error that are less than the cluster distance threshold. It pops up the "Cluster Distance" threshold scroller. Then click on a gene in the microarray image. It then pops up up a window with a list of the similar genes and their expression profile distances to the current gene. Each gene that passes the cluster distance threshold test is indicated in the image with a blue square where the size of the square is proportional to its similarity. It also displays a sorted list of the genes with the cluster distance in the cluster panel that was popped up. On each lines is a series of '*****' - the more stars the higher the similarity to the seed gene. This is a silhouette plot that is used to display a sorted list of similar objects and is described to that described in (Kaufman and Rousseeuw, 1990). Larger squares indicate that more genes are similar. You may change the cluster distance threshold and it will update the display and the list. In addition, the 'edited gene list' is set to the subset of genes that belong to the current cluster.

A) Cluster of similar genes current seed gene in green, similar genes with blue boxes, similarity report on right.

B) Similar genes viewed in scrollable list of expression profile plots ordered by similarity.
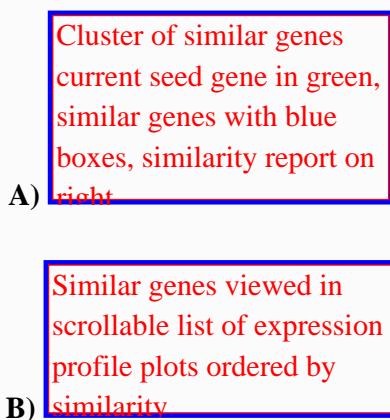
Figure 2.4.5.1 Similar genes clustered to the current gene. This method finds all genes that are similar to the current gene as those defined by their distance between expression profiles being less than the threshold set by the user. Each gene that passes the cluster distance threshold test is indicated in the image with a blue square where the size of the square is proportional to its similarity. This data is from the 38 samples in the MGAP database containing duplicated spots. **A)** Main windows with popup cluster similarity report and cluster distance threshold slider. **B)** Scrollable list of EPplots of similar genes with the red error bars indicating the variation for duplicated spots for each HP sample. The **Err** checkbox may turn the error bar overlays on and off.

2.4.5.2 Cluster counts of similar filtered genes by expression profiles

The **Cluster counts of similar Filtered genes by expression profiles** command analyzes the set of all Filtered genes for the expression profile defined by the HP-E samples. It counts the number of similar genes for each Filtered gene and draws a blue circle whose size is proportional to the number of genes similar to that gene. After it analyses these genes it lists the genes and their counts in the cluster panel. You may change the cluster distance threshold and/or Filter parameters and it will update the display and the list. If you click on a gene with a green circle, it will switch to single gene cluster mode (with the blue squares).

For both of these commands, if you want to view the expression profile plots, click on the **EP plot** button in the cluster window and it pops up the scrollable expression profiles window. If you click on a gene in the image, it will select it as the new current gene and seed gene and recompute the cluster of genes most similar to the new see gene.

For both of these commands, if you want a permanent report, click on the "Cluster Report" button in the cluster window and it will generate a report in the current modality (i.e. scrollable spreadsheet or tab-delimited). You may switch between these two modes by pressing the "Go '...'" button in the report.
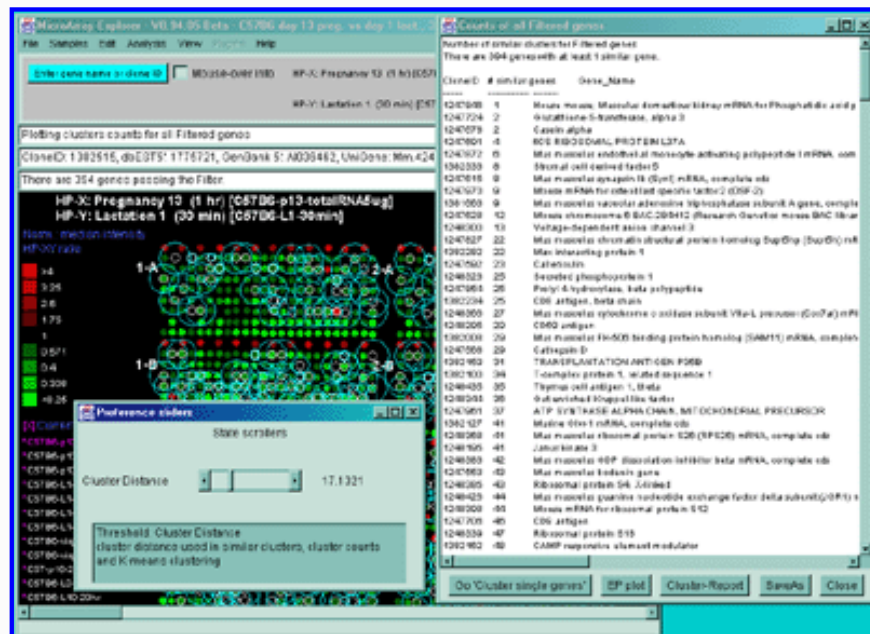
Figure 2.4.5.2 Display of cluster counts for all genes less than the cluster threshold from MGAP 38 sample database. The algorithm counts the number of similar genes for each Filtered gene and draws a blue circle whose size is proportional to the number of genes similar to that gene. That is why there are a larger number of the larger circles.

2.4.5.3 K-means clustering' gene expression profiles for filtered genes

The **K-means cluster gene expression profiles for Filtered genes** command searches the data Filtered gene list for the genes (i.e. primary genes) with the N most orthogonal expression profiles. It will start this recursive computation from the gene with minimum distance to all other genes unless you have selected a "current gene" with the mouse. All Filtered genes are assigned to the nearest K-means primary node. The mean cluster vector is computed and used as the new definition of the cluster center. If you set the "Use median instead of mean for K-means clustering" option in the Clustering submenu, it will compute the center as a median instead of a mean (Bickel, 2001). K-means clustering is described in (Sneath and Sokol, 1973). A new K-means primary gene (i.e. gene for the cluster center) is found that is closest to this new center. Then all of the data Filtered genes are reassigned to

the new cluster centers. The mean+-stdDev of the within-cluster distance to its center is computed. It then pops up a text window with an ordered report of the Filtered genes illustrated by part of a report shown below. [This is part of a report from a 38 sample MGAP database subset of 141 genes from the set of named genes restricted by the CV data filter.] Note that clusters where the "Similarity" data is plotted as a silhouette plot use variable length strings of '****' is about the same for the entire cluster (e.g. cluster #4) contain genes that probably belong together in the same cluster. Clusters that do not (e.g. Cluster 6) probably contain two smaller more robust clusters.
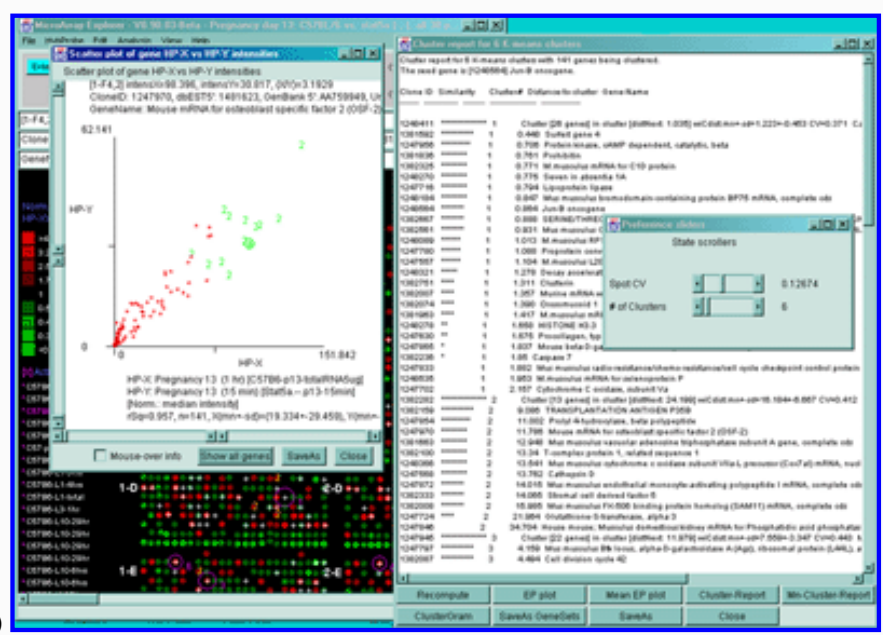
A)
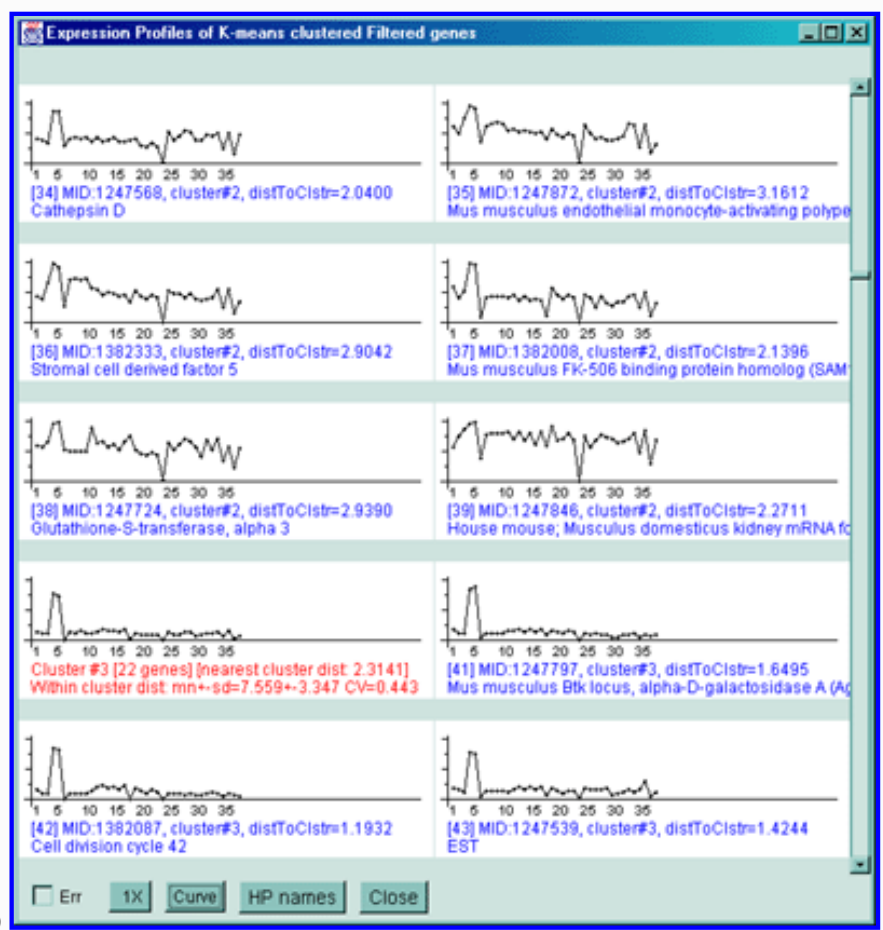
B)
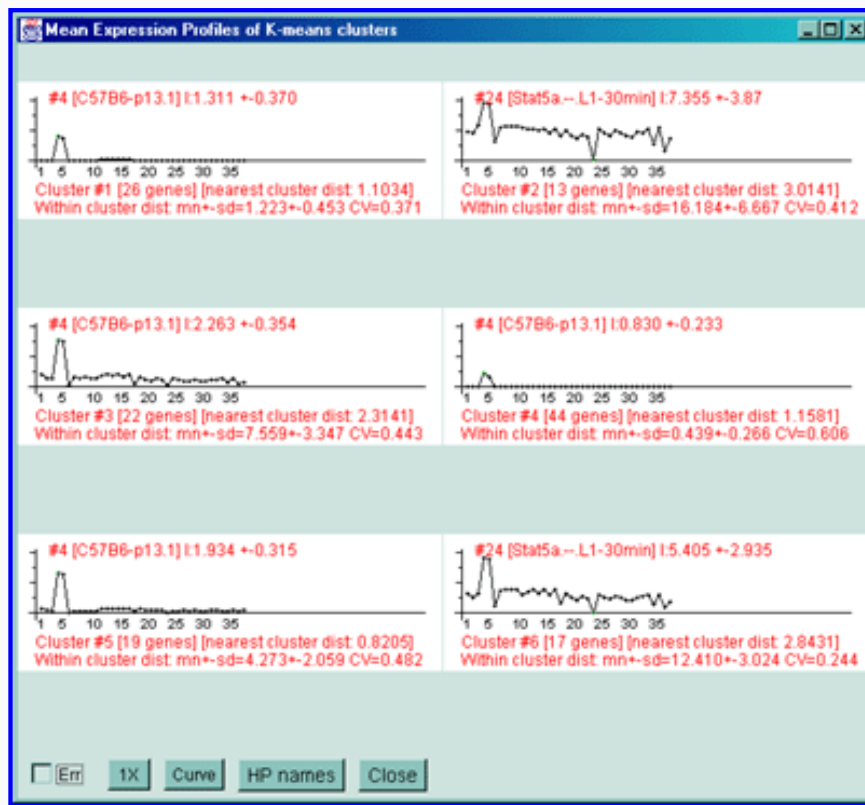
C)

Figure 2.4.5.3 Genes clustered using the K-means cluster method. **A)** Using the current gene as the initial cluster, MAExplorer finds N orthogonal clusters assigning the set of filtered genes to these clusters using the HP-E expression profiles. All genes are iteratively assigned to these clusters. Genes belonging to the current cluster are labeled with a green cluster number both in the array and in the scatter plot. The slider determines the number of clusters (set to 6 here). A 2D scatter plot shows the genes belonging to cluster 6. The K-means cluster report on the right contains a sorted list of the genes in each cluster and has buttons to generate EP plots and reports as well as summary mean EP plots (shown) and mean cluster reports. The detailed list is shown below. **B)** Part of the scrollable EP plots for this data showing genes belonging to both clusters #5 and #6. **C)** The mean EP plots for the 6 clusters.

```
Cluster report for 6 K-means clusters with 141 genes being clustered.
The seed gene is [1248564] Jun-B oncogene.

Clone ID  Similarity      Cluster-#  Distance-to-cluster  Gene-Name
--------  --------------  ---------  -------------------  ----------------

1248411   **************  1          Cluster [26 genes] in cluster [distNext: 1.035] wiCdist:mn+-sd=1.223+-0.453
CV=0.371  Calpactin I light chain
1381592   **********      1          0.448  Surfeit gene 4
1247956   *********       1          0.706  Protein kinase, cAMP dependent, catalytic, beta
1381836   ********        1          0.761  Prohibitin
1382325   ********        1          0.771  M.musculus mRNA for C1D protein
1248270   ********        1          0.775  Seven in absentia 1A
1247716   ********        1          0.794  Lipoprotein lipase
1248184   ********        1          0.847  Mus musculus bromodomain-containing protein BP75 mRNA, complete cds
1248564   *******         1          0.864  Jun-B oncogene
1382667   *******         1          0.888  SERINE/THREONINE PROTEIN PHOSPHATASE PP2A-BETA, CATALYTIC SUBUNIT
1382561   *******         1          0.931  Mus musculus GTP-specific succinyl-CoA synthetase beta subunit (Scs) mRNA,
partial cds
1248089   ******          1          1.013  M.musculus RPS3a gene
1247780   ******          1          1.088  Proprotein convertase subtilisin/kexin type 7
1247557   ******          1          1.104  M.musculus L28 mRNA for ribosomal protein L28
1248321   *****           1          1.278  Decay accelerating factor 1
1382751   ****            1          1.311  Clusterin
1382007   ****            1          1.357  Murine mRNA with homology to yeast L29 ribosomal protein gene
1382074   ****            1          1.390  Orosomucoid 1
1381963   ****            1          1.417  M.musculus mRNA for ribosomal protein L36
1248278   **              1          1.658  HISTONE H3.3
1247630   **              1          1.675  Procollagen, type I, alpha 2
1247865   *               1          1.837  Mouse beta-D-galactosidase fusion protein mRNA, complete cds
```

```
1382236    *                 1         1.85  Caspase 7
1247833                      1         1.882  Mus musculus radio-resistance/chemo-resistance/cell cycle checkpoint control
protein (Rad9) mRNA, complete cds
1248535                      1         1.953  M.musculus mRNA for selenoprotein P
1247702                      1         2.157  Cytochrome C oxidase, subunit Va
1382282    **************    2         Cluster [13 genes] in cluster [distNext: 24.199] wiCdist:mn+-sd=16.184+-6.667
CV=0.412  Max interacting protein 1
1382159    **********        2         9.086  TRANSPLANTATION ANTIGEN P35B
1247854    *********         2         11.002  Prolyl 4-hydroxylase, beta polypeptide
1247970    ********          2         11.786  Mouse mRNA for osteoblast specific factor 2 (OSF-2)
1381663    ********          2         12.948  Mus musculus vacuolar adenosine triphosphatase subunit A gene, complete cds
1382100    ********          2         13.34  T-complex protein 1, related sequence 1
1248366    ********          2         13.541  Mus musculus cytochrome c oxidase subunit VIIa-L precursor (Cox7al) mRNA,
nuclear gene encoding mitochondrial protein, complete cds
1247568    ********          2         13.762  Cathepsin D
1247872    *******           2         14.015  Mus musculus endothelial monocyte-activating polypeptide I mRNA, complete
cds
1382333    *******           2         14.065  Stromal cell derived factor 5
1382008    *******           2         15.985  Mus musculus FK-506 binding protein homolog (SAM11) mRNA, complete cds
1247724    ****              2         21.964  Glutathione-S-transferase, alpha 3
1247846                      2         34.704  House mouse; Musculus domesticus kidney mRNA for Phosphatidic acid
phosphatase, complete cds
1247945    **************    3         Cluster [22 genes] in cluster [distNext: 11.979] wiCdist:mn+-sd=7.559+-3.347
CV=0.443  Mus musculus mRNA for DEDD protein
1247797    **********        3         4.159  Mus musculus Btk locus, alpha-D-galactosidase A (Ags), ribosomal protein
(L44L), and Bruton's tyrosine kinase (Btk) genes, complete cds
1382087    **********        3         4.494  Cell division cycle 42
1247539    **********        3         4.511  EST
1248212    **********        3         5.009  Murine mRNA for integrin beta subunit
1248470    **********        3         5.044  EST
1247521    *********         3         5.299  Mus musculus mRNA for peroxisomal integral membrane protein PMP34
1381808    *********         3         5.924  Mus musculus UDP-GalNAc:polypeptide N-acetylgalactosaminyltransferase-T3
mRNA, complete cds
1381970    *********         3         6.285  Mus musculus thioredoxin mRNA, nuclear gene encoding mitochondrial protein,
complete cds
1382168    *********         3         6.343  N-terminal Asn amidase
1382704    *********         3         6.36  Mus musculus N-myristoyltransferase 1 mRNA, complete cds
1248548    *********         3         6.378  Mus musculus WDR protein mRNA, complete cds
1247564    ********          3         6.652  Erythrocyte protein band 7.2
1248588    ********          3         6.67  M.musculus BAP31 mRNA
1247541    ********          3         6.690  Apolipoprotein D
1248462    ********          3         7.322  Sterol O-acyltransferase 1
1248462    ********          3         7.42  Sterol O-acyltransferase 1
1248521    ******            3         9.121  Mus domesticus nuclear binding factor NF2d9 mRNA, complete cds
1382212    ******            3         10.137  Thyroid autoantigen 70 kDa
1382270    *****             3         10.529  Voltage-dependent anion channel 2
1248152    *****             3         10.541  M. musculus mRNA for MAP kinase-activated protein kinase 2
1247678                      3         19.431  Casein alpha
1247543    **************    4         Cluster [44 genes] in cluster [distNext: 1.035] wiCdist:mn+-sd=0.439+-0.266
CV=0.606  RAS-related C3 botulinum substrate 1
1381923    ************      4         0.158  Prolyl 4-hydroxylase, beta polypeptide
1382052    ************      4         0.209  Trans-acting transcription factor 1
1247882    ***********       4         0.237  Mus musculus AMP activated protein kinase mRNA, complete cds
1248099    ***********       4         0.246  Mus musculus mitogen-responsive 96 kDa phosphoprotein p96 mRNA,
alternatively spliced p67 mRNA, and alternatively spliced p93 mRNA, complete cds
1248351    ***********       4         0.251  Abl-interactor 1
1247540    ***********       4         0.255  Mus musculus mRNA for ZIP-kinase, complete cds
1248316    ***********       4         0.26  Mus musculus proteasome alpha7/C8 subunit mRNA, complete cds
1382671    ***********       4         0.264  Mouse MA-3 (apoptosis-related gene) mRNA, complete cds
1382014    ***********       4         0.277  Transcription elongation factor B (SIII), polypeptide 1 (15 kDa),-like
1247885    ***********       4         0.289  Mus musculus mRNA for ryudocan core protein, complete cds
1248294    ***********       4         0.292  Mus musculus thioredoxin-related protein mRNA, complete cds
1382066    ***********       4         0.306  Inhibitor of DNA binding 2
1248597    ***********       4         0.307  Lipocortin 1
1248591    ***********       4         0.324  Interferon beta, fibroblast
1248445    **********        4         0.333  Mus musculus beta prime coatomer protein mRNA, partial cds
1247775    **********        4         0.34  House mouse; Musculus domesticus male brain mRNA for ARF1, complete cds
1382750    **********        4         0.340  Thymoma viral proto-oncogene
1247905    **********        4         0.341  Monokine induced by gamma interferon
1381668    **********        4         0.351  Mus musculus mitogen-activated protein kinase-activated protein kinase mRNA,
complete cds
1381811    **********        4         0.356  Protein tyrosine phosphatase, receptor type, D
1382031    **********        4         0.358  Protease (prosome, macropain) 28 subunit, beta
1248345    **********        4         0.363  Mus musculus alpha-methylacyl-CoA racemase mRNA, complete cds
1382555    **********        4         0.364  Lysosomal membrane glycoprotein 1
1247820    **********        4         0.367  Tight junction protein 1
```

```
1247598   **********      4           0.374  Retinoblastoma 1
1247595   **********      4           0.378  PROBABLE CALCIUM-BINDING PROTEIN PMP41
1381928   **********      4           0.379  Mus musculus MRJ (Mrj) mRNA, complete cds
1248196   **********      4           0.399  Max protein
1381691   **********      4           0.423  SRY-box containing gene 17
1248225   **********      4           0.434  Mus musculus heat shock transcription factor 1 (Hsf1) gene, partial cds
1248084   **********      4           0.442  Mus musculus Supl15h gene
1247941   *********       4           0.453  Fibroblast growth factor inducible 14
1381623   *********       4           0.468  Stearoyl-coenzyme A desaturase 1
1248202   *********       4           0.473  Mouse mRNA for PAP-1, complete cds
1382115   *********       4           0.512  GLUTATHIONE S-TRANSFERASE GT8.7
1382044   *********       4           0.515  Cartilage derived retinoic acid sensitive protein
1381636   ********        4           0.567  Lymphotoxin B
1381920   ********        4           0.569  Mus musculus mRNA for NEFA protein, complete cds
1247757   *******         4           0.596  Granzyme B
1382094   *******         4           0.609  High mobility group protein 1
1247545   *******         4           0.638  Carbon catabolite repression 4 homolog (S. cerevisiae)
1247607   ***             4           1.188  POLYADENYLATE-BINDING PROTEIN
1247727                   4           1.667  Malate dehydrogenase, mitochondrial
1248244   **************  5                  Cluster [19 genes] in cluster [distNext: 3.473] wiCdist:mn+-sd=4.273+-2.059
CV=0.482  CD80 antigen
1248534   **********      5           1.648  Carbonyl reductase
1247764   **********      5           1.776  H-2 CLASS II HISTOCOMPATIBILITY ANTIGEN, GAMMA CHAIN
1381933   *********       5           2.345  Mouse rpS17 mRNA for ribosomal protein S17, complete cds
1381616   *********       5           2.42   Mus musculus oral tumor suppressor homolog (Doc-1) mRNA, partial cds
1248232   *********       5           2.486  Mus musculus putative glycogen storage disease type 1b protein mRNA,
complete cds
1382644   ********        5           2.717  Cyclin G
1248125   ********        5           2.791  Histocompatibility 2, class II, locus Mb2
1247799   ********        5           2.869  Mus musculus signal recognition particle receptor beta subunit mRNA,
complete cds
1247708   ********        5           3.024  Ephrin A1
1247932   ******         5           4.235  Mus musculus (clone: pMAT1) mRNA, complete cds
1382515   *****          5           4.668  ATPase, Na+/K+ beta 3 polypeptide
1248586   *****          5           4.838  Mus musculus viral envelope like protein (G7e) gene, complete cds
1248198   ***            5           5.874  Mus musculus D9 splice variant 2 mRNA, complete cds
1381623   **             5           6.224  Stearoyl-coenzyme A desaturase 1
1382086   *              5           6.885  Mus musculus (strain C57Bl/6) mRNA sequence
1247887   *              5           7.014  Mouse chromosome 6 BAC-284H12 (Research Genetics mouse BAC library) complete
sequence
1247886                  5           7.810  Cut (Drosophila)-like 1
1248303                  5           8.094  Lipopolysaccharide response
1247621   **************  6                  Cluster [17 genes] in cluster [distNext: 19.157] wiCdist:mn+-sd=12.410+-3.024
CV=0.244  Mus musculus Lsc (lsc) oncogene mRNA, complete cds
1248050   *******        6           7.407  Mus musculus C57BL/6J ribosomal protein S28 mRNA, complete cds
1247698   *******        6           7.571  Adipocyte protein aP2
1248240   *****          6           9.198  Mus musculus mRNA, complete cds
1247862   ****           6           9.844  Mus musculus Nmi mRNA, complete cds
1382162   ****           6           10.330  CAMP responsive element modulator
1248398   ***            6           11.007  Mouse mRNA for ribosomal protein S12
1248281   ***            6           11.143  M.musculus mRNA for histone H3.3A
1247852   ***            6           11.576  Twist gene homolog, (Drosophila)
1381991   **             6           12.809  Prolyl 4-hydroxylase, beta polypeptide
1382753   **             6           13.019  Mus musculus cleavage and polyadenylation specificity factor (MCPSF) mRNA,
complete cds
1248368   *              6           13.639  Mus musculus ribosomal protein S26 (RPS26) mRNA, complete cds
1247639   *              6           13.692  SRY-box containing gene 4
1248435                  6           14.262  Thymus cell antigen 1, theta
1247961                  6           14.75   ATP SYNTHASE ALPHA CHAIN, MITOCHONDRIAL PRECURSOR
1248344                  6           15.217  Gut enriched Kruppel-like factor
1382234                  6           16.351  CD8 antigen, beta chain
```

We call the genes closest to the "center" of the K clusters primary genes and they are reported with additional information. The "Cluster [# genes]" entries in the distance-to-cluster fields indicates that these genes are the center of the clusters (i.e. primary genes). The distNext is the distance from this cluster center to the next nearest K-means cluster center. The number of clusters N (6 in this example) is set in the popup state scroller. If you change the value of N, it will recompute the clusters and the primary genes.

It draws magenta circles around the primary genes in the microarray and the cluster number to the right of the circle. The size of a circle corresponds to the number of genes clustered with that circle. If you click on a gene belonging to any cluster, it defines that cluster as the "current cluster". It will change the labels of the subset of genes that belong to the current gene from red (white)

circle to a green (yellow) cluster number of the current cluster in the intensity (ratio) pseudoarray image. In addition, the 'edited gene list' is set to the subset of genes that belong to the current cluster. If you are also displaying a scatter plot, genes in the current cluster have their red '+' characters changed to the cluster number.

You can click on that gene in the array image to determine its identity. You may also popup an ordered (same as the above report) plot of the clusters expression profiles by clicking on the **EP plot** button. You may plot the mean expression profiles of the N clusters using the **Mean EP plot** button. You may generate a report of all of the clustered genes or of the mean clusters using the **Cluster-Report** or **Mn-Cluster-Report** buttons respectively. If you change the Filter conditions, you may recompute the clusters using the **Recompute Clusters** button. Closing the text window will remove the magenta circles. If you selected the current cluster, the genes that belong to it will still be available in the 'edited gene list' for making reports, saving as a gene subset or for additional gene filtering. If you press the **SaveAs GeneSets** button, then K gene sets are created with the names "Cluster#1", "Cluster#2", ..., "Cluster#K". You can then save or rename the clusters you want and delete the rest. If you press the **ClusterGram** button, it displays the gene sets in a cluster gram order the same way as the cluster report.

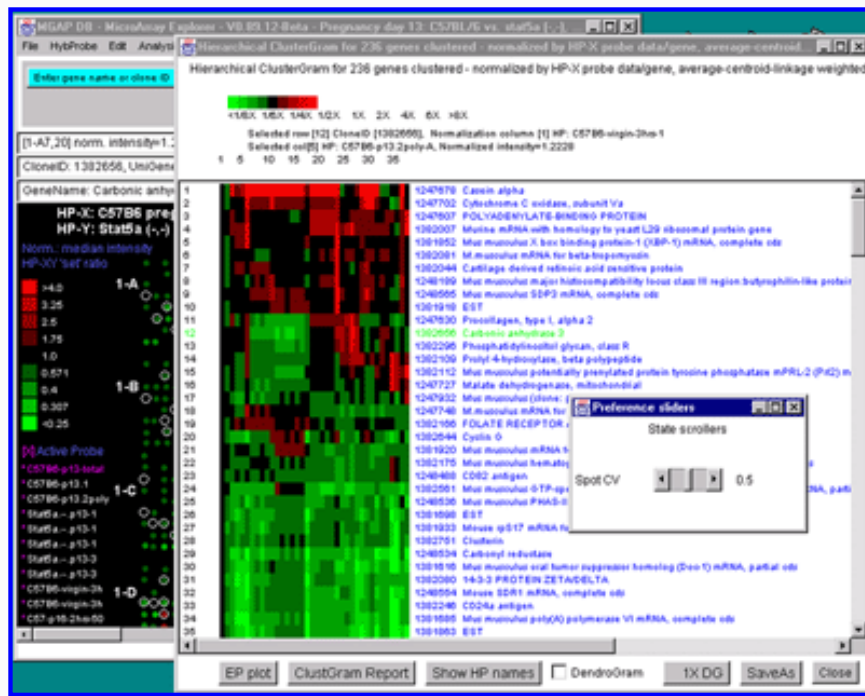## 2.4.5.4 Hierarchical clustering of expression profiles

The **Hierarchical clustering of expression profiles** computes the hierarchical clustering of the expression profiles of data Filtered genes and displays a clustergram and optional dendrogram. Hierarchical clustering is described in (Sneath and Sokol, 1973). The gene data is normalized either by the corresponding HP-X sample data for each gene or the maximum raw intensities for each HP sample in the expression profile set by the **Normalize by HP-X else HP's max intensities** menu toggle. There are three types of clustering linkages: **average-arithmetic-linkage**, **average-centroid-linkage**, and **next minimum linkage**. These may be modified using the **weighted average** that gives equi-weighting to the child clusters in computing the mean of a new cluster, and **un-weighted-average** that weights them by the number of non-terminal clusters. The average-linkage clustering is very compute intensive and takes a while. The next-minimum-linkage is much faster and may result in adequate clustering for some situations.

Clustering is represented by a binary tree and is visualized as an ordered gene clustergram and optional dendrogram sub-plot. This is similar to the methods of (DeRisi, 1996), (Eisen, 1998), and (White, 1999). Currently, MAExplorer does 1-way clustering - not the 2-way clustering of (Weinstein, 1998) and (Eisen, 1998). Each row of the clustergram represents a gene and each column represents a HP in the HP-E list of samples. Each box in a row represents the normalized expression of that gene for the HP represented in that column. The color of the box is one of 9 colors representing the normalized expression ranges and assigned according to the following table:

Table 2.4.5.4. ClusterGram pseudocolor assignments. The colors are assigned to "box" entries in the clustergram corresponding to genes. The color represents data as either the X/Y ratio or X-Y Zdiff relative to the normalizing HP.

bright green				Black	dark red			bright red
<1/8X	1/6X	1/4X 1/2X		1X	2X		4X 6X	>8X

The current gene may be set by clicking on a row that is then highlighted in green. If you click on a colored box, it will also report the HP name for that column and its normalized expression value (highlighting that box with a white circle). If the Web genomic databases are enabled (through the View menu, then it will also popup a Web page for that gene). If you set the current gene in any of the array, scatter plot, gene guesser, etc. displays, it will set it for and position the clustergram at that gene. If the **Dendrogram** checkbox is enabled, then a dendrogram is drawn to the left of the clustergram boxes. Clicking on a region in the dendrogram sets a distance threshold (displayed at the top) and displays all parts of the dendrogram tree in red that have a cluster distance less than what you defined. If the zoom **nnX** button is pressed, then the of dendrogram drawing is magnified by nnnn-fold to make highly similar clusters more visible. Pressing the button repeatedly cycles through: 1X, 2X, 5X, 10X, 20X. Sub-regions of the clustergram may be explored in more detail using the **EP plot** button that pops up a scrollable window of the ordered gene list. You may generate multiple EP-subset plots so as to compare different parts of the clustergram. A report of all of the ordered genes may be created using the **ClustGram Report** button. The **Show HP names** button pops up a numbered list of all samples used in the expression profiles and clustergram. This report has all of the normalized expression profiles on the right side of the report.
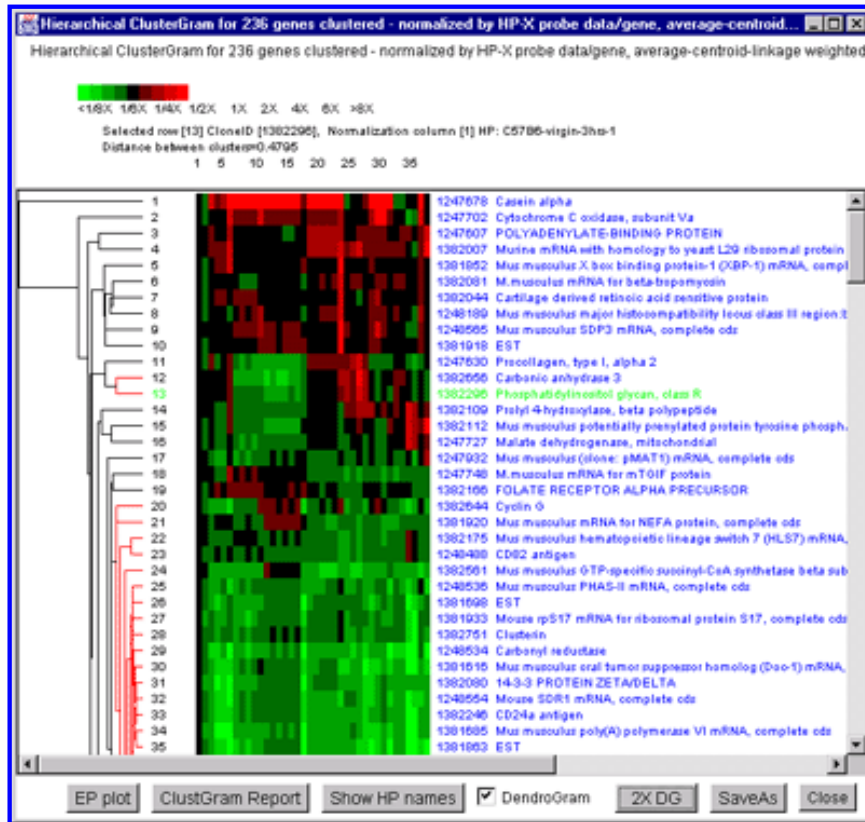
Figure 2.4.5.4 Hierarchical clustering clustergram of genes filtered by ratio histogram bins for 19 samples from the MGAP data set. The hybridized samples are drawn as colored boxes in the 19 columns. Rows of boxes correspond to gene expression profiles. In **A)**, the set of all genes and ESTs was filtered by the CV filter set to 0.387 and the normalization was the Zscore. The gene "Mus musculus D9 spice variant 2 mRNA, complete cds" was selected as the current gene in the clustergram. Data for this gene and the selected HP column is indicated at the top of the clustergram. The list of the 19 samples is shown on the left. **B)** Details of clustergram and dendrogram are shown where the user had selected a cluster distance threshold at "Mouse

mRNA for mitochondrial cytochrome c oxidase subunit Vb" in the dendrogram part of the plot (zoomed by 2X). This selection draws all parts of the dendrogram tree that are less than this distance are drawn in red. **C)** shows the manual selection of genes from the ClusterGram or Dendrogram by clicking on the genes names you wish to capture in the Edited Gene List (EGL) while the Control key is pressed. The zoomed subregion shows three genes in the same cluster that were selected (magenta stars in the right edge of the ClusterGram).

2.4.6 Report menu

Various reports summarizing gene or sample data may be generated and appear in popup tables. These include:

- Data on all of the hybridized samples in the database including links to related Web databases
- Data on additional quantification and descriptive information on the hybridized array samples in the database
- Hybridized sample calibration, Samples vs. Samples correlation coefficient of Filtered genes, and mean & variance tables
- Tables of all named genes, the calibration DNA genes, etc.
- A table of genes in the 'Edited Gene List'
- A table of genes in the current Gene Class
- A table of genes passing the "Filter"
- The N genes with the highest or lowest ratios (HP-X/HP-Y or F1/F2) of the Filtered genes (N defaults to 100, but may be changed as a Preference)
- Tables of additional data for the Filtered genes including expression profile ratios, HP-X vs. HP-Y 'set' statistics if the data is enabled for using HP-X and HP-Y 'set'
- Tables of correlation coefficient statistics of data Filtered genes comparing all samples in the HP-E list against each other.
- Tables are presented as either: **a)** an active "clickable" spreadsheet that may contain links to Web sites and pops up a Web browser window to that site (e.g. histology links or model system in the array report), or display additional gene features and numeric comparisons. **b)** tab-delimited text that may be "cut" and then "pasted" into an Excel type spreadsheet for further analysis.
- Additional expression profile data (from the HP-E samples) and/or statistical data from the HP-X vs. HP-Y 'sets' may be included in gene report tables.
- Clicking on a column name in the dynamic report will sort the report by data in the report in ascending and descending order.

Samples Reports menu

Figure 2.4.6 Reports menu. You may create either dynamic or tab-delimited text reports of either Samples or of subsets of genes.

These may be presented as interactive dynamic tables as well as scrollable text windows capable of being exported to Excel. If Web DB access is enabled, clicking on an entry will bring up a Web browser with access to GenBank data. If the report contains Clone ID as one of the fields, you can click on it to have it define that gene as the current gene and highlight it in the microarray image or scatter plot (if it is being used). The reports are divided into two types - those dealing with lists of arrays (i.e. the sample experimental condition) and those dealing with lists of genes.

The **Report** menu includes:

- **Array reports** - show tables of sample array information.
- **Gene reports** - show tables of sets of genes information.
- -------------------

- **Table format** - generates the table as 'Spreadsheet', 'Tab delimited', or 'Name=Value list'.
- **Table font size** - changes the report font size from default 10 point.

2.4.6.1 Array report menu - hybridized samples global data

You may generate reports of sample array information. The first two menu selections contain descriptive information about specific hybridized microarrays samples. The "Extra Samples info" contains quantitative and extra descriptive information (if available for your database).

The "Samples vs Samples correlation coefficients" computes the correlation coefficients in an upper diagonal matrix for the current set of Filtered genes showing HP samples similarity. Then entries are of the following form where HP:1 and HP:2 correspond to samples listed in the field names of the table and the data is the intensity values using the current normalization method.

```
rSq=0.748, n=1656, HP:1(mn+-sd)=(28991+-19564), HP:2(mn+-sd)=(5044+-9766)
```

The "Calibration DNA summary" table contains the computed means, std-dev, and computed normalization scale factor for all active hybridized samples. The scale factors are used if the 'Calibration DNA' normalization is used.

You must set the Web access checkbox if you want to click on a blue hyperlink in the resulting report to access an associated Web database.

- **Samples info** - show table of all samples accessible in the database.
- **Samples Web links** - show table of Web links for all sample samples accessible in the database.
- **MAE Projects DB** - show table of project databases if the project startup database exists (stand-alone mode).
- **Extra info on Samples** - show table of additional information on all of the samples in the database.
- **Sample vs Sample correlation coefficients** - show table of correlation coefficients for all HP-E selected samples for genes meeting the data Filter criteria.
- **'Calibration DNA' summary** - show a table of means, std-dev, scale factor for all HP-E selected samples.
- **HP mean & variance summary** - show table of means, std-dev, min, max, etc. of raw intensity data for all HP-E selected samples.

A)

B)

C)

Figure 2.4.6.1 Hybridized samples dynamic Report windows. A) Samples Info report. **B)** Sample Web links. Clicking on a blue hypertext link brings up the corresponding genomic Web database entry in a separate Web browser window if the Web access is enabled. The tab-delimited version of the same reports (not shown) may be cut and then pasted into other programs such as an Excel spreadsheet. **C)** HP vs HP correlation table on genes passing the data Filter for all samples in the HP-=E list.

2.4.6.2 Gene reports menu

You may generate gene reports with various additional options. You must set the Web access checkbox if you want to click on a blue hyperlink in the resulting report to access an associated Web database. In addition, specialized gene reports may be generated from some of the cluster plot command windows. These include lists of genes sorted by cluster (K-means cluster #), by hierarchical cluster order, by similarity to a gene, etc. The mean cluster expression values may be reported for K-means clustering.

- **All named genes** - show table of all genes where the gene is named.
- **Genes in 'Edited Gene List'** - show table of genes in the 'Edited Gene List'.
- **Genes in 'Normalization Gene List'** - show table of genes in the 'Normalization Gene List'.
- **Genes in GeneClass** - show table of genes in the current gene class.
- **Filtered gene reports** ▷ - show genes that meet the data Filter criteria.

2.4.6.2.1 Filtered gene reports menu

You may generate gene reports of Filtered genes with various additional presentation options. In the highest/lowest N genes, N defaults to 100 and is set by (Report | Table format | Set max # genes in highest/lowest report) command.

- **Genes passing the Filter** - all genes passing the Filter
- **Highest HP-XY ratios or Zdiff** - top N Filtered genes
- **Lowest HP-XY ratios or Zdiff** - lowest N Filtered genes
- **Highest F1/F2 ratios or Zdiffs** - top N Filtered genes (if duplicates). If ratio data, then highest Cy3/Cy5.
- **Lowest F1/F2 ratios or Zdiffs** - lowest N Filtered genes (if duplicates). If ratio data, then lowest Cy3/Cy5.
- **Expression profiles Filtered genes** - show numeric expression ratios for HP-E data scaled to the first HP in HP-E.
- **HP-XY 'set' statistics of Filtered genes** - show *(mean,stdDev,CV,n)* data for both the HP-X and HP-Y sets.

If Cy3/Cy5 ratio data is being analyzed, then the Highest (Lowest) F1/F2 entries become

- **Highest Cy3Cy5 ratios or Zdiffs** - top N Filtered genes (if duplicates)
- **Lowest Cy3Cy5 ratios or Zdiffs** - lowest N Filtered genes (if duplicates)

A)

B)

Figure 2.4.6.2 Gene Report windows of 50 named genes with highest HP-X/HP-Y 'set' ratios. A) Dynamic gene report of 50 genes with highest HP-X/HP-Y 'set' ratios. A similar report may be generated for the lowest ratios or for single HP-X/HP-Y samples. This type of report may be generated for

the highest or lowest Zdiff values when the Zscore normalizations are used. Clicking on a blue hypertext link brings up the corresponding genomic Web database entry in a separate Web browser window if the Web access is enabled. It also sets the current gene to the gene for that row. **B)** The tab-delimited version of the same report may be cut and then pasted into other programs such as an Excel spreadsheet.

2.4.6.3 Table format menu

The report is presented as a table. However, it may be visualized several different ways. The scrollable spreadsheet includes the ability to click on blue hypertext items and have a Web browser pop up for that item on a Web database (e.g. GenBank, dbEST, UniGene, LocusLink, mAdb Genes, GeneCard, etc). The tab-delimited option enables you to cut the table and paste it into a separate spreadsheet program such as Excel. You may also extend the data in the table to by 'Adding' expression profile ratios and statistics from the HP-X and HP-Y 'set' comparisons.

- ☑ **Spreadsheet** - some cells (indicated by blue) are connected Web databases that are accessed by clicking on them (default).
- ☐ **Tab delimited - suitable for export** - as in Excel, etc.
- --------------------
- ☐ **Set max # genes in highest/lowest report or filter** - sets the number of genes N to report or use in data Filter.
- ☐ **Add EP data to Gene-Reports** - for each of the samples in HP-E.
- ☐ **Use EP data in Gene-Reports** - use raw EP data (under the current normalization) else use data normalized to 0.0 to 1.0 by maximum value for all genes being displayed.
- ☐ **Add HP-X/-Y 'set' statistics data to Gene-Reports** - that includes *(mean,stdDev,CV,n)* for both the HP-X and HP-Y sets.

2.4.6.4 Table font size menu

For wider tables, you can see more information if you use a smaller font to display the table. The font sizes available are:

- ☐ **12pt**
- ☑ **10pt** - (the default size)
- ☐ **8pt**

2.5 View menu

The **View menu** options are used to modify the view of genes visible in the pseudoarray image. Genes may be displayed with additional properties or capabilities including access to Web-based genomic database entries for specific genes. Note that depending on your particular database, if some genomic identifiers are not available then the corresponding "Enable display current gene in *genomic DB* Web browser" will not appear in the menu.
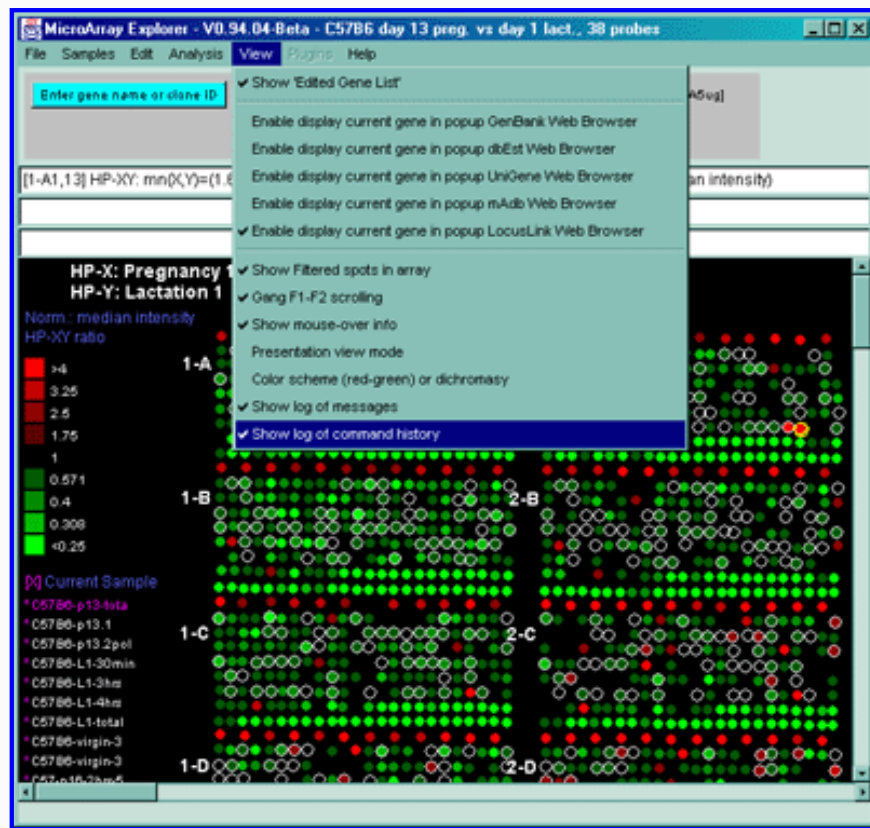
Figure 2.5 View Menu options. These are divided into various options for modifying the presentation as well as recording activity such as the messages or history popup scrollable log windows.

- ☐ **Show 'Edited Gene List'** - toggle showing the EGL as magenta boxes in the pseudoarray image. If enabled, genes set by manual selection or as the result of some filtering operations (see Edit menu, the Filter and Clustering)
- ----------------------
- ☐ **Enable display current gene in GenBank Web Browser** - when click on a gene in microarray image or scatter plot to access NCBI data.
- ☐ **Enable display current gene in dbEST Web Browser** - when click on a gene in microarray image or scatter plot to access NCBI data.
- ☐ **Enable display current gene in Unigene Web Browser** - when click on a gene in microarray image or scatter plot to access NCBI data.
- ☐ **Enable display current gene in mAdb Web Browser** - when click on a gene in microarray image or scatter plot to access NCI/CIT data.
- ☐ **Enable display current gene in LocusLink Web Browser** - when click on a gene in microarray image or scatter plot to access NCBI LocusLink data (by Locus ID if available, and GenBank ID if it] is not).
- ----------------------
- ☑ **Show Filtered spots in array** - if on, show Filtered genes in the microarray image using circle overlays just outside of each spot. Genes not passing the Filter have no circle.
- ☑ **Gang F1-F2 scrolling** - toggle gang scrolling. If on, gang measure F1 and F2 when click on one of them. If off, only measure only the spot that you click on.
- ☑ **Show mouse-over info** - toggle mouse over. If enabled, report HP or gene details when the mouse is moved over the sample names or spots in the array image or genes in scatterplot or expression profile overlay plot.
- ☐ **Presentation view mode** - toggle presentation mode. If enabled, increase the fonts to 12pt and draw darker circles, squares, and "+" so the display details are easier to see when projecting the display or making slides.
- ☐ **Color scheme (red-green) or dichromasy** -toggle between two color schemes red-green and orange-blue for

dichromasy.

- ☐ **Show log of messages** - toggle message logging mode. If on, it pops up a scrollable log of all messages to the three line status area. This is useful for recording measurements and other activity. The messages may be saved in log file.
- ☐ **Show log of command history** - toggle command history logging mode. If enabled, it pops up a scrollable history of all commands issued to MAExplorer. The commands are automatically numbered. This is useful for recording what steps you took during an analysis. The history may be saved in log file.

A) Example of UniGene popup Web browser page for current gene
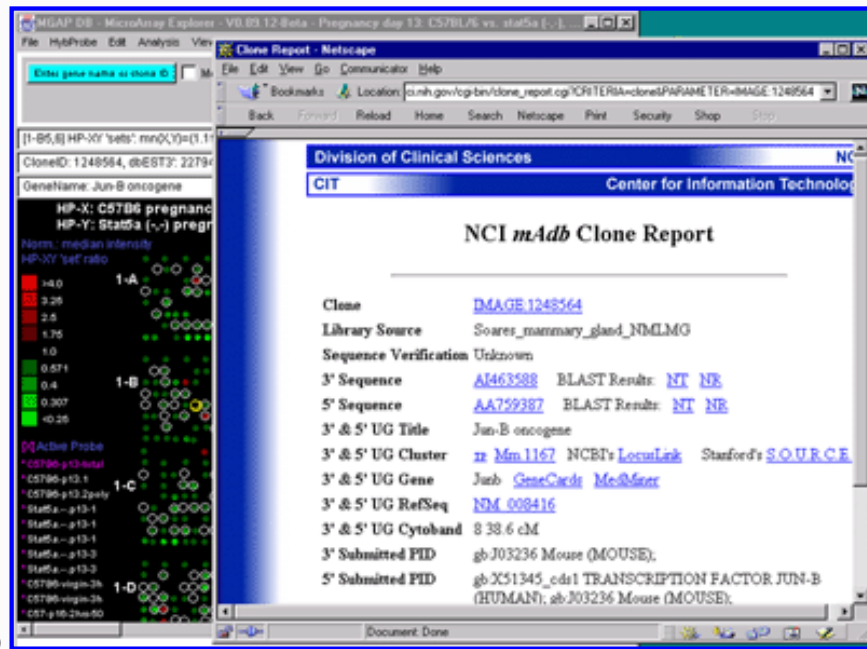
B)
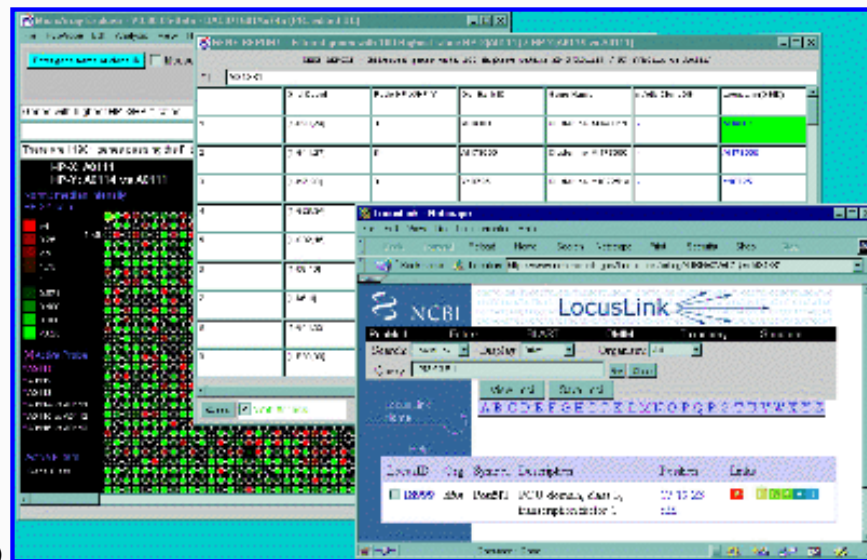
C)

Figure 2.5. Popup genomic browser database page. A) The UniGene Web page pops up in a new Web browser window when the user clicks on a gene in the array image, 2D scatter plot or Report and the view is set to "Display current gene in Unigene Web Browser" toggle was enabled in the View menu. The current gene was "Jun-B oncogene". Alternatively, the **B)** mAdb Gene DB may be selected - as well as GenBank or dbEST genomic databases. **C)** Alternatively, data from the NCBI LocusLink database may be accessed if either the GenBank ID or LocusID is available.

2.5.1 Logging MAExplorer messages

MAExplorer shows various data measurements as well as many other types of information in the three text lines in the status area of the main window. The **Show log of messages** pops up a scrollable log of all messages to the three line status area. This is useful for recording measurements and other activity. The messages may be saved in log file (typically maeMessages.log). Figure 2.5.2 shows an example of the messages popup log window. Clicking on genes in the pseudoarray image or in plots will log the gene data (see Section 3.3) given the current normalization, Samples use (single or multiple), and pseudoarray display mode. The current values of all of the State Threshold scrollers are saved in the message log when the (Edit menu | Preferences | **Adjust all Filter threshold scrollers**) State Thresholds popup window is closed. This is useful for capturing the current settings at any time.

2.5.2 Logging command history

During a datamining session, the user will typically execute many commands from the menu as well as clicking on genes in the pseudoarray image or in plots. It is useful to recording the steps you took during this analysis. The **Show log of command history** pops up a scrollable history of all commands issued to MAExplorer. The commands are automatically numbered. The history may be saved in log file (typically maeHistory.log). Figure 2.5.2 shows an example of the command history popup log window.
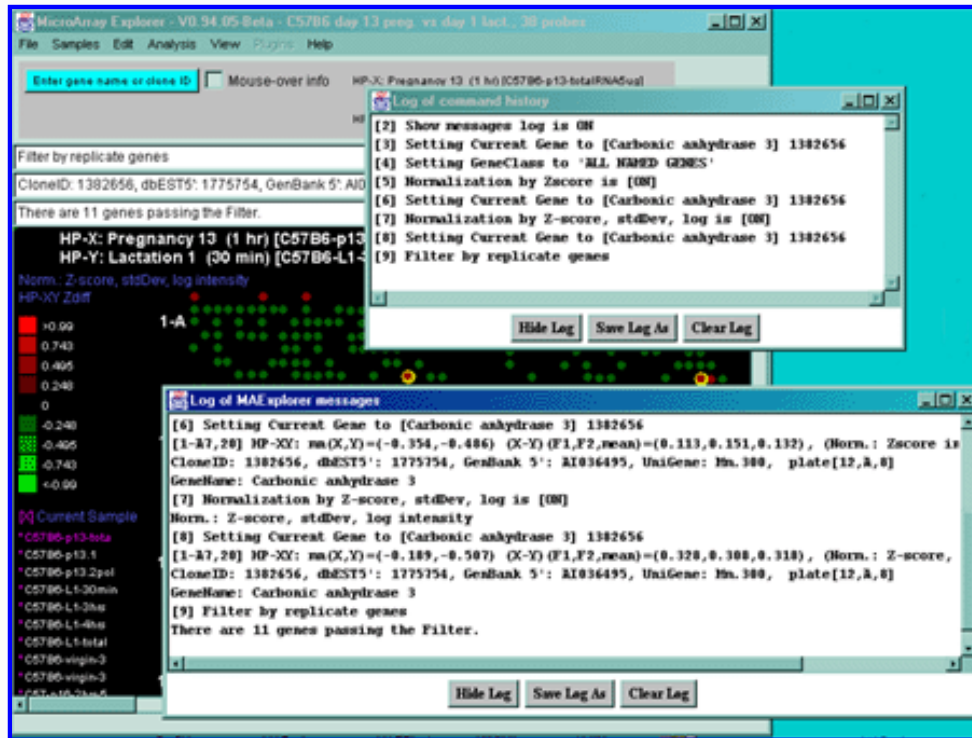
Figure 2.5.2 Examples of messages and command history popup log windows. Measurements and other activity are shown in more detail in the messages window whereas the command history indicates commands (numbered in the order they are executed) in the command history window. Data from either of these windows may be saved in text log files.

2.6 Plugins menu

MAExplorer may be extended by users to use new analysis methods using Java plugins. We call these new methods MAEPlugins which are small Java programs written by users that may be dynamically loaded into MAExplorer and then applied to their data.

These plugins will include plugins written by LECB, those written by academic or commercial groups. See the [MAEplugins](#) for details. If you have a Java compiled plugin in the form of either a Java .class or .jar file, you may load it at run time using the "Load plugin" command in the Plugins menu. If specified in the MAEPlugin, it will be added to the appropriate menu in the MAExplorer menu tree at the end of the specified submenu (see Appendix C. [Table C.5.7](#)). If this submenu "stub" is not specified, it will place in the list of plugins in the Plugins menu (e.g. **plugin #1**, ..., **plugin #n**).
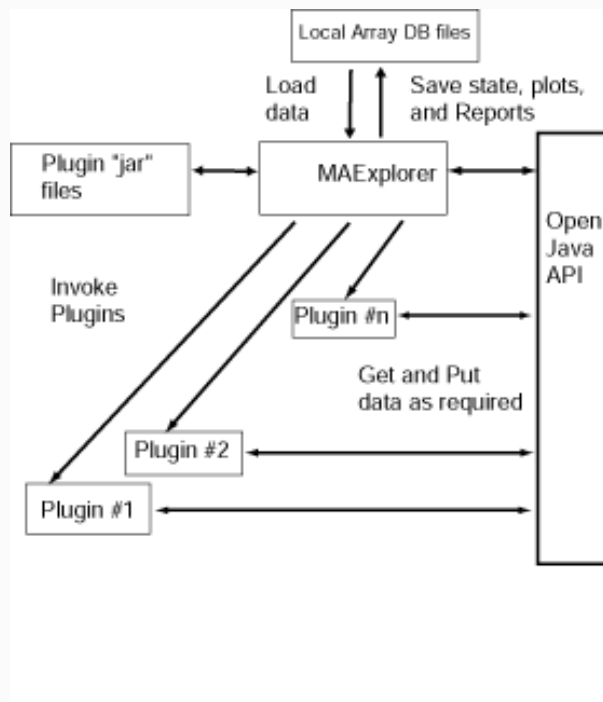
Figure 2.6 MAEPlugins paradigm. If you have a MAEPlugin .jar file, then it may be specified using the "Load plugin" command. When you invoke the command from the menus (or other methods), it accesses data from the current MAExplorer database it may need from the Open Java API.

The **Plugins** menu includes:

- **Load plugin** - popup a browser to load a MAEPlugin jar file
- **Unload plugin(s)** - popup a selector to pick MAEPlugins to unload
- -------------------------------------
-
- **plugin #1** - 1st MAEPlugin added
- **plugin #2** - 2nd MAEPlugin added
- **. . .**
- **plugin #n** - n'th MAEPlugin added

2.6.1 Example of using a Plugin

This shows a short demonstration of what is involved in using a MAEPlugin. The user first load the plugin from the disk. Generally the plugins .jar or .class files are stored in the Plugins/ directory where you have installed MAExplorer. Then they load a particular plugin which installs it in the Plugins pull-down menu. Then they revisit that menu to invoke the particular plugin. You may load any number of plugins (until you run out of computer memory if that should occur).
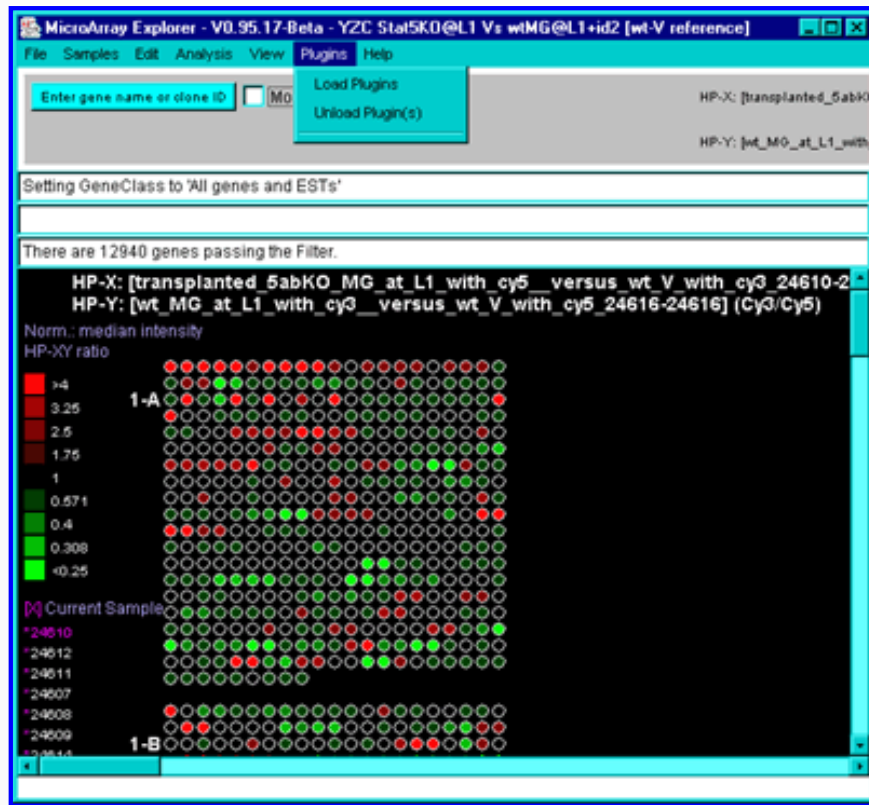
Figure 2.6.1 Loading a MAEPlugin from your file system using the Load Plugins command in the Plugins pull down menu. If you have a plugin .jar or .class file, it may be specified using the "Load plugin" command. This pops up a file browser to let you specify the plugin file.
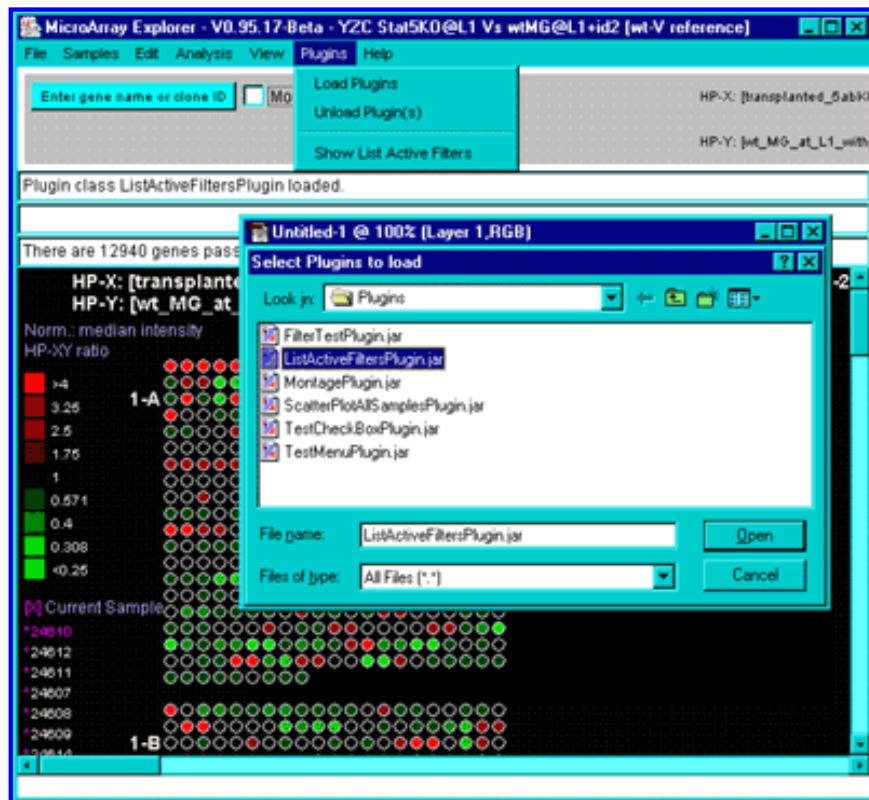
Figure 2.6.2 Executing the new command previously loaded in the Plugin menu. Selecting the new "Show List Active Filters" command that now appears in the Plugins menu invokes the plugin. This pops up a report shown in the next figure.
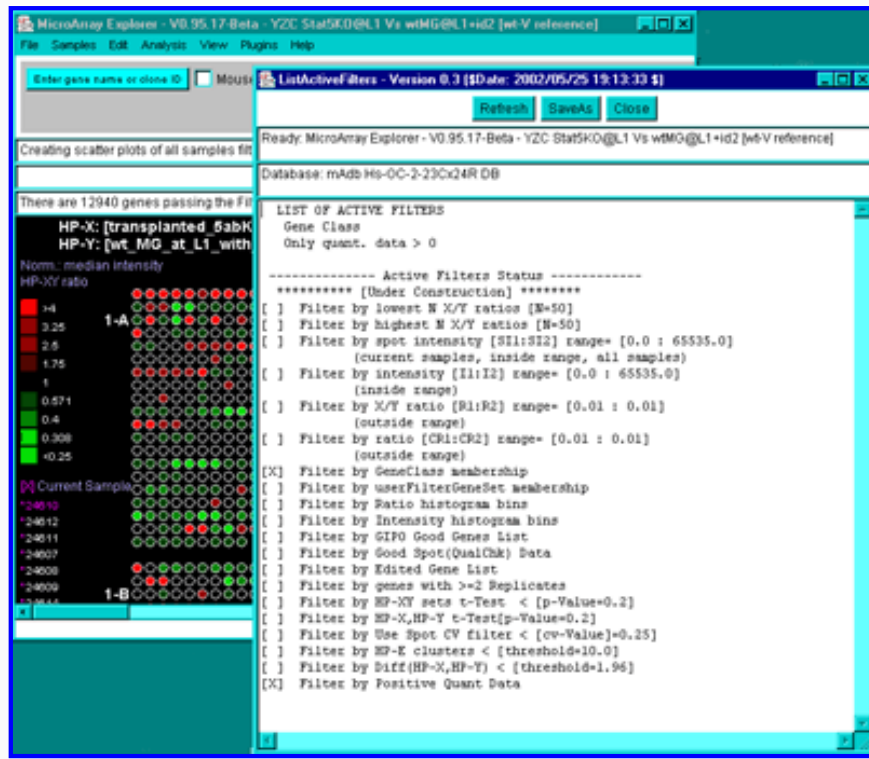
Figure 2.6.3 Popup window from executing the MAEPlugin. This plugin gives a full report on the data Filter status in a new pop up window.
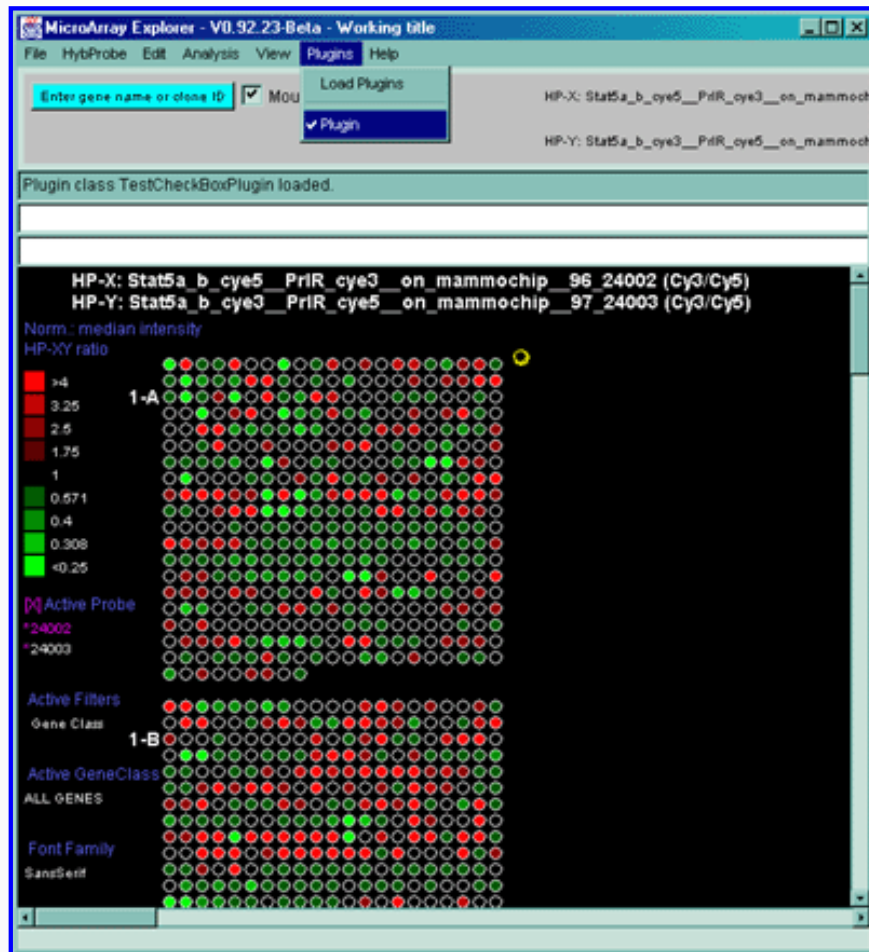
Figure 2.6.2 Plugins menu - executing a previously loaded plugin. Plugins that do not go into particular MAExplorer submenus go into the Plugins menu. Selecting the command will invoke that MAEPlugin.

2.7 Help menu

Various on-line help and documents are available if you are connected to the Internet. These will appear in a separate pop-up Web browser window so you may view them while working with MAExplorer. This includes on-line documentation (including this reference manual), tutorials, and other information. This may be links to other Web pages describing key areas of specific databases. For example, for the MGAP database, the point back to key areas of MGAP including the MGAP Animal Models, Histology atlas, etc. You can then use the browser's "Save as" and "Print" options to save the data to a file or print it.

The **Help** menu includes:

- **Overview** - overview of MAExplorer capabilities
- **Short tutorial** - simple tutorial of things to try in MAExplorer (Appendix A)
- **Advanced tutorial** - advanced tutorial of things to try in MAExplorer (Appendix B)
- **Menu summary** - short summary of the MAExplorer menus.
- **Reference manual** - this Reference Manual document
- **Glossary** - glossary of terms used in MAExplorer Reference Manual
- **Index** - Index of terms used in MAExplorer Reference Manual
- ------------------------------------
  > **Database-specific help menu entries** -
  > entries defined for a particular database (see below)
- ------------------------------------
- **About** - show MAExplorer authors, version and revision date

2.7.1 Adding custom help links to your database to the Help menu

These **Database-specific help menu entries** list of entries are keyed to the database you are using and may be customized by the database maintainer in the configuration file (Section C.5.6) to links relating to the particular database. For example, database specific help for the MGAP database is:

- **List of hybridized samples** - available from MGAP.
- **MGAP animal models** - animal models used in MGAP.
- **MGAP home page** - Mammary Genome Anatomy Program

3. Exploratory Data Analysis - Data Mining

Data mining is the uncovering of relevant patterns of interest in data from a particular problem domain ([Tukey, 1977](#)). Typically this involves using various statistical techniques to identify the patterns including cluster analysis. Researchers across a wide range of fields such as ([Tufte, 1997)](#)) and ([Cleveland, 1985](#)) have suggested that a major aspect of this problem is finding the correct means of graphical presentation to allow humans to be a part of the pattern recognition process. Tufte argues that the proper display of quantitative data in the context of the problem domain can aid in the understanding of complex sets of data. This carries over to the analysis of microarrays with data mining involves having statistical, genomic knowledge database, and graphical components for success. ([Jagota,2001](#)) discusses a number of methods and applications for microarray data analysis and visualization.

This section will briefly address some of the issues to consider. However, a full understanding of the issues involved is beyond the scope of this manual. These issues are covered in other focused statistical methods literature and might also be addressed in consultation with biostatisticians. The Internet has vast resources for microarrays. A few to get you started might include: a microarray citation electronic library http://arrayit.com/e-library/, the National Library of Medicine [PubMed](#) journal search engine, a general microarray Listserv GENE-ARRAYS@ITSSRV1.UCSF.EDU. The MGED group ([Brazma, 2001](#)) has published the MIAME standard which specifies (Minimum Information About a Microarray Experiment). This information is useful in doing an analysis. Also try searching using general Internet search engines.

- Data mining is a pattern discovery activity - use all the tools you have.
- It is open-ended because of the variety of ways data may be partitioned, normalized, pre-filtered, clustered, and viewed. Patterns that are apparent in one view may not be apparent in another.
- When data-mining microarray data, look at correlated genes from the point of view of what relationships might be interesting from a biological view by characterizing genes that cluster together using additional information. I.e. check out the results with various NCBI and PubMed database searches on the resulting genes, by designing other lab experiments to better uncover the cause and effect, etc. Remember correlation does not imply cause and effect.

Organization of Sections in this Chapter

- 3.1 discusses [Objectives in data mining, discovery and analysis](#),
- 3.2 describes the [steps in an analysis](#),
- 3.3 describes how the database may be interrogated for [gene spot intensity & identification](#).
- 3.4 describes [selecting subsets of genes using the data filter](#).
- 3.5 describes [selecting subsets of hybridized sample conditions](#).
- 3.6 describes [setting thresholds using state-slider controls](#).
- 3.7 describes how to [export (i.e. save) report and plot data](#) to your local computer.

3.1 Objectives in data mining, discovery and analysis

There are a number of objectives an investigator has when analyzing a set of data. The types of analyses and how useful they are depends on what they wish to get out of the analyses as well as the type of data.

A good and appropriate experimental design (i.e. the design and setting up of experiments to subsequently be analyzed) is critical for resolving significant differences in gene expression between experimental conditions. We touch on some of the issues here. ([Simon, 2001](#)), ([Dudoit,2000](#)), and Kerr and Churchill ([2001a](#), [2001b](#)) discuss some of the issues of experimental design for microarrays. We do not currently implement the Kerr-Churchill method. However, some of the issues involved in [experimental design](#) based on the types of arrays are discussed in Section 3.1.1 for (Cy3/Cy5)-labeled as well as ^{33}P-labeled samples.

If users are comparing two different types of samples, the analysis would be different than if they were comparing an ordered sequence of samples (e.g. time series, cell cycle, dose-response, tumor-stage, etc.). MAExplorer gives users the ability to:

1. Organize their experiments by sample characteristics allowing them to perform a variety of mining analyses comparing gene expression patterns between sets of different samples or comparing a single sample within an array's set of genes.

2. Explore, compare, and record these analyses and share the results or intermediate data with other investigators.
3. Use graphical direct-manipulation combined with statistical methods, clustering and spreadsheet techniques to gain different insights into the structure of the data.
4. Access public Internet genomic databases for particular genes that are found to be interesting.

Briefly, data mining is the discovery of potentially interesting patterns in the data that were previously unknown. One approaches the analysis of a set of data with minimal expectations. However, some idea of what you are interested in helps focus the search. But beware of the trap of mining the data until you get the results you hope for. The following figure helps illustrate this process.
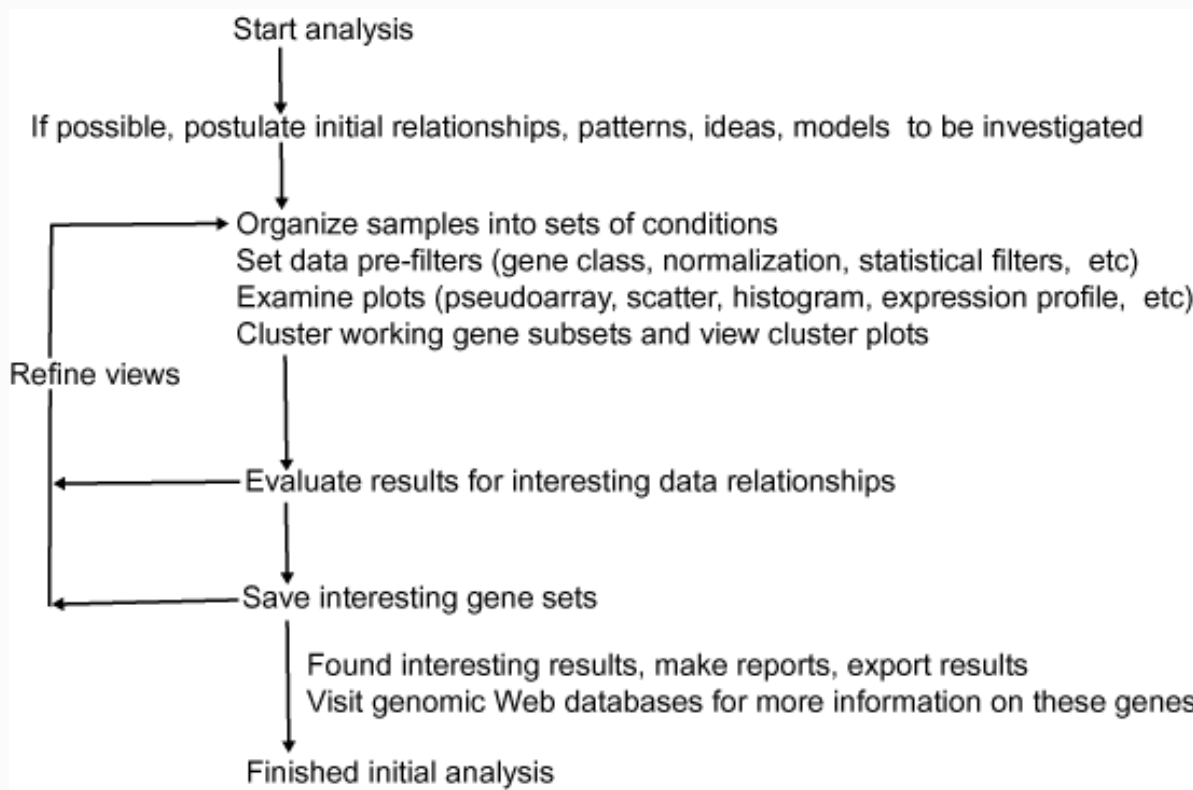
Figure 3.1 Flow chart of a typical data mining session. The user makes some initial decisions on the experimental design such as which hybridized samples to compare, the type and numbers of replicates. They then make initial guesses as to the normalization method to use, and the gene subset (the gene class) to concentrate on when setting the data filter. The data is viewed in various modalities to get a feeling for its inherent dynamic range and where interesting outliers might appear. Clustering and plots helps bring these differences into view. The results are then evaluated and either the process is finished or the views are refined by adjusting data normalization and filter parameters, data subsets to be investigated, clustering methods, plots etc. and the process repeated until the user is able to see the differences between gene subsets more clearly or no significant differences appear to be found.

Obviously, this approach is a first approximation to what is eventually required. But it does capture the flavor of the data-mining process. Typically the user would refine the search using variations of the data filters and might contrast (using gene sets and hybridized sample condition lists operations) results found under one set of conditions with those found under another set of conditions.

Recording the analysis steps during your data mining session - command history

Because of the iterative nature of this process, you might want to keep a record of the commands you have used or the messages and measurements you have made. To do this you need to enable message and command history logging. Go to the View pull-down menu and then select the type of logging you want using the Show log of messages or the Show log of command history commands.

3.1.1 Some experimental design issues of microarray experiments

*** THIS SUBSECTION IS IN THE PROCESS OF BEING UPDATED ***

Proper experimental design of microarray experiments is critical to successful use of microarray data. Several recent reports discuss some of the key issues involved in various aspects of statistical analysis of microarrays: (Radmacher, 2001), (McShane, 2001), (Korn, 2001), (Simon, 2001), (Dudoit,2000).

Comparing HP-X/HP-Y for Cy3/Cy5 data as 'ratio of ratios'

If we have two samples HP-X and HP-Y with a common reference sample P (e.g. $Cy5_P$), then we would be comparing the HP-X "intensity" $Cy3_X/Cy5_X$ against the HP-Y "intensity" $Cy3_Y/Cy5_Y$. Alternatively, you can label Cy3 as the common reference sample P in which case just swap Cy3 and Cy5 in these equations. If you are using a common reference standard (i.e. $Cy5_{X1}$) is the same sample as $Cy5_{Y1}$ eg. a pooled sample $Cy5_P$, then

```
 a)    (Cy3_X/Cy5_X1) / (Cy3_Y/Cy5_Y1)
becomes
 b)    (Cy3_X/Cy3_Y)
```

However, this new comparison is accompanied by additional noise because of use of the two $Cy5_P$ intermediaries.

An alternative method would be to compute $(Cy3_X/Cy5_Y)$ directly. However, this too has its own sources of error and other problems, namely that not all genes are labeled symmetrically with the two dyes since different dyes may have different sequence specific affinities due to a variety of causes. For that reason, dye-swap experiments are often done. I.e. the two samples would be run as $(Cy3_X/Cy5_Y)$ as well as $(Cy3_Y/Cy5_X)$. If one were to plot $(Cy3_X/Cy5_Y)$ against $1.0/(Cy3_Y/Cy5_X)$ and the data were perfectly symmetric (which they are *not*) then one would expect a straight line. That is generally not what you get in practice.

Another issue is that when you have a number of samples A, B, C, D, ..., N and wish to compare them, there are a number of alternate experimental designs you can use with different resulting sets of advantages and problems. If a common pooled $Cy5_P$ sample P were used, then the following experiments would be done:

$$(Cy3_A/Cy5_P),\ (Cy3_B/Cy5_P),\ \ldots\ ,\ (Cy3_N/Cy5_P)$$

This assumes that there is enough of the pooled sample P to be used for *all* of the experiments - otherwise additional sources of error would be introduced. MAExplorer is ideally used with this common reference sample P. It a common pooled sample is *not* used, then the experimental design becomes more complicated - especially if dye-swap experiments are performed for all samples. For N samples taken 2 at a time (i.e. Cy3 and Cy5), then the number of experiments may be impossibly large to perform for other than a very small N. Eg. for N of 3, the number of experiments is 3 and 6 if dye swap experiments are also performed. For N of 4, the number of experiments is 6 and 12. And this is without doing any replicate experiments. If a reasonable number of replicates is added, then this set of experiments becomes even difficult to perform.

MAExplorer is currently not oriented to handling these large combinatoric types of non-pooled sets of experiments. However, you do have the ability to swap (Cy3,Cy5) data on an individual basis so you could compute an average of data from dye-swap experiments - but with the caveats or non-uniform labeling mentioned above.

```
 [(Cy3_X/Cy5_Y) + 1.0/(Cy3_Y/Cy5_X)]/2
```

In general, this is probably not a very good estimate.

3.1.2 Design philosophy of MAExplorer methodology

There are several ways to implement a data mining system on moderate size databases. The first is that all computations are performed on a Web server and the user's Web browser displays the results. The second is download an applet from the Web

server, get the data from the Web server and do computations in the Web browser. A third way is do download data from a Web server and run a local stand-alone program on the data. MAExplorer can be run using both the second and third ways. However, we encourage the use of the stand-alone paradigm as having the best bandwidth and being the most robust. The browser-based computation paradigm (as opposed to server-based) is somewhat unusual. It keeps both the program and data on the server, making user maintenance of the latest versions easier than if they had to constantly upgrade the program or data. This also has the distinct advantage of giving the user instantaneous feedback through rapid visual and tabular views and the ability to more effectively navigate the data since the analysis is done on their desktop computer. Because it is easy to access reference data from other genomic sources (e.g. UniGene, GenBank, NCI/CIT's mAdb clone DB, dbEST, GeneCard, etc.), it can be accessed from their respective Web servers as needed. Complex browser-based computations are used in other data mining or intensive computation domains. With the increased bandwidth of the Internet and compute power and memory of PCs approaching the Cray supercomputers of the previous decade, this paradigm becomes even more feasible. However there are limits to how well it scales because of Web browser limitations. Appendix E.2 discusses these issues in more detail

The major focus of the MAExplorer is interactive data mining with an emphasis on direct graphical and tabular manipulation of the data. The investigator is able to interact with the system by clicking on spots in the array image, points in graphic plots, cells in spreadsheets, by manipulating threshold sliders or typing in gene names/clone Ids. This level of interaction allows investigators to search for and identify patterns of differences with greater ease than with a more static graphic system since it is easier to test ideas by "grabbing onto the data". For example, "what" is the identity of "this" outlier I am pointing to in a scatter plot; "which" genes are best clustered with "this" gene in this clustergram and are perhaps co-regulated; "which" genes have expression ratios within the range of the histogram bins to that I am pointing?

Direct user manipulation of data, as incorporated in MAExplorer, was defined by (Schneiderman, 1997) who defends the position that the direct manipulation of data in data mining is an extremely effective means to amplify human creativity in understanding patterns. Schneiderman's dogma states "overview first, zoom, and then filter details on demand" and favors the use of "shallow search trees, slide controllers, and information-right screens with tightly coordinated panel view of data", (Beardsly, 1999). MAExplorer also uses many of these direct manipulation principles. It was designed to run on the desktop computers with data residing on the same computer and loaded into its memory for rapid direct manipulation - for both the Web browser and stand-alone versions.

3.1.3 Evolution of MAExplorer from earlier proteomic data mining systems

MAExplorer was designed to do flexible exploratory quantitative data analysis of gene data from microarray hybridized sample experiments. Many of the data-mining concepts are derived from a system called GELLAB-II (http://www.lecb.ncifcrf.gov/lemkin/gellab.html) that is a UNIX-based stand-alone exploratory data analysis system for 2D protein gels over multiple experiments (Lipkin and Lemkin, 1981), a review (Lemkin and Lester, 1989) and examples of graphical representations of this type of data (Lemkin, 1995). An on-line GELLAB-II Web-Poster (http://www.lecb.ncifcrf.gov/lemkin/gellab-ep93wd.html) is available showing various screen shots of GELLAB-II in action. Whereas GELLAB works with sets of corresponding spots (i.e. proteins) across sets of 2D gel samples, MAExplorer works with sets of genes (spots in the microarray) across sets of hybridized sample microarrays. With protein gels, one typically has spot alignment problems since gels are generally not superimposable. This is often called the rubber-sheet distortion problem and requires localized alignment of spots based of neighboring spot constellation morphology. We have used Web-based visual methods to visually compare gels including the Flicker (http://www.lecb.ncifcrf.gov/flicker/) image comparison system a Java applet, (Lemkin, 1997), and the 2DWG (http://www.lecb.ncifcrf.gov/2dwgDB/) meta-database of 2D gel images, (Lemkin, 1999a). Since the genes are precisely spotted on the arrays, aligning spots between arrays is not required and greatly simplifies that the data analysis problem.

Part of the Flicker system allows comparison of user 2D gel images with standard images from SWISS-2DPROT for putative identification of unknown spots in the user gels. The user would select a standard 2D gel image from over 20 tissue types, enter their own 2D gel image and align them at spots of interest. They could then switch to a database access mode, click on those spots and generate popup SWISS-2DPROT Web pages for those proteins - similar to Clone reports in MAExplorer. That is accessed at http://www.lecb.ncifcrf.gov/flicker/swissProtIdFlkPair.html.

MAExplorer will have a *groupware* facility similar to what we have done with our [WebGel](http://www.lecb.ncifcrf.gov/webgel/) (http://www.lecb.ncifcrf.gov/webgel/) system described in ([Lemkin et al., 1999b](#)). It is a two-dimensional electrophoresis system for sharing data analyses. In WebGel, users may perform a data-mining analysis and leave the state of the their analysis and accompanying notes to share with their collaborators on a login-protected basis.

3.1.4 Concepts used in data mining with MAExplorer

This section introduces some of the concepts used in data mining microarrays with an emphasis on how they are used with MAExplorer.

Gene data filters - a Boolean AND of gene set tests

A primary MAExplorer concept is that of [gene data filter](#) that selects a *working set* of genes by the conjunction (Boolean AND) of user selectable tests. Each test further restricts the working set of genes to those meeting the test. These criteria include gene membership in particular gene classes, membership in particular user defined or computed gene subsets, and meeting a variety of statistical constraints. Statistics include intra- and inter-array CV, X-Y sets t-tests. Range test criteria include X/Y ratio ranges and histogram bins, intensity ranges and histogram bins. Membership criteria include test if genes are in the current-cluster (derived from cluster-analysis), gene set membership, etc. By selectively including one or more of these filter restrictions, the user can home in on the data that appears to be interest. Of course as in real mining, what appears interesting may not be interesting based on further investigation.

Set operations on gene subsets

Because of the complexity of comparing many different replicated samples, it may be difficult to manually organize the resulting comparisons. MAExplorer offers set-theoretic operations on [sets of genes](#) and [sets of hybridized samples](#) (i.e. intersection, union, difference) to help with this organization (step 9 in [Table 3.2](#)). The results of set operations may be saved and used in subsequent set operations, normalization, as well as with the data filter. This is useful when comparing and documenting procedures, methods, and analyses from several subsets of experiments.

User exploration states

Users needs to be able to save and restore the current state of their explorations of the data and option settings to document and continue at later times. When running in stand-alone mode, the user may save their data mining session on the local disk as in named (.mae file extension) startup files. Clicking on one of a startup file will restart MAExplorer and restore the state to that of the time it was saved. In addition to filter and parameter status, the HP-X, HP-Y, HP-X and HP-Y 'sets', HP-E 'list', the named gene sets and HP condition sets are saved as part of the state

User groupware sharing of exploration states with collaborators

[In the future], these could be saved on a public Web server using multiple named state files. These are protected for the user using a login procedure. A groupware sharing of these intermediate exploratory results is available when they allow another user to access selected states. User states and groupware sharing complete step 11 in the analysis described in Table 3.2.

We now discuss using these tools for analyzing ones data.

3.2 Steps in data mining, discover, and analysis

An analysis scenario may use many methods for viewing the data. A typical sequence of analysis steps is listed below in Table 3.2 in the order they might be performed. Note that this is a ***rough guide for a possible analysis*** and the iteration and backing up for of some of these steps is required for data mining complex sets of conditions, especially in the setting of constraints for the "data filter" (step 4) when the user focuses on subtle patterns of interest (c.f. [Figure 3.1](#)).

Table 3.2 Steps in a data-mining analysis.

1. Select a set of hybridized samples to be analyzed. Additional samples may be added or removed from the HP-X, HP-Y and HP-E sample lists prior to or during an analysis.
2. Organize samples by selecting query type: 2-class (X vs. Y) sample sets or N-sample ordered expression lists.
3. Select normalization (transformation scaling) method appropriate to the type of data and what types of changes you are looking for (patterns or outliers).
4. Restrict search using the "data filter" to subset genes (pre-filter data) by gene-class, gene-subset, ratio range, intensity range, coefficient of variation, t-test, expression profile, cluster membership, etc.
5. Review scatter and expression plots, ratio and intensity histograms to gain insights in overall characteristics of the data range and intensity distributions. Export results using either SaveAs or cut and paste operations.
6. Cluster genes by similar expression profiles and other clustering methods. Review similarity, clustergram and dendrogram plots and reports. Export using either SaveAs or cut and paste operations.
7. Select subset(s) of genes of interest in plots or reports.
8. Access other Web genomic databases for genes of interest from spreadsheets or plots.
9. Save genes of interest in named gene sets and perform set operations if required.
10. Create gene reports for export to Excel using tab-delimited Reports and either SaveAs or cut and paste operations.
11. In the stand-alone version of MAExplorer, save the state of exploratory data analysis for later use or sharing.

In designing a data mining experiment, the first decision to be made is selecting the set of hybridized samples to be compared (steps 1 and 2). This is accomplished by setting the current hybridized sample-X (HP-X) and hybridized sample-Y (HP-Y). In Figure 2.4.4.2 for the scatter plot we selected a single C57B6 pregnancy day 13 and a single Stat5a (-,-) pregnancy day 13 as current HP-X and current HP-Y samples. Changing the normalization changes the view in the scatter plot so that hidden differences may be more apparent (see Figure 2.4.2.3)

The names of the current HP-X and HP-Y samples are displayed at the top of the main window. The current HP-X and HP-Y samples may be changed at any time by clicking on a new sample from a list of samples shown on the left side of the main window or from lists of samples organized by sample population in the Samples menu.

The next decision to be made is selection of the genes to be studied by choosing a subset from the gene class menu list (step 4). Further selection occurs throughout the analysis by clicking on spots in microarray images, points in graphic plots or cells in spreadsheets, by adjusting threshold sliders, or using the text-entry "guesser" to type in gene names, clone IDs, genomic IDs, samples, etc.

The next decision the user must make is to set the intensity data normalization mode (step 3). Normalization of quantitative data is crucial when comparing data between different hybridized microarrays because of spotting, hybridization efficiency, uniformity, and other systematic errors.

Genes of interest may be separated for all of the genes in the database using a cascade of data filters (step 4). Additional filtering options are easily accessible in the (data) Filter menu. Some of the filters require additional parameters. These parameters are set by state scroll bars that pop-up on the screen when data filters requiring them are added to the filter cascade. Changing scroller values causes the data filter to be automatically be reapplied and a new set of genes to be computed.

It is desirable to reduce false-positives found by the data filter by eliminating genes with high quantification variability between duplicate spots on the same sample or spot duplicated in replicate samples. If duplicate genes are available on the array (denoted by Field 1 and Field 2 or F1 and F2 spots), this allows the computation of a coefficient of variation (CV) for the duplicates. This CV may be used in a data filter to reduce potential false-positives. CV is computed as `2|F1-F2|/(F1+F2)` using those spot values for each gene, as $StdDev_{HP}/Mean_{HP}$ for a set of replicate hybridized samples.

Graphical views of the data give the user additional insights into the data. These include spot intensity and ratio or Zdiff

pseudoarray images, [scatter](#) plots, [histogram](#) plots, [expression profile](#) plots, cluster plots showing genes [similar to a specified gene](#), the [number of clustered genes](#) for each clone, divisive clusters for [K-means clustering](#), and [clustergrams](#) and dendrogams for hierarchical clustering.

Scatter plots are useful for visualizing data from two conditions

The scatter plot method (step 5) allows the user to plot the intensity data between two samples, the X-sample and the Y-sample. Gene data may be spot data for two different samples (HP-X and HP-Y), means of two different sets of hybridized samples of replicate samples (sets of HP-X and sets of HP-Y), or the left and right normalized replicate data (F1 vs. F2) for the current hybridized sample. If Cy3/Cy5 data is used, then each sample is the ratio of data from two different hybridized samples. So if we have sample Cy3a and Cy3b then HP-X could be Cy3a/Cy5 and HP-Y could be Cy3b/Cy5 such that we are scaling the Cy3a and Cy3b samples using a common Cy5 normalization sample. Scatter plots are useful for obtaining a better understanding of the outliers when comparing different hybridized samples and determining the reproducibility of spotting when comparing F1 vs. F2 data or replicate sample data.

Filtering genes by histogram plots of ratios, Zdiffs or intensities

Histogram plots may be generated from either X/Y ratios or (X-Y) Zdiffs of two different hybridized samples (single samples or X and Y replicates) or from the F1/F2 intensities of a single hybridized sample. Selecting a bin in a histogram restricts filtered genes to those that are contained in that histogram bin. As an alternate method, data filtering by ratio (Zdiff) or intensity *range* may be used with adjustable range scrollers independent of the histograms. However, histograms and scrollers may be used together. For example, one could filter by the ratio histogram after filtering out genes with low-intensity values that may be considered noise using the intensity sliders. That might help eliminate falsely high ratios resulting from dividing high X values by a very small noisy Y values. Histograms are useful for getting a better understanding of the range and distribution of the gene intensities or ratios.

Expression profile plots (EP-plot) of N conditions for viewing time series, etc.

List HP-E is an ordered list of samples - as different from HP-X and HP-Y that are unordered sets of samples. The expression profile (step 5) of a gene is the plot of its normalized intensity as a function of the samples in the ordered HP-E list. It may be plotted for the current gene in a pop-up window. Selecting a different current gene causes the EP-plot to be displayed for that gene. Multiple EP-plots may be created to view the differences between a few genes you are investigating further. The **HP name** button pops up a window with the ordered list of samples so you can see the details of the sample names being plotted. Selecting a line in a plot displays the intensity data and sample name for that hybridized sample. The data may be plotted as a bar, point or continuous curve and error bars may be turned off to better compare multiple plots.

When there are too many EP-plots to be viewed simultaneously, you might use a scrollable list of expression profile plots that lets you scroll through an arbitrarily large list of genes. However, it is difficult to compare genes that are not sorted in some way (i.e. clustered). Therefore, these are most useful when used after clustering the data and displaying the scrollable EP-plots of the cluster-order data.

Clustering is one way of possibly finding co-expressed genes that exhibit similar expression changes in a set of samples. Genes may show similar co-expression, but that does not prove they are co-regulated at the same point in a pathway - merely that measurements of those genes in a particular set of experiments show similar expression. However, identifying genes with similar expression for which some information is already known about some of the genes may be useful as a starting point to help figure out gene function and pathway using additional experiments and analysis.

There are many methods for doing clustering - each with advantages and disadvantages. We present three methods in MAExplorer and plan on adding a variety of more powerful methods through the MAEPlugin facility under development.

Finding clusters of genes with similar expression profiles: similar, cluster counts, K-means, and hierarchical methods

We may define a cluster of genes as a set of genes whose expression profiles are found to be similar (step 6). The samples used in

computing the expression profiles are specified by the HP-E ordered list. You can scale the list of normalized intensity data for each gene to 1.0 (resulting in finding genes with similar shaped EP-plots). Alternatively, if you don't scale this data it will cluster more on magnitude changes. You can select either the Euclidean distance or the correlation coefficient of the EP lists between two genes as the measure of gene-gene distance. Similarity is 1.0 - normalized distance.

The first cluster method finds a cluster of genes whose expression profiles are similar to that of the currently selected gene. This list of genes is restricted by the constraint that the cluster distance between each of these genes to the selected gene is less than the "Cluster threshold" distance set by the user with a scroll bar. It displays genes that are found both with blue boxes (the larger the box, the higher the similarity) and in a text report window showing the genes and their distances to the current gene. By varying the threshold and observing the results, the user can find a set of highly correlated genes. If the threshold is set to 0.0, no genes are found. If it is set too high, all data filtered genes are found. So it is critical to adjust the threshold to a reasonable level commensurate with the type of data being analyzed and the approximate number of genes expected.

A second cluster method draws blue circles in the array image around all filtered genes meeting the threshold criteria, where the larger the circle the larger the number of similar genes (i.e. passing the threshold) are found to be clustered with that gene. Clicking on a gene toggles between the first and second methods. For both of these methods, it will pop-up a "Cluster Distance" threshold scroller and recomputes the clusters if you change the scroller value or the current gene. It also shows a text report that displays the number of genes similar to each data filtered gene.

A third method called "K-means" clustering K genes (we call primary nodes) whose expression profiles are most orthogonal to each other. It uses the current gene as the first or "seed" node. It then finds the gene furthest from this and assigns it as node 2. Then the gene furthest from both nodes 1 and 2 is assigned to node 3, etc. This process is repeated until all K nodes are assigned. Then the remaining genes are assigned to the closest node. Having defined the initial cluster centers, it recomputes the centroid of each of the clusters. The centroid can alternatively be computed using a median instead of a mean in which case we would be doing K-median clustering (Bickel, 2001). K genes are then reassigned to the nearest new centroids as the new K-means node instances. Finally, the remaining genes are assigned to the nearest centroid. A scrollable K-means cluster text window report pops up with genes sorted by cluster. Clicking on a gene in either the array image or scatter plot assigns all genes in the cluster to which that gene belongs to the "current cluster". Genes in the current cluster are labeled in the array and scatter plot with a small number of the cluster. In addition, genes in the current cluster are copied to the E.G.L. where they can be used in a report, saved in a named gene set, or used for additional filtering. It also pops up a "N-clusters" scroll bar window to let you dynamically adjust the number of clusters. Changing N will recompute the clusters. When the K-means is recomputed, it uses the current gene as the initial seed gene.

The fourth method is a hierarchical clustering method that generates a clustergram and dendrogram similar to that of Eisen's red-black-green clustergram (Eisen, 1998). This was derived from the clustered correlation map (ClusCor) of Weinstein et al. (Weinstein, 1997). The MAExplorer clustergram and dendrogram are dynamic and may be interrogated and used to set the current gene. This means that it may also position a corresponding ordered list of expression profile plots to the same gene so you may view the data as a plot as well. The dendrogram may be zoomed in to explore a part of the dendrogram in more detail. As with the K-means clustering, a report can be made of the ordered genes.

Gene reports: dynamic spreadsheets for Web access or tab-delimited for exporting data to Excel

Pop-up report windows (step 7) may be generated for either individual genes or a global array sample data. Instances of the latter include experimental information and Web links, global statistics, correlation coefficients between array samples, etc. Gene reports may present this data in a number of ways. These include: highest/lowest gene ratios, profiles, parametric and cluster statistics, etc. Reports may be presented as either dynamic Web-interactive spreadsheet tables or as static tab-delimited tables. The latter is useful for exporting data using cut and paste into Excel (step 10). If the user clicks on a blue hyperlinked cell in a dynamic spreadsheet table, it pops up another Web browser window and loads it with data (step 8) from the respective Internet genomic database such as mAdb Clone DB, UniGene, GenBank, dbEST, GeneCard, and MGAP model and histology Web pages.

Collaborative groupware environment

Having immediate access to collaborator's data is a powerful research tool. A collaborative environment is being implemented for

MAExplorer that allows groups of users to share data and intermediate results. These capabilities include: 1) the ability to save and restore exploratory data mining sessions (states) through the Web server including named sets of genes, and 2) to selectively share these states with collaborators. The latter process is sometimes called a groupware environment because if offers a collaborative group the ability to share and interact. These capabilities are modeled after our WebGel system ([Lemkin et al., 1999b](#)). In addition, users can create a custom database Web page as a subset of samples from the entire database. This may be saved on their own computer through their Web browser's "File/Save as" command. This hypertext file could then be used at a later time to access the database or be E-mailed to a collaborator to do the same.

3.2.1 Definition of expression profile

It is helpful to define an expression profile. There may be alternate definitions, but the following is useful for getting an understanding of how it might be computed. An expression profile e_j of an ordered list of N samples (k=1 to N) for a particular gene j is a vector of scaled expression values v_{jk}.

Then, the expression profile is expressed as a list of values:

$$e_j = (v_{j1}, v_{j2}, v_{j3}, \ldots, v_{jN})$$

A difference between two genes p and q may be estimated as a N-dimensional metric "distance" between e_p and e_q. The Euclidean distance is then defined as

$$d_{pq} = (1/N \; \mathrm{SUM}_{j=1:N} \; (v_{jp} - v_{jp})^2 \;)^{1/2}$$

Other distance measures may include correlation coefficient, city-block (or manhatten distance) etc.

For scaled data such that d_{pq} has a maximum value of 1.0 ovger all samples. A *similarity* measure could be computed as 1.0 - distance or

$$s_{pq} = 1 - d_{pq}$$

3.2.2 Clustering Methods

Clusters represent one way to identify similar gene expression across a set of experiment samples. There are many ways to cluster the data, some of which are available in MAExplorer. These include:

1. Finding genes with similar expression
2. K-means clusters where the number of clusters K is fixed prior to clustering and a particular gene is used to start off the cl.
3. Hierarchical clustering where a binary tree hierarchy is created

Other methods include Self Organizing Memory (SOM), fuzzy clustering, Support Vector Machines (SVM), etc.

3.2.2.1 Clustering similar genes

If we have a particular gene s (the "seed" gene), we may want to find a set of all genes $\{g_j\}$ similar to g_s. We can find this set of genes by testing We define a particular gene g_j as similar to seed gene if the distance between genes s and j meets the following criteria.

$$d_{js} < T$$

The threshold T is set by the investigator and in MAExplorer is changed using a slider. Typically, the set of all genes $\{g_j\}$ found is sorted by similarity before being viewed.

3.2.2.2 K-means clustering

K-means clustering finds K clusters of genes with similar expression profiles to a given gene (see [Sneath and Sokol, 1973](#)). Given the number of clusters K, we could use high variance of clusters to determine if they should split into sub-clusters. K-means clustering does not need a distance matrix (see Hierarchical clustering which follows), so it is faster and may cluster large numbers of N genes. However, it is highly dependent on seed selection. It may be useful for getting an initial estimate - especially if other techniques (such as silhouette plots) are also used. The following is a simplified definition of one way to compute a set of K-means clusters of gene expression profile data.

<div style="border:1px solid">

Algorithm:

1. Pick seed gene s and put it into cluster 1 (let $k = 1$)
2. For all clusters $j = 1$ to k, find gene q such that d_{jq} is a maximum
3. Set $k = k+1$. Put gene q into new cluster k
4. For $j = k$ to K, repeat steps 2 and 3 until there are K clusters
5. Then, assign $(N-K)$ remaining genes q into one of the K clusters j with minimum d_{jq}
6. Compute new *virtual gene cluster* centroids as means or medians $\{e_k\}$ for each of K clusters
7. Reassign all N genes q into K new clusters with minimum d_{pq} using virtual genes $\{e_p\}$
8. Variants: use multiple seed genes, range of K values, minimize CV

</div>

3.2.2.3 Hierarchical clustering

Hierarchical clustering of a set of genes will generate a binary tree of clusters with the genes at the terminal ends of the tree and a single cluster of the entire tree at the top (also called the root) of the tree. See ([Sneath and Sokol, 1973](#)) for a discussion on hierarchical clustering. There are many other variants of hierarchical clustering. Hierarchical clustering requires a distance matrix or the equivalent of one [there are more efficient ways to compute it]. For N genes (terminal clusters), it generates $2N-1$ clusters. Distance matrix is upper diagonal matrix D of d_{pq} of size $N(N-1)/2$.

D can get quite large for clustering a large number of genes N [for N=5000, this is > 50 Mbytes!]

The following is a simplified definition of one way to compute a hierarchical clustering of gene expression profile data.

<div style="border:1px solid">

Algorithm:
1. Assign all N genes to clusters 1 to N, set n to N
2. Find two clusters p and q such that d_{pq} is a minimum doing the following:
   2.1 Compute a virtual cluster vector $e_k = average\ (ep,eq)$
   2.2 Set $n = n+1$
   2.3 Assign *virtual* cluster to new cluster
3. Repeat step 2 until $n = 2N-1$.

</div>

3.3 Display of gene spot intensity and identification data measurements

You may select the current gene by clicking in the [pseudoarray image](#) or in the [X-Y scatter plot](#) and MAExplorer [reports](#). The microarray grid coordinates, normalized quantified spot intensity data, plate coordinates, gene name (if known) and associated data for that gene. If you are displaying a pseudocolor ratio (X/Y) or Zdiff (X-Y) image, it will report HP-X/HP-Y or (HP-X - HP-Y)

data respectively. It also sets the gene as the *current gene*. The pseudoarray image coordinates are reported as:

```
     [<field>-<grid name><row#>,<col#>].
e.g.
     [1-A4,3]
```

If there is only one field in the array, it will appear as field 1. In the above example, [1-A4,3] is field 1 grid A row 4 and column 3. Note that the pseudoarray coordinates are for visualization purposes in MAExplorer and may or may not be the same as the coordinates on the actual array. That depends on how the MAExplorer database was defined in the configuration file described in Appendix C.

When the current gene is defined, it will draw a yellow (green) circle around the spot in the ratio (intensity) pseudoarray image and display other features of the gene in the three-line status area near the top of the main window. If background correction is enabled (the "Use background intensity correction" in the Normalization menu), then spot intensity values will appear as `intensity'` (with background intensity subtraction) and `intensity` (without background subtraction).

There are a number of different reporting formats available depending on the array display mode and particular normalization method selected. These include: the pseudoarray image of the intensity of a single sample, the pseudocolor ratio X/Y or Zdiff (X-Y) image (using either HP 'sets' or single samples), or the ratio of Cy3/Cy5 for dual-labeled dyes or F1/F2 for replicate spots for a single sample. In addition, the normalization mode is also displayed in the reporting line. We will present examples of each of these different reporting formats.

You may show the intensity data for a particular spot in the currently displayed pseudoarray image. First select the "Pseudograyscale image" option in the "Show Microarray" submenu in the "Plot menu". If your data has duplicate grids (i.e. fields F1 and F2) then you may look at F1, F2 and mean (F1+F2)/2 data in the reports when you click on a spot. If the "Gang F1-F2 scrolling" switch is disabled in the "View menu", then the `intensity` value is the intensity data value for the gene at that location. If the "Gang F1-F2 scrolling" switch is enabled, then it reports intensity[F1], intensity[F2], and the F1/F2 ratio. These two formats are shown in the following two examples for a C57B6 pregnancy day 13 samples in the MGAP database:

a) Field F1 spot for a single spot in a single sample with the median intensity selected.

```
[1-A4,5] intensity=4.5267, (Norm.: median intensity)
CloneID: 1248228, dbEST3': 2279072, GenBankAcc3': AI463183, UniGene: Mm.13859, plate[5,A,5]
GeneName: Mus musculus ribosomal protein L41 mRNA, complete cds
```

b) Field F1 and F2 replicate spots for a single sample. The top line is shown for each of the different normalization methods.

```
[1-A4,5] intensity[F1]=-0.3067, intensity[F2]=-0.2312, F1-F2=-0.0755, (Norm.: Zscore intensity)
[1-A4,5] intensity[F1]=4.5267, intensity[F2]=6.2408, F1/F2=0.7253, (Norm.: median intensity)
[1-A4,5] intensity[F1]=0.8755, intensity[F2]=1.1457, F1-F2=-0.2701, (Norm.: log median intensity)
[1-A4,5] intensity[F1]=-0.1442, intensity[F2]=-0.0945, F1-F2=-0.0497, (Norm.: Z-score, stdDev, log
intensity)
[1-A4,5] intensity[F1]=-0.1533, intensity[F2]=-0.1004, F1-F2=-0.0528, (Norm.: Z-score, mean
abs.deviation, log intensity)
[1-A4,5] intensity[F1]=630.9911, intensity[F2]=869.9273, F1/F2=0.7253, (Norm.: calibration DNA
intensity)
[1-A4,5] intensity[F1]=1919.9376, intensity[F2]=2646.957, F1/F2=0.7253, (Norm.: scale to max. (65K)
intensity)
CloneID: 1248228, dbEST3': 2279072, GenBankAcc3': AI463183, UniGene: Mm.13859, plate[5,A,5]
GeneName: Mus musculus ribosomal protein L41 mRNA, complete cds
```

If the "Pseudocolor HP-X/HP-Y ratio or Zdiff" option is selected in the "Show Microarray" submenu, data is reported as either

Ratio or Zdiff data depending on the normalization method selected. The data used in the following examples is for C57B6 pregnancy day 13 (HP-X) compared with Stat5a (-,-) pregnancy day 13 (HP-Y).

c) Ratio data for two samples X and Y in separate hybridized arrays. Ratio data for the field F1 and F2 spot data as well as the mnX/mnY ratio is reported. The median normalization was used in this example.

```
[1-A4,5] HP-XY: mn(X,Y)=(5.383,6.834) (X/Y)(F1,F2,mean)=(0.651,0.928,0.787), (Norm.: median
intensity)
CloneID: 1248228, dbEST3': 2279072, GenBankAcc3': AI463183, UniGene: Mm.13859, plate[5,A,5]
GeneName: Mus musculus ribosomal protein L41 mRNA, complete cds
```

d) Zdiff data for two separate samples X and Y. Ratio data for the field F1 and F2 spot data as well as the mnX-mnY Zscore difference is reported. The three Zscore, ZscoreLog, and logMean normalizations were used in this example (first lines are shown).

```
[1-A4,5] HP-XY: mn(X,Y)=(-0.269,0.151) (X-Y)(F1,F2,mean)=(-0.470,-0.370,-0.420), (Norm.: Zscore
intensity)
[1-A4,5] HP-XY: mn(X,Y)=(-0.119,0.051) (X-Y)(F1,F2,mean)=(-0.199,-0.142,-0.170), (Norm.: Z-score,
stdDev, log intensity)
[1-A4,5] HP-XY: mn(X,Y)=(1.010,1.224) (X-Y)(F1,F2,mean)=(-0.362,-0.064,-0.213), (Norm.: log median
intensity)
CloneID: 1248228, dbEST3': 2279072, GenBankAcc3': AI463183, UniGene: Mm.13859, plate[5,A,5]
GeneName: Mus musculus ribosomal protein L41 mRNA, complete cds
```

e) Example of when the "Use dual HP-X & HP-Y Pseudoimage" mode is enabled in the "Show Microarray" submenu of the "Plot" menu. This displays mean data for the HP-X and HP-Y data side-by-side. The median normalization was selected.

```
[1-A4,5] intensity[X]=5.3837, intensity[Y]=6.8342, X/Y=0.7877, (Norm.: median intensity)
CloneID: 1248228, dbEST3': 2279072, GenBankAcc3': AI463183, UniGene: Mm.13859, plate[5,A,5]
GeneName: Mus musculus ribosomal protein L41 mRNA, complete cds
```

Reporting for multiple hybridized samples when using HP-X/-Y 'sets'

If you have enabled MAExplorer to "use HP-X and HP-Y 'sets' of multiple samples" rather than single samples" in the Samples menu, it will report a spot differently using the means (mn), standard deviations (S.D.), coefficient of variations (CV) for the samples in the HP-X and HP-Y 'sets'. For duplicate fields, these are computed using the normalized average of F1 and F2 spots for each gene in each samples. The data used in the following examples is for three C57B6 pregnancy day 13 (HP-X) samples, and five Stat5a (-,-) pregnancy day 13 (HP-Y) samples.

f) Multiple HP-XY 'sets' using median normalization for the pseudoarray image display for the HP-X 'set' of three C57B6 samples.

```
[1-A4,5] HP-X 'set' mean intensity=3.295 stdDev=1.482 CV=0.449 n=3, (Norm.: median intensity)
CloneID: 1248228, dbEST3': 2279072, GenBankAcc3': AI463183, UniGene: Mm.13859, plate[5,A,5]
GeneName: Mus musculus ribosomal protein L41 mRNA, complete cds
```

g) Multiple HP-XY 'sets' using median normalization for the pseudoarray image display for the HP-Y 'set' of five Stat5a (-,-) samples.

```
[1-A4,5] HP-Y 'set' mean intensity=8.180 stdDev=0.986 CV=0.120 n=5, (Norm.: median intensity)
CloneID: 1248228, dbEST3': 2279072, GenBankAcc3': AI463183, UniGene: Mm.13859, plate[5,A,5]
GeneName: Mus musculus ribosomal protein L41 mRNA, complete cds
```

h) Multiple HP-XY 'sets' using median normalization for the pseudoarray image display for the HP-X and HP-Y 'sets' when the "Use dual HP-X & HP-Y Pseudoimage" mode is enabled in the "Show Microarray" submenu of the "Plot" menu.

```
[1-A4,5] HP-XY 'sets': mn(X,Y)=(3.295,8.180) mnX/mnY=0.402 SD(X,Y)=(1.482,0.986)
CV(X,Y)=(0.449,0.120)\
       n(X,Y)=(3,5), (Norm.: median intensity)
CloneID: 1248228, dbEST3': 2279072, GenBankAcc3': AI463183, UniGene: Mm.13859, plate[5,A,5]
GeneName: Mus musculus ribosomal protein L41 mRNA, complete cds
```

i) Multiple HP-XY 'sets' using median normalization for ratio (HP-X/HP-Y) data for the "Pseudocolor HP-X/HP-Y Ratio or Zdiff" display.

```
[1-A4,5] HP-XY 'sets': mn(X,Y)=(3.295,8.180) mnX/mnY=0.402 SD(X,Y)=(1.482,0.986)
CV(X,Y)=(0.449,0.120) \
         n(X,Y)=(3,5), (Norm.: median intensity)
CloneID: 1248228, dbEST3': 2279072, GenBankAcc3': AI463183, UniGene: Mm.13859, platey[5,A,5]
GeneName: Mus musculus ribosomal protein L41 mRNA, complete cds
```

j) Multiple HP-XY 'sets' p-value using median normalization for ratio (HP-X/HP-Y) data for the "Pseudocolor (HP-X,HP-Y) 'sets' p-value display.

```
[1-A7,20] HP-XY: mn(X,Y)=(3.449,0.853) (X/Y)(F1,F2,mean)=(4.09,4.008,4.041), (Norm.: median
intensity)
CloneID: 1382656, dbEST5': 1775754, GenBank 5': AI036495, UniGene: Mm.300,  plate[12,A,8]
GeneName: Carbonic anhydrase 3
```

Reporting for Cy3 and Cy5 channels for a single hybridized sample

k) If you have Cy3/Cy5 data, then you can look at the two channels for a single sample (the current HP sample). For median normalization and the display set to "Pseudocolor Red(Cy5)-Yellow-Green(Cy3) Cy3/Cy5 data" display.

```
[1-A6,11] Cy5/Cy3=0.3588, Cy5=67.324, Cy3=187.622, (Norm.: median intensity)
CloneID: IMAGE:1054189,
GeneName: expressed sequence AW213287
```

Reporting HP-X (Cy3 or Cy5) vs HP-Y (Cy3 or Cy5) data for 2 samples

l) If you want to compare Cy3 or Cy5 in the HP-X sample with a Cy3 or Cy5 value in the HP-Y sample, you do it through the special Cy3,Cy5 scatter plots. There are four types of plots:

1. HP-X Cy3 vs HP-Y Cy3
2. HP-X Cy3 vs HP-Y Cy5
3. HP-X Cy5 vs HP-Y Cy3
4. HP-X Cy5 vs HP-Y Cy5

After the plot is started, clicking on a scatter plot will report data from the point in that plot will print the following data as shown in the following example where HP-X Cy3 is plotted against HP-Y Cy3.

```
[1-A5,16] intensX=4.695, intensY=5.923, (X-Y)=-1.2275, (Norm.: log median intensity)
CloneID: IMAGE:963758,
GeneName: RIKEN cDNA 2410114O14 gene
```

3.4 Selecting subsets of genes using the data Filter

Genes may be selected on a number of criteria specified by the [data Filter](Section 2.4.3) that is a cascade of data tests. The first might be the [gene class](Section 2.4.1) to restrict the set of all genes to a particular subset. Various numeric and statistical data tests might be applied on the remaining genes to exclude those not meeting these tests. For example, genes having a high coefficient of variation between duplicate spots on the same sample or duplicate samples could be eliminated. Then one could select genes that had a (HP-X/HP-Y) ratio greater than 4.0 but less than 8.0, etc. The latter could be done using either the ratio scrollers or by clicking on that bin in the ratio histogram plot. See the Filter menu options and look at one of the [tutorials](Appendix A) for ideas on adjusting the Filter to close in on a particular subset of genes.

3.5 Selecting subsets of hybridized sample conditions

Sets and lists of hybridized sample conditions (HP-X and HP-Y sets, HP-E list) may be selected using various commands from the [Samples Menu](Section 2.2) including pull-down menus, guessing by name or part of a name, or using a ["Chooser"](see Figure 2.2.1) to design your settings for the current (HP-X and HP-Y sets, HP-E list). The Chooser is the easiest way to select these entries. In addition, if you want to change the current HP-X or HP-Y individual sample, you can do this directly [from the array pseudoarray image](#) by clicking on the [X] or [Y] part of the image and then selecting the particular sample to use. Note that if the mouse-over checkbox is enabled, then moving the mouse over the sample names gives you the full sample name. Otherwise, the sample name may be truncated.

3.6 Setting threshold values using the state-scroller sliders

You may filter genes using a variety of thresholding operations (see the [Filter](#) menu (Section 2.4.3) to select any of these). For example, these include a spot intensity (per channel) range [SI1:SI2], gene intensity range [I1:I2], ratio range [R1:R2], Zdiff range [Z1:Z2], Coefficient Of Variation (CV) range [0:1.0], p-value range [0:1.0] for the t-test, etc. Additional threshold scrollers are used with the clustering methods including the number of clusters (default 6), the maximum cluster distance from a gene to a another gene for the latter to be considered in the same cluster, and the absolute difference between HP-X and HP-Y.

For the intensity and ratio threshold filters, the range interpretation may be inside, or outside the specified range. The ratio range [R1:R2] is between 0.01 and 100.0. The Zdiff range [Z1:Z2] and [CZ1:CZ2] are between -4.0 and +4.0. The intensity threshold range [I1:I2] is set to the dynamic range of the min and max intensity for the current normalization method.

A list of possible threshold sliders is shown in the following table. When a Filter is enabled that requires a slider, it pops up the *State Scrollers* window that contains one or more slides. When you disable all filters that use these sliders, the popup window will disappear. The corresponding **Ratio R1[R2]** or **Zdiff Z1[Z2]** sliders are used if you are using a ratio or Zscore normalization - and will change if the normalization changes while the filter is active.

Some of the sliders are implemented with a non-linear scale so that you have more resolution at the low end (eg. p-Value, Spot CV, Diff HP-XY).

Depending on the set of data Filters selected, there may be multiple sliders present in the State Slider popup window (eg. see [Figure 2.4.3](#)).

Table 3.3.1. List of threshold sliders. Sliders are enabled in the State-Scroller popup window when the corresponding data filters are enabled.

Slider name	Associated with operation
Spot Intensity SI1	Filter by spot intensity range per channel
Spot Intensity SI2	Filter by spot intensity range per channel

Percent SI OK	Filter by percent of spots whose spot intensity is in threshold range criteria meets the AT LEAST or AT MOST criteria
Intensity I1	Filter by gene intensity range
Intensity I2	Filter by gene intensity range
Ratio R1	Filter by gene X/Y ratio range
Ratio R2	Filter by gene X/Y ratio range
Zdiff Z1	Filter by gene X-Y Zdiff range
Zdiff Z2	Filter by gene X-Y Zdiff range
Ratio CR1	Filter by Cy3/Cy5 gene X/Y ratio range
Ratio CR2	Filter by Cy3/Cy5 gene X/Y ratio range
Zdiff CZ1	Filter by gene (Cy3-Cy5) X-Y Zdiff range
Zdiff CZ2	Filter by gene (Cy3-Cy5) X-Y Zdiff range
p-Value	Filter by t-Test
Spot CV	Filter by Coefficient of Variation
Cluster Distance	Plot - cluster by expression similarity
# of Clusters	Plot - K-means clustering
Diff HP-XY	Filter by absolute difference (HP-X,HP-Y)
Spot Quality	Filter by continuous spot quality (If data available)

3.7 Exporting report and plot data

Data is typically reported in MAExplorer in report and plot windows. These may be saved using cut and paste if your are using MAExplorer as an Applet or with "SaveAs" buttons on the popup windows if you are running it as an stand-alone application . Reports are then saved as text (.txt extension) files, and plots are saved as GIF (.gif extension) files.

If you are running on a windowing system supporting cut and paste, then you may cut and paste data from reports and plots into applications on your system that allow you to save or print this data. Set the Report menu table-format to "Tab-delimited". Then, in Windows 95/98/NT/2000/XP, cut data from the popup tables (or other text reports) and paste it into Microsoft Excel. In Windows, you can capture (i.e. "cut") the entire screen by pressing the "Prt Sc" or print screen button. To capture a specific window (e.g. a scatter plot), hold the "Alt" key when pressing the "Prt Sc" key. Then go into a Windows imaging application (such as PhotoShop) and paste it into the application. In PhotoShop, in the File menu, select New (or type Control/N). Then when the window is opened, click on the window and paste the MAExplorer screen you had cut into the image window by typing Control/V. In both Excel and PhotoShop you may print the data or save it in a file.

4. Status and Bugs of MAExplorer

This section discusses the status and known bugs in MAExplorer. It also discusses dealing with the reporting of fatal errors so we can resolve them.

Section 4.1 discusses known bugs, Section 4.2 lists the revision notes for older versions known bugs. If you have experienced bugs with an older version of MAExplorer, you might check the revision notes to see if the bug was fixed and download a new version. Section 4.3 discusses problems in using MAExplorer as an applet with Web browsers. Section 4.4 describes handling fatal

4.1 Known Bugs in MAExplorer

Disclaimer: none of our code *ever has bugs...* :-). So despite this, we are working on resolving these bugs and implementing planned functionality. Here is a short non-inclusive list of known problems that we are resolving. We welcome and encourage you to E-mail us with any bugs that you find do exist as well as suggestions for capabilities you would like to see. As the new open-source [MAEPlugins](#) facility evolves, most new (and some old) functionality will migrate to these plugins. Then the user community can help maintain these analytic methods.

If you encounter a fatal error that is detected by MAExplorer, it will popup an [error reporting window](#). Please E-mail this data to us so we can try to resolve the problem.

In the mean time, partially implemented commands are disabled to keep you out of trouble :-) ...

You can help us and get MAExplorer to do more of the things you would like to see. Let us know of problems that you encounter as well as suggestions for changes or new methods you would like to see - [send us E-mail](#).

4.1.1 Browser Applet Bugs

1. When using MAExplorer as a Web browser applet (this does not apply if running as a stand-alone application - and we recommend running it as a stand-alone application not as an applet), there are problems with running MAExplorer on some of the Web browsers on some operating systems. It generally works with Netscape 4.7 or later, on Windows PCs (95/98/NT/2000/XP), and on SUN Unix and with Microsoft Internet Explorer 5.0 on PCs. It may not work as well as an applet on Macintoshes - even with Internet Explorer (although it works fine as a stand-alone application on the Mac) and problems have been reported on SGI systems because of browser problems.

   If you are experiencing Web browser problems using the MAExplorer applet, you might check the discussion of [possible solutions](#).

2. When run as a Web browser applet, you should only **click once** since multiple clicks may cause some browsers to hang. Note that when you click the first time, it starts downloading the applet. Clicking a second time may cause a second request for the applet to be launched. This interferes with the first request and so neither is started correctly.

3. When the MAExplorer applet is being downloaded, some browsers do not indicate to the user that it is being downloaded and so *it may appear* that nothing is going on. Some browsers will not indicate that it is downloading an applet. So *please be patient*. When it starts, the main window will popup and it continues to download data. Wait a while for it to load the applet before giving up. For a 28Kb Internet connection, this could take several minutes. When it is finished, the menu bar

will become active and it will display Ready - click on a gene to query database.

4.1.3 Downloading and Installer Bugs

1. Installer does not start or does not download the data. Normally, when you attempt to download MAExplorer to your computer, the InstallAnywhere software running in your Web browser will detect which OS you are running and suggest the particular download to use in:

```
"Recommended version for your computer
 Download installer for ...your OS..."?
```

Occasionally, we have seen instances where you can not install MAExplorer from within the Web browser. The solution is to explicitly download the particular Platform for your OS in the Available Installers list. And then to follow the instructions on running it.

2. If you have problems downloading this with Netscape 4.7x or later, then try Internet Explorer 5.0. It could be a Mime/type problem with your particular browser not setup correctly.

3. On Solaris, and possibly other Unix systems, you may have problems with the stack limits. Do a "man limit" to read about the command for your particular Unix shell. We have found that the following seems to work. For CSH, add the following to your .cshrc startup file.

```
limit stacksize unlimited
```

4.1.4 Computation speed and display Bugs

1. Because it does a lot of computation, MAExplorer runs best on a fast computer with lots of memory. An optimal system should have at least 128Mb of memory at least 500Mhz CPU speed. For doing hierarchical clustering we recommend 256Mb of memory which will allow you to cluster larger sets of genes without paging the operating system. So you may run into problems with running out of memory on an under-powered computer. To increase the usable memory size to more than 256Mb (assuming your computer has it), then you may need to edit the MAExplorer.lax file that resides where you installed MAExplorer on your computer to increase the heap and stack maximum limit.

2. We have seen instances where the same system will run out of memory with particular Web browsers (again - not a problem when running stand-alone) Netscape 4.7 on a system with a lot of memory. This does not seem to happen with Internet Explorer 5.0. This is because Netscape4.7 on the Windows PC seems to be able to use only about 5.5Mb before it runs out of memory. This is not the case with Internet Explorer 5.0.

3. MAExplorer is difficult to use with a system with a small screen because of the size and number of multiple graphic images and plots. We recommend a screen size of at least 1024x768. As they say these days, more is better :-), generally.

4. For problems with MAExplorer and MacOS, you might also see our MacOS problems FAQ.

5. For problems with MAExplorer and Sun Solaris OS, you might also see our SUNOS problems FAQ.

6. We have on rare occasions seen a data-related problem on MacOS systems. If the data was edited in such a way that the line terminators were Return characters instead of Linefeeds or Return-Linefeeds, the data may become corrupted and MAExplorer has difficulty reading it. The solution is to replace the Return characters with Linefeeds or Return-Linefeeds. One (painful) way this might be done by exporting the data to a system such as a Windows PC; read the data into Excel; save it as an Excel worksheet file format; exit Excel; start Excel again on the new data files; save the data files back into the original MAExplorer directories as tab-delimited text files. MAExplorer should recognize Return as the line terminator. This bug will be resolved in a future release.

4.1.5 User state and login Status

1. User registration and the groupware access functionality are not available yet. However, login is available for collaborator databases on the MGAP MAExlorer Web server.

2. The commands to save and restore user states for a particular user on the back-end server are not available yet. However, in stand-alone mode you may save the state (including all gene and condition sets, parameter conditions, condition sets, etc) on a local computer by doing a (File menu | Databases | Save As DB). The specialized groupware commands that will allow users to selectively share their saved states with particular other users on a public groupware server is not available yet.

4.1.6 Data file names Bug

1. There is a potential problem with using long file names when using MAExplorer with MacOS-8 or MacOS-9. Both of these operating systems limit file names to 32 characters. This could be a problem if importing data from another operating system (Windows, Unix, Linux, etc) that does not have this restriction. The solution is to use short file names. This should not be a problem with MacOS-X since it allows file names with up to 256 characters.

4.1.7 Gene Sets Bugs

1. The full range of GeneClass subsets (Section 2.4.1) is not available yet. However, you may use the gene name guesser to find a set of genes by wildcard into the Edited Gene List (e.g. "*ONCO*" to find all oncogenes or proto-oncogenes, etc). These EGL sets may then be saved as named gene sets and used in the data filter as "Filter by 'User Gene Set'". See the Gene Class ontologies using Gene Set operations (Section 2.4.1) that can achieve a similar effect.

4.1.8 Clustering Bugs

1. The hierarchical clustering method is buggy. It currently clusters genes - but not samples. The latter is under construction. Centroid averaging is implemented and the arithmetic averaging is under construction. Although it generates a clustergram plot, the dendrogram drawing may not be drawn correctly using centroid averaging. Depending on the number of samples, it you do a complete hierarchical cluster for all genes and ESTs on a Netscape browser on a Windows PC, the browser may run out of memory and nothing appears. [When done in the stand-alone application, this is less of a problem since there are no memory restrictions.] You can do three things at this point: 1) switch to Internet Explorer, or 2) decrease the number of genes being clustered (e.g. just "All named genes") or the CV filter. The third option is to disable the clustering cache. This will be done automatically for you and it will continue doing the clustering. You may turn off (or on) the **Use cluster distance matrix cache** in the "Hierarchical Cluster Plots submenu. It will perform the clustering, but will take MUCH longer without the cache.

2. If you are working with very large data sets then if you start a find similar genes or K-Means clustering with a very high threshold, you can not change the threshold until it is done. In these situations, try pre-setting the threshold distance, number of clusters using the (Edit | Preferences | Adjust all Filter threshold scrollers). This will popup the state scroller window with all of the thresholds. You may also select the current gene prior to starting clustering.

4.1.9 Expression profile plots

1. The Expression Profile Overlay display is being improved to offer better graphics and interaction.

2. We have observed some instances of the popup EP plot where data does not seem to correspond to the data. This may be related to the normalization method.

3. When viewing an ordered list of EP plots associated with a sorted list of genes (e.g. clustering genes similar to a seed gene), if there are replicated genes in the clustered genes, they may not show up in the scrollable list of EP plots.

4.1.10 Data conversion problems

1. The Cvt2Mae data file converter application is currently available for beta-testing on the a related MAExplorer [Cvt2Mae Web page](). It enables users to convert their data files (academic one-of-a-kind as well as commercial e.g. Affymetrix, Incyte, GenePix, Scanalyze, etc) to the standard MAExplorer files as described in Appendix C and D.

2. If you are not able to convert your data using Cvt2Mae (which is currently in early Beta testing - see the Cvt2Mae Web page for current status), you can always manually convert the data using the examples in Appendix C.

3. We have seen some problems running the converter with some types of data on MacOS 8-9. The temporary solution is to convert your data using another system such as Windows until we fix the problem.

4. If you are using the NCI-CIT mAdb system to package data for MAExplorer, you must select all samples from multiple projects that you want to compare with MAExplorer while packing the data in mAdb. (See PDF document ["How to download mAdb data For MAExplorer"]()) Currently, you can not easily merge the data after you have downloaded the files to your computer. The solution is to use mAdb to prepackage all data you want to analyze in a single .zip file. You can not currently merge data from separately unzipped files (without manually editing the SamplesDB.txt file). The way you package data using mAdb is to: 1) select ALL of the projects that contain samples that you want (on the mAdb "Top Level Analysis Selection" web page); 2) further select the exact subset of samples that you want in the "Array Selection" table on the "mAdb: Database Retrieval" web page). NOTE: currently, you can only select arrays using the same GAL file so they have compatible geometries. This is on our list of things to change so you can mix and match data from different chips.

4.1.11 Java Plugins bugs

1. Java Plugins are in the process of being implemented and will be made available to the user community to enable them to add their own analysis methods to MAExplorer.

4.2 Revision notes

This section lists the revision history and is useful for deciding whether to upgrade to the most recent release. You may want to check for the latest the [current "Stable release"]() available on the MAExplorer Web site. That may be different than the [Stable release listed in this copy of the Reference Manual.]() The "Beta release" listed below the Stable release in the previous links is experimental and may generally be downloaded as it has more functionality. If you experience problems, you can just reinstall the Stable version.

> Note: An archive of some of the [stable older releases]() is available on the NCI/LECB Web site for a limited period.

- **Version 0.96.11:** Extended MJAeval methods so can more easily invoke a new .mae startup file via a client-server MAEPlugin as well as the invokeCommand method. Modified internal methods and structures to support this.

- **Cvt2Mae Version 0.67:** (08-15-2002) Fixed bug where the generated pseudoarray image was overiding the user specified (grid,row,column) geometry.

- **Version 0.96.10:** Fixed bug in pseudoarray image whereby the current gene icon (yellow circle) sometimes was not reset properly for a database with X,Y coordinate data if the current gene was selected and you switched the current sample.

- **Version 0.96.08:** Fixed bug in pseudoarray image whereby the current gene icon (yellow circle) sometimes was not reset properly if you toggled the mouseover checkbox, selected a different sample using the sample list in the pseudoarray image, or scrolled the pseudoarray.

- **Version 0.96.07:** Fixed bug in font size updates for the pseudoarray image and scatter plot. Added methods to MJAproperty to get font size and font familiy.

- **Version 0.96.06:** Modified data filter property bits for Gene, MJAgene, and Filter to make them more consistent when writing MAEPlugins.

- **Version 0.96.05:** Fixed bug in data filter that caused problems with FilterPlugin's.

- **Version 0.96.04:** Fixed bug when run database with no startup database, it does not refresh spot lists in the PseudoArray image.

- **Version 0.96.02:** Fixed bug when run database with no startup database and no default MGAP demo database. It now lets you search the disk to find a .mae startup file instead of aborting with a Dryrot error. Fixed bug with chooser where it sometimes fails to let you access the last element of a list. The NCI/LECB site and the SourceForge site now both have the same full installers with JDKs (except for AIX and HPUX) and with the MGAP database.
A Cvt2Mae **Version 0.66:** bug has been fixed that makes it easier to automatically find the first row of spot data when scanning the users data input data file. Previously, the user might have to manually enter that starting data arow number in the Edit Layout wizard.

- **Version 0.96.01:** Initial release on SourceForge Logo. Note: The NCI/LECB site will have the same installers., but with JVMs if you want to download a version with the JVM for your computer. The SourceForce does not have the JVMs because of space limitations.

- **Version 0.95.20:** Inserted Open Source License into all code.

- **Version 0.95.19:** Changed default for (File | Save file DB) and (File | SaveAs file DB) so that it saves all samples in the new .mae startup file - not just the samples in HP-X 'set', HP-Y 'set or in HP-E 'list'. Made the three status lines in the main window non-editable.

- **Version 0.95.18:** Maintenance release. The Open Java API has been changed in the mjaGeneList, mjaGene, mjaEval classes.

- **Version 0.95.17:** Maintenance release. The Open Java API has been updated and now includes working updateCurGene(), updateFilter(), updateSlider(), updateLabel(), and close() methods as well as new methods in the MJAgeneList API class. The stable release is now the version with the MaeJavaAPI active. The previous stable release 0.95.04 did not have the API.

- **Version 0.95.16:** Maintenance release.

- **Version 0.95.15:** Maintenance release to fixed FilterPlugins so the FilterTestPlugin example is working. Fixed bug recently introduced that prevented selecting spot in pseudoarray image.

- **Version 0.95.14:** Fixed bug recently introduced that prevented menu commands from working correctly.

- **Version 0.95.12:** Maintenance release. Changes to help support MAEPlugins and additional functions in the Open Java API. Fixed Quant input data parser bug so it now handles strictly lower-case letters for (Grid,Row,Column) data.

- **Version 0.95.11:** Maintenance release. Changes to help support MAEPlugins and additional functions in the Open Java API.

- **Version 0.95.10:** Maintenance release changes to help support MAEPlugins and additional functions in the Open Java API. Menu Plugin command is enabled to load MAEPlugins built using the Open Java API (under construction).

- **Version 0.95.08:** Maintenance release to help support MAEPlugins and additional functions in the Open Java API. Menu Plugin command is enabled to load MAEPlugins built using the Open Java API (under construction). Added Plugins/ folder to distribution that is installed where you install MAExplorer. It contains several alpha-level JAR file plugins that work with the (Plugins | Load Plugin) command. We will be releasing the source code soon as the Open Java API alpha release in put on the MAEPlugins Web site.

- **Version 0.95.05:** Maintenance release. Changes to help support MAEPlugins and additional functions in the Open Java API. Menu Plugin command is enabled to load MAEPlugins built using the Open Java API (under construction). Added (Plugins | Test Plugin Code) to invoke code for a plugin a debugging context (maintenance).

- **Version 0.95.04:** Same as Version 0.95.03 optimized. It was compiled with the optimizing Java compile without debugging symbol tables. All Plugin support codes was removed. This results in a MAExplorer.jar file of 449Kb vs 599Kb for. Since the MGAP Web site Applet version can not use Plugins, this version is optimized for Applets and is to preferred over the Beta 0.95.03 version for now.

- **Version 0.95.03:** Maintenance release. Changes to help support MAEPlugins. Menu Plugin command is enabled, but will only load MAEPlugins built using the Open Java API (under construction).

- **Version 0.95.01:** Increased the number of spots/array that may be used from 16K to unlimited. Maintenance release. Changes to help support MAEPlugins.

- **Version 0.94.15:** Fixed bug where occasionally, it would not load additional samples from the (Samples | Set Samples from Lists | ...) commands.

- **Version 0.94.14:** Maintenance release. Changes to help support MAEPlugins. This release fixes a menu-related bug that occured if you used (File |Open File DB) to switch to a different database or to start the initial database rather than starting directly from the .mae startup file. The problem was that some of the sub menus generated might not correspond to the type of database. For example, if the previous database was an intensity database with duplicates F1 and F2, switching to a Cy3/Cy5 database might have some of the menus still appear as (F1/F2).

- **Version 0.94.12:** Maintenance release. Changed placement of magenta circle marker so it is more visible in ClusterGram when selecting genes in the dendrogram/ClusterGram for the EGL by clicking on a gene with the Control key pressed.

- **Version 0.94.11:** Fixed fatal error introduced ni 0.94.10 that occurs if you start the database with no data. Version 0.94.10 did work if you started it with a particular .mae startup file.

- **Version 0.94.10:** Maintenance release. The "Filter by positive Quant data" toggle checkbox in the (Analysis | Filter menu) is enabled for all databases - not just those requesting it. If the database has 2 channels (F1, F2) or (Cy3,Cy5) each channel is checked. If the background correction is enabled, the background corrected values are tested to see if any of them are negative. Changed the labels in the scatter plot so that instead of displaying the current gene values as **intensX=** and **intensY=** or **intens'X=** and **intens'Y=** , it now displays it as **X=** and **Y=** or **X'=** and **Y'=**. The **'** indicates that background correction is active. A new (Analysis | Plot | Show Microarray | **"Scale pseudoarray image by 1/100 to zoom low-range values**) command has been added to rescale intensity and (Cy3+Cy5) and (HP-X + HY-Y) (Red-Yellow-Green) plots to so low values are easier to visualize.

- **Version 0.94.09:** Maintenance and Reference Manual update. Added copy of "Show 'Edited Gene List'" to Edit EGL

submenu so easier to find.

- **Version 0.94.08:** Changed button layouts on popup plots so it is easier to resize them to remove the right part of the plot if you want to conserve screen space.

- **Version 0.94.07:** Fixed bug recently introduced in the histogram plots which prevent them from working. Fixed another histogram plot bug where the frequency count for "><" thresholding was reported incorrectly.

- **Version 0.94.06:** Added the saving of the current values of all of the State Threshold scrollers when the (Edit menu | Preferences | **Adjust all Filter threshold scrollers**) in the message logging area when the State Thresholds popup window is closed. The Cvt2Mae (Version 0.60) has been extended to now handle Scanalyze data files as well as supporing a separate GIPO input data file.

- **Version 0.94.05:** Added logging of MAExplorer messages to buttons in popup plots and reports. Fixed bug where old fixed gene sets for empty wells were not always restored correctly fro the previous state.

- **Version 0.94.04:** Added logging of MAExplorer messages and command history in popup windows in the View Menu as "Show log messages" that occur during a session and "Show log of command history" respectively. The windows may be saved in log files. Added additional error detection of bad (field,grid,row,column) data when reading the GIPO and Quant files.

- **Version 0.94.03:** Maintenance release for improving MAEPlugin support.

- **Version 0.94.02:** Minor bug fixes related to the renaming of **HybProbe** to **Samples** menus. Also, extensive update of reference manual in both text and figures to showing many of the renamed menus so it is synchronized with the program.

> **Version 0.94.01: Major version release.**

Renamed all previous references in the program to "hybridization probe" or "hybridization sample probe" to the new term "hybridized sample" for clarity. Changed many "HybProbe Menu" to "Samples Menu" and also many menu selections as well as plot and report labels to reflect this change. We are in the process of updating the manual computer screen figures and PDF slide presentations so that "hybridization sample probe" is shown as "hybridized sample". Also fixed other minor problems including fixing the inverted color scale for the "Pseudocolor Red(Cy5)-Yellow-Green(Cy3) Cy5+Cy3 ratio or Zdiff" command.

- **Version 0.93.03:** Minor reorganization of (Plots | Clusters) submenus. This is the last release using the "HybProbe" menu.

- **Version 0.93.02:** Beta-test of JDK-1.3 MAExplorer.jar file and new InstallAnywhere 4.5 installer. This adds JDK-1.3 installers for MacOS-X, IBM AIX, HP-UX, Linux. However, MacOS 8.1-9.x use the MRJ-2.2.5 JDK1.1.8 library. It also fixes a problem recently found running MAExplorer on some Windows-NT systems.

> **Version 0.93.01: Major version release.**

Moved "Cluster Plots" submenu of the "Plots" submenu up one level in the "Analysis" menu.

- Version 0.92.24: Maintenance and added more Plugins support. Top level Plugins menu appears, but is disabled in production version.

- Version 0.92.23: Maintenance and changed names of some commands and some messages. Major edit of Reference Manual including expanded table of contents.

Major version release. Optimized colors in grayscale display for "Pseudocolor (HP-X,HP-Y) 'sets' p-value". Fixed error and optimized t-test computation.

- Version 0.92.21: Changed grayscale display for "Pseudocolor (HP-X,HP-Y) 'sets' p-value" to display it as a color spectrum. Changed data displayed so p-Value is in the front of the list when you click on a spot etc. Added tests to disallow this display unless the data supports it.

- Version 0.92.20: Beta testing new pseudoarray image display "Pseudocolor (HP-X,HP-Y) 'sets' p-value" that displays a gray value proportional to the p-Value in a t-Test of the HP-X 'set' vs HP-Y 'set'.

- Version 0.92.18: Improved EP plot labels so it no longer prints the "MID:" prefix. Improved spot features list so it no longer prints the "plate[null,null,null]" if the plate coordinates are not present in the GIPO file.

- Version 0.92.17: Fixed a bug which incorrectly computed the max and minimum ratio Cy3/Cy5 ratio values if the data allows negative intensity on these channels after background subtraction. This is required to support the GenePix array data format converter in the new version of Cvt2Mae (11-24-2001).

- Version 0.92.15: A new (Edit | Preferences) command **Adjust all Filter threshold scrollers** - popup the state scroller window with all of the thresholds. This is useful when you want to adjust thresholds **before** you enable data Filtering or clustering. A bug was fixed where the state of the previous clustering method was not always cleared if you aborted clustering by clicking on the delete window button (eg. in Windows or the Mac). Click on a row with the Control (Shift) key pressed in the ClusterGram for hierarchical clustering will add(remove) that gene to(from) the Edited Gene List. If the (View | Show 'Edited Gene List') is enabled, then it will draw a magenta '*' before the gene name to indicate that you have selected it. This lets you select a particular subset of the genes from the hierarchical cluster. The Cvt2Mae data converter now has an alpha-release to let you convert non-standard data using the **<User-defined>** data option.

- Version 0.92.14: The name of sample names and sample file names IDs has been made more flexible. If the "Database_File" is longer than 32 characters, there may be problems reading that file with MacOS-8/9. Therefore, you can use the "DatabaseFileID" to specify the "Database_File" (in which case they would be the same). Then, the labels used for the samples are the "Sample_ID" fields. This is upwards compatible with the previous method which used the "Database_File" for both the file name and the label name. The Reports of Highest (Lowest) Cy3/Cy5 data for a single sample were not easily accessible and this has been fixed. Alternate names for automatic detection of ESTs, ESTs similar to named genes were expanded and also includes the "EmptyWell" gene set (i.e. spots with no genes) (see Automatic Gene Class naming based on Gene Name in Appendix C Table C.4.1). A clustering option was added, **Use median instead of mean for K-means clustering**, of mean when computing K-means clustering (Bickel, 2001)).

- Version 0.92.13: Maintenance release. Changed per-sample Good Spot filter algorithm and the CV filter algorithm. Added options to filter by either "HP-X 'set'" or "HP-Y 'set'". Algorithm now uses filters "HP-X or HP-Y" and "HP-X or HP-Y 'sets'" rather than the previous "HP-X and HP-Y" and "HP-X and HP-Y 'sets'". Adding some of the missing state that was not being saved during a "SaveAs DB" operation when creating a checking .mae startup file.

- Version 0.92.12: Maintenance release for Plugins.

- Version 0.92.11: Maintainence release. Increased width of scrollable sample names selection windows so it is easier to see the details on long sample names.

- Version 0.92.10: Maintenance release. Due to a problem that some users are experiencing with the Java Installer, we are reverting to the previous version 4.01 (that does not support Mac-x). We will resolve these problems as soon as possible. Also fixed a problem with the **Pseudocolor Red(Cy5)-Yellow-Green(Cy3) Cy3/Cy5 ratio or Zdiff** which is now resolved.

- Version 0.92.09: Maintenance release. Distribution of Java application with InstallAnywhere 4.5 (previous was version 4.0.1). This now supports MacOS-X, other Unix systems and may fix some other installation problems.

- Version 0.92.08: Fixed bug in **Edit use (Cy5/Cy3) else (Cy3/Cy5) for each HP** command in HybProbe menu that changes were not saved in the state. Fixed bug introduced in **HP Cy3 vs Cy5 intensity** scatter plot.

- Version 0.92.07: Added display of Cy3 and Cy5 channel intensity when click on gene in **Pseudocolor Red-Yellow-Green Cy3/Cy5 ratio or Zdiff** mode. Also added additional support for plugin installation from Configuration DB and .mae startup files.

- Version 0.92.06: The **Pseudocolor Red-Yellow-Green Cy3/Cy5 ratio or Zdiff** - display the Cy3 and Cy5 microarrays as a pseudocolor image as the sum of Cy3 (Red) and Cy5 (Green).

- Version 0.92.04: Internal changes. Also fixed bug in reading Affymetrix data converted with Cvt2Mae in the case where there is no genomic identifier and the "Location" is used as the genomic identifier.

- Version 0.91.03: Made configuration database "swapRowsColumns" default to FALSE. This should probably never need to be true except for the initial MGAP database. If this flag was not set to false in previous configuration DB files, it would swap rows and columns in the pseudoarray image display.

- Version 0.91.02: maintenance revision.

> **Version 0.91.01: Major version release.**

This corrects a few minor bugs including crashing when starting up an empty database (that bug was introduced sometime in the last month). It is the first release with some reorganized code.

- Version 0.90.08: Fixed bug in "Filter by 'Good Spot data'" where the initial "Check spots for Good Spot mode" was not properly initialized.

- Version 0.90.07: Additional fixes in QualCheck mapping now handles the three types: Alphabetic codes, MAExplorer integer property codes, continuous floating point monotonic quality value. If the latter is used, then a "Spot Quality threshold" slider will popup when it is invoked. The MAExplorer Plugins Web page is also available at thei Web site and will track progress with the development of the MAExplorer Plugins Open Java API for investigators adding their own analysis methods.

- Version 0.90.06: Fixed bugs in Genomic database popups. It was ignoring the request to popup a web page for particular genomic IDs.

- Version 0.90.05: Added QualCheck mappings of Affymetrix "Abs Call" of "P" (or "G" or "T") to Good Spot, "A" to Bad Spot, and "M" (or "B" or "F") to Marginal Spot. This allows the "Filter by 'Good Spot data'" to handle this type of data.

- Version 0.90.04: Added experimental GenomicMenu/URL configuation options for future use.

- Version 0.90.03: Added QualCheck data filter for "Filter by 'Good Spot data'". This will be active if the QualCheck data is available in the .quant files.

- Version 0.90.02: Fix bug in some of the State scroller controls when adjusting a scroller latches into a fixed state that can only be changed by restarting MAExplorer. Added LocusLink popup Web pages from LocusID (if it exists) or LocusLink from GenBank identifiers.

(very Beta) Changing convention so use "MasterID" as the master gene index. This makes it more flexible than when used the CloneID as the master gene index. Added LocusLink.

- Version 0.89.41: Added new scatter plots for comparing Cy3/Cy5 ratio data to compare either Cy3 or Cy5 of HP-X against Cy3 or Cy5 for HP-Y. Added additional change limit constraints "Compare channels meeting range" to "Filter by spot intensity [SI1:SI2] sliders". This allows partial filtering by spot intensity.

- Version 0.89.40: Some of the State Scrollers were too sensitive at low values. Therefore, we set them to respond non-linearly with a more precise vernier at the low end.

- Version 0.89.39: Fixed bug t-Test if using multiple samples with arrays with 1 Field.

- Version 0.89.38: Fixed bug in Filter that allows the user to use either "positive only" or "positive or negative" intensity data (used with Affymetrix data).

- Version 0.89.37: Increased the dynamic range of the scrollers for p-value, cluster distance, [Z1:Z2], |Diff XY|. Fixed labels in the Cy3 vs Cy5 plot so they are updated when the current sample changes.

- Version 0.89.36: Fixed error in "Inside Range" for "Filter by Ratio ..." and "Filter by Cy3/Cy5 Ratio ...". Fixed HP-X,Y,E labels in the popup HybProbe menu "Choose HP-X,Y,E ..." command. Also added a new button "<< that removes all entries from the Selected list to the Remainder list. Changed the way the Cy3 vs Cy5 (for ratio data) or F1 vs F2 (if you are using intensity data where all spots are replicated) scatter plots work - the HP-X vs HP-Y scatter plots are not changed. The data Filter still works for all of the plots. Changed the name of the "NPN clustering" method in the menus and reports to "K-means clustering" since that is essentially what it is.

- Version 0.89.35: Improved HP sample ID reporting in various lists, lists, etc. Added HybProbe menu command "Edit use (Cy5/Cy3) else (Cy3/Cy5) for each HP" for use with ratio data. This selectively swaps (Cy3,Cy5) data entries so may use (carefully!) dye-swap data for replicates. Added UP/DOWN buttons in the HP chooser to make it easier to adjust the order of the HP-E list.

- Version 0.89.34: Changed the pseudoarray image generation paradigm so that it uses a fixed size spot and spacing and makes the pseudoarray image size dependent on the number of grids, rows, columns and fields. This results in a more consistent interface.

- Version 0.89.33: Added "Filter by Cy3/Cy5 HP-X ratio or Zdiff sliders" to the Filter menu. This is useful for filtering data from a single sample.

- Version 0.89.31: Added Rename command for Gene Sets and Condition lists. Added "SaveAs GeneSets" to K-means clustering. This saves all of the clusters as named Gene sets ("Cluster #1", "Cluster #2", etc.);

- Version 0.89.30: Fixed problem with popup Web browser for Macintoshes.

- Version 0.89.29: Optimized sorting on startup so it starts much faster when working with large arrays.

- Version 0.89.28: Gene set and Condition set operations now let you select the sets by a pull-down choice menu as well as by typing in names of sets. Additional error checking added for input data files.

- Version 0.89.26: Enhanced configuration file parser and sample name menu default; better error checking on input files to help catch some errors in manually edited user data files.

- Version 0.89.25: Fixed configuration parser errors when reading non-standard data.

- Version 0.89.24a: Fixed data for MAExplorer MGAP demo.

- Version 0.89.24: Fixed bugs in popup browser to use GenBank ID (if it exists) with NCBI server, else it uses the Clone-ID (if it exists) with the nciarray cloneID-to-GenBank server.

- Version 0.89.22: Fixed bugs in ClusterGram, saving current Gene Class.

- Version 0.89.21: Fixed bugs in HP Chooser, Pseudocolor Red-Yellow-Green HP-XY plot menu, HP vs. HP correlation Report.

- Version 0.89.20: Added correct scale for (Plot | Show Microarray | Pseudocolor Red-Yellow-Green HP-XY ratio or Zdiff).

- Version 0.89.19: Added (Plot | Show Microarray | Pseudocolor Red-Yellow-Green HP-XY ratio or Zdiff) command to display the X and Y data as an additive sum of red (HP-X) and green (HP-Y) data.

- Version 0.89.18: New version of download Java installer program that improves portability on different platforms (eg. MacOS, etc). Minor changes (spelling errors in parts of the program). Improved documentation.

- Version 0.89.17: Added commands and functionality to a) find all copies of a named gene in the guesser using the SetE.G.L. button, b) added the "Replicate genes" GeneClass, c) added the "Filter by Genes with replicates" data Filter.

- Version 0.89.16: Fixed several consistency checks when reading non-standard array data.

- Version 0.89.15: Ignore enclosing space characters in SamplesDB, Configuration and .mae startup files.

- Version 0.89.14: Fixed bug in Gene guesser when press the "Done" button. It sets the gene, but did not update the pseudoarray image. Disable genomic browser options in View menu if dbEST3'/5' or GenBankAcc3'/5' is not in the GIPO file database. Since GenBank and dbEST may be accessed via the mAdb Clone report or UniGene report, this should not be a problem. In the Reference Manual a) figures were updated for Genes instead of the older notation of Clone (Version 0.88.*); b) the Short Tutorial (Appendix A) was updated and clarified.

- Version 0.89.12: Added "/Report" directory to hold SaveAs text (.txt) and plot window image (.gif) files. Improved fatal error (called DRYROT errors) reporting. Additional parts of the state saved.

- Version 0.89.11: Added "Filter by positive quant. data" to filter out raw quantified data with negative values which may occur with some types of arrays. Also improved some of the popup window updates when the data Filter, normalization, EGL, or current clone changes.

- Version 0.89.10: Fixed bug in correlation coefficient computation so it works correctly with all Gene Classes and data Filter.

- Version 0.89.09: Fixed bug in reading configuration file that occurs if SubMenus are omitted. Fixed bug recently introduced when updating the finding clones similar to the current clone.

- Version 0.89.08: Fixed bug in auto-update of List of Gene Sets when used with adding(removing) genes using the Control(Shift) clicking on a gene in the array image, scatter plot, etc.

- Version 0.89.07: Fixed bugs in state saving.

- Version 0.89.06: Save the full State when do a (File | SaveAs DB). The full state is restore when you startup the database

you saved. Corrected bug in Gene Set Difference operator. Clarified some menu option names. Can now set # genes to report or Filter in a highest/lowest ratio genes in (Filter | Set max # genes in highest/lowest report). Added Report for (Report | Genes in 'Normalization Gene List').

- Version 0.89.05: Added information fields in HP chooser.

- Version 0.89.04: Corrected bug in Ratio Median correction option if using ratio data (i.e. Cy3/Cy5). Improved scatter plot labeling.

- Version 0.89.03: Optimized the main screen area usage. Corrected data-dependent bug which sometimes occurs with the Use Ratio Median correction option if using ratio data (i.e. Cy3/Cy5).

- Version 0.89.02: Set the Use Ratio Median correction option if using ratio data (i.e. Cy3/Cy5).

Version 0.89.01: Major version release.

Because MAExplorer can be used with both spotted clone arrays and oligo arrays,we renamed clones as genes (except where Clone ID is used) in both the MAExplorer program and in the Reference Manual.

- Version 0.88.08: Enabled saving the name of the last project when you exit so it knows which file to use as a default when you do an "Open disk DB". Fixed the mnX/mnY data in XY-Statistics Report. Fixed title update in main window when open a new database. Also has fix for Linux Java installer (tested on Red Hat Linux version 6.0).

- Version 0.88.07: Disabled the (File | Set Project) command as there is occasionally a problem with reading the maeProject.txt database file which causes it to hang. This should not be a problem for users with multiple databases as you still get the popup startup file browser when you invoke (File | Databases | Open file DB).

- Version 0.88.06: Edit submenu "Sets of Conditions" HP lists is working and is saved/restore with the File menu "Save disk DB"/"Open disk DB".

- Version 0.88.05: Added mouse-over to main window and moved gene locator button to left. Added "Use ratio median correction" option to Normalization menu. This option can be used for rescaling Cy3, Cy5 intensity data if the medians are unequal. Also, the manual has been updated and more discussion on data mining has been added.

- Version 0.88.04: Edit submenu "Sets of HybridProbes" Condition lists are partially working (not working with File menu "Save disk DB"/"Open disk DB", set operations not enabled yet).

- Version 0.88.03 Edit submenu "Sets of clones" and File menu "Save disk DB"/"Open disk DB" Save/Restore clone sets as .cbs clone bit set files in stand-alone.

- Version 0.88.02: cleaned up SamplesDB.txt and MaExplorerConfig.txt specifications and sample files in Appendix C.

- Version 0.87.08: Fixed some bugs in HP-X,-Y,-E chooser. Added additional checking for missing field names in data files. If a field is incorrect, MAExplorer can't continue. It now gives a popup "Dryrot" error message to give us information about which fields are incorrect.

- Version 0.87.06: May now use combined chooser for selecting the active HP-X, HP-Y and HP-E probes.

- Version 0.87.05: May now save all popup text windows as .txt files in stand-alone mode.

- Version 0.87.01: May now save all plot windows as GIF images as .gif files in stand-alone mode.

- Version 0.86.36: Fixed bug in "Filter by spot intensity [SI1:SI2] sliders". Allow simultaneous viewing of HP-X vs HP-Y and Cy3 vs Cy5 (or F1 vs F2) scatter plots. You may now also simultaneously view and HP-X/HP-Y and Cy3/CY5 ratio histograms.

- Version 0.86.35: A new "Filter by spot intensity [SI1:SI2] sliders" mode has been added to the Filter menu. It is useful for filtering individual changes when analyzing Cy3/Cy5 data.

- Version 0.86.34: The new stand-alone distribution is now generated with InstallAnywhere 3.5 instead of 3.5 beta. This corrects some of the problems we had seen with very large core images.

- Version 0.86.34: The handling of "Out of memory errors" seen when clustering large numbers of clones on a small computer has been changed. Previously, it shut off the cache and continued the computation - although VERY slowly. Now, it prints out the following message:

```
There is not enough memory to cluster current filtered clones. Options:
    1. reduce the number of filtered clones and try again, or,
    2. disable cluster-cache (Clustering menu) - will be VERY slow.
```

- Version 0.86.33: A new "Set project" command has been added to the File menu. It selects the current project, if multiple projects have been set up (using the future "New project" command).

- Version 0.86.29: A new "Show mouse-overinfo" mode has been added to the View menu (default is on). It reports HP probe and spot/clone details in the array image, scatter plot, expression profile overlay plot, etc.

- Version 0.86.27: In stand-alone mode, the "Open disk DB" now lets the user open a .mae startup file. The user browses the local file system to select the startup file to use.

- Version 0.86.25: When using ratio (i.e. Cy3/Cy5) data, the menu entries, plot and report legends change the names from (F1,F2) to (Cy3,Cy5) where appropriate.

- Version 0.86.24: Corrected bug in using the ratio and intensity thresholding of mean HP data was previously using the single probe data.

4.3 Web Browser problems when running MAExplorer as an applet

Because MAExplorer is a large system, and there may be occasional problems running it in some Web browsers on some operating systems. We recommend you run MAExplorer as a stand-alone application as it is more robust.

- Normally, MAExplorer works well as an applet with Netscape 4.7 or Windows Internet Explorer 5.0 on Windows PCs and on SUN Unix. It does not seem to work well as an applet on Macintoshs and problems have been reported on SGI systems.

- One solution is install Sun's HotJava browser. The HotJava browser is available from Sun Microsystems . You can download HotJava for the Windows-95/-98/-NT/-2000/-XP and for Solaris (Unix).

- Reguardless of which browser is used, because it is is doing a lot of computation, MAExplorer runs best on a fast machine with lots of memory. This is especially true when running a browser because of the additional browser overhead.

4.4 Handling fatal error reporting (i.e. DRYROT errors)

If you encounter a fatal error that is detected by MAExplorer, it will popup an error reporting window. We call this a "DRYROT"

error (thanks to "S.A.I.L." - Stanford AI Lab) because something is wrong in the program or in the user's data files and from which it can not recover. This type of error should not have happened. Please save and e-mail the report to us so we can try to fix the bug or diagnose the problem. The following figure shows an example of part of a DRYROT error report.

Figure 4.4 Example of a fatal Dryrot Error window. This may occur for a variety of reasons. This window lists the main reason and also lists some of the MAExplorer state information. If you wish, you may save this window (press the "SaveAs" button) and mail it to us. We may try to correct the problem in the next release if it is a problem with MAExplorer. Alternatively, it could be a user data error.
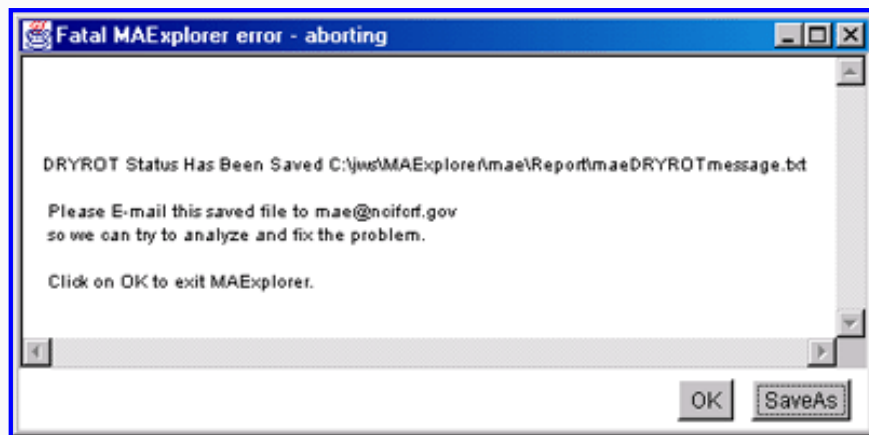
Figure 4.4.1 Example of a fatal Dryrot Error window after SaveAs. This tells you where the saved error message file was saved and the email address to send it to if you wish.

Release Archive for stand-alone MicroArray Explorer on NCI/LECB

This is an archive of some of the older stable versions of the MAExplorer stand-alone application program. These are the full installers which include the Java JDKs for all operating systems as well as the MGAP database. The user reference manual (available as a zip file) specific for that version is also included. After a while, we will remove some of the older releases. To find what the current and beta releases are, see the Install home page. The changes between releases are listed in Section 4.2 Revision Notes.

Release	Release Date	Manual (.zip) for Release
0.96.02	07-02-2002	-
0.95.20	05-31-2002	MaeRefMan.zip (10Mb)
0.95.16	05-24-2002	MaeRefMan.zip (10Mb)
0.95.04	03-22-2002	MaeRefMan.zip (10Mb)

Acknowledgements

Primary contributers to MAExplorer were Peter Lemkin (LECB/NCI), Greg Thornwall (SAIC/FCRDC) in the Laboratory of Experimental and Computational Biology, NCI/NIH, and Jai Evans (DECA/CIT/NIH).

Primary contributers to Cvt2Mae were Peter Lemkin (LECB/NCI), Greg Thornwall (SAIC/FCRDC), Bob Stephens (ABCC/NIH).

We wish to thank the many members of Lothar Hennighausen's Laboratory of Genetics and Physiology (NIDDK) who inspired the initial development of MAExplorer and its continued development. Thanks also to:

Greg Alvord (SAIC/FCRDC),

References to related exploratory data analysis methods and MAExplorer

This short list of references is limited to a few related to exploratory data analysis methods for microarrays as they relate to MAExplorer. It is not meant to be inclusive. More extensive lists of references to many of the array preparation and data mining methods can be found in some of these papers and on the Internet.

1. Beardsly T (1999) *Scientific American*, March 1999, 35-36.

2. Bickel, D (2001) "Robust Cluster Analysis of DNA Microarray Data: An Application of Nonparametric Correlation Dissimilarity. Proceedings of the Joint Statistical Meetings of the American Statistical Association (Biometrics Section, to appear). [http://www.mathpreprints.com/math/Preprint/bickel/20010730/3/]

3. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME) - toward standards for microarray data. *Nature Genetics* **28**: 365-371. [See http://genomics.nature.com/supplementary_info, and http://www.mged.org/)]

4. Cleveland WS (1985) The Elements of Graphing Data. Wadsworth Press, Monterey, CA, pp 1-323.

5. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM (1996) Use of a cDNA microarray to analyze gene expression patterns in human cancer. *Nature Genetics* 1996, **14**: 457-460.

6. Dudoit S, Yang YH, Callow MJ, Speed TP (2000,Aug) Statistical methods for identifying differentially expressed genes in replicated microarrays experiments. TR-578, Dept. of Statistics, Stanford Univ., Stanford, CA.

7. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *PNAS USA* **95**: 14863-14868.

8. Jagota, A (2001) <u>Microarray Data Analysis and Visualization</u>. <u>Bioinformatics By The Bay Press</u>, 101 pp. ISBN-09700297-3-X.

9. Kaufman L, Rousseeuw PJ (1990) <u>Finding Groups in Data: An Introduction to Cluster Analysis</u>. J. Wiley and Sons, Inc., New York.

10. Kerr MK, Churchill GA (2001) Statistical design and the analysis of gene expression microarray data. *Genet Res.* **77**(2):123-8. Review.

11. Kerr MK, Churchill GA (2001) Bootstraping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *PNAS* **98**:8961-8965.

12. Korn EL, Troendle JF, McShane LM, Simon R (2001) Controlling the number of false discoveries: Application to high-dimensional genomic data. BRB, NCI, Bethesda, MD. Aug 22, 2001, Tech Report #3.

13. Lipkin LE, Lemkin PF (1980) Data-base techniques for multiple two-dimensional polyacrylamide gel electrophoresis analyses. *Clin. Chem.* **26**: 1403-1412. (GELLAB-II home page http://www.lecb.ncifcrf.gov/lemkin/gellab.html)

14. Lemkin PF, Lester EP (1989) Database and search techniques for two-dimensional gel protein data: a comparison of paradigms for exploratory data analysis and prospects for biological modeling. *Electrophoresis* **10**: 122-140.

15. Lemkin PF (1995) Representations of protein patterns from 2D gel electrophoresis databases. In: Pickover, C., (Ed) <u>The Visual Display of Biological Information</u>, World Scientific Publishers, River Edge, New Jersey, pp 43-59.

16. Lemkin PF (1993) Xconf: a network-based image conferencing system. Comput. Biomed. Res. 26, 1-27.

17. Lemkin PF (1997) Comparing Two-Dimensional electrophoretic gels across the Internet. *Electrophoresis*, **18**: 461-470. [Extended paper], (Web site http://www.lecb.ncifcrf.gov/flicker/)

18. Lemkin PF, Thornwall G (1999a) Flicker image comparison of 2-D gel images for putative protein identification using the 2DWG meta-database. *Molecular Biotechnology*, **12**:(2) 159-172. [Extended paper and Web site http://www.lecb.ncifcrf.gov/2dwgDB/]

19. Lemkin PF, Myrick JM, Lakshmanan Y, Shue MJ, Patrick JL, Hornbeck PV, Thornwall GC, Partin AW (1999b) Exploratory Data Analysis Groupware for Qualitative and Quantitative Electrophoretic Gel Analysis Over the Internet - WebGel. *Electrophoresis* **20**:(18) 3492-507. (PDF), (also see WebGel server: http://www.lecb.ncifcrf.gov/webgel).

20. Lemkin PF, Thornwall GC, Walton KD, Hennighausen L (2000) The Microarray Explorer tool for data mining of cDNA microarrays - application for the mammary gland *Nucleic Acid Research* **28**:(22) 4452-4459. (PDF).

21. Lemkin PF, Thornwall GC, Walton KD, Hennighausen L (2000) MicroArray Explorer - a tool for data mining of microarrays - Overview. (PDF).

22. Lemkin PF, Thornwall GC, Walton KD, Hennighausen L (2000) MicroArray Explorer - a tool for data mining of microarrays - Examples. (PDF).

23. Lemkin PF, Thornwall GC, Powell J, Asaki E (2000) Using MAExplorer with the NCI/CIT mAdb Web server. (PDF) and [http://nciarray.nci.nih.gov/]

24. Lemkin PF (2001) Introduction to Data Mining of Microarrays using the MicroArray Explorer. ([PDF](#)) or ([PPT](#)).

25. Lemkin PF (2001) Using Cvt2Mae to convert Affymetrix array data for MAExplorer. ([PDF](#))

26. Lemkin PF, Thornwall G, Hennighausen L (2001) MicroArray Explorer - A Java-based Tool For Data Mining Microarrays. Paper presented at the AMS-IMS-SIAM Summer Conference on Statistics in Functional Genomics, June 10-14, 2001. ([PDF](#)).

27. McShane LM, Radmacher MD, Freidlin B, Yu R, Li M-C, Simon R (2001) Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. BRB, NCI, Bethesda, MD. July 1, 2001, Tech Report #2.

28. Radmacher MD, McShane LM, Simon R (2001) A paradigm for class prediction using gene expression profiles. BRB, NCI, Bethesda, Nov 19, 2001, Tech Report #1.

29. Schneiderman B (1997) Designing the Human Interface, 3rd edition. Addison-Wesley Pub. Co., NY, pp 1-638.

30. Schulze A, Downward J (2001) Navigating gene expression using microarrays - a technology review. *Nature Cell Biology* **3**: E190-E195.

31. Simon R, Radmacher MD, Dobbin K (2001) Design of Studies Using DNA Microarrays. BRB, NCI, Bethesda, MD. 2001, Tech Report #4.

32. Sneath PHA, Sokol RR (1973) Numerical taxonomy. W.H. Freeman and Co, San Francisco. pp 1-573.

33. Tufte E (1997) Visual Explanations. Images and quantities, evidence and narrative. Graphics Press, Cheshire, CT, pp 1-156.

34. Tukey J (1977) Exploratory data analysis. Addison-Wesley Pub. Co., Reading, MA, pp 1-688.

35. Vawter MP, Barrett T, Cheadle C, Sokolov BP, Wood WH III, Donovan DM, Webster M, Freed WJ, Becker KG (2000) Application of cDNA Microarrays to Examine Gene Expression Differences in Schizophrenia. Submitted.

36. Weinstein JN, Myers TG, O'Conner PM, Friend SH, Fornace AJ, Kohn KW, Fojo T, Bates SE, Rubinstein LV, Anderson NL, Buolamwini JK, van Osdol WW, Monks AP, Scudiero DA, Sausville EA, Zaharevitz DW, Bunow B, Viswanadhan VN, Johynson GS, Wittes RE, Paull KD(1997) An Inforation-Intensive Approach to the Molecular Pharmacology of Cancer. *Science* **275**: 343-349.

37. White KP, Rifkin SA, Hurban P, Hogness DS (1999) Microarray Analysis of *Drosophila* Development during Metamorphosis. *Science* **286**: 2179-2184.

Appendix A. Short tutorial for MAExplorer

This tutorial is for use with MAExplorer, an exploratory data analysis facility for microarray DNA databases. It may be used with *any* MAExplorer database. As with all tutorials, they are only starting points for getting you started - in this case into understanding the data mining analysis environment. Try out new options on your own, **you can't break anything** :-).

This tutorial lets you

1. Analyze expression of individual genes
2. Analyze expression of gene families and clusters
3. Compare expression patterns in multiple hybridized samples

A.1 Demonstration data

Note that the downloadable MAExplorer stand-alone application includes a subset of 50 hybridized samples from the MGAP database including a number of startup files for that data (see the the list of startup .mae files included in the download installation). If you have access to another MAExplorer database, you can use it.

Using the stand-alone application for the tutorial

These same subsets as well as other subsets of the MGAP data are available in the set of .mae startup files distributed with MAExplorer. To access these files,

1. Start MAExplorer after you have installed it. Eg. in Windows, go to the Windows "Start Menu" and click on MAExplorer. If it is not in your Start Menu, you can go to where you installed it (typically `C:\Program Files\MAExplorer`) and click on `MAExplorer.exe`.
2. Then after it starts, go to the "Files" menu and select "Open disk DB" and select the startup file you want. Alternatively, you can go directly to the list of startup files in `C:\Program Files\MAExplorer\MAE`) and double-click on one of the startup files.

A.2 General instructions:

Throughout this tutorial we refer to condition X and condition Y. These are different hybridized samples in the particular database you have loaded. For example, in the MGAP database X might be lactation and Y might be pregnancy. X and Y 'sets' are multiple samples of these two conditions.

☐ **First, select one of the start up databases.**

- As a stand-alone application, select the startup file entry (files ending with a ".mae" file extension) from a directory of startup files on your local computer. Generally these are in a subdirectory called MAE in a project directory (see Appendix C. Use of MAExplorer with other microarrays).

If the particular samples you want to analyze are not listed in that example, after it starts you will be able to add samples you do want and remove samples you don't want - regardless of which example was intially used if the database "Samples" database contains additional hybridized samples.

When it starts, a main window will pop up. It then downloads a gene database tables and the particular hybridized samples you specified. When it is ready for you to begin interaction, the menu bar will become active and it will display a green Ready - click on a gene to query database message. Depending on your Internet connection speed, it may take a few minutes to set up. If you are running MAExplorer as a stand-alone application and it is getting data from your local disk, startup will be much faster.

☐ **Second,** go to the A.3 instructions for self-guided tutorial below for instructions on what to do next.

HINT: print this tutorial page and then read the following instructions from the printout rather than trying to keep this window visible. You might also print the parts of the MAExplorer Reference Manual for the same reason.

HINT: You might want to keep a record of the commands you have used or the messages and measurements you have made. To do this you need to enable message and command history logging. Go to the View pull-down menu and then select the type of logging you want using the Show log of messages or the Show log of command history commands.

NOTES:. On computers with low resolution (i.e. less than 1024 X 780) you may need to resize the windows and move them to different parts of the screen to view them simultaneously.

A.3 Self-guided tutorial of MAExplorer - notation and examples

The following is a self-guided tutorial (**you** issue the commands) that illustrates some of the data analysis capabilities. In the following examples, the notation **"go to A:B:C"** means go pull-down menu A, then submenu B and, then make selection C. "Selecting a gene" from the microarray image or scatter plot means clicking on a spot in the pseudoarray image or a point in the any of the plots.

A.3.1 Review of types of gene data available in the database

step 1: go to Analysis: GeneClass: All genes
　　　 the array shows *all genes* with white circles.
step 2: go to Analysis: GeneClass: All named genes
　　　 the array shows *named genes* with white circles.
step 3: go to Analysis: GeneClass: ESTs similar to genes
　　　 the array shows *ESTs similar to named genes* with white circles.
step 4: go to Analysis: GeneClass: ESTs
　　　 the array shows *unknown ESTs* with white circles.
step 5: go to Analysis: GeneClass: All genes and ESTs
　　　 the array shows *all named genes and all ESTs* with white circles.
step 6: go to Analysis: GeneClass: Replicate genes
　　　 the array shows *replicate genes* having at least 2 copies in the
　　　 array with white circles.
step 7: go to Analysis: GeneClass: Calibration DNA
　　　 the array shows *calibration DNA* (if present) with white circles.
step 8: go to Analysis: GeneClass: Your plates
　　　 the array shows *clones from user's plates* (if present) with white circles.

A.3.1.1 Analysis of the expression of a single known gene

　　a. ratio between two conditions X and Y (HP-X, HP-Y)
　　b. expression profile of a set of conditions (HP-E) (see Example A.3.1.7)

step 1: click on the blue "Enter gene name" button to pop up a name entry window
step 2: start typing gene name into blue text entry window
step 3: once gene names appear, click on gene of choice
step 4: press "Done" button in pop up window
　　　 A yellow circle will define the gene as the "current gene" in the microarray
　　　 pseudoarray image (info on gene is also provided in the status area above the array).
　　　 If there are replicate grids (left and right fields of repeated genes are denoted
　　　 by F1 and F2) in the array (HP). The mean(HP-X,HP-Y) values and the (HP-X/HP-Y)
　　　 values for the specified gene are reported are reported.
step 5: alternatively, click on an array spot of choice to define any gene
　　　 in the array as the new current gene

A.3.1.2 Find a subset of genes with a common substring (e.g. *ONCO*)

step 1: click on the blue "Enter gene name" button to pop up a name entry window
step 2: start typing "*ONCO*" (without the quotes) into blue text entry window
step 3: once gene names appear, press "Set E.G.L." button in pop up window

Magenta squares will indicate these genes in the pseudoarray image.
These include the 'onco'genes and the proto-'onco'genes

A.3.1.3 Two conditions - scatter plots:

Create a scatter plot of two hybridized samples where condition X data is on the X axis and condition Y data on the Y axis.

step 1: go to Analysis: Plot: Scatter plots: HP-X vs. HP-Y.
    then click on yellow circle in scatter plot to get HP-X/HP-Y ratio for the gene
step 2: click on any point in the scatter plot
    this also alternatively defines any gene in the plot as the new current gene
step 3: zoom in on a region of the plot using the vertical or horizontal scroll bars
step 4: click on another point in the scatter plot to get the HP-X/HP-Y ratio another gene
step 5: press "Close" button to remove pop up window

A.3.1.4 Scatter plot of Cy3 vs Cy5 or replicate spots (F1 vs F2) of one sample

Create a scatter plot of Cy3 vs Cy5 channels or replicate spot F1, F2 data if your database is contains (Cy3,Cy5) ratio data or it contains replicate spot fields (F1,F2).

step 1: go to Analysis: Plot: Scatter plots: Cy3 vs. Cy5
    <u>or</u> go to Analysis: Plot: Scatter plots: F1 vs. F2
    Then, click on green circle in scatter plot to get Cy3/CY5 ratio for the gene
    or F1/F2 ratio for replicate spots for that gene
step 2: click on any point in the scatter plot
    this also alternatively defines any gene in the plot as the new current gene
step 3: zoom in on a region of the plot using the vertical or horizontal scroll bars
step 4: click on another point in the scatter plot to get the HP-X/HP-Y ratio another gene

If you are working with Cy3/Cy5 dye-swap data, you may swap the Cy3/Cy5 channel data to Cy5/Cy3 for any selected subset of samples. This may make it easier to use the data in various ways when data mining. If you do not have this type of data, go to step 7.

step 5': go to Samples: Edit (Cy5/Cy3) else use (Cy3/Cy5) menu
step 6': select the samples you wish to swap and press "Done". This
    enables you to see the swapped results in the scatter plot
step 7: press "Close" button to remove pop up window

A.3.1.5 Filter by expression ratio between two conditions X and Y

step 1: go to Analysis: Plot: Histograms: HP-X/HP-Y
    the histogram shows the ratios
step 2: move pop up plot so you can see it and the array simultaneously
step 3: choose (click on) a ratio bin
    genes filtered by the ratio range of the bin will light up on the array ('+'s)
step 4: click on different bin in the histogram to select another bin
step 5: click on word "Freq" on left in histogram to remove the histogram bin filter

Note of caution: if the signal is close to background the X/Y ratio may be bogus.
You can filter out low intensity genes by

A.3.1.6 Filter by spot intensity range

step 1: go to Analysis: Filter: Filter by spot intensity [SI1:SI2] sliders: Use spot intensity [SI1:SI2] sliders
step 2: adjust intensity lower bound (SI1) to remove low ratio genes
step 3: when done, remove the 'Filter by intensity sliders' by toggling it off (redo step 1 to toggle it off)
step 4: repeat steps 1-3, but this time use Filter : Filter by [I1:I2] sliders :
      Use spot intensity (or Cy3/Cy5) [I1:I2] sliders

A.3.1.7 Multiple conditions - expression profile plots of HP-E data:

step 1: go to Analysis: Plot: Expression profile: Display a gene's expression profile
step 2: after the expression profile window pops up, click on a gene in array to see its profile
step 3: click on a line in the profile plot to see its intensity
step 4: click on a different gene in the array to see its profile
step 5: press "Show HPs" button to see the list of samples used
step 6: press "Close" button to remove pop up windows

A.3.2 Changing the normalization between hybridized samples

You may change the normalization method used to scale data between hybridized samples so they may be compared.

A.3.2.1 Set normalization

step 1: go to Analysis: Normalization: Median intensity
step 2: go to Analysis: Plot: Scatter plots: HP-X vs. HP-Y
      to see the effect of normalization on the scatter plot. Note how outliers appear.
step 3: go to Analysis: Normalization: Zscore of intensity
step 4: go to Analysis: Normalization: Zscore of log intensity, stdDev
step 5: go to Analysis: Normalization: Unnormalized
      this does *not scale* data between samples.
step 6: go to Analysis: Normalization: Median intensity
      this leaves the normalization method in Median mode.

A.3.3 Analysis of the expression profiles of gene classes

You may restrict the set of genes by *Gene Class*. Several built in gene classes are defined. You may also set up additional ones and filter by those (not covered in this short tutorial).

A.3.3.1 Filter by gene class membership

step 1: go to Analysis: GeneClass: All known genes
      the array only shows named genes (additional gene subclasses are being added)
step 2: go to Analysis: Plot: Scatter plots: HP-X vs. HP-Y
      to see the two condition expression of just these genes
step 3: go to Analysis: Plot: Expression profiles: Display Filtered genes expression profiles
      to see the multiple condition expression of just these genes. This may take a
      while if there are many genes
step 4: you can click on a line in any of the plots to see the samples' intensity value for that gene
step 5: when done, press "Close" button in all pop up plot windows

A.3.3.2 Gene Reports

step 1: go to Analysis: Report: Gene reports: Filtered genes: Genes passing Filter
    Clicking on a blue entry will bring up I.M.A.G.E, dbEST, UniGene, or GenBank,
    LocusLink, or mAdb Clone database in pop up Web page
step 2: press "Close" button in report, and close this pop up Web page
step 3: go to Analysis: Report: Table format: Tab-delimited
    to enable creating Excel-compatible reports

A.3.3.3 Exporting Gene Reports to Excel

step 1: repeat step 1 of the **Gene Report**, but this time to make text-formated report
step 2: cut the text from this window and paste it into an Excel window.
    This is useful for exporting data if you are on a Windows PC
step 3: go to Analysis: Filter: all genes
    to restore it to all of the genes from all named genes
step 4: go to Analysis: Report: Table format: Spreadsheet
step 5: press "Close" button in report

A.3.4 Analysis of the expression profile of multiple hand picked genes

Users can manually define a set of genes which are kept in the *Edited Gene List* (E.G.L.). Various operations can then use the EGL to restrict the set of data being analyzed.

A.3.4.1 Define a list of edited genes, then plot all their expression profiles at one time

step 1: go to View: Show 'Edited Gene List'
    this turns on the 'Edited Gene List' magenta square box overlays
step 2: hold CONTROL key and click on genes in array to add a gene
step 2': hold SHIFT key and click on genes in array to delete a gene.
    This lets you edit a list of genes. It also works when clicking in a scatter plot
step 3: go to Analysis: Plot: Scatter plots: HP-X vs. HP-Y
    to see the Edited Gene List in the scatter plot
step 4: try defining (or removing) E.G.L. genes in the scatter plot by holding the
    CONTROL (or SHIFT) key when clicking on points in the scatter plot

A.3.4.2 Filtering by edited gene list

step 1: go to Analysis: Filter: Filter by 'Edited Gene List'
step 2: go to Analysis: Plot: Expression profiles: Display Filtered genes expression profiles
    scroll through the plots to see all of the profiles
step 3: go to Analysis: Filter: Filter by 'edited gene list'
    this turns off the 'edited gene list' filter
step 4: press "Close" button in expression profiles window

A.3.4.3 Report of edited gene list

step 1: go to Analysis: Report: Gene report: genes in 'edited gene list'
    reports edited genes
step 2: press "Close" button in report

step 3: go to Analysis: Filter: Filter by 'edited gene list'
     this turns off the 'edited gene list' filter

step 4: go to View: Show 'edited gene list'
     this turns off the 'edited gene list' squares overlay

A.3.5 Identify a cluster of genes with similar expression profile to the current selected gene

step 1: go to GeneClasses: All named genes and ESTs

step 2: go to Analysis: Plot: Cluster plots: Cluster genes with expression profiles similar to current gene
     this will pop up a cluster summary and cluster distance slider control window.
     Move the summary and slider windows so you can see all 3 windows. The size of
     the cyan boxes on similar genes in the pseudoarray is proportional to the similarity.
     Adjust the cluster distance slider to smaller values and note how the number of genes clustered decreases.
     It should be set for a reasonable number considering the material you are analyzing.

step 3: select (click on) a new current gene
     the genes which belong to that cluster are labeled in the array with cyan boxes
     and are defined as the "current cluster". The current gene you click on has
     a green circle around it

step 4: press "Cluster Report" button in the cluster summary
     this pops up a Gene Report for the clustered genes

step 5: press "Close" button in the report

step 6: press "EP plot" button in the cluster summary
     this pops up a scrollable list of expression profile plots sorted by similarity
     to the current selected gene.

step 7: press "Close" button in the report

step 8: press "Close" button in the cluster summary

A.3.6 Identify clusters of genes with similar expression under various conditions using data mining filters

step 1: go to GeneClasses: ESTs similar to genes

step 2: go to Analysis: Plot: Cluster plots: K-means clustering of gene expression profiles
     this will pop up a cluster summary and slider control window. Move the summary
     and slider windows so you can see all 3 windows. The size of the
     magenta circles in the array is proportional to # genes/cluster

step 3: select (click on) a new current gene
     the genes which belong to that cluster are labeled in the array with tiny green
     numbers are defined as the "current cluster". The current gene you click
     on has a green circle around it

step 4: go to View: Show 'edited gene list'
     genes in the current cluster were also copied to the edited gene list

step 5: go to Analysis: Report: Gene report: genes in 'edited gene list'
     reports genes in the current cluster

step 6: press "Close" button in report

step 7: go to View: Show 'edited gene list'
     this turns off the 'edited gene list' squares overlay

A.3.6.1 Varying the number of clusters

step 1: vary the "# of clusters" slider value from 6 to 10, then 20
note the number of clusters changes and the gene cluster composition also changes

A.3.6.2 Defining a new cluster "seed" to recluster the genes

step 1: select a new current gene in array and press the "Recompute clusters" button
this recomputes the clusters using the current gene as the new seed gene

A.3.6.3 Cluster expression profile plots

step 1: press "EP plot" button and scroll down the list after they appear
the primary nodes for each cluster are indicated with red labels in the set of
profiles, and the other genes are labeled with their cluster number
step 2: press "Mean EP plot" button and scroll down the list after they appear
these are the mean expression plots of the primary nodes clusters.

A.3.6.4 Report of all clusters

step 1: press the "Cluster-Report" button to get a sorted cluster
list scroll the spreadsheet to the right to see the cluster statistics
step 2: press the "Mn-Cluster-Report" button to get a sorted cluster list
scroll the spreadsheet to the right to see the mean expression profiles
step 3: press "Close" button in pop up windows

A.3.6.5 Current cluster in scatter plot

step 1: go to Analysis: Plot: Scatter plots: HP-X vs. HP-Y
step 2: move the plot so you can see both scatter plot and array
step 3: click on a gene in the cluster or on spots in the scatter plot
note that the green cluster numbers are drawn in the scatter plot
step 4: go to Edit: Sets of genes : Save 'Edited gene list' as gene sets
this will pop up a dialog box requesting "Enter new gene set name"
step 5: type "Genes in current cluster class"
this will save the current cluster in a *gene set*. This gene set will
be used in the next example
step 6: press "Close" button in pop up windows
step 7: (optionally) investigate hierarchical cluster with clustergrams and
dendrograms by going to Plot : Cluster Plots : Hierarchical clustering plot for HP-E

A.3.7 User Gene Set operations

You may manipulate sets of genes. Some of these are predefined for you by the database (eg. All named genes, ESTs, etc.). Others are defined by particular operations (E.G.L., clustering, etc.), and lastly others may be defined by you using logical operations on these sets (OR, AND, DIFFERENCE).

A.3.7.1 List of the current gene sets

step 1: go to Edit: Sets of genes : List saved gene sets

this lists the current list of gene sets

step 2: Change the E.G.L. [set of genes](#) and note how the # of E.G.L. genes changes in the list.
You can add (remove) genes to the E.G.L. by clicking on a spot in the array while the
CONTROL (SHIFT) key is held down.

A.3.7.2 Filter by user defined gene set

step 1: go to Edit: Sets of genes : Set 'User Filter Gene Set' (for Filter)
this will request a gene set to use with the Filter in a pop up dialog box.
Enter gene set # for the set for "Genes in current cluster class" which you saved
in the previous example.
then press "Ok" in the dialog box.
step 2: go to Analysis: GeneClass: All genes and ESTs
this resets the filter to look at all genes and ESTs
step 3: go to Analysis: Filter: Filter by 'User Gene Set' membership
this restricts the genes to the saved current cluster in the previous example

A.3.7.3 Gene set operations

step 1: go to Edit: Sets of genes : OR (Union) of 2 gene sets
this will request 3 gene set names in a pop up dialog box.
Enter set # for (All known genes) for the 1st gene set name,
Enter set # for (Genes in current cluster class) for the 2nd gene set name,
Enter "Union of known genes and genes in current cluster" for new gene set name.
then press "Ok" in the dialog box.
this computes the union of the two gene sets into a new gene set
step 2: go to Edit: Sets of genes : Set 'User Filter Gene Set'
this will reset the 'User Filter Gene Set' for the Filter in a pop up dialog box.
Enter the set number or the beginning of the set name 'Union' that is the
set for "Union of known genes and genes in current cluster" just saved.
step 3: try saving other Filtered genes sets and doing other gene set operations.

A.4 Additional tutorials

If you wish to investigate MAExplorer in more detail, try some of the suggested examples in the [advanced tutorial](#) (Appendix B) in
the reference manual.

Appendix B. Advanced tutorial for MAExplorer

There are a number of things you may do in this facility. We wrote this advanced tutorial to help demonstrate some of its
capabilities. A [short tutorial](#) (Appendix A) is also available and we recommend doing it before attempting the advanced tutorial.
Sources of startup data to use with the tutorials are listed in the short tutorial. As with all tutorials, they are only starting points for
getting you into the analysis environment - try out new options on your own, **you can't break anything** :-).

1. Analyze expression of individual genes
2. Analyze expression of gene families and clusters
3. Compare expression patterns in multiple hybridized samples

<div style="text-align: center; border: 2px solid black; display: inline-block;">

NOTE: THIS APPENDIX IS BEING REVISED...

</div>

Here are some things to try

1. You could click on the **Reference manual** in the Help menu. That is *this* reference manual, but it appears in a new Netscape Web browser window so you may read it while working on MAExplorer. Delete the window when you are finished with it.

2. Click on the Glossary in the MAExplorer Help menu. This section describes terms used in MAExplorer. Delete the window when you are finished with it.

3. Click on the Index. It may be useful in finding things that are not in the table of contents.
4. Note: a hybridized sample is a microarray hybridized with cDNA derived from the mRNA from a particular experiment sample (see notation defined in the Overview). Currently, you may start MAExplorer with preloaded set of samples by starting the stand-alone MAExplorer on a .mae file. Alternatively, you may start it with no samples loaded. It the latter case, you would load samples you are interested in from the Samples menu.

When first started, it loads some initial data it needs as well as the particular hybridized samples you specified. After MAExplorer starts, it displays "Ready - click on a gene to query database" and the menus becomes active. Here are some things to try.

1. Click on different genes (i.e. spots in the pseudoarray image). The pseudoarray image may or may not correspond to the actual array - depending on how the data was derived. Notice the data that gets displayed in the three text lines above the image. If the spot you click on is a named gene (e.g. [1-A4,3] at row 4 column 3 in Field 1), it will also print out the GeneName of the gene.

2. Look at the pull-down menus. They consist of sets of commands with similar functionality grouped together in sub-menus. In particular, look at the Analysis menu. It contains an ordered list of submenus that may be thought of as the sequence one might perform an analysis. In reality, an analysis is more complicated and involves iterating various steps (see Figure 3.1).

3. Go to the "Samples" menu. This menu allows selecting hybridized samples for the HP-X, HP-Y, HP-X and HP-Y 'sets', and the HP-E list. You have several ways to do this. The easiest way is to use the "Choose HP-X, HP-Y and HP-E" sample selection wizard. The other alternative is to use one of the set of cascading menus that may be used to change the selected hybridized samples. Note that the HP-X and HP-Y submenus assign samples for subsequent analysis such as X-Y scatter plots. The HP-E submenu sets the list of samples for expression profiles. Note how you may assign a sample either by going through the cascading menus or from an alphabetic list of all hybridized samples in the pseudoarray image. To change the default HP-X (HP-Y) sample, click on the purple "[X]" ("[Y]") box on the left side of the image (above the list of samples) so it is selected. Then click on the desired purple "*" adjacent to the samples listed on the left edge of the pseudoarray image. You may switch between using single and multiple samples (i.e. 'sets') with HP-X and HP-Y. Note the "HP-X:" and "HP-Y:" labels at top left when you switch between single and multiple samples. Go to the HP-X/-Y 'set' and HP-E submenus and list the contents of the respective sets to see what they contain. Note how one may add or remove samples from these sets.

4. Go to the "GeneClass" submenu in the "Analysis" menu. This is set of cascading menus that may be used to change the default Gene Class. Different genes belong to different gene classes and this is a way of sub-setting the data. You may currently set it to All Genes, All Named Genes, ESTs similar to genes, ESTs, Good genes, All named genes and ESTs, Replicate genes (multiple copies on the array), Calibration DNA. Select "ESTs similar to genes". Notice that the display now only shows the red (white) circles on the ESTs similar to genes in the intensity (ratio) pseudoarray image. Look at the

circle overlays on the microarray image - note how they changed. Go back and set the gene class to "All genes" and then "Calibration DNA". The "Calibration DNA" is the set of spots that may be used for normalizing data between microarrays. Check out the "Set gene class subset" submenu. Leave the gene class set to All Genes.

5. If you are connected to the Internet, go to the "View" Menu. Then turn on the switch "Enable display current gene in popup XXXX Web Browser". Depending on your database, XXXX may be GenBank, LocusID, UniGene, dbEST or mAdb Clone DB. Then click on a gene in the image. It pops up a Web browser showing the genomic database Web page for that gene (if any). If you click on a different spot, it will reuse the popup Web browser with new data. If you don't want it to be active, go back to the "View Menu" and click on "Enable display current gene ..." again to disable it.

6. Go to the "Normalization" submenu in the "Analysis" menu. This is used for normalizing data between microarrays so they may be compared. The default is to scale the data to the median value for each array. Investigate the other normalization methods. If your quantified data contains spot background data, you may also enable background correction that subtracts a microarray specific overall background intensity value from each intensity value in each array.

7. Go to the "Report" submenu in the "Analysis" menu. You may generate a report in several formats "Spreadsheet" or "tab-delimited", the latter being cut and paste compatible with Excel. In the "Samples Report" sub-submenu in the "Report" submenu, click on "Hybridized Samples", then switch to the other Report format and do it again. Click on "Hybridized Sample Web links". In the "Gene Reports" submenu, try "All named genes". Then try the "Highest HP-X/HP-Y ratio" in the "Filtered gene reports" submenu. If you are using a list of HP-E or HP-X/-Y 'sets' of samples, you might try looking at the expression profile ratios or statistical data respectively though other Reports.

8. Review the data " Filter" submenu in the "Analysis" menu. Select the "Filter by expression sliders" [I1:I2]. Expression is intensity in the case of ^{33}P or biotin labeled, or (Cy3/Cy5) in the case of ratio data. This pops up a state sliders window. The I1 scroll bar (lower limit of the gray value intensity) to about 100. Look at the genes that were eliminated because they are out of the range. If you select the "Filter by spot intensity sliders" [SI1:SI2], it will filter genes by spot intensity in *either* F1 or F2 duplicate (if you have duplicate spots), or the Cy3 or Cy5 spots. You can select which sets of HPs to use (current HP, HP-X and HP-Y, HP-XY sets, or HP-E). You might disable the [I1:I2] filter while you experiment with [Si1:SI2] filter. If you have (F1,F2) or (Cy3,Cy5) data, put up the F1 vs F2 or Cy3 vs Cy5 scatter plot and you can visualize how the thresholding works.

   In the Filter menu, add the "Filter by ratio or Zdiff sliders". Then the [R1:R2] ratio range sliders are added to the state slider window and may be used for filtering genes. If the normalization method is one of the Zscore methods, it filters by the difference of the Zscores otherwise by the ratio and the [Z1:Z2] range is used. Note that the genes that pass the filter will appear to have a red (white) circle in the pseudoarray intensity (ratio) grayscale (pseudocolor), or red "+" in the scatter plots so you might try moving the controls while in those plot modes. Try some of the other filters. The spot CV test removes genes where replicate spot values (F1 and F2 in the case of a single sample or replicate samples in the case of HP-X and HY-Y 'sets' or the HP-E' list of genes) are not well correlated. The t-Test filter may be used with sets of X and Y samples to find genes with a *p*-value less than the specified threshold.

9. Go to the "Plot" submenu in the "Analysis" menu. Then got to the "Show Microarray" submenu, try pseudocolor ratio (or Zscore) modes and finally leave it in the pseudograyscale image intensity mode. Try using the "Scatter Plots" and "Histograms" in the Plot menu. Vary the normalization methods and see how it affects the array image and scatter plots. When you are in a scatter plot, click on a point. It will display data similar to when clicking on an image. If UniGene data is available in your database, Go to the Views menu and set "Enable display current gene in popup UniGene Web browser" and click on a point again. This pops up another Web browser window and lookup that gene in the UniGene database. Change the threshold slider and notice how points appear and disappear. Click on a bin in the ratio histogram, it will filter the genes so that only the genes that have ratios in that bin are displayed.

10. Go to the "Plot" submenu in the "Analysis" Menu and then to the "Expression profile" submenu. Then select the "Display gene expr. profile for HP-E". It pops up a window with two buttons "Show HP names" and "Close". Then click on a gene in the image. It will draw the expression profile for that gene. Move the mouse so it is over one of the vertical bars in the plot

to get the data for that particular HP. If you click on a different spot in the image it will display the new expression profile in the same window. Click on the "Show HP names" button to popup a window with the list of HPs matching the numbers in the expression profile plot. Now create a scatter plot of two HPs and then click on a red "+" in the plot. It will update the expression profile plot with this gene. If you want to compare several expression profile plots, repeated create the "Display gene expression..." windows from the View menu. Move the windows close to each other so they are easier to compare. Only the last one you created allows you to change the gene. If you don't want a popup window anymore, click on its "Close" button. You may save the scatter plot as a GIF file by pressing the "SaveAs" button which will save it in the Report subdirectory associated with the startup data..

11. Next we look at gene clustering. Go to the "Plot" submenu in the "Analysis" menu, and then to the "Cluster plots" submenu, try "Cluster genes with expression profiles similar to current gene". This pops up a slider with the cluster threshold. Then click on a gene in the image. It will then popup a text window with the genes whose cluster distance is less than the cluster threshold. It is sorted by minimum cluster distance. Notice that some genes in the image have different size blue boxes around them. The larger the box, the smaller the cluster distance and the more similar. Move the cluster threshold slider. This will change the clustering as seen in both the image and in the cluster popup window. Click on the "Report" button in the text window. This will popup a report on these genes sorted by minimum cluster distance. Then click on the "EP plot". This pops up a scrollable list of expression profile plots for the genes you have filtered by this test on so you can review the actual expression profiles. Close this window and set the Filter to a small number of genes such as with "ESTs similar to genes". Note that the genes passing this filter are saved in the E.C.L. and may be saved as a gene subset or part of the data Filter.

Turn on one or more Filters to reduce the number of genes to say under 100 (e.g. t-test or spot CV filters). Then press the "Go 'Cluster all genes'" button in the cluster window. This is equivalent to invoking the "Cluster counts of Filtered genes by expression profiles" command from the "Cluster plots" submenu. Notice the Filtered genes has blue circles of different sizes. The larger the circle, the more genes there are that are similar to that gene. Move the cluster threshold slider and note that the number of similar genes changes, the size of the blue circles will change. As with the other cluster mode, you may generate a report of sorted cluster counts. Click on a gene with the largest green circle. This will then switch you back to single gene clustering mode where you can investigate that gene in more detail.

12. Next we look at K-means gene clustering. Go to the "Plot" menu and in the "Cluster plots" submenu, try 'Display K-means gene expression profiles for Filtered HP-E' (similar to K-means clustering). This pops up a scroller for "N-clusters", the number of clusters desired, with a default of 6 clusters. It will then popup a K-means report window with various controls. The centers of the N clusters are indicated with magenta circles where the size corresponds to the number of genes in the cluster. If you click on a gene in the array, it will draw all members of its cluster as green numbers (try this with the scatter plot present as well). Press "EP plot" to scroll through a list of expression profiles of the genes sorted by cluster. Press "Mean EP plot" to scroll through the summary expression profiles of the clusters. Press "Cluster-Report" and "Mn-Cluster-Report" to generate reports for these clusters. You may change the seed gene by changing the current gene (click on a different gene in either the array image or the EP plot. Then and press "Recompute" to recompute the clusters. Pressing the "Show HP names" pops up a list of the samples used in the expression profile. Now add a HP-X vs HP-Y scatter plot. If you select a particular gene, it puts all genes associated with the cluster for that gene in the "Current Cluster" and colors them with a green cluster number instead of a "+".

13. Next we look at hierarchical gene clustering. Go to the "Plot" menu and in the "Cluster plots" submenu, try "Hierarchical clustering of expression profiles". Then select "Display clustergram of gene expr profiles". This will compute the clustergram for the Filtered genes with the data normalized by the assigned HP-X sample. It will then popup a clustergram window with various controls. The clustergram is scrollable. You may optionally add the dendrogram with the "Dendrogram" checkbox. Clicking on a row will show data for that gene. Clicking on a box in the clustergram will show data for the particular sample for that gene. You may zoom the dendrogram by repeatedly clicking on the "xxxX DB" zoom button. Press "EP plot" to scroll through a list of expression profiles of the genes sorted the same as the clustergram. Press "ClusterGram Report" to generate a report of the expression profile data sorted the same as the clustergram. Pressing the "Show HP names" pops up a list of the samples used in the expression profile. You may save the entire clustergram - dendrogram image in a GIF file as before by pressing the "SaveAs" button.

14. You may perform operations on sets of genes. For example, merge sets of genes found under two different experiments or conditions. Go to the "Sets of genes" submenu in the Edit menu and pick the "List saved gene sets" selection. This lists the default gene sets. Then select "Assign 'User Filter Gene Set'". This will request a gene set to use with the Filter in a pop up dialog box. Select the set for "Genes in current cluster class" that you saved in the previous example. Then press "Ok" in the dialog box. Then select "All genes" in the GeneClass menu. This resets the filter to look at all genes. Select "Filter by 'User Gene Set' membership" in the Filter menu. This restricts the genes to the saved current cluster in the previous example.

15. Gene set operations may be performed on pairs of gene sets. Select "Union of 2 gene sets" entry from the "Sets of genes" submenu in the Edit menu. This will request 3 gene set names in a pop up dialog box. Select 'Similar ESTs' for the 1st gene set name, select "Genes in current cluster class" for the 2nd gene set name, Enter "Union of similar ESTs and genes in current cluster" for new gene set name. Then press "Ok" in the dialog box. This computes the union of the two gene sets into a new gene set. Then select "Filter by 'User Gene Set' membership" as before. This will reset the 'Use Gene Set' for the Filter in a pop up dialog box. Select the set for "Union of similar ESTs and genes in current cluster" just saved. Try saving other Filtered genes sets and doing other gene set operations.

16. Go to the "Help" menu. The MAExplorer documentation and glossary as well as other MGAP documents are available and will appear in a new popup Web browser.

17. If you are running MAExplorer as a stand-alone application, you may save the state of your analysis (but not the gene subsets at this time). Go to the "File" menu and then "Databases" submenu. Select "Save as file DB". Enter a file name to save your startup state. Then you can restart MAExplorer on this data set at a latter time by clicking on this file (in a windows based system). Saving the database will also save the state of the data Filter, the gene sets and the condition lists. You can see this by listing them in the Edit menu sub-options after you have restarted a previously saved database.

Appendix C. Use of MAExplorer with user's microarray data

This section discusses the use of MAExplorer to convert microarray data from a variety of sources including various types of labeling ^{33}P-labeled, biotin-labeled, or Cy3/Cy5 ratio-labeled spotted membranes or glass slides or oligo-chips of different geometries and numbers of duplicate spots/gene.

Note: This appendix contains a "computerese" description on how to use MAExplorer with your array data. The user-friendly "wizard" tool Cvt2Mae makes it much easier for most molecular biologists to use. Cvt2Mae is available for download on the MAExplorer home page. This appendix gives some examples below of some of the required file data for those of you who want to understand the data file formats or want to manually convert your data or use your own conversion program on your data. Note that you do not need to read this chapter to use MAExplorer if you use the Cvt2Mae converter, use some other conversion software (eg. the NCI/CIT mAdb array database server), etc.

MAExplorer requires a specification of array geometry and quantification information. These are defined in a **configuration** startup file. The startup file contains the initial list of hybridized samples to be loaded, and other parameters such as the name of the configuration file (if it is different from the default name). A stand-alone application causes the .mae startup file (or the PARAM list in the case of an applet) to be read when it is started. The configuration file contains various defaults. If any of these are specified in the configuration file, the override the built in default values. Values from the .mae startup or applet PARAMs will override the configuration file values. These configuration parameters may be overwritten by arguments in the stand-alone .mae startup files or PARAMs in the Applet startup specifications.

A few additional files are required and are defined in the configuration file. These include: a Gene-In-Plate-Order or **GIPO** file; a **samples** database file listing names of the samples available for loading; and a **gene class** names file. An optional (but deprecated) **extra array information** file may be specified to access additional data about samples. Quantified hybridized sample

array spot data (**Quant** files) from each array is put into a separate data file. Note that *all* data files are tab-delimited files such as may be generated with Excel, relational databases or directly from array spot quantification software.

Hybridized sample arrays must be scanned and then spots quantified using other software. MAExplorer does not do spot quantification from scanned image files. However, MAExplorer can use spot data from a variety of array image quantification programs that generate tab-delimited data files. The data needs to be converted to the MAExplorer schema described in this Appendix.

The derivation of quantified spot data files from hybridized sample arrays is discussed later in this section as are in the quant file data format.

The configuration file is created once for each new array GIPO geometry and database of hybridized samples. It is independent of the number of samples. Configuration parameters include array geometry (# of grids, # of duplicate spots/gene, etc), whether the data is intensity or ratio data (e.g. Cy3/Cy5), etc. The configuration file may also include labeling, quantification dynamic range, default analysis thresholds, mapping of used data file table-field names to expected MAExplorer names for the GIPO and quantification files, additional database-specific pull-down menu plugins, names of gene sets and sample condition lists, etc.

The GIPO file is independent of the number of array samples and describes the mapping between spot position in an array and its gene identification as well as corresponding data such as original plate number, row and column; UniGene ID, GenBank ID, dbEST ID, etc. These files will be described in more detail including how one can create the necessary database files that MAExplorer requires for use with other microarrays.

Directory (i.e. folder) structure of stand-alone databases

When running as a stand-alone application, MAExplorer assumes that data from a local computer has a specific directory structure. The required and optional directories (also called "folders" on some operating systems) and files they contain are diagramed here from a database project directory in your file system. The notation "/folder-name" indicates that "folder-name" is a folder inside of the project.

```
(specific database directories and files they contain)
           / Cache
                   / (copies of any data files saved from Web DB access)

           / Config
                   / MaExplorerConfig.txt
                   / SamplesDB.txt
                   / GIPO-db.txt

           / MAE
                   / (set of startup database files).mae

           / Images
                   / (set of original or sampled array .jpg images) (optional)

           / Plugins
                   / (optional set of .jar or .class MAEPlugin files)

           / Quant
                   / (set of spot quantified data files).quant

           / Report
                   / (set of .txt and .gif report files generated using SaveAs

           / State
                   / (set of gene set files).cbs and
                   / (set of condition list files).hbl generated using Save DB
```

Figure C.1 Directory structure of stand-alone databases required by MAExplorer. The "/Config", "/Quant", and "/MAE" directories are required. The /MAE directory is only used with the stand-alone version with <u>.mae files</u>, not for the applet. [When used with an applet, the main path is the path of the download JAR file and .mae files are not used.] The "/Report", and "/State" directories are created by MAExplorer as needed and the user need not create them prior to running MAExplorer. The text reports and plot GIF images are saved in the /Report folder when you "Save" a report or plot. When you "Save" the current database session (File | Databases | Save ...), the gene sets and sample lists are saved in the /State folder for use when you restart MAExplorer on the .mae startup file. The optional "/Cache" directory is only used (and then, only optionally) when downloading data from a Web server. The optional "/Image" directory is only used in there are JPEG images of the arrays provided and their resolution and alignment must correspond to the (X,Y) spot data in the Quant files. The "/Plugins" directory is where the <u>MAEPlugins</u> packaged with MAExplorer are normally kept and where MAExplorer looks when you attempt to load a plugin. Since you can browse your file system, they do not have to appear here.

Examples of some of the database files required by MAExplorer

These could be used as examples that could be used in creating your own database files. When the MAExplorer converter tool, Cvt2Mae, is released it will eliminate the need for manually editing your database files.

Sample MGAP database configuration, quantification data and startup files are available for use as examples with which to make your own files or for inspection.

- http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database/Config/
- http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database/Quant/
- http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database/MAE/

In addition, examples of the (Config/, Quant/ and MAE/) files needed for various types of arrays are available at:

- Example-P33_F1F2
- Example-NAME_GRC
- Example-Cy3Cy5

Additional directories used at run-time

When running MAExplorer as a stand-alone application, you may save data on the disk Text reports and plot graphics windows are saved as ".txt" text and ".gif" image files when the user uses the "SaveAs" button in the respective popup windows. These files are saved in the "Report" subdirectory.

Similarly, when the entire database is saved (File | Databases | SaveAs ...DB) into a .mae startup file, the set of gene set files are saved as ".cbs" files and the set of condition list files are saved as ".hbl" files in the "State" subdirectory. These are automatically reloaded into MAExplorer when the .mae startup file is used to restart MAExplorer.

If your array data has JPEG or similar images of the original arrays, the should be saved in the "Images" directory. For example, the NCI-CIT mAdb database server allows you to download sampled images for your data in an "Images" subdirectory at the same time you download the other MAExplorer data files. The images can then be used by various MAEPlugin programs. If your quantified data converted to .quant files has (X,Y) coordinates corresponding to spots in these images, then you may be able to use the Montage MAEPlugin to show where the current spots are in sub-regions of all of the input images. This plugin will be available on the MAEPlugin Web site when we release the MAEPlugin facility for Beta-testing.

Tools for automating the construction a local stand-alone database

Software tools for aiding the construction of local stand-alone databases from vendor supplied GIPOs and spot quantification files are not available at this time, but will be made available in the future.

Manually constructing a local stand-alone database

Although the Cvt2Mae converter tool can convert many files, you could alternatively build these files manually. We suggest using Excel or your favorite RDBMS system to manipulate the data. At the end, save the data into files with tab-delimited fields with the above file extensions (i.e. .txt, .quant, .mae). The layout of these files and what is optional and what is not is described in detail (maybe too much!) below. You could use an ASCII file text editor instead of Excel (such as Wordpad, Emacs, etc.) - **but be careful not to add or delete tabs since this will destroy the integrity of the database tables**. Be consistent in your file names; avoid spaces; use ASCII characters in file names that are system independent (i.e. A-Z, a-z,0-9, "-", "+", "_"); Use either "-" or "_" or both.

For a specific database (db), make sure the names of the configuration files in /Config directory are entered in the MaExplorerConfig-db.txt file for that database. You may have multiple databases in the same /Config, /Quant and /MAE directories if the file names do not conflict. The trick is to have the .mae startup file in the /MAE directory point to the specific configFile to be used. Since MAExplorer reads the MaExplorerConfig-db.txt file when it first starts up, it discovers the names of the other database files. If there is no name conflicts, then there is no problem mixing data.

Each spot data (.quant) sample file has a name which must be entered in the Database_File field of the Samples-db.txt row entry for a new sample. The Sample_ID field is a descriptive name of that sample.

Often GIPO files supplied by array vendors have additional fields not currently used by MAExplorer. You can leave them in (they will be ignored) or take them out (loading a database is faster).

If the *field headings in the various user's tables are not the same as that required by MAExplorer*, you can easily fix this by adding (Table,Field) mapping entries to your version of the MaExplorerConfig-db.txt file (see mapTF for examples).

Note that the optional Menu_Source_Name entry in the Samples-db.txt file specifies the sub-menu, if any, that the sample will appear in the **Samples** menu **By Source** sub-menu.

If the optional extra sample information file is used, then make sure the sample names and database file names are the same, and that there are corresponding rows in each table.

C.1 Creating quantified spot data files from hybridized sample arrays

Quantified spot data from images scanned from hybridized sample arrays may be created using a variety of software programs. Discussion of these is beyond the scope of this manual. However, several of these including Pathways 2.01, ImageQuant-NT, and others generate tab-delimited text files. These files may be used directly as the quantified spot files required by MAExplorer, or simplified first (by removing unused or redundant data fields). Typically, the files are named (or renamed) to that of the sample to distinguish them from each other and a .quant file extension assigned instead of the .txt file extension. Other programs generate tab-delimited files that could be mapped to our .quant file formats. (For example, the NCI/CIT mAdb system generates such a mapping for GenePix$^{(TM)}$, and ScanArray formated data.)

C.1.1 Color and prefix notation for the following tables: (req), (opt), (future)

The following tables list **parameter**s and some typical values that might be included in the configuration and quantification files. These examples illustrate the variety of parameters or fields with examples of values that might be used. **Required** parameters are in black with "(req)" prefix. **Optional** parameters are indicated in blue with a "(opt)" prefix. Optional parameters are not normally specified and are generated in the .mae file when you save the state of a data exploration. Parameters that might be used with Cy3/Cy5 ratio data are indicated in magenta with a "(ratio)" prefix. **Future** options not currently used are indicated in green with a "(future)" prefix. Alternative options are indicated in red with an "(alt)" prefix.

C.2 Table of samples that can be loaded into MAExplorer

The samples available to be analyzed in a database are listed in a samples database table. This lists *all* samples that *could* be loaded. The user will then select a subset of these to be analyzed. The selection is done either in preset Web startup pages, or with

the stand-alone application `.mae` startup files, or at run-time by selecting new entries from the Samples pull-down menus. Extra information may be provided to MAExplorer for each sample through this table and will be available for the <u>Sample Array report</u> in Section 2.4.6.1.

A typical sample database table might look like:

```
Sample_ID      Project       Database_File
control 1      breastCancer  control1
control 2      breastCancer  control2
control 3      breastCancer  control3
tumor 1        breastCancer  tumor1
tumor 2        breastCancer  tumor2
tumor 3        breastCancer  tumor3
```

You may optionally include a Database_ID field. For example:

```
Sample_ID      Project       Database_File  Database_ID
control 1      breastCancer  control1       270314
control 2      breastCancer  control2       270315
control 3      breastCancer  control3       270316
tumor 1        breastCancer  tumor1         270317
tumor 2        breastCancer  tumor2         270318
tumor 3        breastCancer  tumor3         270319
```

The Database_ID may be useful if there are file length problems on some systems (i.e. MacOS 8-9), we offer the option of using the Database_ID as the file name for the .quant (Quant/ directory) and .jpg (Images/ directory) rather than the Database_File name. For example one could specify "Quant/270314.quant" and "/Images/270314.quant" rather than the default "Quant/control1.quant" and "/Images/control1.quant" names.

The Samples database table includes some required as well as optional fields (see Table C.2.1.1):

Table C.2.1 List of Samples data file table fields. The Samples table lists hybridized samples that are accessible to the user and may be loaded into a database session if they wish. (See <u>Section C.1.1</u> for option notation.)

Field	Description
(req) Sample_ID	descriptive name of the sample, **free text**. [Note: an older depricated name is "Membrane_ID"]
(req) Project	that the sample belongs. Used for login protection and grouping of samples
(req) Database_File	name of the `.quant` spot database file, no spaces. This is the **file name** for the sample.
(opt) DatabaseFileID	database file ID corresponding to Database_File and Sample_ID. For use with RDBMS Web databases (e.g. experiment id #). NOTE: if you are encoding auxillary data files using this identifier, e.g. sampled array images in the `Images/` directory, then this field is required if you want to access those images.

Table C.2.1.1 List of optional Samples data file table fields. These fields may be used for some additional operations. If they are not in the Samples DB table, then the operations will not be available. (See <u>Section C.1.1</u> for option notation.)

(opt) Menu_Source_Name	Sample `SubMenu` j that this sample belongs. You could use the word "Default" or leave out this entry if you do not want to use sub menus.
(opt)Orig_File_Name	if applicable. The original file name and sample name if the data was split out from a multiple hybridized sample file.

(opt)Strain	if applicable
(opt) Source	if applicable
(opt) Probe	if applicable
(opt) Stage	if applicable (eg, developmental stage, dose, time point, etc)
(opt) Login	(optional) TRUE if login required with a Web server else blank. This is used primarily with the Applet when interacting with a Web server
(opt) GeneCard_URL	GeneCard ID if applicable
(opt) Histology_URL	(e.g. MGAP) histology DB Web page if applicable
(opt) Model_URL	(e.g. MGAP) mouse model database Web page if applicable
(opt) BGLow	global low value of array background intensity
(opt) BGAvg	global average value of array background intensity
(opt) BGRms	global root-mean-square value of array background intensity

Table C.2.1.2 List of optional Samples data file table fields. These fields are not currently used in any computations but are returned in the Sample Array report in Section 2.4.6.1.

(opt) Contributor	name of researcher submitting the sample
(opt) Contrib_Institute	researcher's organization
(opt) Submission_Date	when submitted
(opt) Exposure	minutes or hours of radiolabel or fluorescent exposure
(opt) Sample_Nbr	internal sample number
(opt) FilterType	name of the array layout
(opt) FilterType_Description	additional description of array layout
(opt) Comments	details describing sample
(opt) Researcher	researcher performing the hybridization
(opt) SampleGrid	serial number of the array or grid or internal laboratory numbering. (Useful if reusing arrays etc)

C.3 Quantified spot data file formats

MAExplorer has been designed to be able to read quantified spot data from a variety of spot analysis software packages. So the data file format is very flexible. Essentially, a data file contains one or more spot intensity values per gene in each row of the data file. A spot location is specified by a GIPO (field#, grid#, grid column#, grid row#) 4-tuple with the field value optional. Note: a "grid" is sometimes called a "block" or a "patch". If the field specification is omitted and there are duplicate spots in multiple fields of grids, then it is defined implicitly. In that case, the corresponding spot intensity data for each field for a gene is specified as separate columns going from left to right. The (grid#, column#, row#) part of the specification may be encoded several ways: a) explicitly as (grid#, column#, row#) or b) NAME_GRC.

Alphanumeric mappings for Grid, Grid Row and Grid Column data

Most quantitative data formats use integers for (grid,row,col) values. However, some formats use letters [A:Z] for the first 26 (i.e. [1:26]), and [a:z] for the next 26 (i.e. [27:52]) values. Sometimes only lower case letters are used - in which case we must map [a:z] to [1:26]. When MAExplorer first sees a letter in reading the data, it checks to see if it is an upper or lowercase letter and generates the offset needed to generate the mapping.

The Molecular Dynamics Name_GRC numbering mapping for Grid, Grid Row and Grid Column data

Alternatively, the Molecular Dynamics ImageQuant GIPO coordinates represented by NAME_GRC with entries of the form "Grid - <grid#>R<row#>C<column#>" (e.g. "Grid -3R6C8") may be used with or without the replicated field entry to replace the entries (grid, grid col, grid row) in the GIPO table and Quant spot data files. For [G grids, R rows and C columns], this would cover a set of spots in the range [1,1,1] through [G,R,C].

Some examples of typical quantified spot data files might look like:

1. Single spot/gene intensity data.

grid	grid col	grid row	RawIntensity	Background
1	1	1	2226.8	32.6
1	1	2	1234.8	25.6
. . .				
10	25	28	3333.8	23.6

2. Double spots/gene intensity data contained in two fields of duplicate spots.

grid	grid col	grid row	RawIntensity1	Background1	RawIntensity2	Background2
1	1	1	2226.8	32.6	2345.9	39.4
1	1	2	1234.8	25.6	1245.9	39.4
. . .						
10	25	28	3333.8	23.6	3345.9	25.4

3. Double spots/gene intensity data contained in two fields of duplicate spots.

field	grid	grid col	grid row	RawIntensity	Background
1	1	1	1	2226.8	32.6
1	1	1	2	1234.8	25.6
. . .					
1	10	25	28	3333.8	23.6
. . .					
2	1	1	1	2226.8	39.4
2	1	1	2	1234.8	39.4
. . .					
2	10	25	28	3333.8	25.4

4. Double spots/gene intensity data using the Molecular Dynamics' NAME_GRC notation.

NAME_GRC	RawIntensity1	RawIntensity2
GRID- 1-R1C1	2126.500	3662.350
GRID- 1-R2C1	2311.430	3306.290
GRID- 1-R3C1	3696.470	5780.310
GRID- 1-R4C1	3167.450	5245.440
. . .		

5. Cy3/Cy5 spot/gene ratio data.

grid	grid col	grid row	Cy3	Cy3Bkgd	Cy5	Cy5Bkgd
1	1	1	2226.8	32.6	2345.9	39.4
1	1	2	1234.8	25.6	1245.9	39.4
. . .						
10	25	28	3333.8	23.6	3345.9	25.4

The basic Quant spot data file table includes entries listed in Table C.3.1:

Table C.3.1 List of Quant data file table fields. This specifies the spot quantification data. There may be one or more spots, corresponding to the same gene, on each row. (See Section C.1.1 for option notation.)

Field	Description
(opt) field	field for duplicate genes if using single 'RawIntensity' value/Row
(req) grid	grid name (either A,B,C,... or 1,2,3,...)
(req) grid col	column with in a grid
(req) grid row	row within a grid
(opt+alt) NAME_GRC	(alternative specification of "grid, grid col, grid row").
(req) RawIntensity1	intensity value for field 1. Use this form if there is more than 1 intensity value/row.
(req) RawIntensity2	intensity value for field 2 (required if it exists and for Cy3, Cy5 data)
(req+alt) RawIntensity	intensity value for field 1, if only one field used
(opt) Background1	background intensity value for field 1
(opt) Background2	background intensity value for field 2 (if it exists for F1,F2 data or Cy3, Cy5 data)
(opt+alt) Background	background intensity value for field 1, if only one field used
(opt) QualCheck	(future) quality check for data indicating "bad" spots or genes. Current codes are listed in the Table C.4.2 of QualCheck semantics

Note: If NAME_GRC is specified (eg. for use with ImageQuant-NT data), then the explicit (grid, grow row, grid col) fields are not required. Note: For [G grids, R rows and C columns], this would cover a set of spots in the range [1,1,1] through [G,R,C].

Note: If Cy3/Cy5 double fluorescent labeling is used, then the RawIntensity1 and RawIntensity2 fields may be replaced with Cy3RI and Cy5RI names and the (RawIntensity1, RawIntensity2) fields mapped to (Cy3RI, Cy5RI) in the configuration file mapTF entries (table C.5.4 below). (See Section C.1.1 for option notation.)

Field	Description
(req) Cy3RI	RawIntensity1 value for Cy3
(req) Cy5RI	RawIntensity2 value for Cy5
(opt) Cy3Bkgrd	Background1 value for Cy3
(opt) Cy5Bkgrd	Background2 value for Cy5
(opt) Cy3	RawIntensity1 value for Cy3
(opt) Cy5	RawIntensity2 value for Cy5

C.4 The GIPO table database file format

The gene-in-plate-order (GIPO) table used to make the connection between a spot on a microarray and the plate well corresponding to a gene. We are working on extending the format so that it will more easily handle GIPO tables from a variety of sources.

Data is extracted from a table created from the gene-in-plate-order (GIPO) gene coordinate table. This links spots in a microarray to these Genomic "gene ID"s and gene names. This table may contain Clone ID, GenBank, dbEST, UniGene IDs, LocusID corresponding to these Master Gene IDs. An optional table of Clone IDs and Gene Classes the gene belongs to may also be defined.

A typical GIPO database table might look like:

```
 Location   grid   grid col   grid row   plate   plate row   plate col   Clone ID   GenBankAcc GeneName
   . . .
 39        A      2          15          2       1           3           1247601    AA763423    "Mus musculus
A kinase anchor protein (AKAP-KL) mRNA, alternatively spliced isoform 1, complete cds"
 40        A      2          16          2       1           4           1247553    AA763380    Mus musculus
bodenin gene
 41        A      2          17          2       1           5           1247865    AI465019    "Mouse beta-D-
galactosidase fusion protein mRNA, complete cds"
   . . .
```

The basic GIPO table includes the following fields:

Table C.4 List of GIPO data file table fields.

These fields define the mapping between a spot's grid coordinates on the array and its genomic identifier, gene name, its plate, etc.

Field	Description
(opt) field	array field for duplicate genes
grid	array grid name (either A,B,C,... or 1,2,3,...)
grid col	array column within a grid (either A,B,C,... or 1,2,3,...)
grid row	array row within a grid (either A,B,C,... or 1,2,3,...)
(opt+alt) NAME_GRC	alternative specification to "grid, grid col, grid row". It is generated by the Molecular Dynamics spot quantification software.
(opt) **Master Gene ID**	This is the **master gene identifier** used in MAExplorer. It must be one or more of the identifiers listed in Table C.4.3. One of these will be selected as the Master Gene ID (MID)
(req) Gene Name	**Master Gene Name**. The GeneName options are listed in Table C.4.1. These alternative GeneClasses are automatically recognized from the Gene Name.
(opt) plate	plate name for original gene. If this is not specified, it uses the grid value.
(opt) plate row	plate row name for original gene. If this is not specified, it uses the grid row value.
(opt) plate col	plate column name for original gene. If this is not specified, it uses the grid col value.
(opt) QualCheck	quality check for data indicating "bad" spots or genes. Current codes are listed in the Table C.4.2 below

Table C.4.1 List of possible Master Gene Name

The Master Gene Name must be define as one the following identifiers:

Field	Description
(opt) GeneName	Gene name
(opt) Unigene cluster Name	alternative for GeneName if the latter is not specified.

Automatic Gene Class naming based on Gene Name

Some Gene Classes are automatically recognized from the Gene Name including:

1. Unknown ESTs - the Gene Name is "EST", "ESTs", "expressed sequence", "unknown"
2. ESTs similar to known genes - the Gene Name is "EST, ...", "EST ...", "expressed sequence ..."
3. Calibration DNA - the Gene Name is the 'calibDNAname' value defined in the Configuration database table.
4. User's Plates - the Gene Name is the 'your plate' value defined in the Configuration database table.

5. Empty Well - the Gene Name is the 'EmptyWell' value defined in the Configuration database table.
6. Known genes - if the non-null Gene Name does not fit into any of the above categories and it is not an empty well, it is assumed to be a known gene.
7. Good genes - normally set by the QualCheck field in the GIPO file, but if the gene name is "EmptyWell", it is flagged as a bad gene.

Alternative Grid,Row,Column encoding scheme: NAME_GRC

Some quantification programs (e.g. Molecular Dynamics "ImageQuant-NT) specify "grid, grid_col, grid_row" by a single symbol we denote NAME_GRC coded as follows

```
GRID-   grid#-Rrow#Ccol#
```

For example, if grid #, row# and column# are (8,12,11), then it codes it as

```
GRID-   8-R12C11
```

Table C.4.2 List of QualCheck codes and their semantics

The data filter "Filter by 'Good Spot data'" may be used in eliminating bad spot data on a per-gene set basis. This uses the "QualCheck" field in the quantified data table is present. It maps either an 1) integer numeric code (see Appendix C of the Reference Manual), 2) an alphabetic code (e.g. Affymetrix "Abs Call") of "P" (or "G" or "T") to Good Spot, "A" (or "B" or "F") to Bad Spot, and "M" to Marginal Spot, or 3) a continuous quality value. In this latter case, QualCheck may be a continuous monotonically increasing floating point value (e.g. 0.0 to 100.0, or 0.0 to 1.0, -100.0 to +100.0, etc.) in which case a "Spot Quality" State threshold slider will popup when the filter is invoked. Additional property value codes may be added in the future.

Status	QualCheck value	Semantics
Good gene	2	the spot data is "Good" (some systems report this by a NULL quality measure). It has a good gene name. Alternatively, letter codes may be used "P", "G", "T".
Bad gene	4	the spot data is bad, a good gene name.
Bad spot	8	is a non-analyzable spot (eg. marker, or "Bad", "Not Found", "Empty". etc.) Alternatively, letter codes may be used "A", "B", "F".
Duplicate spot	16	is duplicate of another gene on array
Marginal spot	256	is a marginally quantified spot. Alternatively, letter codes may be used "M".

Table C.4.3 List of possible Master Gene Identifiers

Additional data is used to point to data in external genomic databases by specifying the identifier. This may be used to dynamically link genes in the MAExplorer database to Web database servers to bring up Web pages from these databases. Note the Master ID needs to be specified and may be any one of the following identifiers. The appropriate genomic Web browser access will be enabled depending on the genomic Master ID specified. (See Section C.1.1 for option notation.) The fields include:

Field	Description
(opt) Location	alternate spot identifier. E.g., Affymetrix 'probe_set', or Incyte 'IncyteID', etc. This may be numeric or alphanumeric
(opt) Clone ID	I.M.A.G.E. consortium database clone ID. It may have a "IMAGE:" or "ATCC:" prefix
(opt) Unigene cluster ID	NCBI UniGene database ID

(opt) dbEST3'	NCBI dbEST database
(opt) dbEST5'	NCBI dbEST database
(opt) GenBankId	NCBI GenBank database
(opt) GenBankId3'	NCBI GenBank database
(opt) GenBankId5'	NCBI database
(opt) LocusID	NCBI LocusLink database
(opt) SwissProtID	Swiss-Prot database

Table C.4.4 Extending Genomic IDs and associated URLs

MAExplorer allows you to define your own gene identifiers that will map to external genomic databases. You add the following entires in sets of 4 to the Configuration database or to the .mae startup file. These entries will be added to the View menu where you may select the external genomic database to visit when you activate MAExplorer to launch a brower on clicking on a gene. The following table shows the 4 required fields for 2 entries. There may be any number of external genomic IDs. (See Section C.1.1 for option notation.)

Parameter	Value	DataType	Comments
(opt) **GenomicMenu1**	GenBank	String	Name of the database. This will appear in the View menu
(opt) GenomicURL1	http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=2&form=1&term=	String	URL to which one adds the 'GenomicIDreq' value
(opt) GenomicURLepilogue1		String	epilogue of the URL if any
(opt) GenomicIDreq1	GBID	String	Name of the GenomicID required and that is specified in the GIPO file as one of its fields
(opt) **GenomicMenu2**	UniGene	String	Name of the database. This will appear in the View menu
(opt) GenomicURL2	http://www.ncbi.nlm.nih.gov/UniGene/query.cgi?ORG=Mm&CID=	String	URL to which one adds the 'GenomicIDreq' value
(opt) GenomicURLepilogue2		String	epilogue of the URL if any
(opt) GenomicIDreq2	UID	String	Name of the GenomicID required and that is specified in the GIPO file as one of its fields

C.5 Configuring MAExplorer for use with other arrays

MAExplorer has been designed so that it may be reconfigured for different array dependencies including: geometries, number of replicate fields, scanner dynamic ranges, labeling, etc. When first started, MAExplorer reads this configuration file and then uses this information to handle reading different types of array data files that are subsequently loaded. To make it easier to understand, the entire table is presented as several sub-tables - however, MAExplorer reads it as a single table (the default being called `MaExplorerConfig.txt`). Note that optional parameters are for the most part - optional. Many of these may be set from the MAExplorer menus once the program is started. The reconfigured state may then be saved (File | Database | SaveAs ... DB) with these and other state values retained for the next time the particular startup database is used.

We are developing tools for creating and editing the configuration file. In the mean time, edit the file with Excel and save the finished table as a tab-delimited text file with the name `MaExplorerConfig.txt` in the `Config` sub-directory) in the directory where your database is stored.

Table C.5 List of Configuration data file table fields.

Parameter subset	Function of these parameters
1. Array content & geometry	Describes the content and geometry of the arrays (required)
2. Threshold defaults	Describes the threshold defaults (optional)
3. Array database files	Describes the array specific database files (required)
4. Table field mapping	Describes "mapTF" table,field mapping. This maps user defined names to names required by MAExplorer and is only required if the user names are different from the names MAExplorer expects.
5. URL genomic databases	Describes base addresses of genomic Web DBs (optional). If you do not specify these, default values are supplied from the program.
6. User menus	Describes user-specific menus (optional)

The following sub-tables list the configuration **parameter**s and some typical values that might be included. These examples illustrate the variety of parameter options with *examples* of values that might be used. Required entries are listed at the tops of the tables.

A typical MAExplorer minimal configuration database table might look like:

```
Parameter            Value              DataType    Comments
MAX_FIELDS           1                  int         # replicate grids/array
MAX_GRIDS            2                  int         # grids/field
MAX_GRID_COLS        38                 int         # columns/grid
MAX_GRID_ROWS        27                 int         # rows/grid
usePseudoXYcoords    true               boolean      use pseudoarray XY coord image - no XY data
gipoFile             GIPO.txt           File        name of GIPO file from
samplesDBfile        SamplesDB.txt      File        name of Samples DB file
dataBase             demo               String      default name of project database
dbSubset             demo1              String      default database subset name
useRatioData         false              boolean     treat duplicate(F1,F2) data as ratio (F1/F2) - eg.
Cy3/Cy5
EditDate             Tue Aug 21 15      String      demo
```

Table C.5.1 List of array database-specific <u>content</u> and <u>geometry</u> configuration (Parameter,Value) entries

This table lists most of the options that the use could define. If they define an option, it will override the default set by MAExplorer. The values are shown for some typical databases. (See (A HREF="#optTblNotation">Section C.1.1 for option notation.)

A) Array geometry parameters

Parameter	Value	DataType	Comments
(req) MAX_FIELDS	2	int	# duplicate grids (blocks, patch, etc.) of spots for each gene in the array (i.e. F1, F2, etc.). Note that Cy3 and Cy5 data for each spot count as one field.
(req) MAX_GRID_COLS	24	int	# cols/grid in the array
(req) MAX_GRID_ROWS	9	int	# rows/grid in the array
(req) MAX_GRIDS	8	int	# grids in the array

(opt) ignoreExtraFields	FALSE	boolean	if there are additional fields of data in the GIPO or .quant files, then ignore them. Only use the first rawIntensity field. Note: this option is not normally used.
(opt) reuseXYcoords	FALSE	boolean	Reuse XY coordinates from first sample for rest of the samples
(opt) SpotRadius	7	int	(2 to 20 pixels) 50 microns, scroller. Note: this should be set to about 4 or 5 for a 10000 gene DB.
(opt) swapRowsColumns	FALSE	boolean	set if swap rows and columns in the array (used with our particular Research Genetics arrays)
(opt) usePseudoXYcoords	FALSE	boolean	use pseudoarray XY coordinates image if there is no explicit no XY spot position data generated by the quantification software
(future) FIELD_LAYOUT	LtoR	String	fields are Left to Right
(future) FIELDS_ARE_NUMBERED	TRUE	boolean	Data files contain field number. Otherwise field is extrapolated
(future) GRID_LAYOUT	Horizontal	String	Grids are Left To Right in the array
(future) GRID_PER_ROW	4	int	# grids per row in each field of the array

B) Ratio and background parameters

Parameter	Value	DataType	Comments
(ratio) fluorescentLbl1	Cy3	String	name of dye for fluorescent label 1
(ratio) fluorescentLbl2	Cy5	String	name of dye for fluorescent label 2
(ratio) useRatioData	TRUE	boolean	set if data is Cy3/Cy5 ratio data otherwise it assumes intensity data for each spot
(opt+ratio) useRatioMedianCorrection	FALSE	boolean	when using ratio data mode (Cy3/Cy5), use ratio median correction as the default
(opt) useBackgroundCorrection	FALSE	boolean	use background correction as the default when startup
(future) useCy5/Cy3	FALSE	boolean	compute Cy5/Cy3 ratios instead of Cy3/Cy5 ratios

C) Names of database, etc.

We indicate example values by italics.

Parameter	Value	DataType	Comments
(opt) calibDNAname	*mouse genomic DNA*	String	name for calibration DNA if available - replacing cloneID in the case where the clones are not yet in the I.M.A.G.E. database. The particular clone is located using the Plate(grid,row,col) reported when selecting the current gene.
(opt) classNameX	HP-X 'set'	String	default name of HP-X samples 'set'
(opt) classNameY	HP-Y 'set'	String	default name of HP-Y samples 'set'
(opt) dataBase	*MGAP DB*	String	name of the database project
(opt) dbSubset	*Preg 13 vs Lact 1*	String	name of the subset of data from the database
(opt) maAnalysisProgram	*Research Genetics Pathways 2.01*	String	name of spot quantification program
(opt) yourPlateName	*your plate*	String	name of researcher's clones if available - used in the cloneID data field in the case where the clones are not yet in the I.M.A.G.E. database. The particular clone is located using the Plate(grid,row,col) reported when selecting the current gene. (See Table 2.4.1)
(opt) emptyWellName	*empty wells*	String	what you called empty wells if there are any in the database. (See Table 2.4.1)
(opt) EditDate	06-19-00, Lemkin	String	comment why changed

D) Display Views

Parameter	Value	DataType	Comments
(opt) gangSpotFlag	TRUE	boolean	set gang spot display on startup for database with duplicate spots
(opt) presentationViewFlag	FALSE	boolean	start MAExplorer with larger fonts and graphics symbols suitable for live presentations
(opt) showEGLflag	FALSE	boolean	show EGL genes on startup from previously saved database that had EGL genes selected.
(opt) showMouseOver	TRUE	boolean	show mouse-over info when move mouse in windows
(opt) useDichromasy	FALSE	boolean	use orange-blue else use red-green color scheme
(opt) viewFilteredSpotsFlag	TRUE	boolean	view Filtered spots the array pseudoimage. If it is off, it shows just the pseudoarray image without spots passing the filter or MAExplorer state information.

Note that there are many other parameters reflecting the state of MAExplorer that are saved in the .mae startup file when doing a (File | Database | SaveAs...DB) operation. These are reviewed and set from the MAExplorer menus. These parameters are not listed here - although they could be used in setting up an initial .mae startup file.

Table C.5.2 List of default <u>threshold</u> database-specific configuration (Parameter,Value) entries

Some of the default thresholds and sizes may be defined here as it may be useful to vary them with different types of data.

Parameter	Value	DataType	Comments
(opt) CanvasHorSize	1100	int	pixels, horizontal size of microarray image **DEPRICATED**
(opt) CanvasVertSize	1100	int	pixels, vertical size of microarray image **DEPRICATED**
(opt) fontFamily	SansSerif	String	default text font family. See Font Family for other fonts. Some fonts look better with some operating systems.
(opt) clusterDistThr	10	float	default cluster similarity threshold in [0.0 : 100.0], scroller
(opt) maxGenesReported	50	int	max # of genes in highest/lowest gene report
(opt) maxPreloadImages	4	int	max # HP samples to initially load
(opt) nbrOfClustersThr	6	int	default # clusters for K-means clustering
(opt) pValueThr	0.2	float	default p-value for statistical tests
(opt) spotCVthr	0.25	float	default spot Coefficient of Variation value
(opt) allowNegQuantDataFlag	FALSE	boolean	set if .quant file data has negative intensity values otherwise it clips the negative values to 0.0
(opt) usePosQuantDataFlag	TRUE	boolean	Filter out genes where .quant file data has negative intensity values otherwise it uses the negative data

Table C.5.3 List of array specific auxiliary database files (Parameter,Value) entries

This lists the names of the database-specific auxiliary files. Note that the names of these files may change with the database but the name of the initial configuration file containing these names (i.e. `MaExplorerConfig.txt` does not change. Optional **Parameter**s are indicated with a "*" prefix. (See Section C.1.1 for option notation.)

Parameter	Value	DataType	Comments
(req) gipoFile	GIPO-DB.txt	File	Composite Gene-In-Plate-Order (GIPO) file containing the spot print order, Clone-IDs, gene names, GenBank IDs, plate coordinates, etc. (See Appendix C.4)
(req) samplesDBfile	Samples-DB.txt	File	list of hybridized samples in the database. [Note: an older depricated name was "membranesDBfile"]. (See Appendix C.2)
(opt) quantFileExt	.quant	String	alternate quantification spot file name extension to use instead of ".quant". (You might set it to ".txt") (See Appendix C.3)

Table C.5.4 List of optional (Table,Field) mappings to configure specific user's data types

Sometimes user data tables contain the proper data required by MAExplorer, but the names of the columns (i.e. fields) are different. MAExplorer can map user (table,field) names to the internal names it uses. This allows users to maintain their tables in the names they choose. The following *mapTF* entries are not required if the fields in the corresponding tables already have the *MAE field name*. The entries use the mapping where

- **[TableName]** is the name of the table (repeated twice in the following specification).
- **[MAE field name]** is the *internal MAExplorer* name of the field in the table.
- **[User field name]** is the external name of the field in table in the *user's file* that corresponds to the internal MAExplorer

field.

```
[TableName],[MAE field name],[TableName],[User field name]
```

The following table fields <u>may</u> be mapped. Note: mapping is required *only when the table field names of your data files are different than the internal MAExplorer table field names*.

1. **GipoTable** - GIPO table for entire database
2. **SamplesTable** - list of all samples in the database
3. **QuantTable** - each `.quant` spot data file

The following is an example of some of the parameters that might be added to the Configuration file to perform field name mappings. Note: these mappings are only required if the data field names are non-standard. This shows some typical field name mappings. It will not be the same for your data. (See <u>Section C.1.1</u> for option notation.)

Parameter	Value	DataType	Comments
(opt) mapTF	GipoTable,grid,GipoTable,SA	String	GIPO table grid name (numbers or letters)
(opt) mapTF	GipoTable,grid row,GipoTable,R	String	GIPO table row of grid name (numbers or letters)
(opt) mapTF	GipoTable,grid col,GipoTable,C	String	GIPO table column of grid name (numbers or letters)
(opt) mapTF	GipoTable,plate,GipoTable,RG Pl	String	GIPO table plate where clone came from
(opt) mapTF	GipoTable,plate row,GipoTable,RG row	String	GIPO table row of plate where clone came from
(opt) mapTF	GipoTable,plate col,GipoTable,RG col	String	GIPO table column of plate where clone came from
(opt) mapTF	GipoTable,Clone ID,GipoTable,Clone id	String	GIPO name of Clone ID
(opt) mapTF	GipoTable,GeneName,GipoTable,Gene name	String	GIPO table map gene name
(opt) mapTF	GipoTable,Unigene cluster ID,GipoTable,ucid	String	GIPO table UniGene cluster id (if available)
(opt) mapTF	Unigene cluster name,GipoTable,ucn	String	GIPO table UniGene cluster name (if available)
(opt) mapTF	GipoTable,GenBank 3',GipoTable,gb3'	String	GIPO table GenBank 3' id (if available)
(opt) mapTF	GipoTable,GenBank 5',GipoTable,gb5'	String	GIPO table GenBank 5' id (if available)
(opt) mapTF	GipoTable,dbEST 3',GipoTable,est3'	String	GIPO table dbEST 3' id (if available)
(opt) mapTF	GipoTable,dbEST 5',GipoTable,est5'	String	GIPO table dbEST 5' id (if available)
(opt) mapTF	QuantTable,grid,QuantTable,SA	String	Quant table array grid name (numbers or letters)
(opt) mapTF	QuantTable,grid row,QuantTable,R	String	Quant table row of grid name (numbers or letters)
(opt) mapTF	QuantTable,grid col,QuantTable,C	String	Quant table column of grid name (numbers or letters)
(opt) mapTF	QuantTable,RawIntensity,QuantTable,Intensity	String	Quant table RawIntensity data
(opt) mapTF	QuantTable,Background,QuantTable,BkgrdIntens	String	Quant table background intensity
(opt) mapTF	QuantTable,RawIntensity1,QuantTable,Cy3RI	String	Quant table RawIntensity1 Cy3 data
(opt) mapTF	QuantTable,RawIntensity2,QuantTable,Cy5RI	String	Quant table RawIntensity2 Cy5 data
(opt) mapTF	QuantTable,Background1,QuantTable,BkgrdCy3RI	String	Quant table background intensity for Cy3
(opt) mapTF	QuantTable,Background2,QuantTable,BkgrdCy5RI	String	Quant table background intensity for Cy5

Table C.5.5 List of configuration genomic database URLs (Parameter,Value) entries

These entries are base addresses of genomic and other Web servers that are used for accessing gene or hybridized sample specific data in external databases. Note that these entries are hardwired in MAExplorer and do not need to be specified in the Configuration file unless you wish to override these defaults.

Parameter	Value	DataType	Comments
(opt) dbEstURL	http://www.ncbi.nlm.nih.gov/irx/cgi-bin/birx_doc?dbest+	String	NCBI dbEst server by dbEST ID. You may use an alternative server.

(opt) GenBankAccURL	http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=2&form=1&term=	String	NCBI GenBank server by GenBankAcc ID. You may use an alternative server.
(opt) GenBankCloneURL	http://www.ncbi.nlm.nih.gov/irx/cgi-bin/submit_form_query? TITLE=dbEST+Retrieval+Output&INPUTS=1& BRACKETS=NONE&ADDFLAGS=-b&DB=dbest& NDOCS=10&Q1=	String	NCBI GenBank entry by Clone_ID server. You may use an alternative server.
(opt) GenBankCloneURLepilogue	[clin]	String	Epilog added after Clone_ID. You may use an alternative server.
(opt) IMAGE2GenBankURL	http://nciarray.nci.nih.gov/cgi-bin/UG_query.cgi? ORG=Mm&ACC=IMAGE:	String	lookup GenBank from CloneID server. You may use an alternative Image to GenBank server. The "ORG=Mm" should be changed to reflect the proper species, eg. "ORG=Hs" for human, etc.
(opt) IMAGE2GIDURL	http://nciarray.nci.nih.gov/cgi-bin/UG_query.cgi? ORG=Mm&GID=IMAGE:	String	NCI/CIT lookup GenBank GID from CloneID server. You may use an alternative CloneID to GenBank GID server. The "ORG=Mm" should be changed to reflect the proper species, eg. "ORG=Hs" for human, etc.
(opt) IMAGE2unigeneURL	http://nciarray.nci.nih.gov/cgi-bin/UG_query.cgi? ORG=Mm&CLONE=IMAGE:	String	NCI/CIT lookup UNIGENE from CloneID server. You may use an alternative CloneID to UniGene server. The "ORG=Mm" should be changed to reflect the proper species, eg. "ORG=Hs" for human, etc.
(opt) unigeneURL	http://www.ncbi.nlm.nih.gov/UniGene/query.cgi? ORG=Mm&CID=	String	NCBI UNIGENE by Clone ID server. You may use an alternative UniGene server. The "ORG=Mm" should be changed to reflect the proper species, eg. "ORG=Hs" for human, etc.
locusLinkURL	http://www.ncbi.nlm.nih.gov/LocusLink/list.cgi?ORG=&V=0&Q=+	String	NCBI LocusLink by GenBank ID server. The LocusLink server is accessed by GBIG
gbid2LocusLinkURL	http://www.ncbi.nlm.nih.gov/LocusLink/list.cgi?SITE=104 &V=1&ORG=Hs&ORG=Mm&ORG=Rn&ORG=Dr&ORG=Dm&Q=	String	NCBI LocusLink by LocusID server. The LocusLink server is accessed by LocusID
(opt) swissProtURL	http://www.expasy.ch/cgi-bin/get-sprot-entry?	String	SwissProt by SwissProt ID
(opt) pirURL	http://pir.georgetown.edu/cgi-bin/iproclass/iproclass?choice=entry&id=	String	PIR ProClass database by SwissProt ID
(opt) GeneCardURL	http://bioinfo.weizmann.ac.il/cards-bin/carddisp?	String	GeneCard DB server. You may use an alternative server.
(opt) histologyURL	http://mammary.nih.gov/models/	String	E.g NIDDK MGAP histology DB server. If you have an alternative histology model server, put it here.
(opt) modelsURL	http://mammary.nih.gov/models/	String	e.g. NIDDK MGAP mouse models DB server. You may use an alternative models server.
(opt) proxyServer	http://www.lecb.ncifcrf.gov/cgi-bin/maeProxySvr?	String	NCI/LECB proxy server to access servers outside of the Java "sandbox". If you set up MAExplorer on your local server, then] this should point to a proxy server on your system.

Table C.5.6 List of configuration database-specific userHelp menu (Parameter,Value) entries

When creating a specific MAExplorer database, the Help menu may be configured for specific links to related databases. Any number of additional Help entries may be used (including none). The following shows the entries for MGAP.

Parameter	Value	DataType	Comments
(opt) HelpMenu1	List of hybridized samples	String	Help sub menu URL
(opt) HelpMenu2	MGAP animal models	String	Help sub menu URL
(opt) HelpMenu3	MGAP home page	String	Help sub menu URL
(opt) HelpURL1	http://www.lecb.ncifcrf.gov/mae/maeHybridizations.html	String	Help sub menu URL
(opt) HelpURL2	http://mammary.nih.gov/models/	String	Help sub menu URL

(opt) HelpURL3	http://mammary.nih.gov/	String	Help sub menu URL

Table C.5.7 List of configuration database-specific user Plugin menu (Parameter,Value) entries [Future]

When creating a specific MAExplorer database, the Plugin menu entries may be added to different parts of the menu tree. Any number of additional unique Plugin entries may be used (including none). The following table illustrates some possible plugin specifications that can be loaded (or not) at startup time or loaded when invoked from a menu.

Parameter	Value	DataType	Comments
(opt) **PluginMenuName1**	New Cluster plot	String	Plugin sub menu string
(opt) PluginMenuStubName1	PlotMenu:cluster	String	name of Plugin menu stub to add menu entry
(opt) PluginClassFile1	NewClusterPlot.jar	String	Name of class file
(opt)sPluginCallAtStartup1	InstallInMenu	String	handling plugins at startup: "InstallInMenu", "RunOnStartup", "NoInstall"
(opt) **PluginMenuName2**	New sample report	String	Plugin sub menu string
(opt) PluginMenuStubName2	ReportMenu:sample	String	name of Plugin menu stub to add menu entry
(opt) PluginClassFile2	NewSampleReport.jar	String	Name of class file
(opt)sPluginCallAtStartup2	InstallInMenu	String	handling plugins at startup: "InstallInMenu", "RunOnStartup", "NoInstall"
(opt) **PluginMenuName3**	Client-server	String	Plugin sub menu string
(opt) PluginMenuStubName2	-none-	String	name of Plugin menu stub to add menu entry
(opt) PluginClassFile2	ClineServerMAE.class	String	Name of class file
(opt)sPluginCallAtStartup2	InstallInMenu	String	handling plugins at startup: "InstallInMenu", "RunOnStartup", "NoInstall"

List of acceptable Menu stub names for: PluginMenuStubName

When MAEPlugin's are available, you will be able to insert them into various parts of the MAExplorer menu. If the menu stub is not found, it will install them in the generic "Plugin" pull-down menu.

- "FileMenu"
- "FileMenu:Databases"
- "FileMenu:State"
- "FileMenu:Groupware"
- "HPmenu"
- "GeneClassMenu"
- "NormMenu"
- "EditMenu"
- "EditMenu:EGL"
- "EditMenu:GeneSet"
- "EditMenu:CondList"
- "EditMenu:Preferences"
- "FilterMenu"
- "PlotMenu"
- "PlotMenu:EPplots"
- "PlotMenu:Histogram"
- "PlotMenu:PseudoArray"
- "PlotMenu:ScatterPlots"

- "ReportMenu"
- "ReportMenu:Genes"
- "ReportMenu:Samples"
- "ClusterMenu"
- "ClusterMenu:ClusterFlags"
- "ViewMenu"
- "PluginMenu"
- "HelpMenu"

C.6 Using the Cvt2Mae 'wizard' tool to convert your array data for use with MAExplorer

In order to use MAExplorer on your data, you must convert your data files into the data formats described in Appendix C and Appendix D. Although we and others have done this by editing user's data files into the required formats, it is a non-trivial process.

Therefore we have created a **C2M** Java conversion tool called Cvt2Mae to automate these conversions. You may
download and install Cvt2Mae on your computer and use it to convert your array data to MAExplorer data format. Figure C.6.1 shows Cvt2Mae array data converter.

Cvt2Mae is a "Wizard" driven process designed for use by molecular biologists. It handles commercial chips such as Incyte, Affymetrix, GenePix, Scanalyze, etc. or one-of-a-kind academic chips. It asks you questions to describe your chip and your data. We call the chip description the "Array Layout". After you have created or edited an array layout, you may save it for use in future conversions. [The array layouts are kept in a subdirectory "ArrayLayout" in the directory where you installed Cvt2Mae.] Since an ArrayLayout is a file, you could mail it to a collaborator. After you have answered the questions, you then run the converter and it generates the proper set of converted data files. In the case of user defined array layouts, we denote the latter as <User-defined> where the user assigns a name to that layout as part of the description. Essentially, the array layout contains a set of "rules" for describing the user's array data so Cvt2Mae knows how to read it. We plan to add the MAGE-ML standard to Cvt2Mae as one of the array layouts so it should be able to handle a wider variety of data.
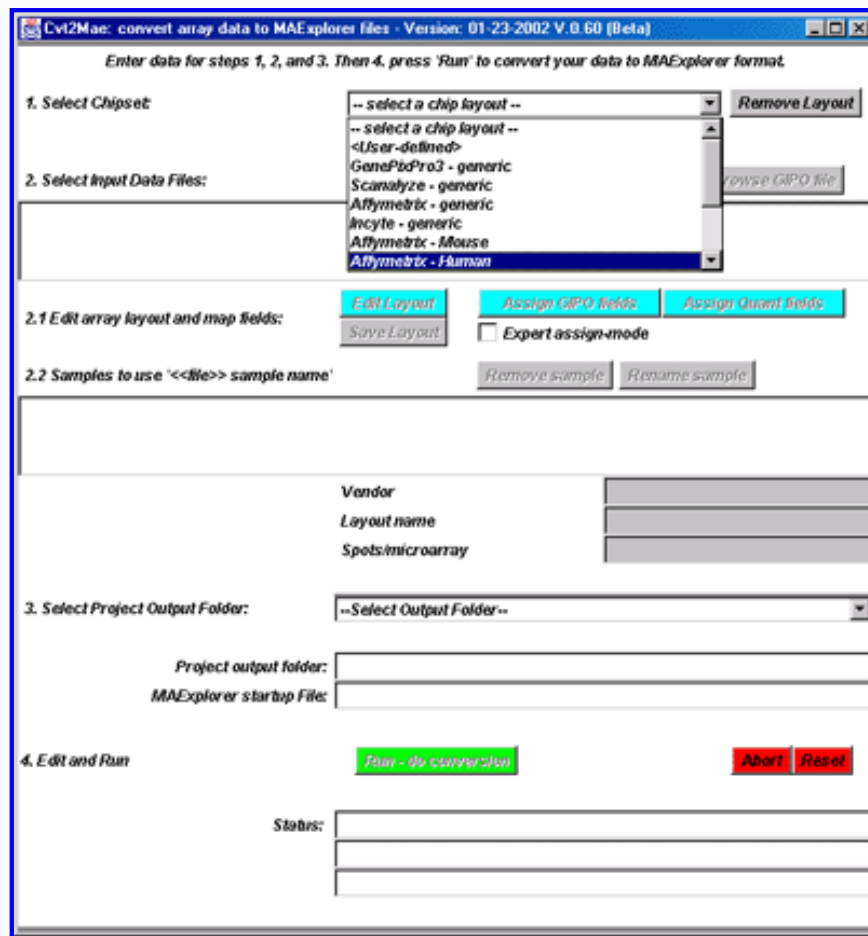
Figure C.6.1 The Cvt2Mae array data converter. Selecting a Chipset Array Layout. The built-in array layouts are shown for the various chip types. User-defined layouts may be added by selecting the <User-defined> layout then editing the layout using the Edit Layout, Assign GIPO fields, and Assign Quant fields. These options are described in more detail in the Cvt2Mae home page..

Converting data for known chip "Array Layouts" or lists of quantified spots

Assuming the desired array is in the list of chip array layouts, follow the eight step process below with steps 3 and 5 omitted. If the user must describe their own array data using the <User-defined> chip array layout, then they would do step 3. If your chip is one of the chips listed in the chip Array Layouts list, then you may be able to do an "Edit Layout" to modify the description without having to define the chip layout from scratch - in which case do step 3.

1. Select the desired Array Layout.
2. Select the set of input files to be converted.
3. If the array layout needs to be edited, use the Edit Layout, Assign GIPO fields, and Assign Quant fields wizards
4. Select the project output folder (i.e. directory) to place the converted data
5. (Optionally) save the edited Array Layout in case you want to use it again in the future.
6. Press "Run" to convert the data.
7. Press "Done" when the conversion is finished.
8. Go to the project directory and then to the MAE sub-directory (listed after step 4).
   Click on the "Start.mae" file to start MAExplorer on the next data. This assumes that you have previously installed MAExplorer.

The details on Cvt2Mae including more description, PDF examples of conversions for several different types of arrays, the

Appendix D. Use of MAExplorer as a stand-alone application

The MAExplorer program is used primarily as a stand-alone Java Application. The same Java program may also be used as a Java Applet that runs within a Web browser. Both the applet version and the stand-alone version may be run on a stand-alone computer without an Internet connection if the microarray database data files reside on that machine. The stand-alone Java version is more robust and has some important advantages discussed in the Introduction and Appendix E.2.

This section discusses the installation of MAExplorer as a stand-alone application on a variety of computers. Since Java is portable between Microsoft Windows (95/98/NT/2000/XP), Macintoshes, Linux, Solaris, etc., it is possible to freely download and install MAExplorer and Cvt2Mae on your computer and run it as an application program.

There is a discussion on using it with other arrays (Appendix C) that requires editing data files for use with MAExplorer. An array data conversion tool is being constructed which will automate this process in the future.

1. Stand-alone applications can utilize huge amounts of memory (for example, we were able to use >80Mb on a 128Mb Windows NT system). This means that an analysis of large numbers of hybridizations and complex statistical computations could be done locally.

2. The stand-alone version can do local disk I/O. This means that data-mining sessions may be saved and restored locally to continue a data-mining session at a latter time. Plots may be saved as GIF (.gif) files, and text windows as ASCII text (.txt) files. Hybridized sample quantification files slowly downloaded from a microarray database Web server may be cached on the local machine - saving time when the data session is resumed at a later time. Users could then add GIPO conforming hybridized sample quantification files on their desktop and perform a data-mining analysis of hybridized data from multiple sources - mixing data from public servers as well as private data. This latter option would be most useful using commercial arrays.

3. Users may download a self-extracting Java application. This is available on the MAExplorer Web site for a variety of computer systems and includes a Java Virtual Machine (JVM) for your operating system. [The JVM is only available on the installers on the LECB/NCI web site, it is not available on the SourceForge Web site because of space considerations. However, if by some chance your computer does not have a JVM, this should not be a problem since the JVMs (called JDK's) are available at various Web sites.] This JVM does not interfere with any other JVMs you may already have installed. However, it guarantees optimal operation of MAExplorer. This makes the stand-alone application more robust than the Applet as well as being simple to install.

4. Having more memory implies allows taking advantage of more sophisticated statistical, clustering and data-mining methods, as well as working with large numbers of samples with larger numbers of spots. Some Web browsers will not give you access to large amounts of memory - even if it is available on your computer.

5. Having the stand-alone version on the user's computer allows them to configure a local database for their own data.

D.1 Installing MAExplorer as stand-alone application

It is possible to download MAExplorer for data-mining either stand-alone on an your personal computer or in setting up a Web site for publishing your array data (like the MGAP site).
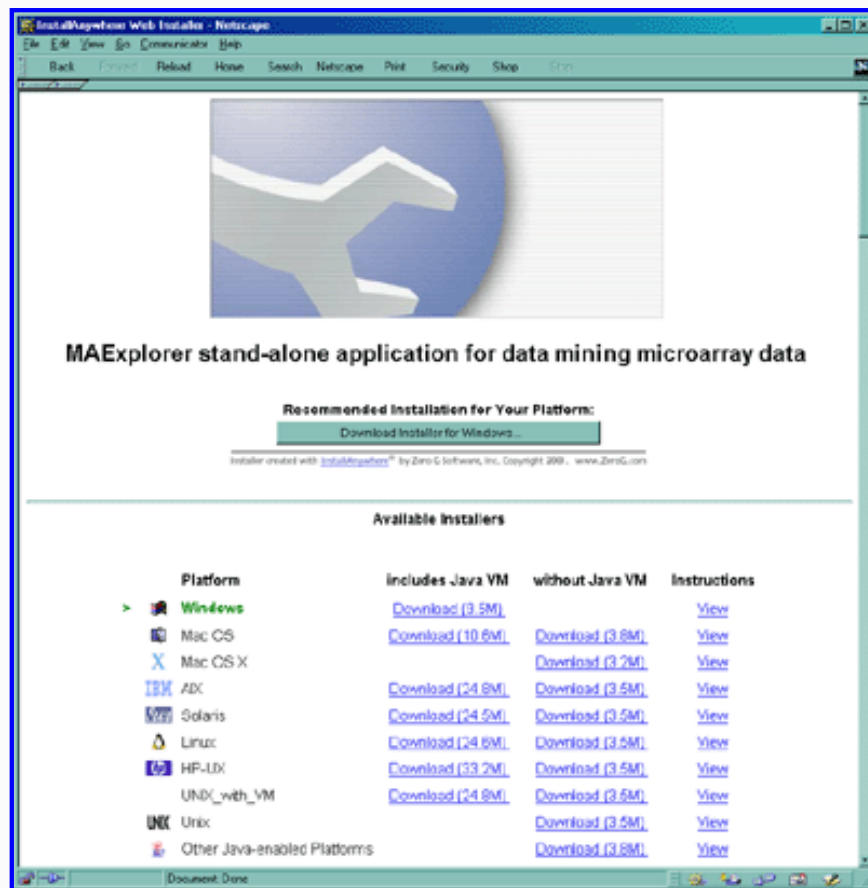
Figure D.1 Web page showing options for installing MAExplorer as a stand-alone application. Installers are available for Windows95/98/NT/2000/XP, Mac OS, Solaris, Linux, Unix, and other Java enabled platforms.

D.2 Downloading MAExplorer for stand-alone use with other arrays

After MAExplorer is installed, you may start it by clicking on a startup icon for one of the saved databases (Windows) or running the program (Unix, Macintosh). In Windows, you may also start it from the Window's "Start" menu. This stand-alone version enables you to save data on your local machine and to run MAExplorer independent of an Internet connection as well as other features that are discussed throughout this manual.

You first need to download MAExplorer for your particular type of operating system. These include Windows 95/98/NT/2000/XP, MacOS for Power PC, Sun Solaris, HP-UX, other Unix versions (e.g. Linux, etc). The Windows, MacOS and Solaris versions include a Java run-time (Java Virtual Machine) that works with MAExplorer. We recommend you download the full distribution for your computer (which includes a recent Java virtual machine (JVM) if it exists). This insures proper operation of MAExplorer and does not interfere with other Java applications you might have installed or will install.

This installation process uses a commercial "Java Installer" (InstallAnywhere(TM) by ZeroG Inc.) that requests you "Grant" it permission to save the installation on your computer. It will suggest where to install it or you can install it wherever you want. For example, in Windows it may suggest saving it in `C:\Program Files\MAExplorer\` - you can specify an alternative directory if the default disk does not have much free space left.

1. Go to download MAExplorer to begin the installation.
2. With Netscape, Click on "Grant" to questions "Starting programs stored on your computer" and "Reading, modification, or deletion of nay of your files". With Internet Explorer, click on "Yes" in response to the popup window "Do you want to install and run ...". [All data files are kept in the same directory tree where the MAExplorer program is kept. MAExplorer does not use any other directories on your computer and does not affect other Java programs on your computer.]
3. Either click on the default "Download ..." button that was automatically determined for your machine or scroll down and click on a "download" hypertext link for the operating system you want to download. If it is the latter, it pops up a "Save as"

window for you to select a temporary directory in which to put the "installMae.exe" (Windows), "installMae.bin" (MacOS and Unix systems), etc. In Windows this could be "C:\Temp" or in Unix "/var/tmp", etc. After you have installed MAExplorer, you can delete this file.

4. After the file was downloaded on your machine, go to that directory and start that file (e.g. click on it on a Windows or MacOS system). [If you wish, you may exit your Web browser at this point since it is no longer needed].
5. When the installer starts, it pops up a window "Preparing to Install" and gives you the option of selection a language. Click on the "OK" button.
6. It then pops up an "Introduction" window. Click on "Next".
7. It then pops up a "Choose Install Folder". You have the option of specifying a different directory. E.G. (on Windows) if it says: `C:\Program Files\MAExplorer` and you don't have space on disk C: but do on disk E: and want to put it into its own directory, you could specify it as `E:\MAExplorer`. Then click on the "Install" button.
8. It then pops up a new window with "Install Complete". Click on the "Done" button. You are now ready to run MAExplorer.

D.2.2.1 Subsequent updating of only the MAExplorer JAR file from the MAExplorer server

Generally, you can greatly shorten the subsequent download of new versions of MAExplorer. This will work most of the time as most often this is the only change in the distribution (not including the documentation). If this causes problems, it may mean that the other files have changed and you may have to reinstall the complete distribution as in Appendix D.2 above.

1. Download just the MAExplorer.jar file.
2. Use this to replace the previous version of the MAExplorer.jar file on your system with this file.

Additional instructions for Installing MAExplorer on Sun Solaris

These instructions are valid for use with both TWM and CDE window managers. If you have problems with the Sun installer, you may need to update your Solaris OS system patches using a recent patch set.

It may involve more than a single patch. It is the latest Recommended Patch Cluster from Sun. We **STRONGLY** recommend having your System Admininistrator do this for you if you have not done this before. Point your Web browser to:

```
http://sunsolve.Sun.COM/pub-cgi/show.pl?target=patches/patch-access
```

and choose the appropriate patch set for the version of Solaris (2.6, 7, or 8, etc.) that you are running. Do not choose any of the x86 versions unless you are running Solaris x86. Click on either the Download HTTP option or Download FTP option, and click the GO button to download the patch set.

---

1. Download Solaris version of MAExplorer by clicking on the correct link.
2. After downloading, open a shell or `xterm` window then, `cd` to the directory where you downloaded the installer.
3. At the prompt type: "`sh ./installMAExplorer.bin`". this will run Install Anywhere.
4. The InstallAnywhere program should now be running, follow the instructions to install MAExplorer.
5. After MAExplorer is installed click on the Exit key to end the Install Anywhere program. Then cd to the directory where it was installed and type "MAExplorer&" to run MAExplorer.
6. To uninstall MAExplorer: `cd` into `/UninstallerData` subdirectory and at the Unix prompt type `Uninstall_MAExplorer`. This will bring up a window to prompt you to uninstall MAExplorer. Click on uninstall button.

If you requested it, note that a Java virtual machine is included with this download. It will run automatically when you run the shell script.

D.2.2 Enabling MAExplorer to fetch data from a microarray Web server

When run as a stand-alone application, MAExplorer can be set up to read data from a MAExplorer enabled Web database server

and to cache the data to your local computer. The following lists the three variables that should be set in the `.mae` startup database file (see full table in discussion on setting up the configuration file). These variables may be set from the Edit menu Preferences submenu **Use Web DB** and **Web DB data caching** toggle switches, and the URL of the database by the **Set Web DB** command.

Table D.2.2 Parameters specifying Web database access. These optional parameters may be used for control access to a microarray Web database. The example values are shown for the MGAP database.

Parametery	Value	DataType	Comments
(opt) enableFIOcaching	TRUE	boolean	enable caching data files from Web server on local compute
(opt) saCodeBase	http://www.lecb.ncifcrf.gov/mae/	String	Web database to use to get the data
(opt) useWebDB	TRUE	boolean	set get data from a Web database

D.2.3 Enabling MAExplorer cache to save Web data on local computer

When run as a stand-alone application, MAExplorer can be set up to cache data from a MAExplorer enabled Web database server to your local computer. You can enable caching for future access using the Edit menu Preferences submenu **Web DB data caching** toggle switch.

D.3 Starting MAExplorer by clicking on a .mae file

Once the MAExplorer stand-alone application has been installed and registered with the operating system, then for Windows (and other operating systems when you can register files for startup) just click on the icon for any file with a **.mae** file extension to start MAExplorer on that data. For Unix systems you can start MAExplorer on a .mae file by:

```
(installation path)/MAExplorer MAE/(some startup file).mae
```

You can let your Unix system find MAExplorer by putting it in your `path` variable in your login or shell startup script.

```
set path = ($path <MAExplorer installation path>)
```

Then, you would start it by specifying the startup file residing in the MAE/ subdirectory as:

```
MAExplorer MAE/(some startup file).mae
```

There is a set of sample MGAP `.mae` files in the MAE subdirectory in the downloaded installation.

D.4 The data file format for .mae startup files

When MAExplorer is used as a stand-alone application, it first reads a (tab-delimited) startup file. This file contain the names of the hybridized samples to be loaded as well as some of the additional parameters listed here. Note most of these parameters may be specified in the configuration file as defaults and therefore do not need to be included in the .mae file unless you wish to override the configuration values. (See Table C.5 tables for a list of these parameters).

The `.mae` startup files are simply tab-delimited ASCII files with a .mae file extension. These could be created or edited either manually (e.g. using Microsoft Excel and saving the file as a tab-delimited file) or by various database programs (eg. the NCI/CIT mAdb program, the MAExplorer Cvt2Mae program being developed). They may also be generated by MAExplorer (File:Database:Save as file DB). The .mae file form consists of two tab-delimited columns containing fields **Name** and **Value**. These field names appear in the first row. This is followed by instances of the various parameters. A simple .mae file is shown in the following table Table D.4 using a 4 sample Lactation database subset from the MGAP database. Although any of the configuration file values can be specified in the .mae file, we list some of the more common optional parameters are indicated in

Table D.4 The Minimum data required entries for .mae startup files. These entries are shown with example values from some of the data in the MGAP demonstration database. Each sample is specified by an **image*i*** name.

Name	Value
image1	C57B6-L1-30min
image2	C57B6-L3-1hr
image3	C57B6-L10-29hrs-1
image4	Stat5a.--.L1-30min

Table D.4.1 Some of the common optional entries for .mae startup files. These entries are shown with example values for the 4 samples in Table D.4. See (Appendix C.5 for lists of many other options.

Name	Value	DataType	Comments
(opt) maxPreloadImages	4	int	overide the number of samples (called images) to actually load. This may be less than the number of image entries.
(opt) configFile	MaExplorerConfig-MGAP.txt	String	name of Configuration file if *not* MaExplorerConfig.txt
(opt) dataBase	MGAP DB	String	name of this specific database
(opt) dbSubset	Pregnancy 13 days: C57BL/6 vs. stat5a (-,-), 8 samples	String	title for this subset name of the database
(opt) Xlist	1,2,3	String	hybridized samples for initial HP-X 'set'. Corresponding to image1, image2, etc. Empty if not defined - may be defined using the Choose HP-X(Y,E) in the File menu.
(opt) Ylist	4	String	hybridized samples for initial HP-Y 'set'. Corresponding to image1, image2, etc. Empty if not defined - may be defined using the Choose HP-X(Y,E) in the File menu.
(opt) Elist	1,2,3,4	String	hybridized samples for initial HP-E 'list'. Corresponding to image1, image2, etc. Empty if not defined - may be defined using the Choose HP-X(Y,E) in the File menu.
(opt) classNameX	C57B6 lactation (days 1,3,10)	String	Experimental class name for the HP-X 'set' of hybridized samples
(opt) classNameY	Stat5a (-,-) lactation day 1	String	Experimental class name for the HP-Y 'set' of hybridized samples
(opt) noMsgReporting	TRUE	boolean	If set TRUE, used with Applet only to not send loading status message.
(opt) reuseXYcoords	FALSE	boolean	If set TRUE and the quantified data files have the (x,y) coordinates for each spot, then use the same coordinates for all subsequent data files so that the arrays can be superimposed (for Flickering two HPs).
(opt) usePseudoXYcoords	FALSE	boolean	If set TRUE, force MAExplorer to generate pseudoarray (X,Y) spot coordinates and ignore (X,Y) data in the quantified spot files if it exists. This will be set to TRUE automatically if there are no (X,Y) data fields in the quantified spot files.

D.5 Using MAExplorer as an Applet on your computer

It is possible to create a Web site to publish users data using the MAExplorer.jar file to support your private microarray database Web site. (Note that you can also get the MAExplorer.jar file from the directory where you installed MAExplorer on your computer). You might choose to mimic the way we did the http://www.lecb.ncifcrf.gov/mae MGAP Web site or organize it differently. You need to do the following:

1. Create a a sub-directory *dir/* in your htdocs/ Web server tree (where *dir/*is what you called the directory).
2. Create *dir/Config* and *dir/Quant* sub-directories.
3. Copy the *MAExplorer.jar* file into *dir/*.
4. Edit configuration files and copy them to *dir/Config* sub-directory.

5. Copy quantification files into the *dir/Quant* sub-directory.
6. Add HTML Web pages containing <APPLET> code (see example below).

The following is a simple example of HTML code containing an applet which will invoke MAExplorer. You may add other options with PARAMs in the Applet (or for that matter in the the .mae startup file) that overide any options normally specifed in the Configuration file (See Appendix C.5).

```
<HTML>
<HEAD>
<TITLE>MAExplorer Startup: C57B6 Pregnancy vs Lactation</TITLE>
</HEAD>

<BODY>
<H2>MAExplorer Startup: C57B6 Pregnancy vs Lactation</H2>

This startup database will start the MAExplorer. It contains a subset
of the database consisting of four C57B6 mammary development
hybridized samples (HP): two each for pregnancy and lactation.

<APPLET CODE=MAExplorer.class  ARCHIVE=MAExplorer.jar
        WIDTH=10 HEIGHT=10 ALIGN=absmiddle>
  <PARAM NAME=configFile VALUE=MaExplorerConfig-MGAP.txt>
  <PARAM NAME=dbSubset VALUE="C57B6 pregnancy vs lactation">
  <PARAM NAME=image1 VALUE=C57B6-p13.1>
  <PARAM NAME=image2 VALUE=C57B6-L1-30min>
  <PARAM NAME=image3 VALUE=C57B6-p13.2poly-A>
  <PARAM NAME=image4 VALUE=C57B6-L1-total>
  <PARAM NAME=Xlist VALUE=1,3>
  <PARAM NAME=Ylist VALUE=2,4>
  <PARAM NAME=Elist VALUE=1,3,2,4>
  <PARAM NAME=classNameX VALUE=Pregnancy>
  <PARAM NAME=classNameY VALUE=Lactation>
(Sorry, you need a Java-capable browser to view this.)
</APPLET>
</BODY>
</HTML>
```

D.6 List of startup .mae files included in the download installation

This is a list of the .mae startup files from the MAExplorer MGAP database. These are included with the distribution for use in the tutorials, etc. They include data from 50 hybridized samples of mouse mammary breast tissue including normal and some knockout samples. There are about 1700 duplicate clones on the arrays which are membranes printed by Research Genetics and hybridized by the MGAP group. See the MGAP home page for more information on these samples. The .mae files are available as separate files http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database/. The data is also packaged as a zip file http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database.zip, and a Unix tar file http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database.tar.

.mae file name	X vs Y comparison	# of hybridizations
C57vsDevModels-15probes.mae	HP-XY is C57B6 vs developmental models	15
C57vsDevModels-15probes-cache.mae	HP-XY is C57B6 vs developmental models (with cache)	15
C57vsDevModels-38probes.mae	HP-XY is C57B6 vs developmental models. HP-E is all samples	38
Lact1-C57vsStat5a-38probes.mae	HP-XY is C57B6 Lactation day 1 vs Stat5a (-,-). HP-E is all samples	38
Lact1vs10-10probes.mae	HP-XY is C57B6 Lactation day 1 vs Lactation day 10. HP-E is all samples	10
Lact1vs10-38probes.mae	HP-XY is C57B6 Lactation day 1 vs Lactation day 10. HP-E is all samples	38
Lact-C57vsStat5a-5probes.mae	HP-XY is C57B6 Lactation day 1 vs Stat5a (-,-)	5

Lact-C57vsStat5aCEBPnull-19probes.mae	HP-XY is C57B6 Lactation day 1 vs Stat5a (-,-) and CEBP-null, HP-E has samples of other tissues	19
MAEstartupDefault.mae	none	none
Preg13day-C57vsStat5a-19probes-cache.mae	HP-XY is C57B6 Pregnancy day 13 vs Stat5a (-,-). HP-E has samples of other tissues (with cache)	19
Preg13day-C57vsStat5a-19probes.mae	HP-XY is C57B6 Pregnancy day 13 vs Stat5a (-,-). HP-E has samples of other tissues	19
Preg13day-C57vsStat5a-38probes.mae	HP-XY is C57B6 Pregnancy day 13 vs Stat5a (-,-). HP-E is all samples	38
Preg-C57vsStat5a-4probes.mae	HP-XY is C57B6 Pregnancy day 13 vs Stat5a (-,-)	4
Preg13VsLact1-18probes.mae	HP-XY is C57B6 Pregnancy day 13 vs Lactation day 1. HP-E is all samples	18
Preg13VsLact1-38probes.mae	HP-XY is C57B6 Pregnancy day 13 vs Lactation day 1. HP-E is all samples	38
Preg-C57vsStat5a-8probes.mae	HP-XY is C57B6 Pregnancy day 13 vs Stat5a (-,-)	8
Preg13day-Stat5aVsCEBP-null-38probes.mae	HP-XY is C57B6 Lactation day 1 vs Stat5a (-,-) and CEBP-null, HP-E is all samples	38
reuseXY-Preg13day-C57vsStat5a-38probes.mae	HP-XY is C57B6 Pregnancy day 13 vs Stat5a (-,-). Use XY coords of first probe for remainder for flickering. HP-E is all samples	38
reuseXY-Preg-C57vsStat5a-8probes.mae	HP-XY is C57B6 Pregnancy day 13 vs Stat5a (-,-). Use XY coords of first sample for remainder for flickering	8
MGAP-50samples.mae	C57B6 day 13 preg. vs day 1 lact., 50 samples	50

Appendix E. Design issues

This appendix addresses a number of key design issues on the implementation of MAExplorer and the implications they have on its efficiency. The ordinary user of MAExplorer need not be concerned with any of these issues. A PowerPoint presentation describing the class structure of the "Software design of the MAExplorer data mining tool" is available as either an Adobe Acrobat file (PDF) or a PowerPoint file (PPT).

E.1 Internal data structures design to facilitate direct manipulation

MAExplorer was constructed using a number of fundamental data objects including clones (genes), hybridized samples (membranes or glass arrays), tables, etc. organized using an object-oriented methodology enforced by Java. Sets of genes are implemented as bit sets for efficiency in both storage and set-theoretic operations. With a set being implemented as 64-bits/word, a set intersection, union or difference can be performed on 64 genes in parallel in one logical (i.e. AND, OR, XOR) computer instruction. This makes the data filter quite efficient when computing the intersection of many gene sets. When ordered gene lists are required, memory and compute intensive lists are used - but only when needed. Tab-delimited ASCII is used as the basic I/O file type for all types of data. This simplifies I/O and allows data to be prepared with a variety of systems including Excel, array quantification programs, relational database systems, etc.

Another major decision was to use multiple pop-up windows for 2D plots, histograms, expression profiles, clustergrams, reports, dialog boxes, etc. rather than sharing a single window. These windows are maintained by a special pop-up registry that handles many of the bookkeeping chores involved with tracking and updating multiple windows viewing the same underlying data. Whenever an event occurs which may change the set of data filtered genes, the current gene or the current cluster set of genes, the registry is notified. Some of the events are the current clone changed, the Filter parameters changed, the sample labels changed, the normalization method changed, etc. It in turn notifies all relevant active plots, tables and reports - requesting them to update themselves if necessary. This object-oriented design greatly simplifies the process of synchronizing the various data presentations with changes in the database.

E.2 Approaches to data mining: client-centric and server-centric models

There is a range of approaches for performing data mining of microarray data over the Internet. However, all assume rapid access to underlying databases and the ability to transform data from one presentation mode to another where differences might be easily observed. One extreme is the server-centric model using CGI or Applets in Web browser. This assumes that all data search and analysis is performed on a back-end server and graphic or tabular results from the server are sent back to the researcher over the Internet. The server-centric model has the advantage of keeping all user data up-to-date, but the disadvantage of performing all computations and graphics generation on the back-end server. Relying so much on the server for major computations and graphics generation can result in significant delays if the networks or servers are heavily loaded. The other extreme is the client-centric model. Here all of the data being analyzed is copied to a user's computer and computationally expensive analyses are done there. This has the disadvantage for the user of possibly not having the most up-to-date data to analyze as well as setup time overhead. However, it does distribute the computational load, allowing more effective data mining with many alternate views and avoiding excessive delays during a data mining session. In both the Web browser applet and the stand-alone application, data is downloaded to MAExplorer. The difference being access to the local file system with some additional capabilities in the case of the latter.

A good intersection of the server-centric and client-centric methods is to distribute the computation and data to the systems where they can be handled most effectively. Because Java enables computation in a Web browser, PCs currently available have enormous power and memory, and high-speed Internet connections are readily available, it is now possible to distribute some of the data and computations to the desktop. If high-speed direct manipulation methodology is to be made available on the Internet for microarray data mining, then it must be brought to the user's desktop browser or local computer rather than residing solely on the back-end server. This is the approach taken in designing the MAExplorer.

Table E.2 Comparison of client-centric vs. server-centric data mining. The table shows a comparison of some of the features of client-centric and server-centric (using CGI and/or Applet) data mining analysis methods. The client-centric approach presented here primarily uses Java with data downloaded to the client's computer. A server-centric approach might use a mix of HTML, CGI, servlet and Java. However, even a client-centric approach may take advantage of server support for additional functionality (e.g. accessing genomic servers to gain additional information about specific genes or sets of genes).

Approach	Advantage (+) disadvantage (-)	Feature
Client-centric a)	+	Java programs run (pretty much) on all operating system platforms as either stand-alone or applets (in browsers)
Client-centric b)	+	handles rapid response required for direct manipulation on the new generation of very fast desktop computers
Client-centric c)	+	stand-alone version may be restarted quickly from local data or data cached from the Web server
Client-centric d	+	size limitations are not a problem with stand-alone Java applications
Client-centric e)	+	Java plug-ins allows prototyping new local and Web DB analysis method functionality by any group of users
Client-centric f)	-	for the applet version, there is slow startup because the program and all data has to be downloaded each time it is run
Client-centric g)	-	difficult to build large stable Web-applets handling very large data sets. However, stand-alone applications don't have this problem
Client-centric h)	-	for the stand-alone application version, it must be installed on client's computer where there nmight be some level of incompatibility

Approach	Advantage (+) disadvantage (-)	Feature
Server-centric a)	+	may have better resources for very large data sets but with dependence on server
Server-centric b)	+	faster startup than downloaded applet since minimal GUI is required and data does not have to be loaded before computation requests may be made to the server
Server-centric c)	+	may be easier to prototype and distribute new functionality using third party software such as RDBMS, S-plus, etc. using centralized CGI or servlets where only one copy is required on the server
Server-centric d)	-	susceptible to Internet traffic bandwidth problems for large numbers of users
Server-centric e)	-	susceptible to server-load dependencies for large numbers of users
Server-centric f)	-	difficult to get very rapid response for direct manipulation for data mining

E.3 Conversion of microarray data files to MAExplorer format using Cvt2Mae

A tool is being developed that converts microarray data files, both commercial and one-of-a-kind research data to a complete MAExplorer data format. Input data will be tab-delimited, although it may be possible to use XML data at some point. When the tool becomes available, it will be announced on the MAExplorer home page and in this manual.

Cvt2Mae data converter

Because it is difficult to manually edit user's microarray quantified data files, we constructed the Cvt2Mae data converter program (also see Appendix C.6). The idea is to create array layouts for known array chips and to let the user define their own for specialized arrays. These user-defined layouts may then be saved and used in subsequent data conversions. The basic problem of data conversion is that of "field picking" to map user data fields to those required by MAExplorer, and of setting the appropriate options in the MAExplorer configuration files. User-interactive wizards query the user and then does this information to perform the conversion generating the output data files that are ready to use with MAExplorer. Cvt2Mae then generates the directory tree of required data files described in Appendix C.

E.4 Extending MAExplorer functionality using Java Plugins

We are adding the ability for users to add their own Java Plug-in Extensions to MAExplorer. These will extend capabilities of the core MAExplorer program to other analysis methods by users. The MAEPlugins Web site will be an Open Java API, open-source Java code examples, our plugins and donated plugins, links to plugins at other Web sites. Typical plug-ins include: normalization, Filters, PCA, clustering, client-server, Web-server functional analysis of cluster results, etc. We group these into three types of new analytic functionality:

1. Using Java code to implement the complete plugin. This means it will be portable across all systems.
2. Accessing local programs written in any language (eg. the R statistics package). This method may not be portable across all systems.
3. Web-CGI or client-server to specialized genomic DBs Plug-ins. This method should be portable across all systems.

The MAExplorer Open Java API (Applications Programming Interface) will allow users to get at all data structures without understanding the details of the system. The specialized application classes are derived from the GatherScatterAPI class which can access all of the internal MAExplorer data structures. This allows us to improve and change the internal data structures without causing problems with plugins using those data structures.

The following figures show the top level plugin design.
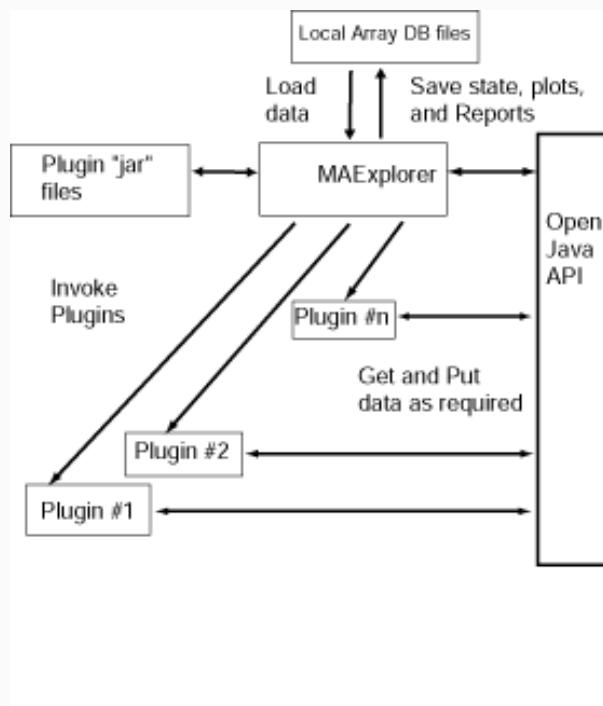
Figure E.4.1 Overall MAEPlugin design for MAExplorer. Plugins are dynamically loaded into MAExplorer where they may be invoked from a menu entry or by various other means such as startup, normalization, etc.
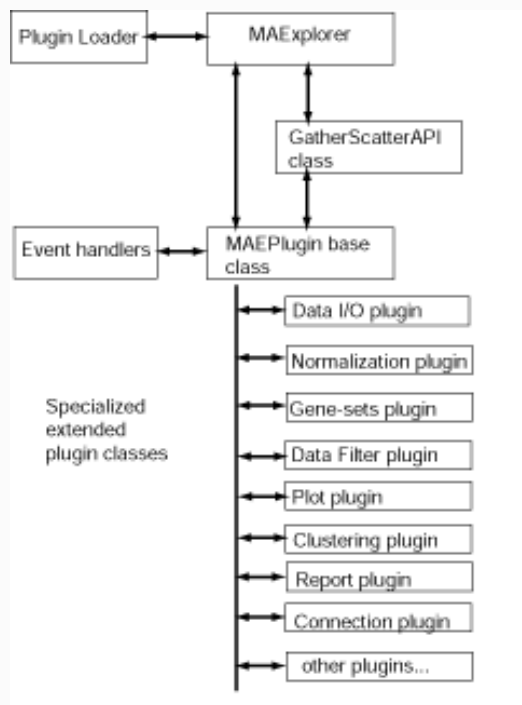
Figure E.4.2 Open Java API for MAEPlugins. Each type of application could be derived from specialized Java classes that contain most of the access methods required for that type of analysis.

E.5 Web database server design

Although MAExplorer can be run stand-alone on a user's computer, additional capabilities may be made available with support

from the back-end Web database server. This server design, used with the MGAP database, includes several distinct functions (Figure 1). The primary one is the hosting of login-protected microarray quantitative data and auxiliary flat files required to support basic MAExplorer operations. These "flat files" could be synthesized on the fly from searches on a relational database server that is part of the microarray database Web server. The public database does not require a login while the collaborator subset of the database does.

In support of the MGAP server, additional software was written to automate the pre-processing of the microarray quantitative data from Research Genetics' Pathways array quantification analysis program and perform compression and Web server updates for this data. The Web server also hosts several common gateway interface (CGI) programs. These include user login support, a Web proxy server (to access other genomic Web sites from the Java applet), support of login-protected user state file access, custom database creation, user state files, and "groupware" user-access support.

Downloading the stand-alone MicroArray Explorer (current release)

You may freely [download and install the current stable release](#) of the stand-alone version of the MAExplorer program. You are free to use or redistribute MAExplorer ([see disclaimer](#)). We also include a subset of 50 Mammary Genome Anatomy Program [(MGAP)](#) hybridized sample data to run stand-alone on your computer platform. These are may also be downloaded directly as:

- individual files http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database/
- a ZIP file http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database.zip
- a Unix Tar file http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database.tar

This data may be used for learning about MAExplorer with the tutorials and for investigating some of the stages of normal and mouse-model mammary development. The MAExplorer reference manual may be [viewed](#) in your browser from the Web from this Web site. Alternatively, you may download the full manual as a Acrobat [MaeRefMan.pdf](#) PDF file (> 5Mb).

If you have problems with the installation, then you might want to read the rest of this section and also the part of the manual which discusses installation ([Appendix D](#)) and using it with your arrays ([Appendix C](#)). The latter requires editing your data files for use with MAExplorer. The [Cvt2mae](#) is a "wizard" array data conversion tool automates this process.

- An archive of some of the [stable older releases](#) of MAExplorer is available for a limited period on the LECB/NCI server.

- For more details on specific releases, see the [revision notes](#) (Section 4.2).

If you have previously installed MAExplorer and you want to update just the [MAExplorer.jar](#) file, you can do this as described in [Section 1.3](#).

After initially, installing MAExplorer (or the Cvt2Mae for that matter), you can simply [download the latest .jar file](#) and overwrite the previous version you had when you installed the program. The MGAP demo data can be [downloaded separately](#).

SourceForge [Download MAExplorer Installer](#) Version: 0.96.11, 08-19-2002
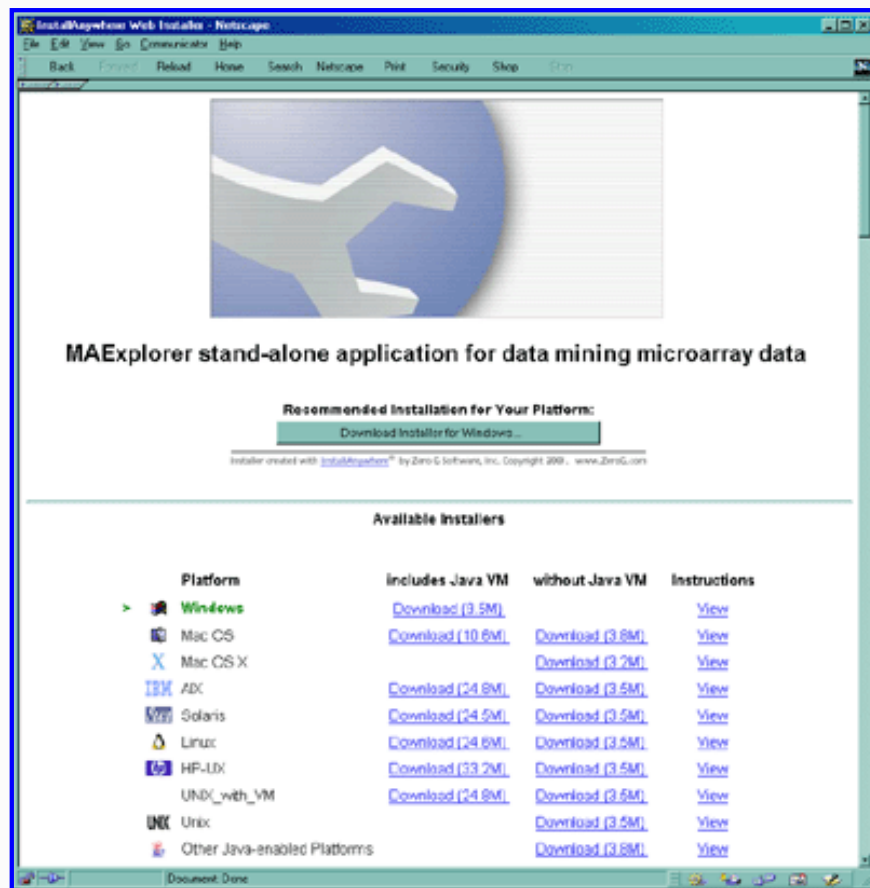
Figure. Web page showing options for installing MAExplorer as a stand-alone application. Installers are available for Windows95/98/NT/2000/XP, MacOS-8/9, MacOS-X, Solaris, HP-UX, Linux, Unix, and other Java enabled platforms. [Click on the figure to see a high resolution version.] NOTE: the MacOS installer is currently not available. If you have problems with the Sun installer, you may need to update your Solaris OS system patches (see below).

Distribution contents

1. We recommend **including** the Java Virtual Machine (JVM) for a more robust installation.
   This will not affect any of your other Java applications or Web browsers as it is used only with MAExplorer.
2. The distribution includes:
   o The MAExplorer Java stand-alone application,
   o A set of 50 hybridized samples data from the http://mammary.nih.gov/mgap DB. These may be accessed after you have installed MAExplorer by clicking on a ".mae" file in the /MAE/ directory in the installation directory. These date are not on the SourceForge Web site but are available at http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database.zip.
   o Support files for your operating system - possibly including the JVM which you may optionally download (JVMs are only on the NCI/LECB site).

1. Procedure for downloading and installing MAExplorer on your computer

1. Click here to select the current installer for your operating system. This Web page allows you to select the operating system you are using. If you have problems downloading the installer with Netscape 4.7x or later, then try Internet Explorer 5.0. It could be a Mime/type problem with your browser setup.

2. You start the download process when you click on the installer for your computer platform. (You may alternatively use the default installer discussed below.) Follow the directions it provides as you download the installer. It also provides instructions in the "View" hyperlink adjacent to the operating system you selected that tells you what to do after you finished the download. Part

of the installation consists of telling the installer where you want to 1) put the executable installer (a temporary directory where you have lots of room is a good choice), and 2) the "installation" directory where you will typically leave the distribution after the installer unpacks it.

We use the commercial [InstallAnywhere](TM) program to create the installers. It provides installers for:

- Windows 95/98/NT/2000/XP
- Mac OS (OS8 and OS9, OS-X)
- Solaris
- HP-UX
- Linux
- Unix
- Other Java enabled platforms

Other systems will be added as installers become available through InstallAnywhere (www.ZeroG.com).

1.1 The Default Installer

Alternatively, you can use the default installer that is selected for your computer. ***If you want to control where the files are saved on your computer, then use the explicit installer for your particular platform described above***.

The default installer will put the installer executable in a fixed directory and the installed MAExplorer files in another fixed directory.

- For Windows, the installer will go into `C:\InstallAnywhere_Installers\` and the program files into `C:\Program Files\MAExplorer\`.

- For Unix systems, it will put them at `$HOME/InstallAnywhere_Installers` and the program files into `$HOME/MAExplorer/`.

- For MacOS, it will put them on the desktop.

1.2 Installation Notes

Currently, the Windows and Linux installers are robust. We have had mixed success with Mac OS and Solaris.

Note that the installers (where possible) will include a copy of a recent Java Virtual Machine (JVM) from InstallAnywhere(TM) to make running MAExplorer on your computer more robust. This is used locally and *only* affects the running of MAExplorer. It will *not* affect any other Java applications on your computer. In the case of Mac OS, if you have an older version of the MRJ JVM, it will ask you if you want to upgrade to the newer version (MRJ-2.4.5) - however you do not have to unless you want to.

The [MAExplorer Reference Manual](#) describes the details of MAExplorer as well as showing a number of screens illustrating various data-mining operations. Several tutorials are available and are discussed in the Reference Manual.

1.3 Downloading just the MAExplorer.jar file after initial install

If you have previously done an installation. you may avoid a complete re-installation download by getting just the latest Java [MAExplorer.jar](#) file. You should replace the old version of this file on your system with the one you are downloading. This will work if the new MAExplorer.jar file does not depend on any new entries in the configuration files (which generally the case - try it and see what happens).

2. Description of Sample MGAP Datasets in the Distribution

The MGAP data is supplied in three directories. The installer leaves other directories and files in the "installation" directory needed to run MAExplorer as a stand-alone application. These include `MAExplorer.jar, MAExplorer.exe` (`.bin` if UNIX or Mac), `Uninstall-MAExplorer.exe` (`.bin`), etc. The `lax` files, and `jre, resource` directories are used by InstallAnywhere and you do not need to be concerned with their contents. The MGAP data directories are listed here and are discussed in detail in Appendix D of the Reference Manual:

1. `/Config` - contains database configuration and samples tables for this database
2. `/Quant` - contains 50 quantified hybridized sample spot lists
3. `/MAE` - contains MAExplorer ".mae" startup files

2.1 List of MGAP demo data MAExplorer ".mae" Startup Files in the /MAE Directory

The table lists the startup files provided for the MGAP database. Some good sets of data to try initially are the Lact1vs10-38probes.mae, Preg13VsLact1-38probes.mae, and Preg13day-C57vsStat5a-38probes.mae startup files.

.mae startup file	Data set contents
Lact-C57vsStat5a-5probes.mae	5 probes. (X,Y) is lactation day 1 (C57B6, Stat5a(-,-))
Lact-C57vsStat5aCEBPnull-19probes.mae	19 probes. (X,Y) subset is lactation day 1 (C57B6, Stat5a(-,-) + CEBP-null)
Lact1-C57vsStat5a-38probes.mae	38 probes. (X,Y) subset is lactation day 1 (C57B6, Stat5a(-,-))
Lact1vs10-38probes.mae	38 probes. (X,Y) subset is C57B6 lactation day (1,10)
MAEstartupDefault.mae	No initial samples loaded
Preg-C57vsStat5a-4probes.mae	4 samples. (X,Y) is pregnancy (C57B6, Stat5a(-,-))
Preg-C57vsStat5a-8probes.mae	8 samples. (X,Y) is pregnancy (C57B6, Stat5a(-,-))
Preg13VsLact1-38probes.mae	38 samples. (X,Y) subset is pregnancy (C57B6, Stat5a(-,-))
Preg13day-C57vsStat5a-19probes-cache.mae	19 samples from MGAP Web server. (X,Y) subset is pregnancy (C57B6, Stat5a(-,-))
Preg13day-C57vsStat5a-19probes.mae	19 samples. (X,Y) subset is pregnancy (C57B6, Stat5a(-,-))
Preg13day-C57vsStat5a-38probes.mae	38 samples. (X,Y) subset is pregnancy (C57B6, Stat5a(-,-))
Preg13day-Stat5aVsCEBP-null-38probes.mae	19 samples. (X,Y) subset is pregnancy (Stat5a(-,-),CEBP-null)
reuseXY-Preg-C57vsStat5a-8probes.mae	Same as other startup, but uses XY coordinates of 1st sample
reuseXY-Preg13day-C57vsStat5a-38probes.mae	Same as other startup, but uses XY coordinates of 1st sample
C57vsDevModels-15probes-cache.mae	15 samples from MGAP cache. (X,Y) subset is (C57B6, knock-outs)
C57vsDevModels-15probes.mae	15 samples. (X,Y) subset is (C57B6, knock-outs)
C57vsDevModels-38probes.mae	38 samples. (X,Y) subset is (C57B6, knock-outs)
MGAP-50samples.mae	50 samples. All of the public samples sorted alphabetically

2.2 Starting MAExplorer Using a ".mae" Startup File

If you are on Windows 95/98/NT/2000/XP system, simply click on the .mae file you want to use. Hint: you might put a short-cut to the *installation-directory*`\MAE\` directory on your desk-top to make it more convenient to find the files.

If you are on a Macintosh system, then start MAExplorer and then run the startup .mae file you want by going to the File menu and then the Databases submenu. Use the "Open disk DB" option to browse your disk and then open up the startup file of interest.

If you are on a Unix system, then you supply the MAE file explicitly in the command line. You might consider adding the "installation" directory to your UNIX $PATH or $path variable to have UNIX automatically find the executable binary.

```
cd installation-directory/
MAExplorer.bin MAE/Preg13VsLact1-38probes.mae
```

2.3 The MAExplorer Error Log File

Each time you run MAExplorer, it creates or overides the previous error log file called `MAEerr.log` in the *installation-directory*. If you are experiencing major problems, this file is useful to us in helping figure out what is wrong. Otherwise, just ignore it.

2.4 Problems installing MAExplorer on some operating systems

1. The MacOS installer is available, but may not work with older versions of MacOS. In addition, there may be problems if your file names are longer than 32 characters. For now, the solution is to use short file names. There may also be problems if your data files have embedded carriage returns in addition to line feeds. For now, the solution is to strip the CRs out of the data file.

2. On Solaris, and possibly other Unix systems, you may have problems with the stack limits. Do a "man limit" to read about the command for your particular Unix shell. We have found that the following seems to work. For the Unix C-shell (csh), add the following to your .cshrc startup file.

```
limit stacksize unlimited
```

   In addition, we have set the default stack size that MAExplorer uses to 256Mbytes. If your computer has less physical memory, it will page. You may also increase this number as well if you have more memory and want to use it. The solution is to edit the MAExplorer.lax file found where you installed MAExplorer. Change the two instances of memory allocation from 256000000 to a smaller number that is less than your actual memory size.
3. On Solaris, if you download the version with the JVM, unless your Solaris system has been updated recently, it may not be able to find the libCrun.so.xxx version required by the JVM. Try downloading the non-JVM version or update your Solaris system.
4. If you have problems with the Sun installer, you may need to update your Solaris OS system patch set. It is not a single patch. It is the latest Recommended Patch Cluster from Sun. We **STRONGLY** recommend having your SysAdmin do this for you if you have not done this before. Point your Web browser to:

```
http://sunsolve.Sun.COM/pub-cgi/show.pl?target=patches/patch-access
```

   and choose the appropriate patch set for the version of Solaris (2.6, 7, or 8) that you are running. Do not choose any of the x86 versions unless you are running Solaris x86. Click on either the Download HTTP option or Download FTP option, and click the GO button to download the patch set.

2.5 FAQ of problems using MAExplorer on Mac OS for NCI/CIT mAdb users

Q: How many characters can I use in array names for data to be donwloaded to MAExplorer?

A: For Mac-X, with 256 character file names, this is not a problem. For MacOS 8 and 9 with 32 character file names it may be a problem. Because MAExplorer uses file extensions (eg. ".quant"), you are currently limited to 25 characters or less. We will be modifying the system to remove this limit.

Q: I tried unsuccessfully to open NCI/CIT mAdb data (nciarray.nih.gov) on a Mac OS system. I generated a .zip file using mAdb "BETA Formatted Array Data Retrieval Tool" , then decompressed this .zip file using "Stuffit Expander" on my Mac. The Start.mae file could not be opened by MAExplorer, what can I do to fix this?

A: Stuffit Expander (default settings) removes a form feed character from decompressed text files, this prevents the Start.mae (and other text files used by MAExplorer) to be read by MAExplorer. To fix this you need to set Stuffit Expander so that it will keep the form feed characters when it decompress text files:

```
Open Stuffit Expander by double clicking its icon
Click on menu File -> Preferences
Click on "Cross Platform"
Click on "Never" button of 'Convert text file to Macintosh format:'
```

Your .zip will be decompressed properly and the text files from your mAdb data can now be open by MAExplorer.

Q: How do I start MAExplorer on my data automatically by double-clicking a Start.mae file on my Mac.

A: There is no easy way to do this at this time. Use the File menu, Databases, Open Disk DB browser to specify the Start.mae file.

2.6 Sun Solaris (or other Unix system) Memory Problems

We have on occasion seen the following types of memory errors. This discusses how to handle them.

MAExplorer Stack size Memory Error on Sun Solaris

Running MAExplorer on a Solaris (or other Unix system) may produce this error:

```
% MAExplorer

Stack size of 97664 Kb exceeds current limit of 8192 Kb.
(Stack sizes are rounded up to a multiple of the system page size.)
See limit(1) to increase the stack size limit.
```

If the Sun (under Solaris) is slow in loading MAExplorer or has memory errors (shown above) one should first see what the memory limits are set to on your machine using the "limit" command. If they are too small they should be increased or set to "unlimited" (see in 2.4 above

MAExplorer LAX file

If the problems persist, one might have to edit the MAExplorer.lax file found in the MAExplorer directory (see example below). The default memory settings in the MAExplorer.lax file (found in the installation directory) should be no larger than the total memory of the machine or paging problems will occur. For instance, if you have 192Mb of memory in your Sun, edit the **"lax.nl.java.option.native.stack.size.max"** and **"lax.nl.java.option.java.heap.size.max"** options to be under 192Mb. You can use any text editor to do this. More memory may be needed to be installed on your Sun to run MAExplorer with very large datasets.

Default Lax settings
The Lax file is a startup file generated by InstallAnywhere when we packaged MAExplorer. It is used when MAExplorer starts up on your computer. We currently set the memory limits to 256Mbytes. If you have more memory, you can edit the Lax file to have it use more memory.

```
#    LAX.NL.JAVA.OPTION.JAVA.HEAP.SIZE.MAX
#    -------------------------------------
lax.nl.java.option.java.heap.size.max=256000000
```

```
#   LAX.NL.JAVA.OPTION.NATIVE.STACK.SIZE.MAX
#   ----------------------------------------
lax.nl.java.option.native.stack.size.max=256000000
```

The MicroArray Explorer MAEPlugins Home page

We have added a Java plugin extension facility to MAExplorer. These plugins are called "MAEPlugins". MAEPlugins allow investigators to extend the core capabilities of MAExplorer program by writing special programs to implement new analysis methods and access data from a MAExplorer database. The design of this extension enables users to write these new methods and have them added to the MAExplorer menus or be invoked when MAExplorer starts up. In addition, default MAExplorer functionality may be changed by replacing existing MAExplorer methods with user defined methods.

> The Open Java API is available for beta-testing. Not all of the functionality in the Open Java API has been implemented or thoroughly tested.
>
> • **NEW** Keep checking this Web page for the current status of the API as well as the MAExplorer Revision Notes which gives a history of new features and changes to both MAExplorer and the API.
>
> MAExplorer is open source with a Mozilla 1.1 general public license. However, we have made the MAEPlugins *public domain* (a secondary license that is even less restrictive) with no restrictions on their use. This will enable the research community to modify and help improve MAExplorer and the MAEPlugins as required. We are dividing the plugins into those that are donated and those that require interaction with the supplier. We hope that most plugin developers will make them available as open source, but that is not a requirement. If you are interested in writing a plugin or working with us on this **open source project** please contact us.

1. The MAEPlugins home page

This Web page includes documentation, an Open Java API (Application Programming Interface) with javadoc documentation, open source Java source code and jar file examples, and lists of MAEPlugins. It will contain donated MAEPlugins and links to MAEPlugins offered at other Web sites. Typical MAEPlugins could include: normalization, distance metrics, data Filters, PCA or other visualization tools, graphic plots, clustering, classifiers, array sample I/O data conversion, client-server, Web-server functional analysis of cluster results, etc.

1.1 MAEPlugins are grouped into three types of implementations

These allow various degrees of portability and server independence.

1. Using 100% Java code to implement the complete plugin. This means it will be portable across all systems.
2. Accessing local programs written in any language (eg. the R-statistics package). This method may not be portable across all systems.
3. Web-CGI or client-server to specialized genomic database plugins. This method should be portable across all systems.

2. Open Java API (Applications Programming Interface)

The MAExplorer Open Java API (Applications Programming Interface) allows users to access all data structures without having to understand the low level internal details of the MAExplorer system.

The Open Java API is included in the regular .jar file distributed when you download MAExplorer. The current MAExplorer JAR file containing the Open Java API may be downloaded from MAExplorer.jar. You also will also download the jar file when installing MAExplorer.

Note: the Open Java API is undergoing alpha-level testing. During this process some of the methods may change and move between MJAxxxx classes as we optimize the API. New methods are also being added to further reflect the underlying MAExplorer capabilities and data structures.

3. Distribution of MAExplorer Plugins

The distribution system for MAEPlugins is very flexible. There are several options for distributing Plugins on this Web site including:

1. MAEPlugins with Java source files, ready-to-run Jar files, and documentation. These are developed by various groups and contributed to the Open Source Web site. You may maintain these yourself or contract the authors or contact the Open Source development group for MAExplorer.
2. Links to other Web sites where you may obtain the Jar files and (or) Java source and documentation for MAEPlugins from that provider. No MAExplugins will be kept on the this server unless the source code is included.

4. Lists of MAEPlugins being made available

- List of MAEPlugins sorted alphabetically
- List of MAEPlugins by analysis method
- List of MAEPlugins by links to other Web sites to contact for the plugins

5. How to write a MAEPlugin

- The Java Plugin paradigm
- Open Java API documentation
- How to write a MAEPlugin
- Java MAEPlugin code examples

Design of MAEPlugins

This document discusses the paradigm how MAEPlugins are used with MAExplorer and the design used to give them access to MAExplorer data. The first part discuss the top level design and the second part gives an example of using a plugin. The details on the internals for MAExplorer itself are described in a Design doc (PDF) or (PPT). However, an understanding of the MAExplorer internals is not required to write a MAEPlugin.

1. Overview of MAEPlugin design

The MAExplorer Open Java API (Applications Programming Interface) allows users to access almost all data structures without understanding the details of the system. Specialized interfacing classes (MJAxxxx), organized by function, are accessed from the MaeJavaAPI class. The MJAxxxx classes map internal data to user data in a protected manner. Users do not have direct access to internal MAExplorer data structures. However, MAEPlugins do have access to relevant data. This allows us to improve and change the internal data structures without causing problems with plugins using those data structures. The following figures show the top level plugin design.
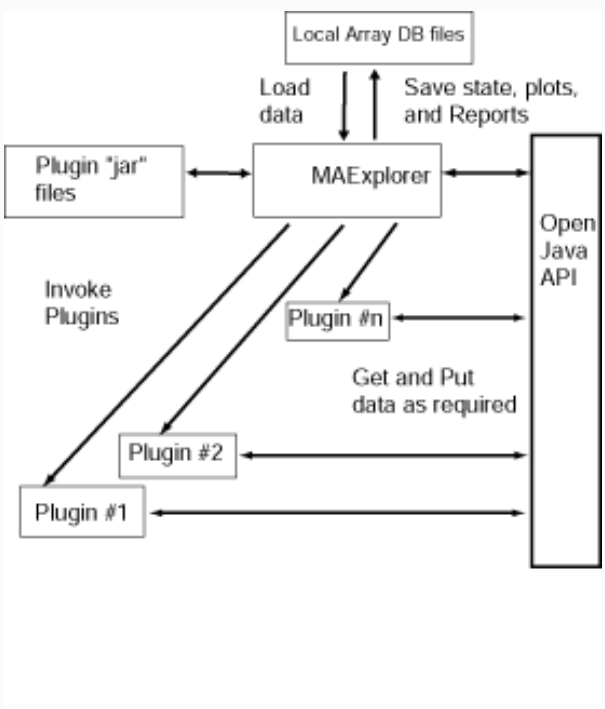
Figure 1. Overall MAEPlugin design for MAExplorer. Plugins are dynamically loaded into MAExplorer where they may be invoked from a menu entry or by various other means such as startup, normalization, data filtering, etc. Any number of plugins may be loaded simultaneously. They may be loaded and unloaded dynamically, and saved for automatic loading when the current database is saved.
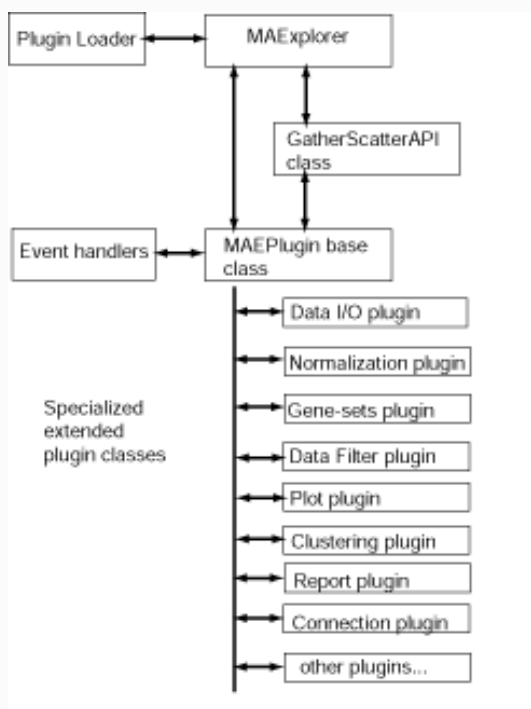
Figure 2. Open Java API for MAEPlugins. Each type of application could be derived from specialized Java classes that contain most of the access methods required for that type of analysis. The Gather - Scatter API is a means of "gathering" data from MAExplorer internal data structures for the plugin. When a plugin wants to store data back into MAExplorer, it is "scattered" back into the internal data structures. This is implemented using the MaeJavaAPI and MJAxxx classes described in the Open Java API.

2. Example of using a Plugin

This shows a short demonstration of what is involved in using a MAEPlugin. The user first load the plugin from the disk. Generally the plugins .jar or .class files are stored in the Plugins/ directory where you have installed MAExplorer. Then they load a particular plugin which installs it in the Plugins pull-down menu. Then they revisit that menu to invoke the particular plugin. You may load any number of plugins (until you run out of computer memory if that should occur).
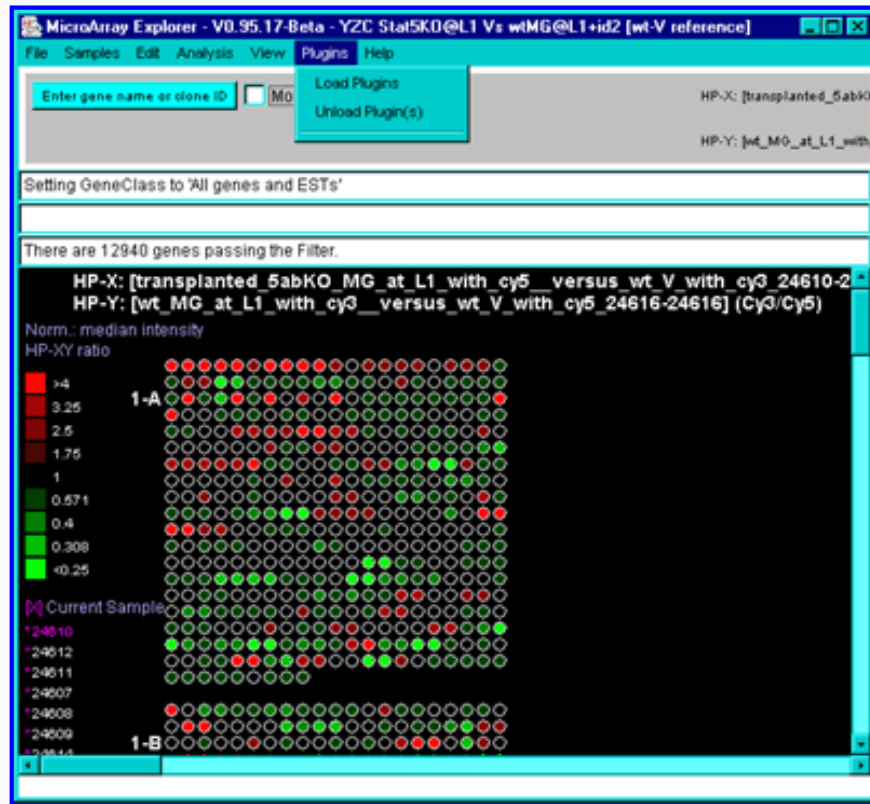
Figure 3. Loading a MAEPlugin from your file system using the Load Plugins command in the Plugins pull down menu. If you have a plugin .jar or .class file, it may be specified using the "Load plugin" command. This pops up a file browser to let you specify the plugin file.
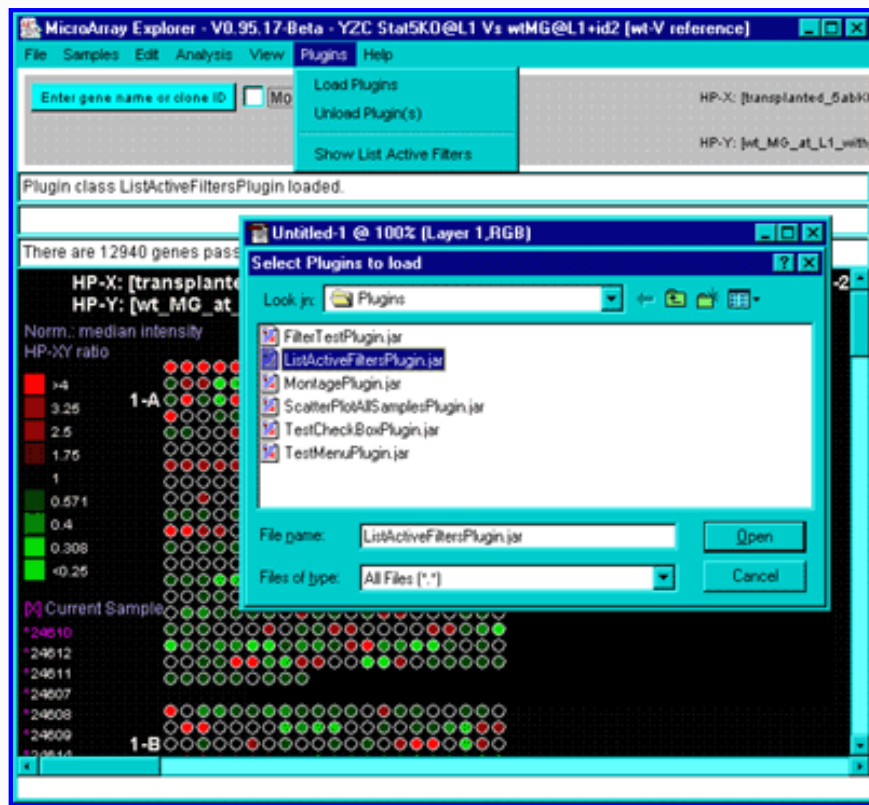
Figure 4. Executing the new command previously loaded in the Plugin menu. Selecting the new "Show List Active Filters" command that now appears in the Plugins menu invokes the plugin. This pops up a report shown in the next figure.
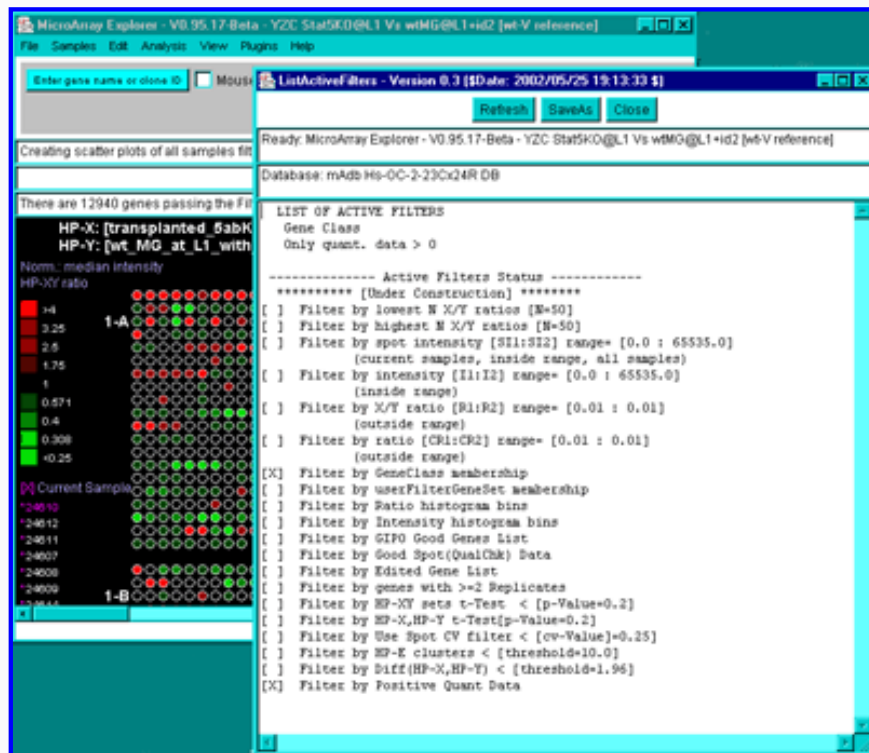
Figure D.5. Popup window from executing the MAEPlugin. This plugin gives a full report on the data Filter status in a new pop up window.

MAEPlugins Open Java API

MAEPlugin home | Design | **Open Java API** | MJA classes | MJA javadocs | Open Java API javadocs |
Plugin Tutorial Examples | List of Plugins | Developing a Plugin | Installing Plugins |
MAExplorer home | MAExplorer revision notes | Help desk

This document describes the alpha-level MAEPlugins Open Java API (Applications Programming Interface) to enable researchers to write their own MAEPlugins for use with MAExplorer. The Open Java API (or API) is presented here as two javadoc trees.

The Open Java API is automatically included in the MAExplorer.jar file. Although it wastes some space, we are exporting the symbol tables with the files in MAExplorer.jar so that you could use it with a debugger (such as Forte for Java CD) to develop a MAEPlugin.

Notes:

1. The Open Java API is undergoing alpha-level testing. During this process some of the methods will not work, may change and may move between MJAxxxx classes as we optimize the API. New methods are also being added to further reflect the underlying MAExplorer capabilities and data structures.

2. The NormalizationPlugin and ConnectionPlugin are being debugged.

3. **If there is functionality need but can't find in the Open Java API, make a suggestion or better still - help develop the API code.**

- docsOJAPI/ is the entire Open Java API javadoc tree
- docsMJA/ is the data access MaeJavaAPI MJAxxx classes subset javadoc tree

The first, docsOJAPI/, is the entire API accessible to the plugin writer including the MAEPlugin classes required to extend your plugins. However, many of the MAEPlugin.xxx methods are not normally called explicitly by the plugin writer. Instead, a subset of classes, docsMJA/ constituting the MaeJavaAPI set of classes, is the library of access methods that the plugin writer normally uses.

1. List of MaeJavaAPI (MJA) classes

The MJA classes are organized by function. For example, if you want to access data and methods on samples, then go to the MJAsample or MJAsampleList classes. See the javadocs for the Open Java API for details. The detailed descriptions of these classes are available in the docsMJA javadocs.

```
MJAxxxx Class        Objects and method access
 ------------         ------------------------
 MJAbase             base class and constants used by other MJA classes
 MJAcluster          cluster data structures and methodst
 MJAcondition        condition lists of samples and ordered lists of condition lists
 MJAeval             command interpreter to invoke MAExplorer commands
 MJAexprProfile      expression profiles data
 MJAfilter           data filters
 MJAgene             single gene data
 MJAgeneList         lists of genes and get sets
 MJAgenomicDB        genomic databases on the Internet
 MJAgeometry         array geometry, spot to gene maps, etc.
```

[MJAhelp](#)	popup browser help methods
[MJAhistogram](#)	histogram plots
[MJAmath](#)	built-in math functions
[MJAnormalziation](#)	normalization data and methods
[MJAplot](#)	scrollable 2D plot support [Future]
[MJAproperty](#)	get and put individual properties
[MJApropList](#)	get lists of properties
[MJAsample](#)	get and put single sample top-level data
[MJAsampleList](#)	get lists of samples top-level data
[MJAscrollablePlot](#)	scrollable 2D plot support [Future]
[MJAsort](#)	built-in sort methods
[MJAstatistics](#)	built-in statistics methods
[MJAstate](#)	get and save state, get additional state info
[MJAutil](#)	built-in utility methods

How To Write a MAEPlugin using the Open java API

This document describes how to write MAEPlugins using the Open Java API (Application Programming Interface). It discusses key issues to be addressed when writing a MAEPlugin and describes in sufficient detail to enable researchers to write their own MAEPlugins for use with MAExplorer.

1. Using a Java development environment to develop and debug a Plugin

We have designed the MAExplorer.jar file so that it contains both MAExplorer and the Open Java API. All MAExplorer classes are compiled with the symbol table so that it may be used in a debugger. We use the Sun Forte for Java (Community Edition) which is a free development environment (IDE) available over the Internet. Forte (and most other IDEs) allows you "mount" a jar file. So to create a new plugin you would:

1. Create a new project (e.g. see [ExamplePlugin](#) source code example and discussion).
2. Copy some example MAExplugin code into this project and rename the modules and classes to correspond to the names of your new plugin.
3. Mount the MAExplorer.jar file (you can mount it from the directory where you installed MAExplorer (eg. typically C:/Program Files/MAExplorer/MAExplorer.jar for Windows, etc.)
4. After you have compiled your plugin and want to test it, you create a new Jar file with the name of the plugin (e.g. ExamplePlugin.jar). Forte lets you do with with a jar packager.
5. To test it you first open the MAExplorer.jar file tree you have mounted. Then you select "MAExplorer". At this point you can execute MAExplorer or run the debugger.
6. After MAExplorer is running, you select "Load Plugin" from the MAExplorer Plugins pull-down menu, and then enter the name of the plugin (e.g. ExamplePlugin.jar).
7. At this point you may run the plugin by going back to the Plugins menu and select the entry corresponding to your plugin.
8. If you want to make a change in your plugin and try again, you do not need to restart MAExplorer. Instead first unload your plugin using the "Unload Plugin" command in the Plugins menu. Then rebuild your plugin and try again.

2. Installing your MAEPlugin in a working MAExplorer environment

1. After you are happy (or somewhat happy) with your plugin, copy the plugin jar file (e.g. ExamplePlugin.jar) to the installation Plugins/ directory where you can access on any of your data (e.g. C:/Program Files/MAExplorer/Plugins for Windows, etc.)
2. To use the plugin on any database, just start MAExplorer and then load and run the plugin as above.

Tutorial Examples of MAEPlugins

This document gives a simple tutorial example of MAEPlugins source code.

1. Example of a simple PopupPlugin and how it uses the Open Java API

The following code illustrated how to create a simple popup plugin ExamplePlugin.java using the Open Java API by extending the PopupPlugin. It passes the MaeJavaAPI mja instance to the actual workhorse, Example.java, that then retrieves and saves any MAExplorer data it requires. We show very simple examples of this code to give the flavor of the procedures required and how it interfaces with the API.

1.1 Example of plugin class that is loaded into MAExplorer

By convention, we will name the class that is loaded into MAExplorer XxxxxPlugin.java and the subsequent primary body of the plugin class Xxxxx.java where Xxxxx is some particular class. In our following example, Xxxxx is "Example", but it might be "MyNewClusterMethod" etc. We first show ExamplePlugin.java that serves as the interface between MAExplorer and the primary body class Example.java.

1.1.1. You must import the two class definitions:

```
import MAEPlugin.popup.PopupPlugin;
import MAEPlugin.*;
```

If you are writting other types of plugins, you need to import those instead (eg. MAEPlugin.analysis.NormalizationPlugin, MAEPlugin.analysis.FilterPlugin, etc).

1.1.2 The XxxxxPlugin.java class must have the following methods as a minimum:

1. XxxxxPlugin() - is the constructor for the class (here it is ExamplePlugin).
2. pluginMain() - the method end-users must implement to use the API.
3. updateCurGene() - update any data since current gene has changed.
4. updateFilter() - update any dependent data since Filter has changed.
5. updateSlider() - update any dependent data since a threshold slider may have changed.
6. updateLabels() - update any dependent data since global labels may have changed.

7. close() - close the plugin

The XxxxxPlugin() method is called at the time the plugin is loaded. Any particular actions that may be required can be performed at that time. In this example, we merely set the name of the plugin as it is to appear in the Plugins pull-down menu.

The pluginMain() method is called at the time the plugin is invoked by selecting the menu entry.

The four special event handling methods updateCurGene(), updateFilter(), updateSlider(), and updateLabels() are invoked by the MAExplorer PopupRegistry when any of these events occurs. If you are doing nothing with the events, they may be no-ops. However, if you want to take action on these events, you would normally implement the actual event handling code in your Xxxxx.java class.

```java
/** File: ExamplePlugin.java */

import MAEPlugin.popup.PopupPlugin;
import MAEPlugin.*;

/**
 * This class invokes the ExamplePlugin plugin.
 */

public class ExamplePlugin extends PopupPlugin implements MAEUpdateListener
  {
     /** The current instance of a plugin called "Example".
      * The instance may be non-null if run previously and is needed to kill
      * a previous instance when new instances are created.
      */
     private Example
       eObj= null;

     /**
      * ExamplePlugin() - this is the constructor end-users must implement
      * to use the API. It is called at the time the plugin is loaded.
      */
     public ExamplePlugin() throws PluginException
       { /* ExamplePlugin */
         /* Note: "Example plugin" is a string that appears in the
          * Plugin menu.
          */
         setMenuLabel("Example plugin");

         MJApopupRegistry
           pr= MAExplorer.mja.mjaPopupRegistry;
         int
           propBits= (pr.PRPROP_CUR_GENE | pr.PRPROP_FILTER | pr.PRPROP_LABEL |
                      pr.PRPROP_SLIDER | pr.PRPROP_UNIQUE);
         pr.addUniquePopupWindowToReg(this, "ShowListActiveFilters", propBits);
       } /* ExamplePlugin */


     /** pluginMain() - the method end-users must implement to use the API.
      * It is invoked when the user selects the plugin in a menu.
      */
     public void pluginMain()
       { /* pluginMain */
```

```
        MaeJavaAPI
          mja= MAExplorer.mja;     /* Open Java API library access */

        if(eObj==null)
           eObj= new Example(mja);
        else
          { /* re-rerun Example on new data */
            eObj.dispose();
            eObj= null;
            System.gc();
            mja.mjaUtil.maeRepaint();
            eObj= new Example(mja);
          }
     } /* pluginMain */

/** updateCurGene() - update any data since current gene has changed.
  * This is invoked by the MAExplorer PopupRegistry.
  * @param mid is the MID (Master Gene ID) that is the new current gene.
  */
public void updateCurGene(int mid)
   {
     if(eObj!=null)
       eObj.updateCurGene(mid;
   }

/** updateFilter() - update any dependent data since the data Filter
  * has changed. This is invoked by the MAExplorer PopupRegistry.
  */
public void updateFilter()
   {
     if(eObj!=null)
       eObj.updateFilter();
   }

/** updateSlider() - update any dependent data since a threshold slider
  * has changed. This is invoked by the MAExplorer PopupRegistry.
  */
public void updateSlider()
   {
     if(eObj!=null)
       eObj.updateSlider();
   }

/** updateLabels() - update any dependent data since global labels
  * have changed. This is invoked by the MAExplorer PopupRegistry.
  */
public void updateLabels()
   {
     if(eObj!=null)
       eObj.updateLabels();
   }

/**
  * close() - close the plugin. This will be called if you
  * had specified the plugin as PRPROP_UNIQUE since previous
```

```
    * instances will be closed before the new instance is started.
    * @param preserveDataStructuresFlag to save data structures
    */
  public void close(boolean preserveDataStructuresFlag)
    {
      if(eObj!=null)
        eObj.close();
    }

} /* end of class ExamplePlugin*/
```

1.2 Example of the main body of plugin code

The main body of code the plugin writer generates is illustrated here showing how one might access data and methods from the Open Java API. We illustrate this with a very simple example, Example.java, showing the entry point a retrieving a few data structures from the Open Java API. In this example, we will popup a new Frame and add Action and Window listeners (code not shown to support the Frame since that is not the point of this example). However, any Java code could be used.

```
/** File: Example.java */

public class ListActiveFilters extends Frame
        implements ActionListener, WindowListener, etc.
{

  /** Example() - Constructor
   */
  public Example(MaeJavaAPI mja)
    { /* Example */
      /* [1] Access Open Java API required through MaeJavaAPI instances
       * of these MJA classes.
       */
      MJAfilter
        mjaFilter= mja.mjaFilter;          /* Open Java API library */
      MJAgeneList
        mjaGeneList= mja.mjaGeneList;       /* Open Java API library */
      MJAproperty
        mjaProperty= mja.mjaProperty;       /* Open Java API library */
      MJAsampleList
        mjaSampleList= mja.mjaSampleList;   /* Open Java API library */

      /* [2] Get the data  */
      String
        sR= "Example of some data accessed from MAExplorer\n",
        maePrjPath= mjaProperty.getMaeCurProjectPath(),
        maeBrowserTitle= mjaProperty.getMaeBrowserTitle(),
        maeDatabase= mjaProperty.getMaeDatabaseTitle(),
        maeDbSubset= mjaProperty.getMaeDbSubsetTitle();
      String
        sActive[]= mjaFilter.getListFilterNames();
      int
        nActive= sActive.length;
      sR += "  LIST OF ACTIVE FILTERS\n";
      for(int i=0;i<nActive;i++)
```

```
      if(sActive[i]!=null)
        sR += "    " + sActive[i] + "\n";

    int
      nSamples= mjaSampleList.getNbrHPsamples();
    String
      sampleNames[]= mjaSampleList.getHP_Elist_SampleNames();
    sR += "  LIST OF SAMPLES\n";
    for(int i=0;i<nSamples;i++)
      sR += sampleName[i] +"\n";

    int
      filteredMIDlist[]= mjaGeneList.getMIDindicesForFilterGeneList(),
      nFilteredGenes= filteredMIDlist.length;
    String
      filteredGeneNames[]=
          mjaGeneList.getGeneFieldDataFromGeneList("workingCL", "GeneName");
    sR += " LIST OF FILTERED GENES\n";
    for(int i=0;i<nSamples;i++)
      sR += "Gene ["+filteredMIDlist[i]+"] = "+filteredGeneNames[i]+"\n";

    System.out.println(sR);          /* print to java console */
  } /* Example */

 /* In this example, no actions are taken on popup registry events.
  * However, the methods must exist in the code.
  */
 public void updateCurGene(int mid) { }
 public void updateFilter() { }
 public void updateSlider() { }
 public void updateLabels() { }
 public void close() {this.destroy(); }

} /* end of class Example.java */
```

List of All MAEPlugins By Origin and Analysis Method

This document lists all of the MAEPlugins alphabetically, by analysis method, and also links to MAEPlugins available on other Web sites. The MAEPlugins include those donated to the MAExplorer Open source Web site. All plugins distributed from this Web site will have the Java source code, JAR file and documentation.

If you want to use the jar file plugins directly: (1) install MAExplorer from the list of Jar files on this Web site, (2) get the jar file(s) from the plugins below and save them in the Plugins/ directory where you installed MAExplorer, (3) run MAExplorer and use the (Plugins | Load Plugin) menu command to load the plugin. After it is loaded, just use it as you would any other menu command. The Plugins-jar.tar file is available with all of the MAEPlugin jar files. Simply unpack the directory using Unix tar or a Windows unzip program into a directory you can access when running MAExplorer. To let MAExplorer go directly to these files when you do a (Plugins | Load plugins) menu command, copy the .jar files into the Plugins/ directory where you previously installed MAExplorer.

If you want to use these plugins as a basis for developing your own plugins, see **developing a plugin** and other resources available on this Web site. The source code for each plugin is available beow. We encourage, but don't require, plugin writers to donate their new plugin analytic methods to the MAExplorer Open Source Web site for others to use.

1. List of all MAEPlugins sorted alphabetically

1.1 Alpha-level MAEPlugins

(not fully developed)

1. **FilterTestPlugin** is an example Filter plugin to halve the number of genes accepted by the other filters. Available as source code and as jar file.
2. **ScatterPlotAllSamplesPlugin** generates a scatter plot of filtered genes for all samples (very primitive at this point). source code and as jar file.

1.2 Stable MAEPlugins

1. **ListActiveFiltersPlugin** lists the state of the data filter options, modes and threshold values. source code and as jar file.
2. **MontagePlugin** shows a montage composite image of subregions around current gene "cut" out of the original images (if the data available to MAExplorer supports it. The NCI/CIT mAdb exported data does support this feature). source code and as jar file.
3. **TestCheckBoxPlugin** is a very simple plugin to toggle a checkbox in the menu. It is meant to be used as a simple example of a checkbox plugin. source code and as jar file.
4. **TestMenuPlugin** is a very simple plugin to be invoked when selected from the menu. It is meant to be used as a simple example of a menu plugin. source code and as jar file.

2. List of MAEPlugins by analysis method

2.1 Connection to servers

2.2 Clustering methods

2.3 Data filtering methods

1. **ListActiveFiltersPlugin** lists the state of the data filter options, modes and threshold values. source code and as jar file.

2.4 Normalization methods

2.5 Plot methods

1. **ScatterPlotAllSamplesPlugin** generates a scatter plot of filtered genes for all samples (very primitive at this point). source code and as jar file.

2.6 Report methods

2.7 Visualization methods

1. **MontagePlugin** shows a montage composite image of subregions around current gene "cut" out of the original images (if the data available to MAExplorer supports it. The NCI/CIT mAdb exported data does support this feature). [source code](#) and as [jar file](#).

2.8 Other methods

1. **TestCheckBoxPlugin** is a very simple plugin to toggle a checkbox in the menu. It is meant to be used as a simple example of a checkbox plugin. [source code](#) and as [jar file](#).
2. **TestMenuPlugin** is a very simple plugin to be invoked when selected from the menu. It is meant to be used as a simple example of a menu plugin. [source code](#) and as [jar file](#).

3. List of MAEPlugins by links to other Web sites

This list contains links to other Web sites where you may obtain the Jar files and (or) Java source and documentation for MAEPlugins from that provider. No MAExplugins will be kept on the this server unless the source code is included.

Cvt2Mae Data Converter

[Install Cvt2Mae on your computer]

Cvt2Mae Home | [Description](#) | [Download](#) | [FAQ](#) | [Appendix](#) | [Example](#) | [Revisions](#) | [MAExplorer home](#) | [Help Desk](#)

Cvt2Mae Basics

In order to use the [MAExplorer](#) data-mining tool on your cDNA or oligo tab-delimited array data, you must convert your data files into the data formats described in [Appendix C](#) and [Appendix D](#) of the [MAExplorer reference manual](#). Although this maybe done by editing user's data files by hand into the required formats, it is a non-trivial process. Therefore we have developed a "wizard" conversion tool called Cvt2Mae to automate these conversions.

Cvt2Mae is a Java program designed to make it easier for use by researchers to use MAExplorer by helping them convert their data into the MAExplorer format. Cvt2Mae handles commercial chips such as Affymetrix, as well as other standard formats such as GenePix and Scanalyze or one-of-a-kind custom academic chips (<User-defined>). In addition, you may specify the fields of interest for the "Print file" or (GIPO or Gene-In-Plate-Order) file, and the fields containing the quantified data.

The Cvt2Mae converts specific chip information you entered into what we call an "Array Layout". This Array Layout file may be edited and saved for use in future conversions and shared with collaborators. Essentially, the Array Layout contains a set of "rules" for converting the user's data. After you have filled out the forms in Cvt2Mae, it will generate the set of converted data files and directories to be used directly with MAExplorer.

Cvt2Mae Data Conversion Steps and Tutorials

There is a detailed description on using the Cvt2Mae converter that provides the level of detail you need to use it effectively. In addition, a step by step example is provided converting Affymetrix data.

There are several slide shows describing how to use the Cvt2Mae to convert various data sets. They consist of a series of screen shots from Cvt2Mae that go through each of the steps on how to set up the parameters and convert your data. There are two for Affymetrix data, one is a downloadable PDF and other an extensive online version.

Tutorials

1. Affymetrix Data with a full description (HTML)
2. Affymetrix Data (PDF) or (PPT)
3. GenePix Data (PDF) or (PPT)
4. Scanalyze Data (PDF) or (PPT)
5. <User-defined> Data (PDF) or (PPT)
6. Incyte Data PDF

Downloading latest Cvt2Mae Version

You may freely download and install the current release of the Cvt2Mae stand-alone application. You are free to use or redistribute Cvt2Mae. You may want to review the revision history.

[download] Download and Install Cvt2Mae.

Instructions on downloading and installing Cvt2Mae.

Frequently Asked Questions (FAQ)

Here are some questions you might have about the Cvt2Mae data converter.

FAQ

If Additional Help is Needed

Before emailing us for help, please read these Cvt2Mae Web pages to ensure that you have set the parameters correctly and have the raw data in the correct format. You might also read the Appendix C of the MAExplorer manual. MAExplorer and Cvt2Mae also create log files that might be of helpful in troubleshooting.

If you then are still having problems email the help desk. Please include:

- A detailed description of the problem including error messages
- Information about the computer system (memory, operating system etc.)
- What array chip you are using, describe in detail if it is not one of the default Array Layouts.

Cvt2Mae Data Conversion Steps Description

The Cvt2Mae "Wizard"

You should use the Cvt2Mae program to convert your data to the MAExplorer format. Cvt2Mae has a multi-step process (wizard) that allows you to create an Array Layout that describes your data. One could edit the raw data files by hand but this is tedious and prone to errors. NOTE: The step titles below are links to in depth descriptions.

Step 1 Select Array Chip (Array Layout)

Cvt2Mae has several predefined Array Layouts (Affymetrix, GenePix, Scanalyze and others) available from a pull down menu. One can also create, edit and save their own custom Array Layouts using the <user-define> Array Layout.

Step 2 Select input file(s)

Some arrays may have several data files, such as a separate GIPO file or separate spot quantification files. Some files have multiple samples within each file. You can also pick multiple data files to convert. Cvt2Mae has the flexibility to handle complex data files.

Step 3 Edit Array Layout

For <user-defined> arrays, you must first define some parameters such as array geometry, which row contains the fields, if it is Cy3/Cy5 data or intensity data etc. before converting data. Also, if you are using one of the predefined Array Layouts and your data is slightly different you will have to edit the Array Layout to correct these differences. These parameters are described in detail in Appendix A. The Array Layout can then be saved and used many times for other data files with the same data format.

Step 4 Choose output folder/directory

This is the project directory where the converted data files will be saved and used with MAExplorer.

Step 5 Convert Data

The last step is to click the "Run" button which starts the data conversion. Several folders are created in the project directory to hold the converted data. Once this is completed, the "Run" button will turn into a "Done" button which you press to exit the program. You can now go the newly created MAE sub directory and click on the Start.mae file (assuming you have installed MAExplorer) to start MAExplorer on your converted data.

Status Window and Help

There are 3 message areas at the bottom of the Cvt2Mae window that are used for reporting error and status messages. If certain parameters are not consistent, error messages will appear in the message area along with suggestions on how to correct the problem.

The Edit Layout wizard also has its own information area that is used for reporting. When you hold the mouse over the a field on the left side of the wizard window, information about that parameter will appear in the lower message area.

Example of Using Cvt2Mae to convert Some Affymetrix data for MAExplorer

This detailed example shows how one might convert Affymetrix data for use with MAExplorer. The example is presented as a series of computer screen shots. Similar screen shots are available as PDF documents for other types of chip array layouts. The example is divided into three parts: specifying the input data files, editing the array layout, and generating the output data files for use with MAExplorer.

1. Specifying the input data files

Figure 1. shows the Affymetrix tab-delimited data in Excel. Figure 2. Initial state of the Cvt2Mae Program. Figure 3. Selecting a Chipset Array Layout. Figure 4. Selecting one or more user input data files by pressing the "Browse input file name" button. Then select a user input data file using the file browser. Figure 5. Files selected by user and samples "discovered" in the data file.

2. Editing the array layout

Figure 6. Edit Layout Wizard for name of the Array Layout with **A)** original and **B)** the new layout name. Figure 7. Edit Layout Wizard for Grid Geometry. Figure 8. Edit Layout Wizard for Starting Data Rows. Figure 9. Edit Layout Wizard for Ratio or Intensity data. Figure 10. Edit Layout Wizard for optional (X,Y) spot coordinates available in the input data. Figure 11. Edit Layout Wizard for optional Genomic ID values available in the input data. Figure 12. Edit Layout Wizard for optional Gene Names available in the data. Figure 13. Edit Layout Wizard for optional calibration DNA available in the data and UniGene species prefix. Figure 14. Edit Layout Wizard for optional user names for Project, Database, Sub-database, etc. Figure 15. Edit Layout Wizard for optional HP-X and HP-Y 'set' experimental class (i.e. condition) names. Figure 16. Edit Layout Wizard for changing the default data filter threshold slider values.

2.1 Specifying the mapping between your data file fields and those required by MAExplorer

There are two special wizards for specifying the mapping array layout GIPO and Quant input data field names. These mappings allow the converter to take your data specified in some columns (i.e. Fields) of your data input file and use it to generate standard MAExplorer output files. Figure 17. shows the Edit Layout Wizard for "Assign GIPO fields" used to generate the MAExplorer GIPO data file. Figure 18. shows the Edit Layout Wizard for "Assign Quant fields" used to generate the MAExplorer Quant files (one for each hybridized sample). Figure 19. shows saving modified Array Layout if you have made changes.

3. Generating the output data files for use with MAExplorer

Finally, the array layout has been defined and we can run the converter. Figure 20. Selecting the output folder in which to save the converted files. Figure 21. Browse to select the output folder in which to save the converted files. Figure 22. shows the interface after selection of the output file folder using a file browser. Figure 23. shows the conversion being performed after the user pressed the RUN button. Figure 24. shows the conversion summary instructions after the conversion is finished. Figure 25. shows the files that are generated by Cvt2Mae for use by MAExplorer. Figure 26. Starting MAExplorer on the converted data by clicking on Start.mae file. Note that the location of the "MAExplorer startup file:" in Figure 8. Go to that file and click on it to start MAExplorer. Alternatively, start MAExplorer and do "File | Open Disk DB" and open that file to start it up.

Figure 1. shows the Affymetrix tab-delimited data in Excel. (after missing fields have been edited as described above).

Figure 2. Initial state of the Cvt2Mae Program. The user must select an array layout or define one in order to analyze the input data file or files.
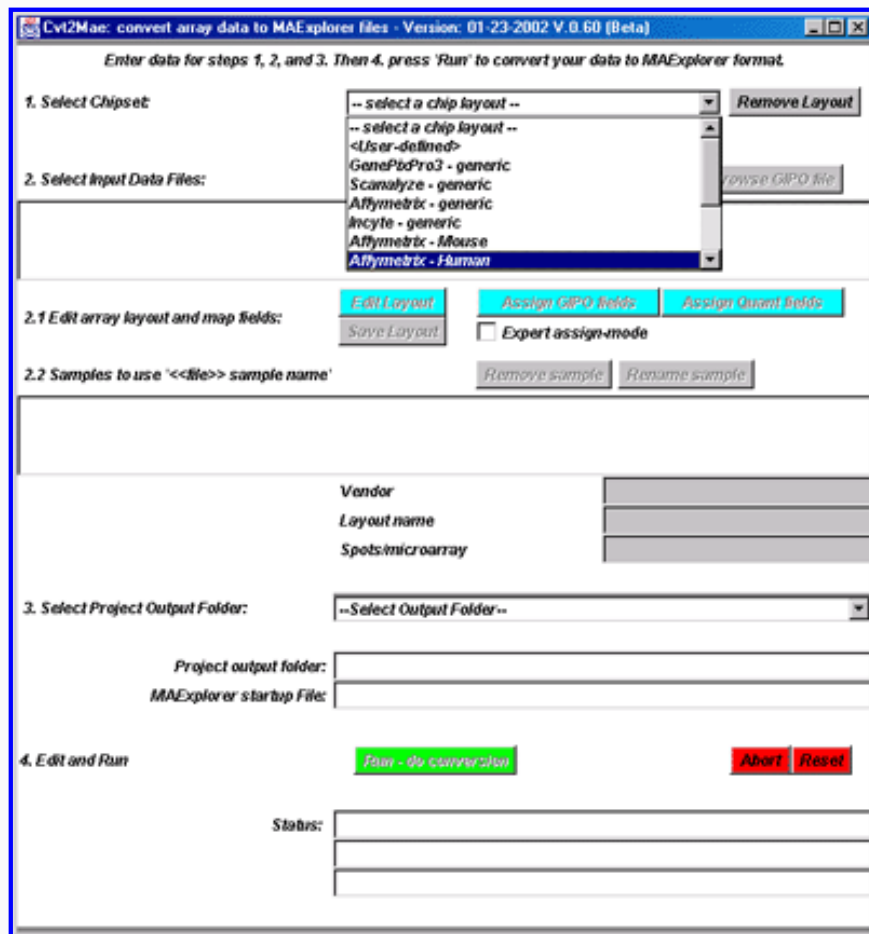
Figure 3. Selecting a Chipset Array Layout. The built-in array layouts are shown for the Incyte and Affymetrix. User-defined layouts would be added by selecting the <User-defined> layout.

Figure 4. Select one or more user input data files by pressing the "Browse input file name" button and then pick a file. If the layout indicates that it may contain more than one hybridization, it will attempt to find the data. You can subsequently rename individual samples which may be necessary if you are reading several files with the same sub-sample names. After the file browser pops up, select a user input data file. If you are using a file that contains all of your samples, then you only need to specify one file. If you have several files, then repeat this step until you have added all of the files you want.
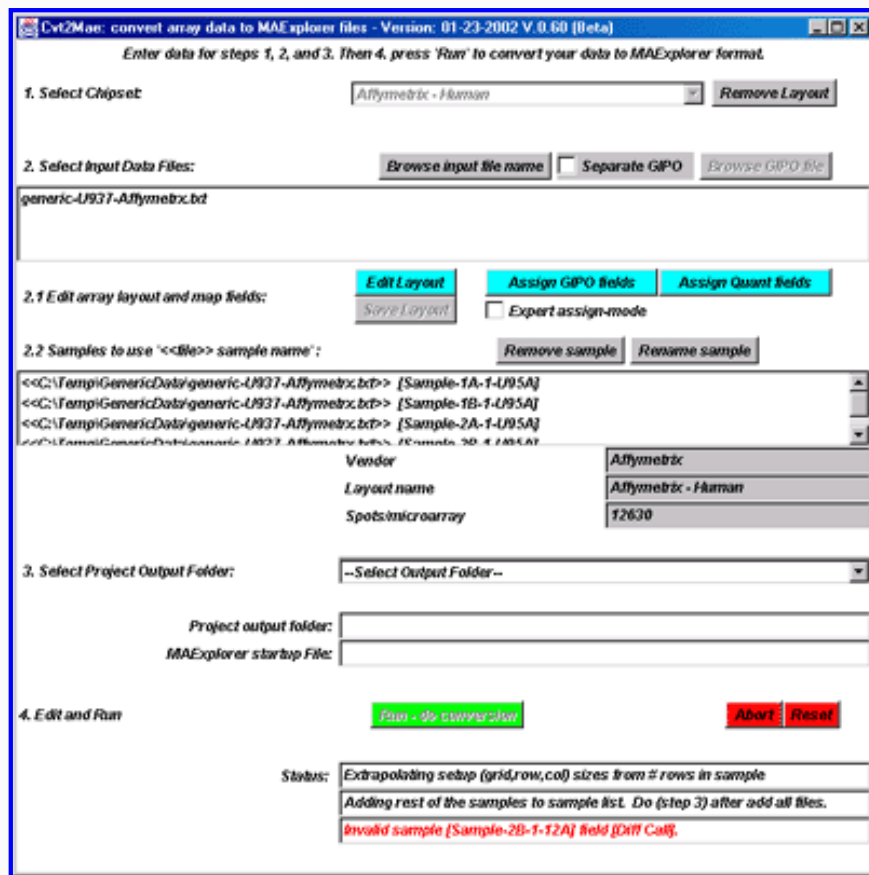
Figure 5. Files selected by user and samples "discovered" in the data file. Each input file is analyzed to determin if it has] multiple samples and if so they are added to the list of input files at below step 2.1 in the window. You may remove any samples which may be necessary for bad data. You may rename any sample which may be necessary if you have the same sample name occuring in several different data files (they are actually different samples).
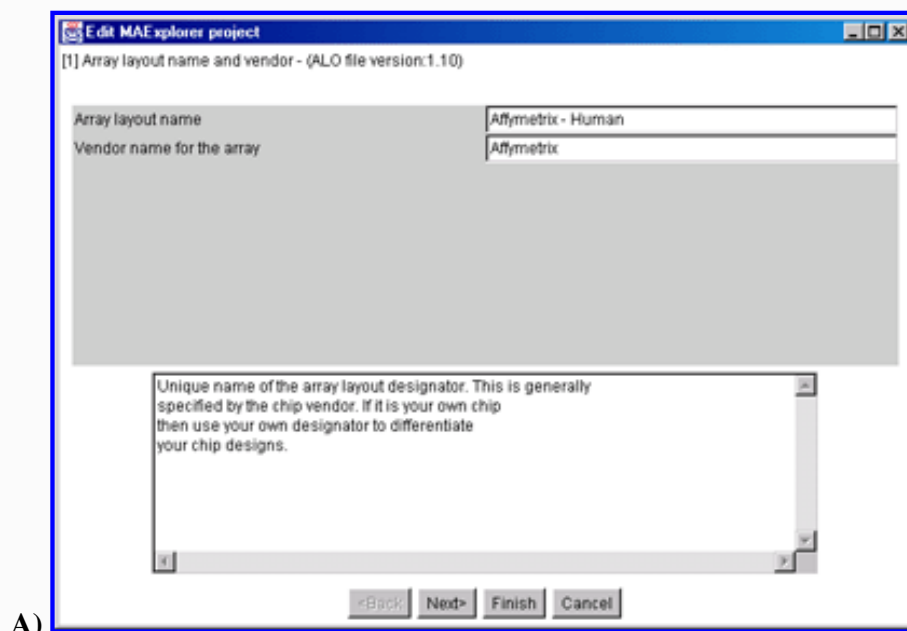
A)
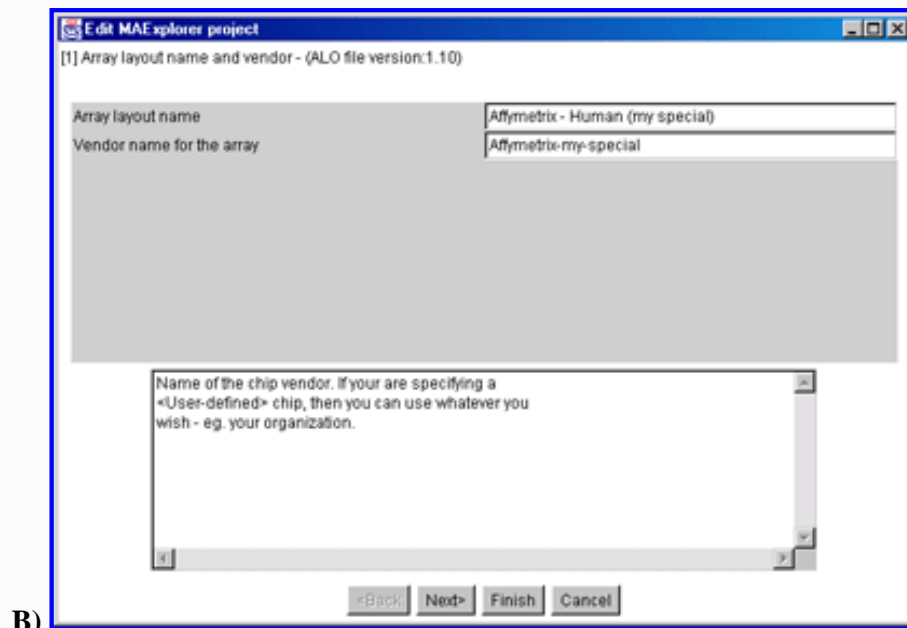
B)

Figure 6. Edit Layout Wizard for name of the Array Layout. A) is the original array layout frome the database. **B)** Since we may want to edit it, we will rename the vendor and Array layout name. This will enable us to save the changed layout if we wish. You may not overide system defined layouts, but you may overide your own layouts or save a system layout under a new name (as is shown here).

Figure 7. Edit Layout Wizard for Grid Geometry.

Figure 8. Edit Layout Wizard for Starting Data Rows.

Figure 9. Edit Layout Wizard for Ratio or Intensity data.

Figure 10. Edit Layout Wizard for optional (X,Y) spot coordinates available in the input data.

Figure 11. Edit Layout Wizard for optional Genomic ID values available in the input data.

Figure 12. Edit Layout Wizard for optional Gene Names available in the data.

Figure 13. Edit Layout Wizard for optional calibration DNA available in the data and UniGene species prefix.

Figure 14. Edit Layout Wizard for optional user names for Project, Database, Subdatabase, etc.

Figure 15. Edit Layout Wizard for optional HP-X and HP-Y 'set' experimental class (i.e. condition) names.

Figure 16. Edit Layout Wizard for changing the default data filter threshold slider values.

Mapping Array Layout GIPO and Quant input data field names

In addition to the above global defintions, additonal Array Layout parameters need to be define. These are mapping of input file data field names for GIPO and Quant data to the names required by MAExplorer. There are two wizards for helping define these mappings. For the predefined Array Layouts these are already setup but may need to be defined or edited for user-defined data.
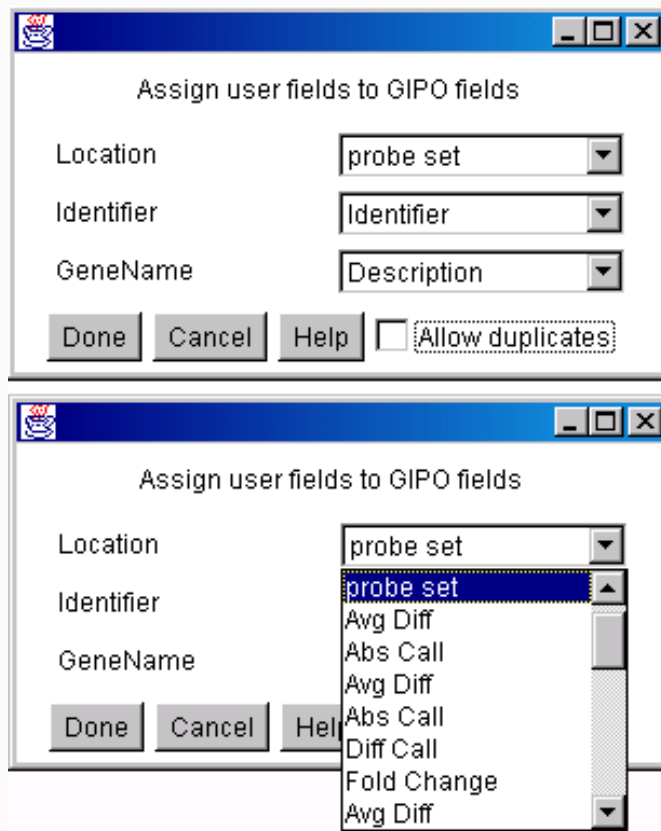
Figure 17. Edit Layout Wizard for Assign GIPO fields. These Gene-In-Plate-Order data field mappings should only be changed if required for additional data fields you may have added to your input file. All fields should be defined. (it is required for <User-defined> data). In general, it may be ok to have some non-critical genomic ID fields undefined.
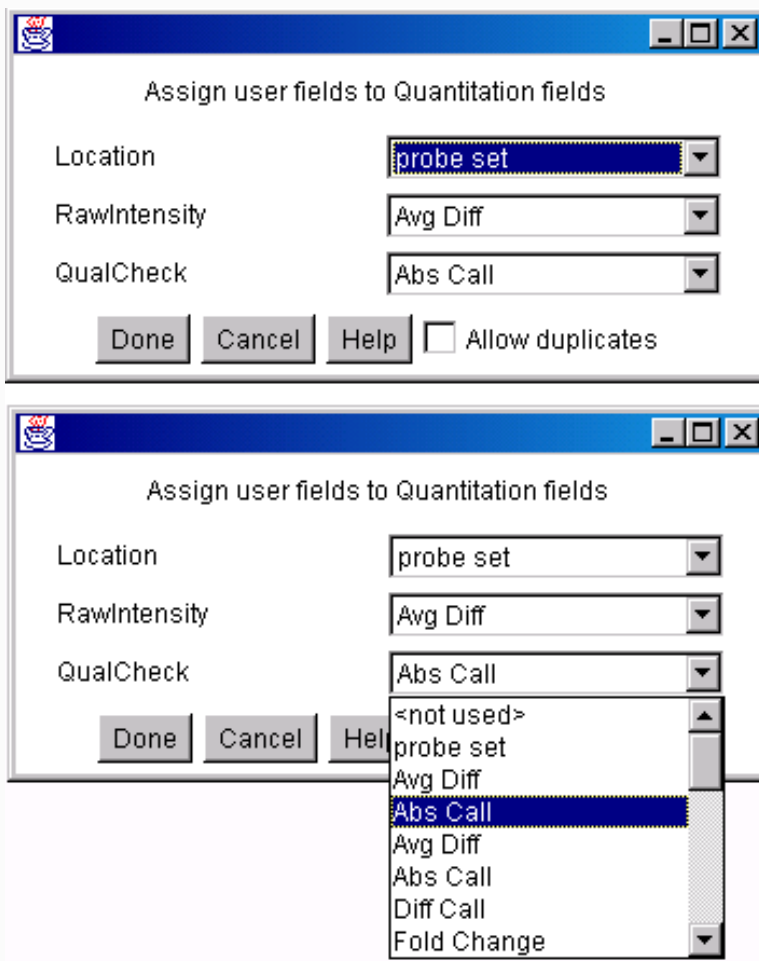
Figure 18. Edit Layout Wizard for Assign Quant fields. These [Quantification data field](#) mappings should only be changed if required to define all fields (it is required for <User-defined> data).
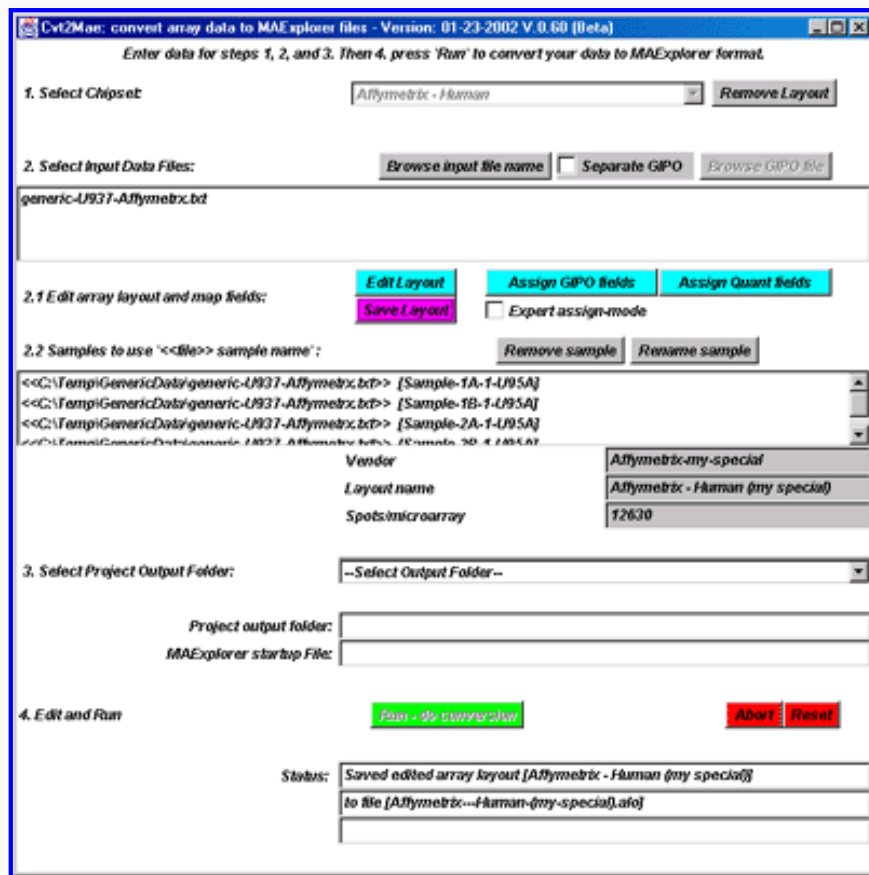
Figure 19. Saving modified Array Layout if you have made changes. This is useful if you have changed the array layout with "Edit Layout", "Assign GIPO fields", or "Assign Quant fields" so that you can use it another time.
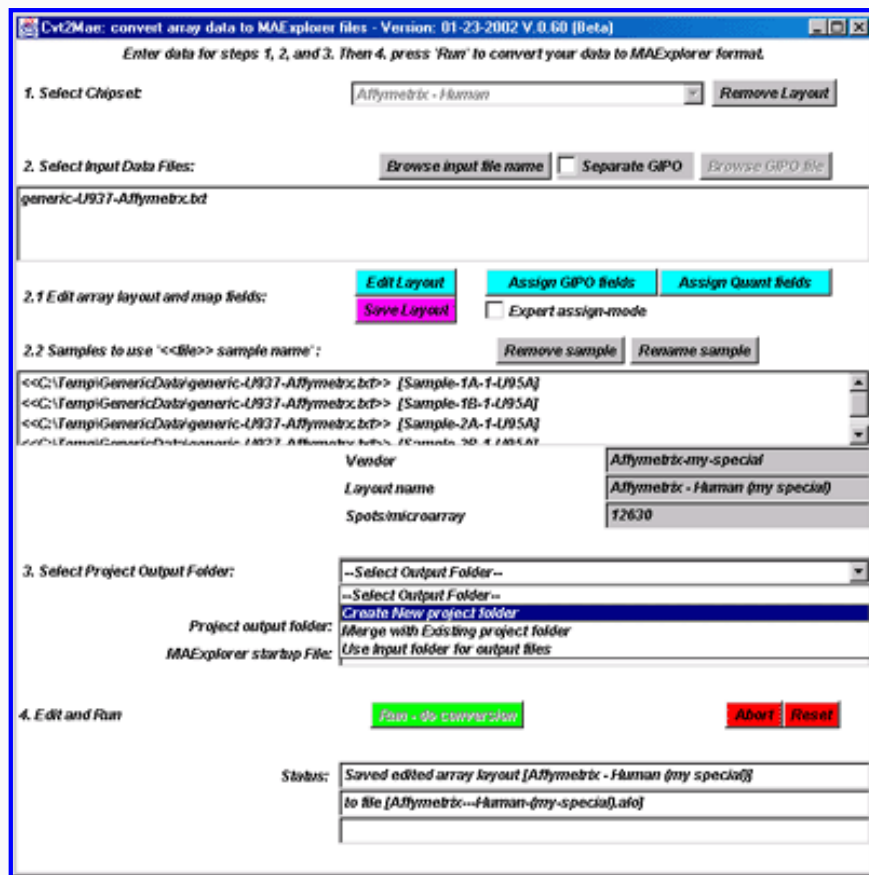
Figure 20. Selecting the output folder in which to save the converted files. The Magenta "Save Layout" button means that you may save the edited array layout if you wish. You now need to create an output folder to put the converted data. You may create a New Folder, use an Existing Folder or use the Same Folder that contained the input files. We selected the "New Folder" option.
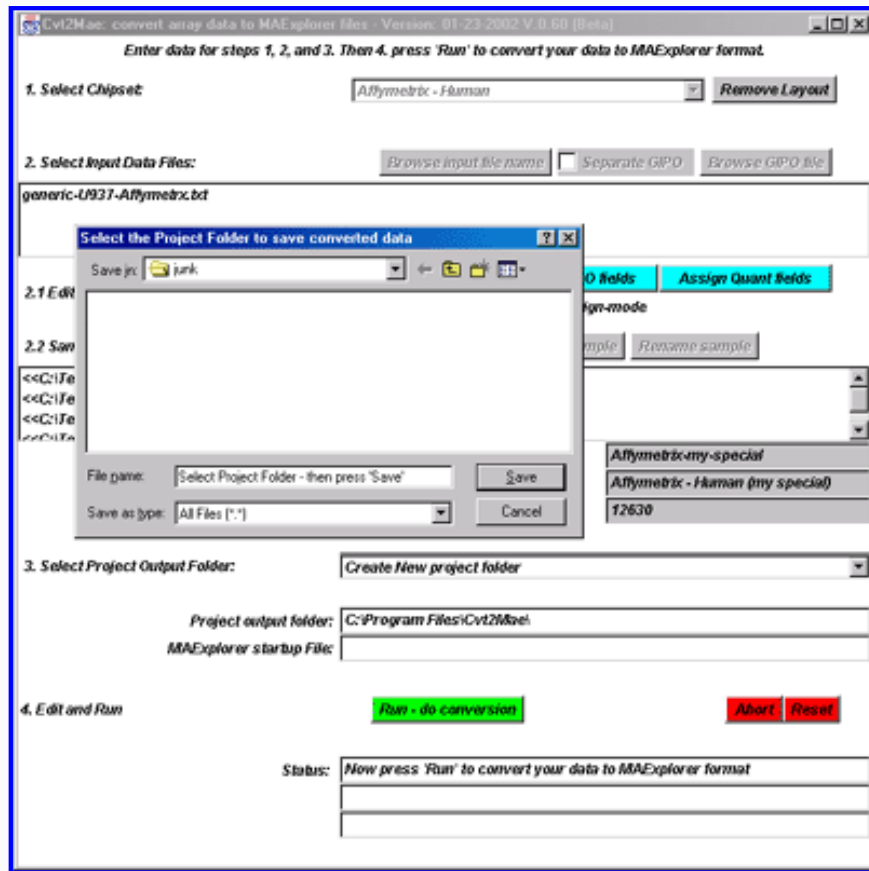
Figure 21. Browse to select the output folder in which to save the converted files. You may create a new folder here. Select the "name" of the folder - don't go into the folder.

Figure 22. shows the interface after selection of the output file folder using a file browser. Notice that the current project directory is now displayed in the interface as well as the location of the MAExplorer Start.mae file that will be generated. The data will be created when the Run button is pressed.
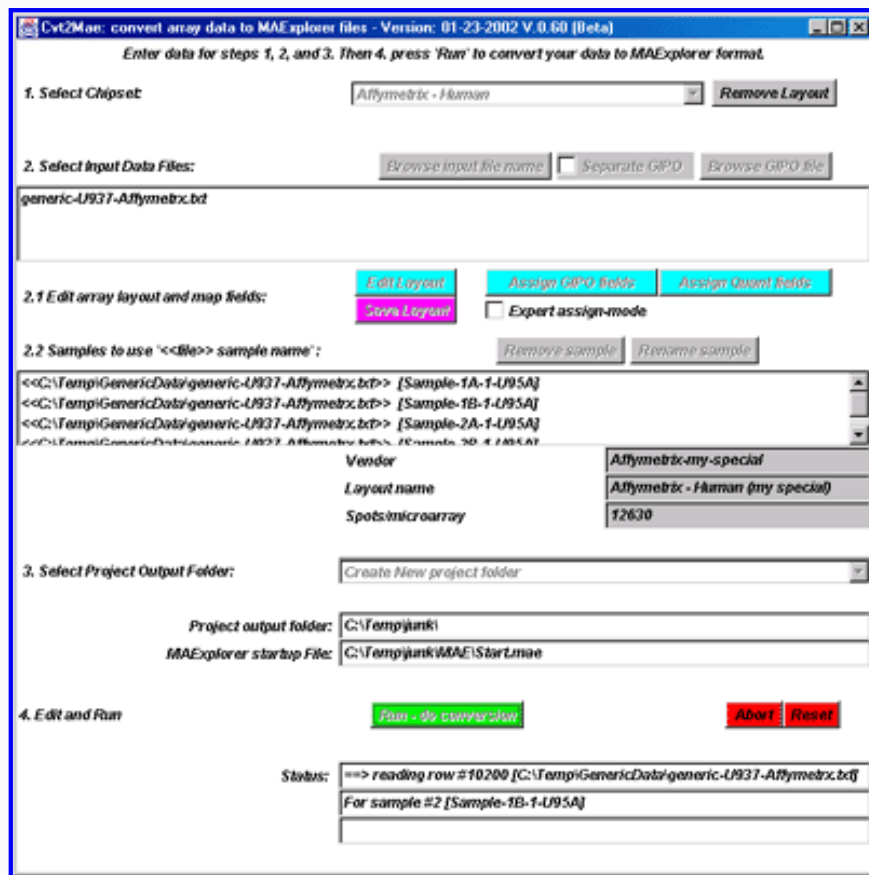
Figure 23. shows the conversion being performed after the user pressed the RUN button. This process takes a minute or so depending on the speed of the computer and the complexity of the data.
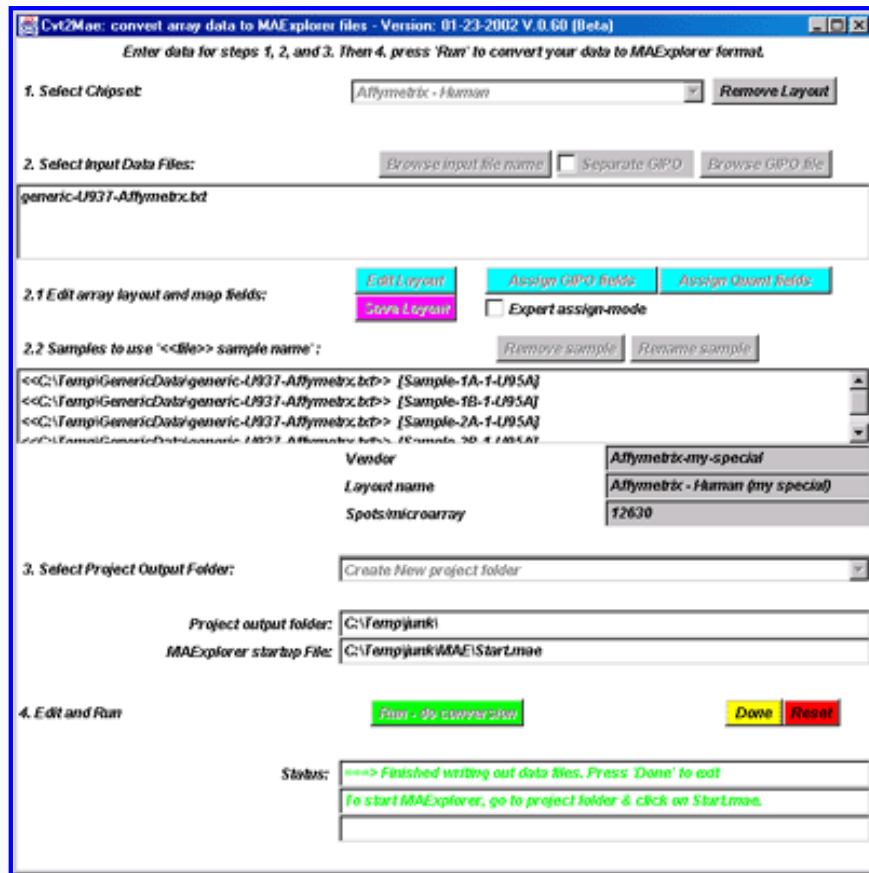
Figure 24. shows the conversion summary instructions after the conversion is finished. At this point press the DONE button to exit the converter.

Figure 25. shows the files that are generated by Cvt2Mae for use by MAExplorer. The generated data consists of several directories that are described in the Reference Manual Appendix C.
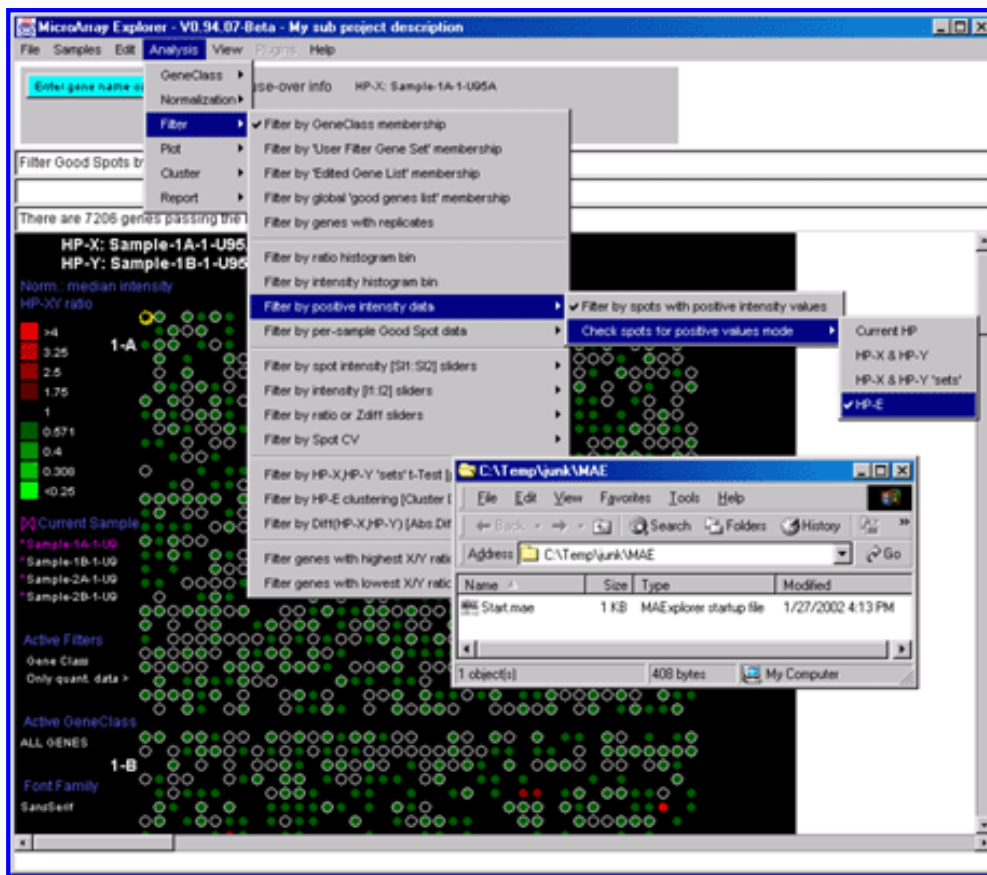
Figure 26. Starting MAExplorer on the converted data by clicking on Start.mae file. Alternatively, Note that the location of the "MAExplorer startup file:" is specified. Go to that file and click on it to start MAExplorer. Alternatively, start MAExplorer and do "File | Open Disk DB" and open that file to start it.

MAExplorer [MAExplorer home | Cvt2Mae home | Help desk | LECB/NCI/FCRDC]

Downloads for The MicroArray Explorer Project

You may download program installers for your particular computer for both MAExplorer and Cvt2Mae as well as the MAEPlugins. If you are interested in the source code, that is also available. Individual files are available in the CVS directories listed in the table below (see instructions on using CVS to access these files directly with CVS). Gzipped tar archived packages of the source code are also available.

If you want to upgrade your installation to the latest JAR files, simply download the JAR files and save them wherever you have installed the programs replacing the previous jar files. A Plugins-jar.tar file is available with all of the MAEPlugin jar files. Simply unpack the directory using Unix tar or a Windows unzip program into a directory you can access when running MAExplorer. To let MAExplorer go directly to these files when you do a (Plugins | Load plugins) menu command, copy the .jar files into the Plugins/ directory where you previously installed MAExplorer.

You may also download the Mammary Genome Anatomy Program (MGAP) public data set that can be used with MAExplorer. There is a list of of PDF documents describing MAExplorer and the Cvt2Mae data conversion wizard that may be downloaded.

See the Revision notes for more information on what changes have been made and what new features are available.

Click on the entries to download the files.

1. Access of MAExplorer, MAEPlugins, Cvt2Mae from either Web server

Program	Version	Program installers	Information on installing	Source	Jar file(s)
MAExplorer	0.96.11	MAExplorer	installing MAExplorer	source code	MAExplorer.jar
MAEPlugins	-	(not required)	Using MAEPlugins	source code	List of MAEPlugins
Cvt2Mae	0.67	Cvt2Mae	installing Cvt2Mae	source code	Cvt2Mae.jar

2. Download the Mammary Genome Anatomy Program (MGAP) public data set

The Mammary Genome Anatomy Program (MGAP) using mouse models has available a of public data set of 50 samples that may be downloaded and used with MAExplorer or other types of analysis. The hybridized samples data consists of tab-delimited files (no images) for about 1700 duplicate spots/membrane. There is a list of startup .mae files included in the download. You may download it several different ways.

Download method	Web address
As separate files	http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database/
A single zip file	http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database.zip
A single tar file	http://www.lecb.ncifcrf.gov/mae/MGAP-Array-database.tar

Javadocs documentation views of the MAExplorer Project

Java documentation is useful for writing MAEPlugins as well as understanding the MAExplorer code. Because of size restrictions the docsFull and docsAllPublic directories are currently available on the NCI/LECB web server.

However, you can generate your own javadocs for the code using the Unix script CreateMAExplorerJavaDocs.do for MAExplorer and CreateCvt2MaeJavaDoc.do .

View	Javadoc folder
Full javadocs (public+private) for MAExplorer	docsFull
Full javadocs (public only) for MAExplorer	docsAllPublic
Open Java API javadocs for MAExplorer	docsOJAPI
MaeJavaAPI (MJA) javadocs for MAExplorer	docsMJA
Full (public+private) javadocs for Cvt2Mae	javadocs

Mozilla Public License 1.1 (MPL 1.1)

1. Definitions.

1.0.1. "Commercial Use" means distribution or otherwise making the Covered Code available to a third party.

1.1. ''Contributor'' means each entity that creates or contributes to the creation of Modifications.

1.2. ''Contributor Version'' means the combination of the Original Code, prior Modifications used by a Contributor, and the Modifications made by that particular Contributor.

1.3. ''Covered Code'' means the Original Code or Modifications or the combination of the Original Code and Modifications, in each case including portions thereof**.**

1.4. ''Electronic Distribution Mechanism'' means a mechanism generally accepted in the software development community for the electronic transfer of data.

1.5. ''Executable'' means Covered Code in any form other than Source Code.

1.6. ''Initial Developer'' means the individual or entity identified as the Initial Developer in the Source Code notice required by **Exhibit A**.

1.7. ''Larger Work'' means a work which combines Covered Code or portions thereof with code not governed by the terms of this License.

1.8. ''License'' means this document.

1.8.1. "Licensable" means having the right to grant, to the maximum extent possible, whether at the time of the initial grant or subsequently acquired, any and all of the rights conveyed

herein.

1.9. ''Modifications'' means any addition to or deletion from the substance or structure of either the Original Code or any previous Modifications. When Covered Code is released as a series of files, a Modification is:

    A. Any addition to or deletion from the contents of a file containing Original Code or previous Modifications.

    B. Any new file that contains any part of the Original Code or previous Modifications.

1.10. ''Original Code'' means Source Code of computer software code which is described in the Source Code notice required by **Exhibit A** as Original Code, and which, at the time of its release under this License is not already Covered Code governed by this License.

1.10.1. "Patent Claims" means any patent claim(s), now owned or hereafter acquired, including without limitation, method, process, and apparatus claims, in any patent Licensable by grantor.

1.11. ''Source Code'' means the preferred form of the Covered Code for making modifications to it, including all modules it contains, plus any associated interface definition files, scripts used to control compilation and installation of an Executable, or source code differential comparisons against either the Original Code or another well known, available Covered Code of the Contributor's choice. The Source Code can be in a compressed or archival form, provided the appropriate decompression or de-archiving software is widely available for no charge.

1.12. "You'' (or "Your") means an individual or a legal entity exercising rights under, and complying with all of the terms of, this License or a future version of this License issued under Section 6.1. For legal entities, "You'' includes any entity which controls, is controlled by, or is under common control with You. For purposes of this definition, "control'' means (a) the power, direct or indirect, to cause the direction or management of such entity, whether by contract or otherwise, or (b) ownership of more than fifty

percent (50%) of the outstanding shares or beneficial ownership of such entity.

2. Source Code License.

2.1. The Initial Developer Grant.
The Initial Developer hereby grants You a world-wide, royalty-free, non-exclusive license, subject to third party intellectual property claims:

(a) under intellectual property rights (other than patent or trademark) Licensable by Initial Developer to use, reproduce, modify, display, perform, sublicense and distribute the Original Code (or portions thereof) with or without Modifications, and/or as part of a Larger Work; and

(b) under Patents Claims infringed by the making, using or selling of Original Code, to make, have made, use, practice, sell, and offer for sale, and/or otherwise dispose of the Original Code (or portions thereof).

(c) the licenses granted in this Section 2.1(a) and (b) are effective on the date Initial Developer first distributes Original Code under the terms of this License.

(d) Notwithstanding Section 2.1(b) above, no patent license is granted: 1) for code that You delete from the Original Code; 2) separate from the Original Code; or 3) for infringements caused by: i) the modification of the Original Code or ii) the combination of the Original Code with other software or devices.

2.2. Contributor Grant.
Subject to third party intellectual property claims, each Contributor hereby grants You a world-wide, royalty-free, non-exclusive license

(a) under intellectual property rights (other than patent or trademark) Licensable by Contributor, to use, reproduce, modify, display, perform, sublicense and distribute the Modifications created by such Contributor (or portions thereof) either on an unmodified basis, with other Modifications, as Covered Code and/or as part of a Larger Work; and

(b) under Patent Claims infringed by the making, using, or selling of Modifications made by that Contributor either alone and/or in combination with its Contributor Version (or portions of such combination), to make, use, sell, offer for sale, have made, and/or otherwise dispose of: 1) Modifications made by that Contributor (or portions thereof); and 2) the combination of Modifications made by that Contributor with its Contributor Version (or portions of such combination).

(c) the licenses granted in Sections 2.2(a) and 2.2(b) are effective on the date Contributor first makes Commercial Use of the Covered Code.

(d) Notwithstanding Section 2.2(b) above, no patent license is granted: 1) for any code that Contributor has deleted from the Contributor Version; 2) separate from the Contributor Version; 3) for infringements caused by: i) third party modifications of Contributor Version or ii) the combination of Modifications made by that Contributor with other software (except as part of the Contributor Version) or other devices; or 4) under Patent Claims infringed by Covered Code in the absence of Modifications made by that Contributor.

3. Distribution Obligations.

3.1. Application of License.
The Modifications which You create or to which You contribute are governed by the terms of this License, including without limitation Section **2.2**. The Source Code version of Covered Code may be distributed only under the terms of this License or a future version of this License released under Section **6.1**, and You must include a copy of this License with every copy of the Source Code You distribute. You may not offer or impose any terms on any Source Code version that alters or restricts the applicable version of this License or the recipients' rights hereunder. However, You may include an additional document offering the additional rights described in Section **3.5**.

3.2. Availability of Source Code.

Any Modification which You create or to which You contribute must be made available in Source Code form under the terms of this License either on the same media as an Executable version or via an accepted Electronic Distribution Mechanism to anyone to whom you made an Executable version available; and if made available via Electronic Distribution Mechanism, must remain available for at least twelve (12) months after the date it initially became available, or at least six (6) months after a subsequent version of that particular Modification has been made available to such recipients. You are responsible for ensuring that the Source Code version remains available even if the Electronic Distribution Mechanism is maintained by a third party.

3.3. Description of Modifications.
You must cause all Covered Code to which You contribute to contain a file documenting the changes You made to create that Covered Code and the date of any change. You must include a prominent statement that the Modification is derived, directly or indirectly, from Original Code provided by the Initial Developer and including the name of the Initial Developer in (a) the Source Code, and (b) in any notice in an Executable version or related documentation in which You describe the origin or ownership of the Covered Code.

3.4. Intellectual Property Matters

(a) Third Party Claims.
If Contributor has knowledge that a license under a third party's intellectual property rights is required to exercise the rights granted by such Contributor under Sections 2.1 or 2.2, Contributor must include a text file with the Source Code distribution titled "LEGAL'' which describes the claim and the party making the claim in sufficient detail that a recipient will know whom to contact. If Contributor obtains such knowledge after the Modification is made available as described in Section 3.2, Contributor shall promptly modify the LEGAL file in all copies Contributor makes available thereafter and shall take other steps (such as notifying appropriate mailing lists or newsgroups) reasonably calculated to inform those who received the Covered Code that new knowledge has been obtained.

(b) Contributor APIs.
If Contributor's Modifications include an application programming interface and Contributor has knowledge of patent licenses which are reasonably necessary to implement that API, Contributor must also include this information in the LEGAL file.

(c)   Representations.
Contributor represents that, except as disclosed pursuant to Section 3.4(a) above, Contributor believes that Contributor's Modifications are Contributor's original creation(s) and/or Contributor has sufficient rights to grant the rights conveyed by this License.

3.5. Required Notices.
You must duplicate the notice in **Exhibit A** in each file of the Source Code. If it is not possible to put such notice in a particular Source Code file due to its structure, then You must include such notice in a location (such as a relevant directory) where a user would be likely to look for such a notice. If You created one or more Modification(s) You may add your name as a Contributor to the notice described in **Exhibit A**. You must also duplicate this License in any documentation for the Source Code where You describe recipients' rights or ownership rights relating to Covered Code. You may choose to offer, and to charge a fee for, warranty, support, indemnity or liability obligations to one or more recipients of Covered Code. However, You may do so only on Your own behalf, and not on behalf of the Initial Developer or any Contributor. You must make it absolutely clear than any such warranty, support, indemnity or liability obligation is offered by You alone, and You hereby agree to indemnify the Initial Developer and every Contributor for any liability incurred by the Initial Developer or such Contributor as a result of warranty, support, indemnity or liability terms You offer.

3.6. Distribution of Executable Versions.
You may distribute Covered Code in Executable form only if the requirements of Section **3.1-3.5** have been met for that Covered Code, and if You include a notice stating that the Source Code version of the Covered Code is available under the terms of this License, including a

description of how and where You have fulfilled the obligations of Section **3.2**. The notice must be conspicuously included in any notice in an Executable version, related documentation or collateral in which You describe recipients' rights relating to the Covered Code. You may distribute the Executable version of Covered Code or ownership rights under a license of Your choice, which may contain terms different from this License, provided that You are in compliance with the terms of this License and that the license for the Executable version does not attempt to limit or alter the recipient's rights in the Source Code version from the rights set forth in this License. If You distribute the Executable version under a different license You must make it absolutely clear that any terms which differ from this License are offered by You alone, not by the Initial Developer or any Contributor. You hereby agree to indemnify the Initial Developer and every Contributor for any liability incurred by the Initial Developer or such Contributor as a result of any such terms You offer.

3.7. Larger Works.
You may create a Larger Work by combining Covered Code with other code not governed by the terms of this License and distribute the Larger Work as a single product. In such a case, You must make sure the requirements of this License are fulfilled for the Covered Code.

4. Inability to Comply Due to Statute or Regulation.

If it is impossible for You to comply with any of the terms of this License with respect to some or all of the Covered Code due to statute, judicial order, or regulation then You must: (a) comply with the terms of this License to the maximum extent possible; and (b) describe the limitations and the code they affect. Such description must be included in the LEGAL file described in Section **3.4** and must be included with all distributions of the Source Code. Except to the extent prohibited by statute or regulation, such description must be sufficiently detailed for a recipient of ordinary skill to be able to understand it.

5. Application of this License.

This License applies to code to which the Initial Developer has attached the notice in **Exhibit A**

and to related Covered Code.

6. Versions of the License.

6.1. New Versions.
Netscape Communications Corporation
(''Netscape'') may publish revised and/or new
versions of the License from time to time. Each
version will be given a distinguishing version
number.

6.2. Effect of New Versions.
Once Covered Code has been published under a
particular version of the License, You may always
continue to use it under the terms of that
version. You may also choose to use such Covered
Code under the terms of any subsequent version of
the License published by Netscape. No one other
than Netscape has the right to modify the terms
applicable to Covered Code created under this
License.

6.3. Derivative Works.
If You create or use a modified version of this
License (which you may only do in order to apply
it to code which is not already Covered Code
governed by this License), You must (a) rename
Your license so that the phrases ''Mozilla'',
''MOZILLAPL'', ''MOZPL'', ''Netscape'', "MPL",
''NPL'' or any confusingly similar phrase do not
appear in your license (except to note that your
license differs from this License) and (b)
otherwise make it clear that Your version of the
license contains terms which differ from the
Mozilla Public License and Netscape Public
License. (Filling in the name of the Initial
Developer, Original Code or Contributor in the
notice described in **Exhibit A** shall not of
themselves be deemed to be modifications of this
License.)

7. DISCLAIMER OF WARRANTY.

COVERED CODE IS PROVIDED UNDER THIS LICENSE ON AN
"AS IS'' BASIS, WITHOUT WARRANTY OF ANY KIND,
EITHER EXPRESSED OR IMPLIED, INCLUDING, WITHOUT
LIMITATION, WARRANTIES THAT THE COVERED CODE IS
FREE OF DEFECTS, MERCHANTABLE, FIT FOR A
PARTICULAR PURPOSE OR NON-INFRINGING. THE ENTIRE
RISK AS TO THE QUALITY AND PERFORMANCE OF THE
COVERED CODE IS WITH YOU. SHOULD ANY COVERED CODE
PROVE DEFECTIVE IN ANY RESPECT, YOU (NOT THE
INITIAL DEVELOPER OR ANY OTHER CONTRIBUTOR)
ASSUME THE COST OF ANY NECESSARY SERVICING,

REPAIR OR CORRECTION. THIS DISCLAIMER OF WARRANTY CONSTITUTES AN ESSENTIAL PART OF THIS LICENSE. NO USE OF ANY COVERED CODE IS AUTHORIZED HEREUNDER EXCEPT UNDER THIS DISCLAIMER.

8. TERMINATION.

8.1. This License and the rights granted hereunder will terminate automatically if You fail to comply with terms herein and fail to cure such breach within 30 days of becoming aware of the breach. All sublicenses to the Covered Code which are properly granted shall survive any termination of this License. Provisions which, by their nature, must remain in effect beyond the termination of this License shall survive.

8.2. If You initiate litigation by asserting a patent infringement claim (excluding declatory judgment actions) against Initial Developer or a Contributor (the Initial Developer or Contributor against whom You file such action is referred to as "Participant") alleging that:

(a) such Participant's Contributor Version directly or indirectly infringes any patent, then any and all rights granted by such Participant to You under Sections 2.1 and/or 2.2 of this License shall, upon 60 days notice from Participant terminate prospectively, unless if within 60 days after receipt of notice You either: (i) agree in writing to pay Participant a mutually agreeable reasonable royalty for Your past and future use of Modifications made by such Participant, or (ii) withdraw Your litigation claim with respect to the Contributor Version against such Participant. If within 60 days of notice, a reasonable royalty and payment arrangement are not mutually agreed upon in writing by the parties or the litigation claim is not withdrawn, the rights granted by Participant to You under Sections 2.1 and/or 2.2 automatically terminate at the expiration of the 60 day notice period specified above.

(b) any software, hardware, or device, other than such Participant's Contributor Version, directly or indirectly infringes any patent, then any rights granted to You by such Participant under Sections 2.1(b) and 2.2(b) are revoked effective as of the date You first made, used, sold, distributed, or had made, Modifications made by that Participant.

8.3. If You assert a patent infringement claim against Participant alleging that such Participant's Contributor Version directly or indirectly infringes any patent where such claim is resolved (such as by license or settlement) prior to the initiation of patent infringement litigation, then the reasonable value of the licenses granted by such Participant under Sections 2.1 or 2.2 shall be taken into account in determining the amount or value of any payment or license.

8.4. In the event of termination under Sections 8.1 or 8.2 above, all end user license agreements (excluding distributors and resellers) which have been validly granted by You or any distributor hereunder prior to termination shall survive termination.

9. LIMITATION OF LIABILITY.

UNDER NO CIRCUMSTANCES AND UNDER NO LEGAL THEORY, WHETHER TORT (INCLUDING NEGLIGENCE), CONTRACT, OR OTHERWISE, SHALL YOU, THE INITIAL DEVELOPER, ANY OTHER CONTRIBUTOR, OR ANY DISTRIBUTOR OF COVERED CODE, OR ANY SUPPLIER OF ANY OF SUCH PARTIES, BE LIABLE TO ANY PERSON FOR ANY INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES OF ANY CHARACTER INCLUDING, WITHOUT LIMITATION, DAMAGES FOR LOSS OF GOODWILL, WORK STOPPAGE, COMPUTER FAILURE OR MALFUNCTION, OR ANY AND ALL OTHER COMMERCIAL DAMAGES OR LOSSES, EVEN IF SUCH PARTY SHALL HAVE BEEN INFORMED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION OF LIABILITY SHALL NOT APPLY TO LIABILITY FOR DEATH OR PERSONAL INJURY RESULTING FROM SUCH PARTY'S NEGLIGENCE TO THE EXTENT APPLICABLE LAW PROHIBITS SUCH LIMITATION. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OR LIMITATION OF INCIDENTAL OR CONSEQUENTIAL DAMAGES, SO THIS EXCLUSION AND LIMITATION MAY NOT APPLY TO YOU.

10. U.S. GOVERNMENT END USERS.

The Covered Code is a ''commercial item,'' as that term is defined in 48 C.F.R. 2.101 (Oct. 1995), consisting of ''commercial computer software'' and ''commercial computer software documentation,'' as such terms are used in 48 C.F.R. 12.212 (Sept. 1995). Consistent with 48 C.F.R. 12.212 and 48 C.F.R. 227.7202-1 through 227.7202-4 (June 1995), all U.S. Government End Users acquire Covered Code with only those rights

set forth herein.

11. MISCELLANEOUS.

This License represents the complete agreement concerning subject matter hereof. If any provision of this License is held to be unenforceable, such provision shall be reformed only to the extent necessary to make it enforceable. This License shall be governed by California law provisions (except to the extent applicable law, if any, provides otherwise), excluding its conflict-of-law provisions. With respect to disputes in which at least one party is a citizen of, or an entity chartered or registered to do business in the United States of America, any litigation relating to this License shall be subject to the jurisdiction of the Federal Courts of the Northern District of California, with venue lying in Santa Clara County, California, with the losing party responsible for costs, including without limitation, court costs and reasonable attorneys' fees and expenses. The application of the United Nations Convention on Contracts for the International Sale of Goods is expressly excluded. Any law or regulation which provides that the language of a contract shall be construed against the drafter shall not apply to this License.

12. RESPONSIBILITY FOR CLAIMS.

As between Initial Developer and the Contributors, each party is responsible for claims and damages arising, directly or indirectly, out of its utilization of rights under this License and You agree to work with Initial Developer and Contributors to distribute such responsibility on an equitable basis. Nothing herein is intended or shall be deemed to constitute any admission of liability.

13. MULTIPLE-LICENSED CODE.

Initial Developer may designate portions of the Covered Code as "Multiple-Licensed". "Multiple-Licensed" means that the Initial Developer permits you to utilize portions of the Covered Code under Your choice of the NPL or the alternative licenses, if any, specified by the Initial Developer in the file described in Exhibit A.

LEGAL file - MAExplorer Software under the Mozilla Public License (V 1.1)

MAExplorer Software under the Mozilla Public License (version 1.1)
Date: 5-31-2002

This document comprises the LEGAL File pursuant to Articles 3.4 and 4 of the Mozilla Public License (version 1.1) stating the intellectual property and other limitations associated with the use of MAExplorer under this License. Dr. Peter Lemkin as an employee of The National Cancer Institute (NCI), an agency of the United States Government, is the Initial Developer of MAExplorer (the Original Code). As such, the following limitations apply to this License:

1. This is a work of the United States Government and as such there is no copyright associated with it. Notwithstanding herein, the Government does not provide or grant any automatic license to any copyright or patent.

2. Neither the NCI, the National Institutes of Health or the United States Government endorse MAExplorer or any product. Express or implied endorsement by these entities is prohibited.

3. The NCI, as an agency of the United States Government, is forbidden by statute to indemnify a third party. Thus, no indemnification for any loss, claim, damage or liability is intended or provided by NCI under this License. The NCI, as an agency of the United States Government, assumes liability only to the extent provided under the federal Tort Claims Act, 28 U.S.C. 2671 et seq.

4. This License shall be construed and governed in accordance with Federal law as applied by the Federal courts in the District of Columbia. In case of conflict in law between Federal law and the laws of any other jurisdiction, Federal law as applied by the Federal courts in the District of Columbia shall prevail.

List of Figures

All figures are available in low (default) and high resolution. Click on the image and it will bring up the high resolution version.

List of Tables

Glossary of terms used in MAExplorer

- **Abbreviations:**
  - **E.C.L.** is Edited Gene list
  - **G.C.** is Gene Class
  - **H.P.** is hybridized sample
  - **HP-E** is list of expression profile hybridized samples
  - **HP-X** is list of X-axis hybridized samples
  - **HP-Y** is list of Y-axis hybridized samples

- **applet** - a java program that runs in a Web browser and is downloaded from the Web server each time it is run. Applets do not require installing any software on your computer. However, they may not read or write any files on your computer. MAExplorer was initially used used as an applet when accessing microarray data on the original MGAP http://www.lecb.ncifcrf.gov/maemicroarray Web server.

- **background** - the background intensity measured near a spot in a microarray image. Spot intensity measurements from images with high background or where background varies over the image because of non-specific hybridization in parts of the array need to be corrected. Some data sets have this background available as a single average value for the entire array and others may have it on a per-spot bases. MAExplorer can optionally subtract this background if it is available.

- **cDNA** - complementary DNA immobilized as spots on the target microarray are hybridized with mRNA of the hybridized sample. Alternative methods of constructing arrays attach synthetic oligonucleotides to the arrays rather than cDNA from clones.

- **clone** - consisting of many copies of a particular cDNA. Each spot in the microarray target represents cDNA from a different clone. In MAExplorer we use the term gene to refer to both clone spotted arrays and oligonucliotide arrays.

- **Clone ID** - The unique identifier assigned to the clone by the I.M.A.G.E. consortium. This may be used to lookup the clone and related data in the NCBI dbEST, GenBank, UniGene, NCI/CIT mAdb clone databases. The Clone ID is reported in all results from MAExplorer searches.

- **Cluster** - a set of genes with similar hybridized samples expression profiles. It may also be a set of HPs with similar expression profiles for a specific set of genes. Different clustering methods define "similarity" in different ways. See Cluster Plots in the Reference Manual.

- **current gene** - the particular gene being analyzed in MAExplorer. This is set by the user by clicking on a gene in the array, scatter plot or Clone ID instance in a report or typing the name or Clone ID in the popup text window. It is indicated with a green circle in the array and scatter plots, and by a darkened background in the Reports.

- **current cluster** - the particular subset of genes belonging to a cluster of genes found using the K-means clustering method. These genes belong to the cluster to which the current gene belongs. Therefore, to change the current cluster - change the current gene. If K-means clustering is active and the current cluster is defined, the genes which belong to it will be indicated with a tiny green cluster number in the array image and in the scatter plot (if one is present). When the current cluster is defined, the genes in the current cluster are copied to the 'edited gene list'.

- **dbEST** - Expressed Sequence Tag (EST) data from the NCBI dbEST database. The database may be indexed by clone Id and may have entries for the 3' and/or 5' sequenced genes.

- **DRYROT error** - is a fatal error that is detected by MAExplorer, it will popup an error reporting window. We call this a "DRYROT" error (thanks to "S.A.I.L.") because something is wrong in the program or data files and from which it can not recover. This type of error should not have happened.

- **duplicate genes** - (notation) genes that are spotted multiple times (i.e. F1, F2) for *all* of the genes on the same array. This is different than replicated genes that *may not* have copies for all genes.

- **Edited Gene List** - (E.G.L) is a subset of genes in the MAExplorer database. It may be defined and edited manually or may be set from various MAExplorer operations (e.g. set to the current cluster when doing similarity clustering, K-means clustering, etc).

- **EST** - Expressed Sequence Tag of a particular gene fragment which is expressed in a particular tissue.

- **Expression Profile** - (or EP) is the vector of the spot intensity values or expression profile for the same gene for the list of hybridized samples being analyzed. This is used in computing similarity between genes. Alternatively, it may be a vector of gene intensities for a hybridized sample. This is used in computing similarity between samples.

- **F1,F2** - are the left (F1) and right (F1) replicate fields of the microarray (if the array supports it). Some arrays duplicate the gene in corresponding grids of spots in the microarray (Research Genetics, NIA neuroarrays, etc.). This lets us compute a coefficient of variation (CV) for the gene. If the CV is low, then the measurement is more reliable.

- **Filter** - (data filter) is a sequence of restrictions on the set of all genes used to determine a subset list of genes. These restrictions include gene membership in a "Gene Class" (eg. oncogenes); quantitative tests including: gene class

membership, spot-CV-value, ratio or intensity range or membership in histogram bins, statistical and clustering tests, etc. These are then used to pre-filter the set of all genes to a *working* subset which may then be used for viewing, reporting, or saving in a gene subset. The subset may then be used in subsequent MAExplorer operations or saved for later use.

- **gene** - in MAExplorer we use the generic term gene to refer to data for both gene spotted arrays and oligonucliotide arrays.

- **Gene Class** - (G.C.) the sets of particular genes with similar properties, e.g. oncogenes, heat shock, milk proteins, etc. These can be used to help partition the set of all genes as part of the data filtering process. The MAExplorer Gene Name Guesser (Section 2.3.1)

- **gene expression profile** - (or EP) the vector of normalized quantitative values for the corresponding gene in the set of hybridized sample (HP-E). This vector represents the expression of the gene across a set of experimental conditions such as developmental stages, dose-response, time-series, etc.

- **gene subset** - a set of genes defined by the current filter may be saved and used at a later time as part of the Filter. In addition, there are a number of set operations (union, intersection and difference) which may be used to derive new sets of genes.

- **grid** - a rectangular matrix of spots on the microarray. Generally the spots are laid out in a series of grids (also called "blocks" or a "patches"). They are indexed within the grid as a "grid row" and a "grid column" If the grids are replicated, they are indicated by multiple fields of duplicated spots.

- **grid coordinates** - The genes are layed down in a microarray in rows and columns of a grid. There are multiple grids which may have a space between them. For example, in the MGAP microarray database there are 8 grids (named A through H) to a field with a space between grids. Finally, there may be two fields (left and right named 1 and 2 or F1 and F2) which are duplicates of the grids so as to allow us to get an estimate of the hybridized sample variance by spotting the same gene multiple times in the array.

- **hierarchical clustering** is a clustering method that constructs a binary tree where the leaf objects are the set of genes (or conditions) being clustered. The algorithm is described in Section 3.2.2.3 describing the K-means clustering. It is invoked by the Hierarchical clustering menu entry in Section 2.4.5.4.

- **Hybridized Sample** - (or HP) a particular hybridized sample's quantified data. It may be a hybridized microarray of cDNA clones (or alternatively oligonucleotides) are hybridized with ^{33}P radio-labled or Cy3 and Cy5 flourescent-labeled mRNA total sample probe. We abbreviate this hybrized sample in MAExplorer as H.P. or HP.

- **HP-X** - the individual H.P. sample assigned to the Y-axis when doing an analysis. The H.P-Y 'set' is a set of samples used to compute mean intensity for each gene.

- **HP-Y** - the individual H.P. sample assigned to the Y-axis when doing an analysis. The H.P-Y 'set' is a set of samples used to compute mean intensity for each gene.

- **HP-E** - an ordered list of hybridized samples that is used in computing the gene expression profile.

- **image** - the microarray image associated with a hybridized sample with a microarray obtained by scanning the radioactive or flourescent labeled array on a phosphoimager or densitometer (if ^{33}P labeling was used to expose film) at about 50 or 100 microns/pixel. [see particular brands and models for details.] The image is then quantified to obtain individual spot intensity values using an image analysis program (the image analysis step is *not* part of MAExplorer).

- **I.M.A.G.E.** - I.M.A.G.E. clone consortium is a public collection of sequenced clones which offers the ability to locate and acquire specific cDNA clones by assigned ID numbers. MAExplorer uses the Clone ID as the primary gene identifier.

- **intensity** - refers to the integrated density over a spot and corresponds to the level of that hybridized sample's labeled mRNA to the genes cDNA immobilized on the array. Also see background density.

- **K-means** is a clustering method that assigns data Filtered genes to a fixed number of orthogonal clusters. The algorithm is described in Section 3.2.2.2 describing the K-means clustering. It is invoked by the K-means clustering menu entry in Section 2.4.5.3.

- **microarray** - a nylon membrane, glass slide, or other subtrate with thousands of cDNA clones or oligonucleotides "spotted" in unique positions on the array. The cDNA are typically PCR products of approximately 0.6-2.4kb representing specific genes and typically 100-500 microgram/ml (Duggan, *Nature Genetics Supplement*, **21**, Jan, 1999, pg 11). For example, the initial MGAP database array created by Research Genetics is a nylon membrane (3x7 cm) with 3500 spots (1500 duplicated clones in cDNA and EST inserts). The array may be "hybridized" with labeled samples to quantify how much of each cDNA is present in the sample. See (Schulze, 2001) for a nice summary of the current technology.

- **mAdb** - the NCI/CIT MicroArray Database Program [http://nciarray.nci.nih.gov/] is a repository of microarray data from the NCI ATC (Advanced Technology Center) microarray spotting facility.

- **mAdb Clone DB** - a clone database integrating hypertext links to various genomic databases for a clone. This is part of the mAdb.

- **MGAP** - Mammary Genome Anatomy Program. MAExplorer was initially developed for this Program. This is part of the Biology of the Mammary Gland [http://mammary.nih.gov/] Web site.

- **normalization** - the method for scaling all spot quantitation measurements on a set of hybridized samples to the same relative range so as that data from the differerent samples may be compared.

- **probe** - (MAExplorer convention) total sample labled mRNAs used to test hybridization with target cDNA immobilized on the array. Synonymous with hybridized sample in MAExplorer.

- **pseudoarray image** - a MAExplorer image representing the quantified spot data in the database. It may or may not correspond to the actual layout of grids, grid-rows and grid-columns in the original array. It is useful for getting the gestalt of the overall changes in the data. There are many representations of the data (see Section 2.4.4.1 Show Microarray for discussion and examples of these representations.

- **quantification** - the analysis of a scanned microarray image for the purpose of quantifying each spot on the array. This process typically consists of first finding the extent of a spot and then integrating the intensity. If multiple dyes are used (e.g. Cy3, Cy5) then the process is performed for each color-filtered image.

- **ratio** - in the context of microarrays and MAExplorer, consists of computing the ratio of Cy3/Cy5 or HP-X/HP-Y sets of samples. You need to see the particular context of its use to determine which meaning is implied.

- **replicate genes** - (notation) genes that are spotted multiple times for *some* of the genes on the array. This is different than duplicated genes that *do* have copies for all genes.

- **reports** - are popup window reports of ordered sets of genes meeting various search criteria may be generated. These consist of the quantitative measure, Clone Id, dbEST id, GenBank Id, Unigene link, etc. These are created in popup windows in MAExplorer as either scrollable dynamic spreadsheets or tab-delimited text windows which can be exported by cutting and pasting into Excel. The default dynamic spreadsheet option, contains active hyperlink fields which point to other web databases (such as GeneCard, dbEST, GenBank, Unigene, mAdb clones, etc.).

- **sample** - same as hybridized sample in the context of MAExplorer.

- **SaveAs** - is the command or button used in many of the plot windows to save the current plot as a full resolution GIF file specified by the user in a popup file browser window. It is available only in stand-alone mode and not in the applet version. These files could then be used for documenting a data-mining experiment or for publication.

- **silhouette plot** - is way to graphically display a sorted list of similar objects. In the way it is used in MAExplorer, each object in the descending sorted list (e.g. of genes) has a line (or set of "******") proportional to the similarity to the object at the top of the list. It has the largest number of stars since its similarity is 1.0. See (Kaufman and Rousseeuw, 1990) for discussion on silhouette plots. It may be visualized in the following example as:

```
Object   Graphic   Similarity
A        *******      1.00
B        *****        0.83
C        *****        0.80
D        **           0.30
E        *            0.17
F                     0.06
```

- **similarity measure** - is a scalar measure computed from two vectors of normalized measurements for two objects indicating how similar these two objects are. In MAExplorer, the vector is the expression profile (see Section 3.2.1) of the same gene for the list of hybridized samples being analyzed. It may alternatively be the expression profile of a list of genes for a given hybridized sample. So similarity measures indicate how similar two different expression profiles are. In MAExplorer, you may use either an Euclidian distance or correlation-coefficient as the scalar metric.

- **similarity cluster** is a clustering method that finds the subset of objects (eg. genes or conditions) that is most similar to the selected object. The list of genes is reduced to those whose "distance" between the selected object and the candidate genes is less than some threshold. The algorithm is described in Section 3.2.2.1 describing the similarity clustering. It is invoked by the similarity cluster menu entry in Section 2.4.5.1.

- **spot** - the spot in a microarray image corresponds to a clone or oligonucleotide attached to the array substrate. The intensity of the spot corresponds to the level of hybridized sample for that gene to the DNA in the spot which correlates to mRNA transcription expression.

- **startup file** - when using MAExplorer as a stand-alone application, this is the file (with the file name ending with a ".mae" extension) that invokes MAExplorer. On systems where you launch applications by double clicking on them, you would double click on this file. On systems such as UNIX where you specify arguments to a command line interface, you would type "MAExplorer *startup .mae file name*".

- **target** - (MAExplorer convention) cDNA or oligonucleotide of a gene immobilized as spot on a microarray. The array contains thousands of these hybridized sample spots.

- **UniGene** - the NCBI UniGene database "is an experimental system group for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location."

- **Zdiff** - statistical measure of difference between two samples (eg. HP-X and HP-Y) when using the Zscore or Zscore of logs of the data normalizations. If samples x,y have Zscore values for gene c in samples HP-X and HP-Y, or HP-X 'sets' and HP-Y 'sets', then the Zdiff is computed as:

```
Zdiff(x,y,c) = Zscore(x,c) - Zscore(y,c)
```

- **Zscore** - HP data normalization scaling method to set the mean value to 0.0 and the maximum and minimum values of the

scaled values to about +3.0 to -3.0 (standard deviations). It only makes sense if the data is roughly a gaussian distribution. If it is not, you could use the Zscore of the logs of the data to make it more gaussian. If c is a gene in hybridized sample p with $mean_p$ and $stdDev_p$, then it is computed as:

```
Zscore(p,c) = (c - mean_p)/stdDev_p
```

Index

This index is designed to be used with a Web browser. Since the page numbers in a "paper" copy of the document depends on the font and browser window sizes we do not include page numbers. Click on a hyperlink in an index entry to jump to that entry in the reference manual. Unfortunately, the lack of page numbers makes it difficult to use the index with a paper copy of the reference manual.