MicroArray Explorer - A Java-based Tool For Data Mining Microarrays
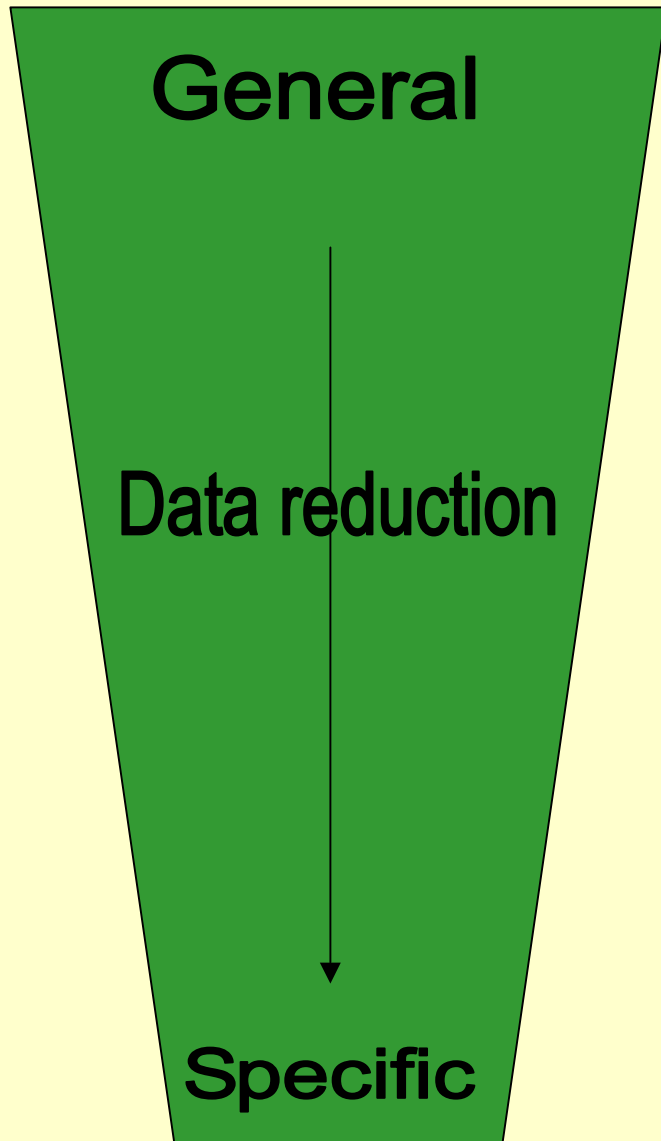
http://www.lecb.ncifcrf.gov/MAExplorer

Peter F. Lemkin[1], Gregory Thornwall[2], Lothar Hennighausen[3]

[1]LECB, NCI/FCRDC, Frederick, MD 21702; [2]SAIC/FCRDC;
[3]LGP, NIDDK, Bethesda, MD

Data Mining - Finding Putative Relevant Patterns

General

Data reduction

Specific

Results

Quantified array spot data for multiple samples

Organize by sample, gene expr, gene sets

Change views, normalization, data filters

Visualize and query: plots, cluster, reports
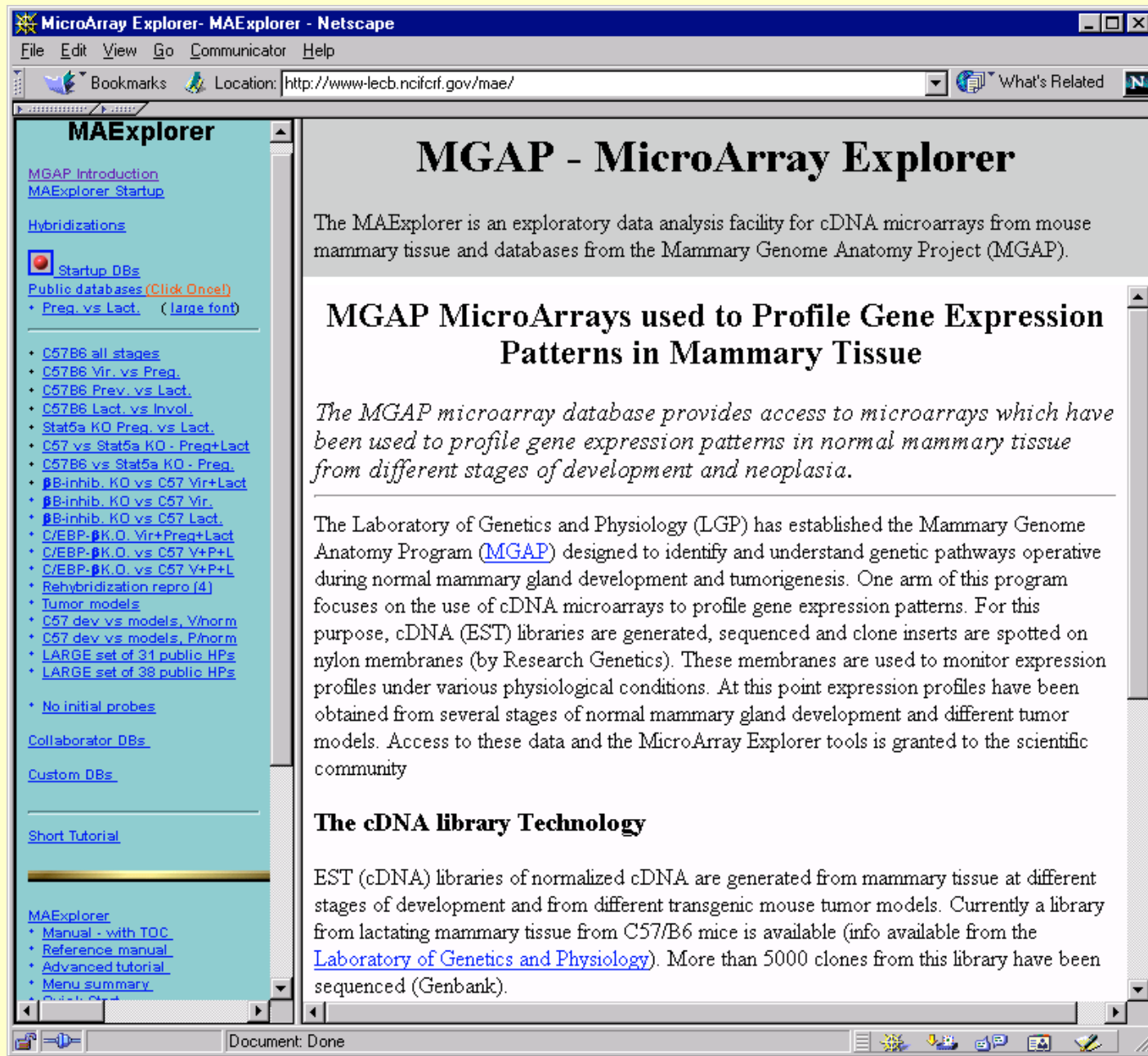
Explore external genomic databases

What is the MicroArray Explorer?

- **MAExplorer is a flexible Java stand-alone (off-line) or applet (Web-based) microarray data-mining tool**

- Handles multiple cDNA arrays with replicate spots
- Handles intensity or ratio (Cy3/Cy5) quantified array data
- Analyzes data for 2-conditions and expression profiles
- Data-filters gene sets by statistics, clustering, gene set membership
- Direct data manipulation in graphics, spreadsheets
- Accesses genomic Web servers from plots and reports
- Currently oriented for cDNA, <u>extending to protein arrays</u>

MAExplorer was Developed for the Mammary Genome Anatomy Program

- MGAP models, histology, expression
  http://mammary.nih.gov/

- Mouse models of
  1. normal mammary development C57B6
  2. Knock-outs: Stat5a (-,-) & (+,-), CEBP null, TGFα etc.
  3. Tumor models: WAP taq, p53 mutant, etc.
  4. Transgenics: WAP Int3, BRCA1, etc.

- Initially, MAExplorer was a Java applet for analyzing MGAP arrays   **http://www.lecb.ncifcrf.gov/mae**

- MAExplorer identified genes in preferentially expressed during lactation, ***Nucleic Acids Res.*** **(2000) 28:4452**

Mammary Geneome Anatomy Program
MAExplorer http://www.lecb.ncifcrf.gov/mae

MicroArray Explorer- MAExplorer - Netscape

File Edit View Go Communicator Help

Bookmarks   Location: http://www-lecb.ncifcrf.gov/mae/   What's Related

MAExplorer

MGAP Introduction
MAExplorer Startup

Hybridizations

Startup DBs
Public databases (Click Once!)
• Preg. vs Lact.   (large font)

• C57B6 all stages
• C57B6 Vir. vs Preg.
• C57B6 Prev. vs Lact.
• C57B6 Lact. vs Invol.
• Stat5a KO Preg. vs Lact.
• C57 vs Stat5a KO - Preg+Lact
• C57B6 vs Stat5a KO - Preg.
• βB-inhib. KO vs C57 Vir+Lact
• βB-inhib. KO vs C57 Vir.
• βB-inhib. KO vs C57 Lact.
• C/EBP-βK.O. Vir+Preg+Lact
• C/EBP-βK.O. vs C57 V+P+L
• C/EBP-βK.O. vs C57 V+P+L
• Rehybridization repro (4)
• Tumor models
• C57 dev vs models, V/norm
• C57 dev vs models, P/norm
• LARGE set of 31 public HPs
• LARGE set of 38 public HPs

• No initial probes

Collaborator DBs

Custom DBs

Short Tutorial

MAExplorer
• Manual - with TOC
• Reference manual
• Advanced tutorial
• Menu summary

MGAP - MicroArray Explorer

The MAExplorer is an exploratory data analysis facility for cDNA microarrays from mouse mammary tissue and databases from the Mammary Genome Anatomy Project (MGAP).

MGAP MicroArrays used to Profile Gene Expression Patterns in Mammary Tissue

The MGAP microarray database provides access to microarrays which have been used to profile gene expression patterns in normal mammary tissue from different stages of development and neoplasia.
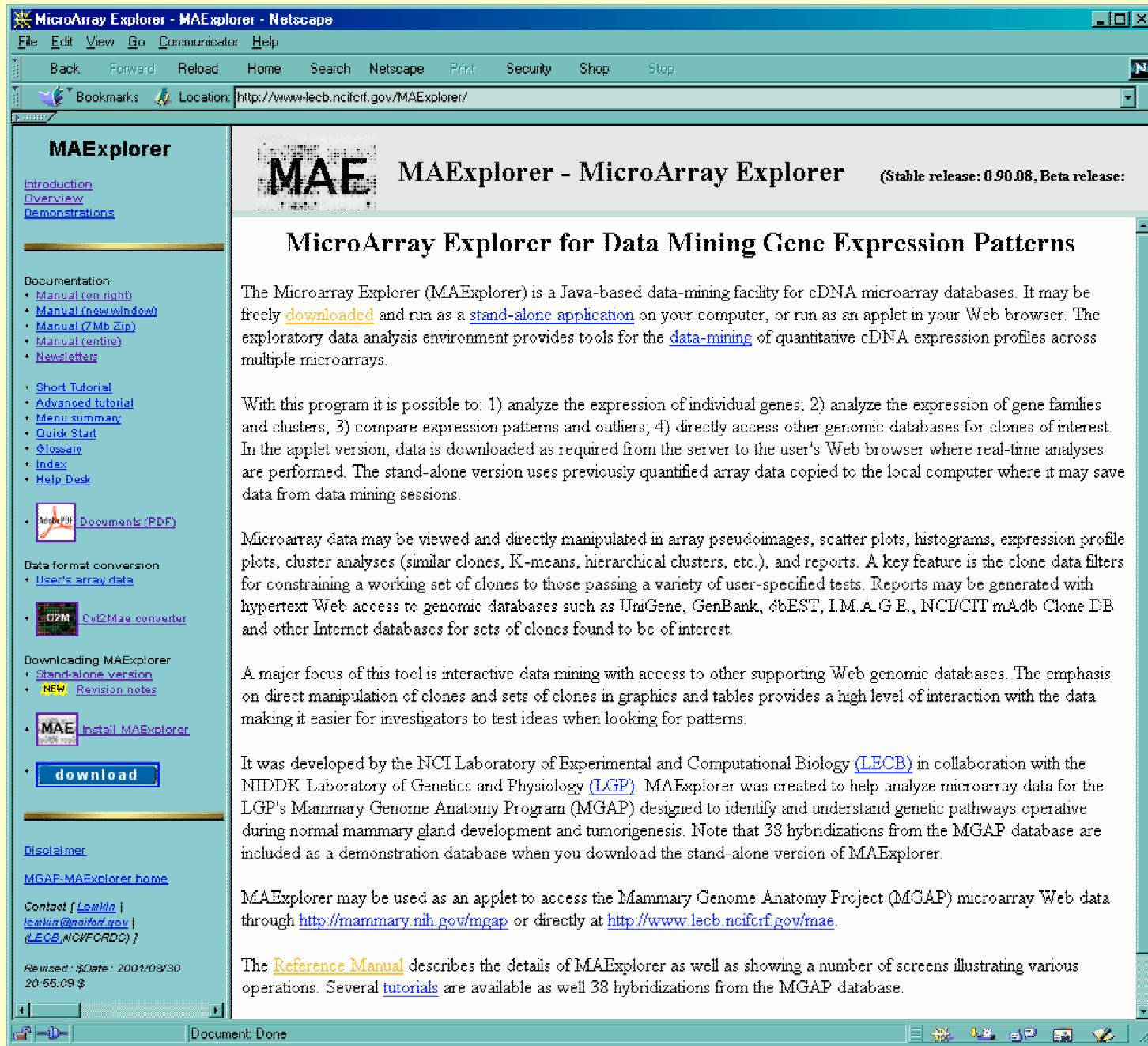
The Laboratory of Genetics and Physiology (LGP) has established the Mammary Genome Anatomy Program (MGAP) designed to identify and understand genetic pathways operative during normal mammary gland development and tumorigenesis. One arm of this program focuses on the use of cDNA microarrays to profile gene expression patterns. For this purpose, cDNA (EST) libraries are generated, sequenced and clone inserts are spotted on nylon membranes (by Research Genetics). These membranes are used to monitor expression profiles under various physiological conditions. At this point expression profiles have been obtained from several stages of normal mammary gland development and different tumor models. Access to these data and the MicroArray Explorer tools is granted to the scientific community

The cDNA library Technology

EST (cDNA) libraries of normalized cDNA are generated from mammary tissue at different stages of development and from different transgenic mouse tumor models. Currently a library from lactating mammary tissue from C57/B6 mice is available (info available from the Laboratory of Genetics and Physiology). More than 5000 clones from this library have been sequenced (Genbank).

Document: Done

MAExplorer Home Page
http://www.lecb.ncifcrf.gov/MAExplorer

MAExplorer

Introduction
Overview
Demonstrations

Documentation
- Manual (on right)
- Manual (new window)
- Manual (7Mb Zip)
- Manual (entire)
- Newsletters

- Short Tutorial
- Advanced tutorial
- Menu summary
- Quick Start
- Glossary
- Index
- Help Desk

- Documents (PDF)

Data format conversion
- User's array data

- Cvt2Mae converter

Downloading MAExplorer
- Stand-alone version
- NEW Revision notes

- Install MAExplorer

download

Disclaimer

MGAP-MAExplorer home

Contact [Lemkin]
lemkin@ncifcrf.gov]
(LECB, NCI/FCRDC)]

Revised : $Date : 2001/08/30
20:55:09 $

MAE   **MAExplorer – MicroArray Explorer**   (Stable release: 0.90.08, Beta release:

MicroArray Explorer for Data Mining Gene Expression Patterns

The Microarray Explorer (MAExplorer) is a Java-based data-mining facility for cDNA microarray databases. It may be freely downloaded and run as a stand-alone application on your computer, or run as an applet in your Web browser. The exploratory data analysis environment provides tools for the data-mining of quantitative cDNA expression profiles across multiple microarrays.

With this program it is possible to: 1) analyze the expression of individual genes; 2) analyze the expression of gene families and clusters; 3) compare expression patterns and outliers; 4) directly access other genomic databases for clones of interest. In the applet version, data is downloaded as required from the server to the user's Web browser where real-time analyses are performed. The stand-alone version uses previously quantified array data copied to the local computer where it may save data from data mining sessions.

Microarray data may be viewed and directly manipulated in array pseudoimages, scatter plots, histograms, expression profile plots, cluster analyses (similar clones, K-means, hierarchical clusters, etc.), and reports. A key feature is the clone data filters for constraining a working set of clones to those passing a variety of user-specified tests. Reports may be generated with hypertext Web access to genomic databases such as UniGene, GenBank, dbEST, I.M.A.G.E., NCI/CIT mAdb Clone DB and other Internet databases for sets of clones found to be of interest.

A major focus of this tool is interactive data mining with access to other supporting Web genomic databases. The emphasis on direct manipulation of clones and sets of clones in graphics and tables provides a high level of interaction with the data making it easier for investigators to test ideas when looking for patterns.

It was developed by the NCI Laboratory of Experimental and Computational Biology (LECB) in collaboration with the NIDDK Laboratory of Genetics and Physiology (LGP). MAExplorer was created to help analyze microarray data for the LGP's Mammary Genome Anatomy Program (MGAP) designed to identify and understand genetic pathways operative during normal mammary gland development and tumorigenesis. Note that 38 hybridizations from the MGAP database are included as a demonstration database when you download the stand-alone version of MAExplorer.

MAExplorer may be used as an applet to access the Mammary Genome Anatomy Project (MGAP) microarray Web data through http://mammary.nih.gov/mgap or directly at http://www.lecb.ncifcrf.gov/mae.

The Reference Manual describes the details of MAExplorer as well as showing a number of screens illustrating various operations. Several tutorials are available as well 38 hybridizations from the MGAP database.

MAExplorer GUI-based User Interface
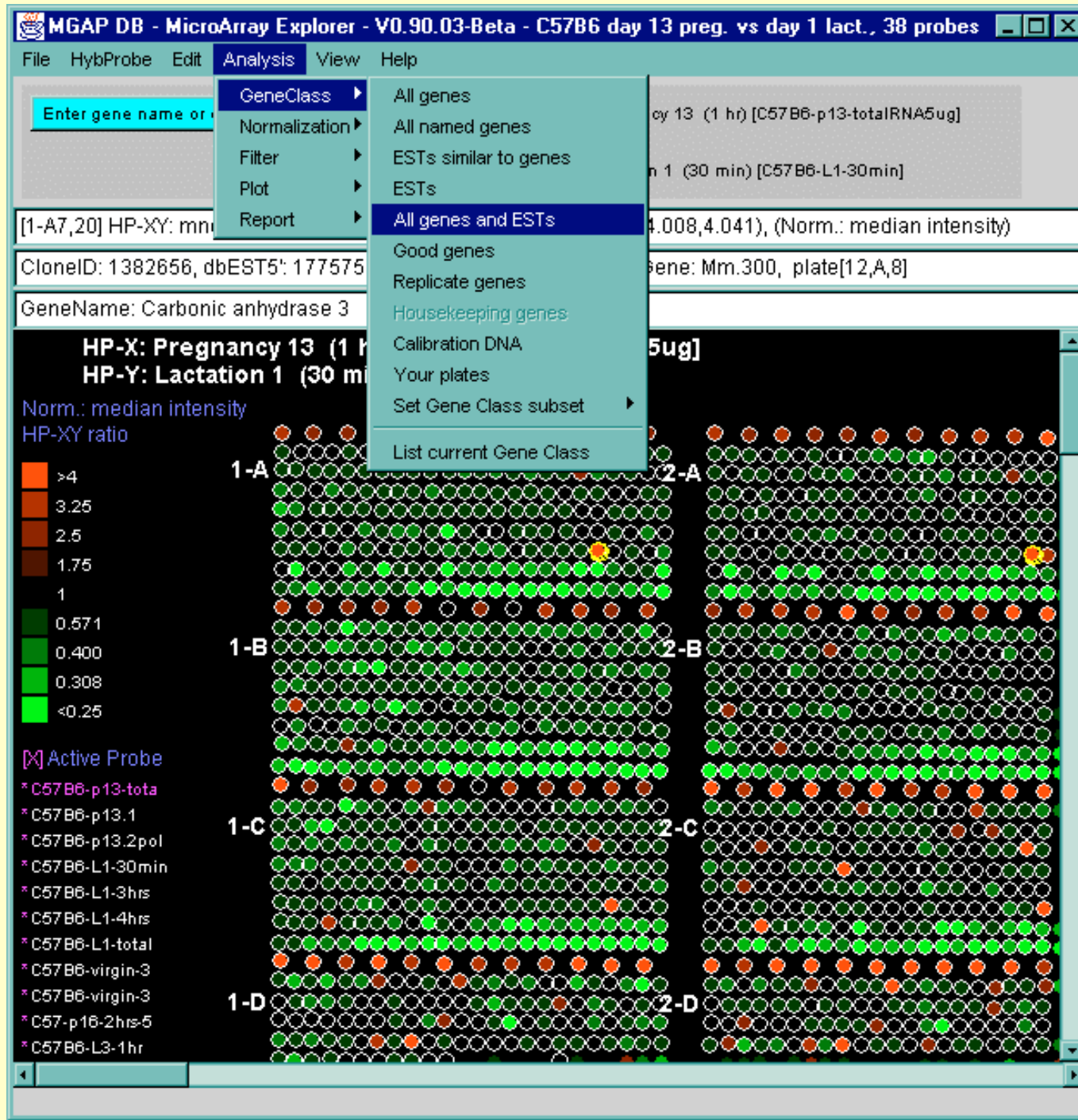
Data Sources

- Schema: tab-delimited experiment data files:
  1. GIPO (array print file)
  2. List of samples in DB
  3. Configuration
  4. Separate spot quantification data files

- Use with any experiment data that uses this schema

- Files are read from:
  1. Local disk for stand-alone application or
  2. Web server for Web-browser applet - optionally cached locally

- Used with data from NCI/CIT-ATC mAdb Web server

- The Cvt2Mae tool is used to convert generic data to this schema

Lists of Hybridized Probe (HP) Samples

- <u>Multiple-sample database</u> may consist of many conditions with replicate samples

- <u>Organize samples as lists of conditions:</u>
    1. X and Y pairs of individual samples
    2. Sets of X and Y replicates (X-set, Y-set)
    3. Ordered expression profile list (E-list)

- Manipulate <u>named lists</u> of sample conditions

- <u>Save and restore named sample lists</u> during analysis and between sessions

Operations on 2- & N-conditions of Samples[1]

Set of **HP-X** replicate samples	Set of **HP-Y** replicate samples

Operations on <u>sets of replicate samples</u>- e.g. *t*-test, CV statistical tests

Order list of **HP-E** samples

Operations on <u>order lists of samples</u>: e.g. clustering, Expression Profile plots

[1] "HP" is hybridization probe i.e. the sample hybridized to the chip

Choosing HP-X, HP-Y Sets and HP-E Lists

Gene Data Filter is Intersection of Tests

Gene subset tests
*Gene class
*Edited Gene List
*Good genes
*User named gene set

Spot Intensity and ratio tests
*Ratio or Zdiff range
*Intensity range
*Ratio histogram bin
*Intensity histogram bin
*Highest/Lowest ratios
*Abs |X-Y| difference

Statistical tests
*Coeff. of Variation
* t-Test of X vs Y sets
*Clustering distances
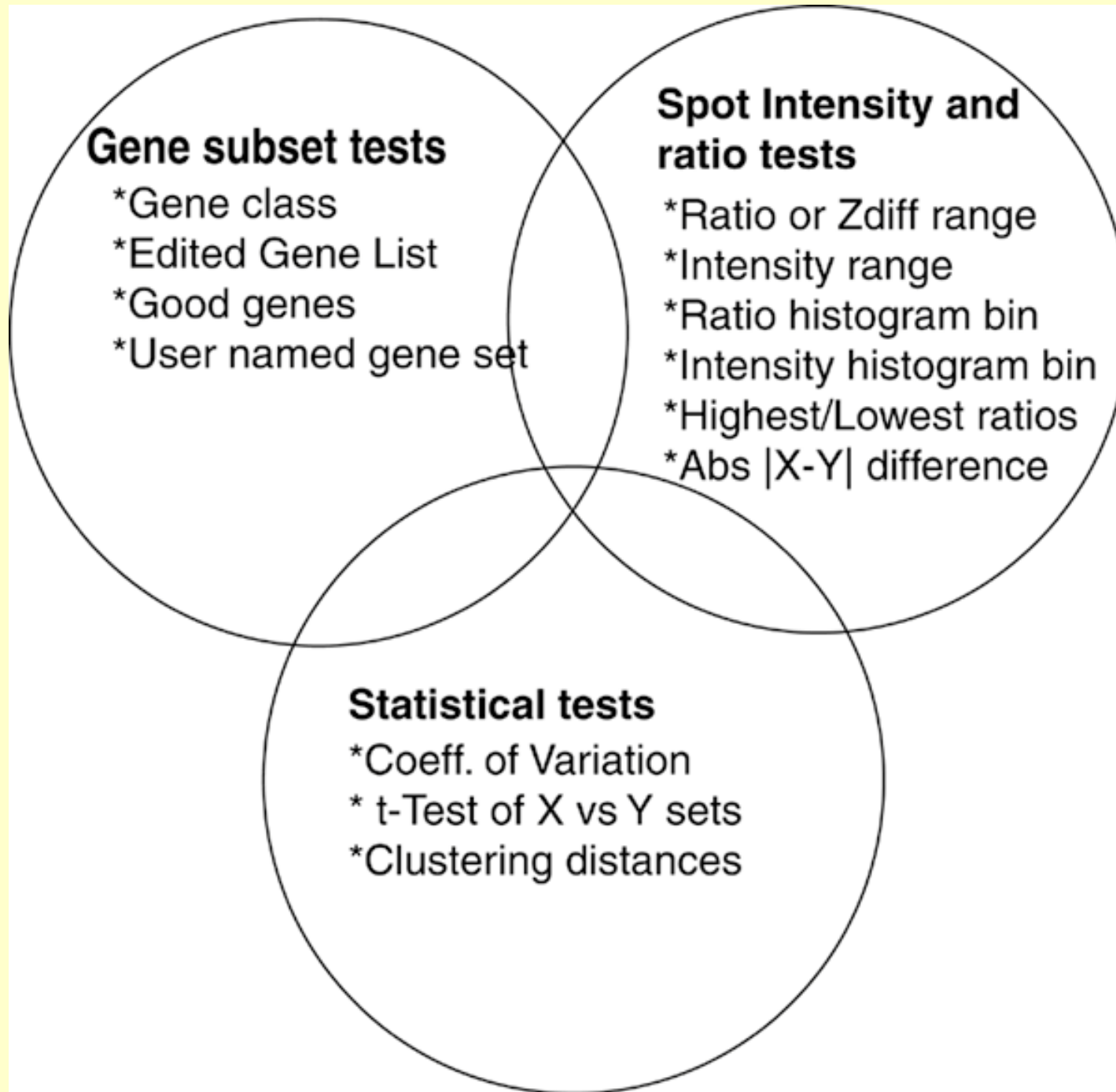
MAExplorer User Interface - Filter Menu

Plots

Plots allow visualization and direct manipulation of gene data

1. <u>Pseudo array image</u> - intensity, ratio (X/Y)

2. <u>Zoomable scatter plots</u> - X vs Y, Cy3 vs Cy5, duplicate spots

3. <u>Histograms</u> - ratio and intensity

4. <u>Expression profiles</u> - individual genes and overlay plots

5. <u>Silhouette plots</u> - similarity clusters, K-means clustering

6. <u>Hierarchical clustering</u> - clustergram, dendrogram

Scatter Plots of Two Conditions

- X-Y scatter plot of two 13-day pregnancy samples: C57B6 vs Stat5a (-,-) [MGAP]
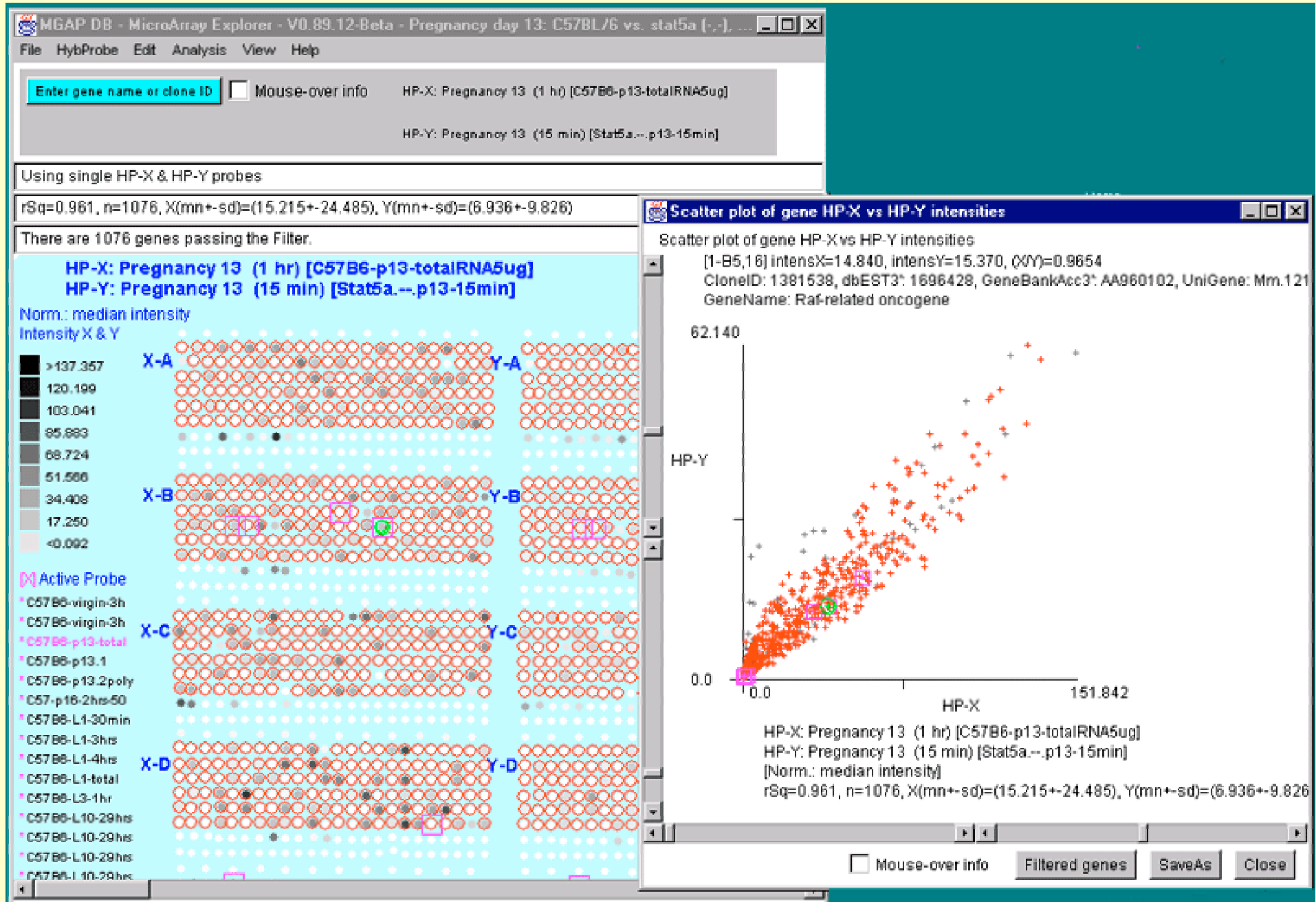
Zoomed X-Y Scatter Plot

- Zoomed-in on Raf-related oncogene (green circle) using scrollbars
- Genes <u>not</u> passing Filter may be grayed out in the plot
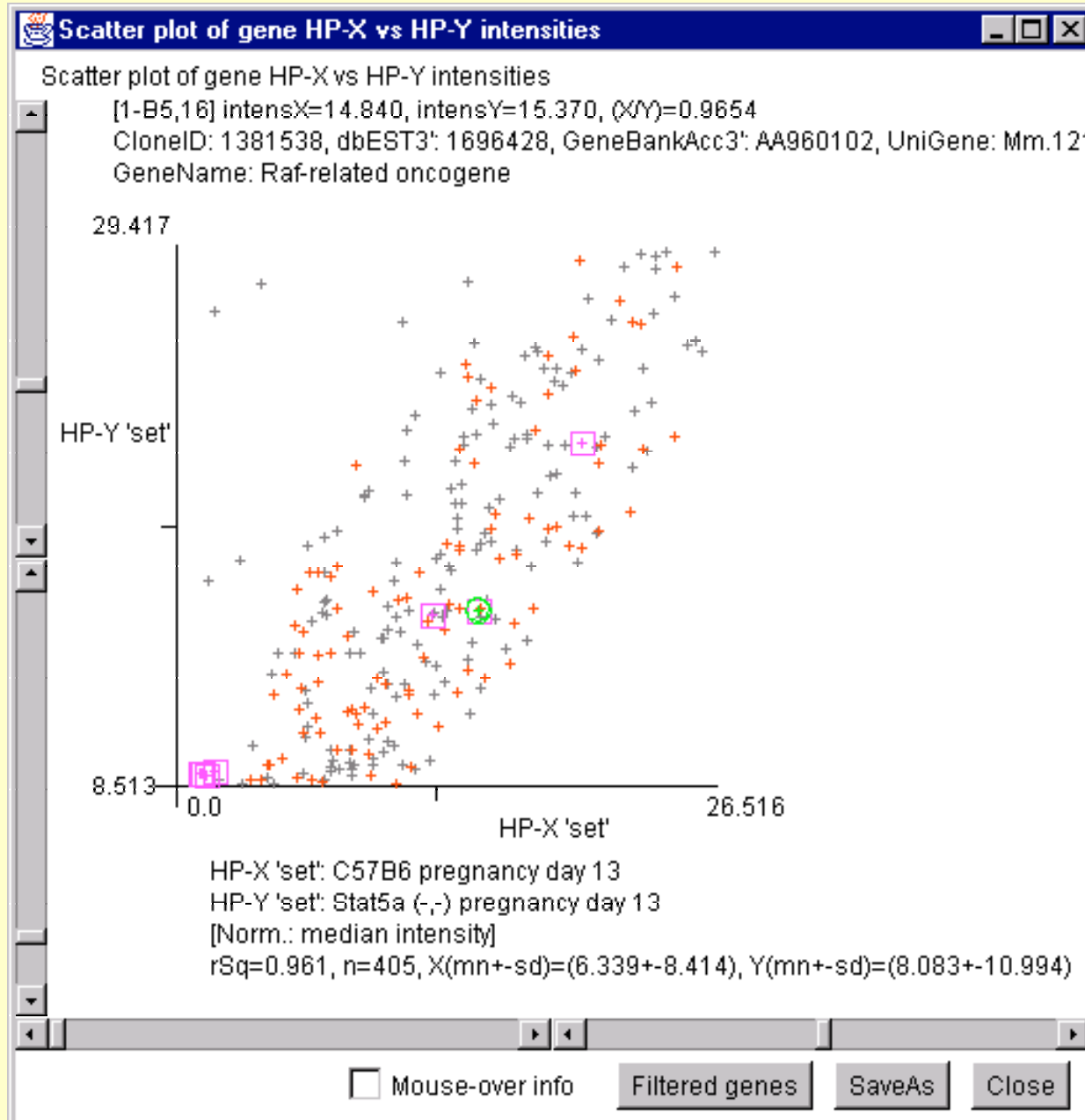
Genes Filtered by Intensity-Histogram Bin
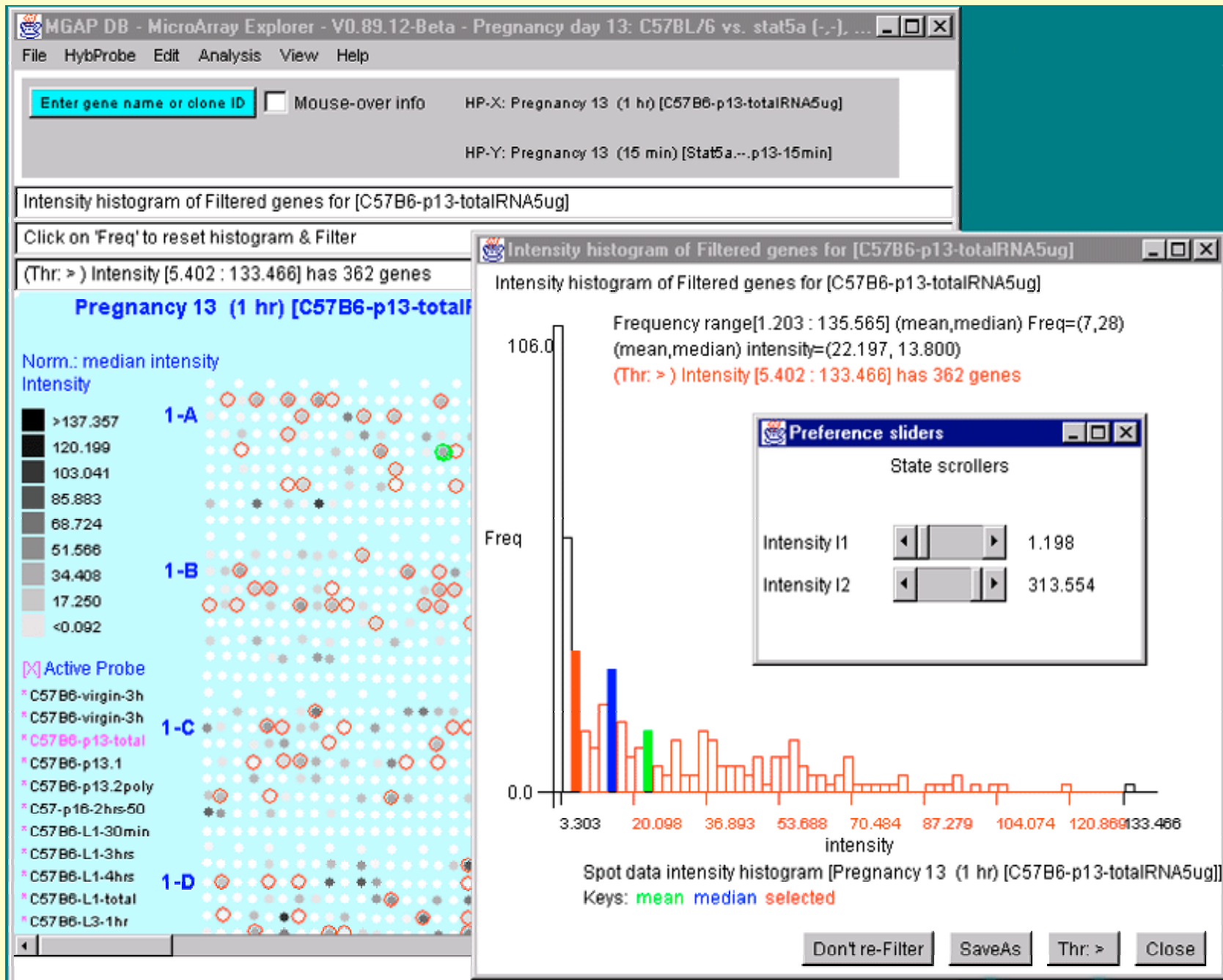
Expression Profile Plots of N-conditions

Cluster Genes With Similar Expression

K-means Clustering of Expression Data

List of Expression Profiles Plots of Clusters

- Scrollable list of EP plots showing genes from clusters #1, #2, #3
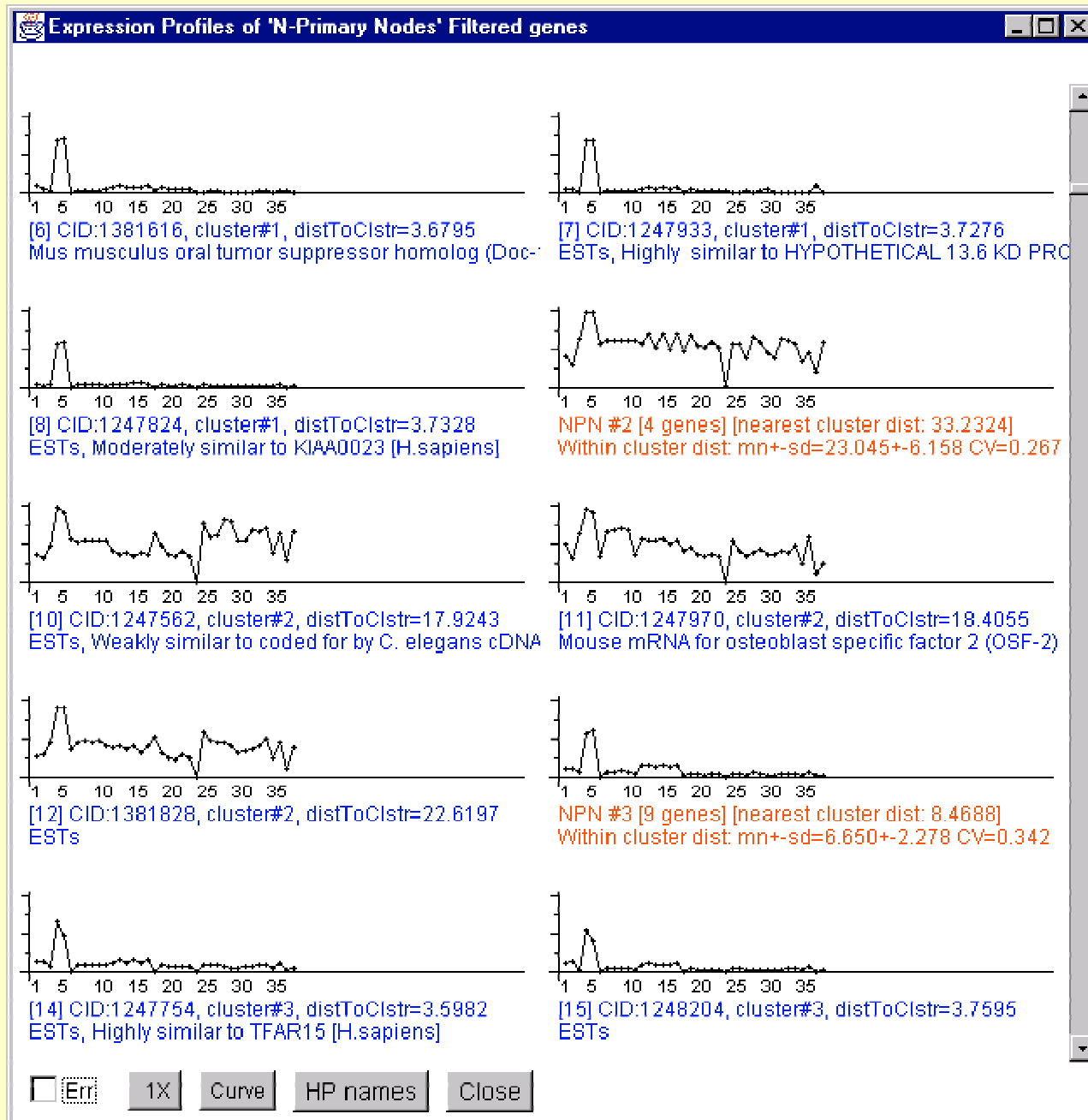
Mean Expression Profiles of K-means Clusters

Mean Expression Profiles of 'N-Primary Nodes' clusters
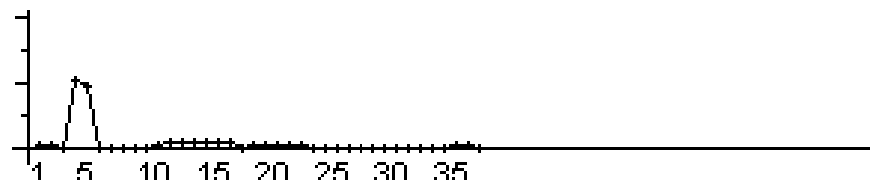
NPN #1 [8 genes] [nearest cluster dist: 1.8891]
Within cluster dist: mn+-sd=2.152+-1.254 CV=0.582
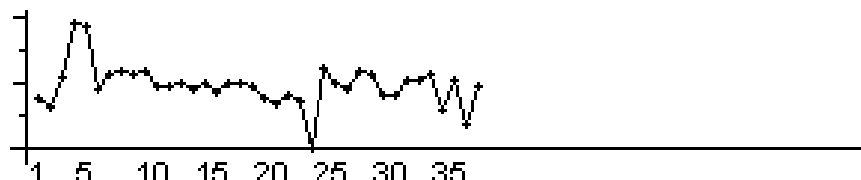
NPN #2 [4 genes] [nearest cluster dist: 33.2324]
Within cluster dist: mn+-sd=23.045+-6.158 CV=0.267
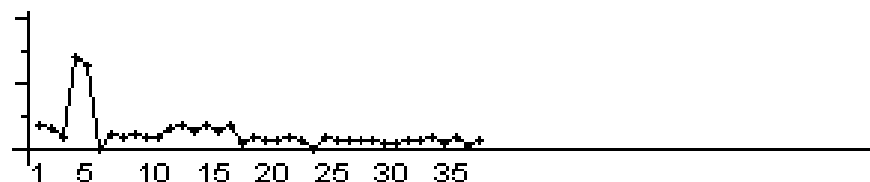
NPN #3 [9 genes] [nearest cluster dist: 8.4688]
Within cluster dist: mn+-sd=6.650+-2.278 CV=0.342
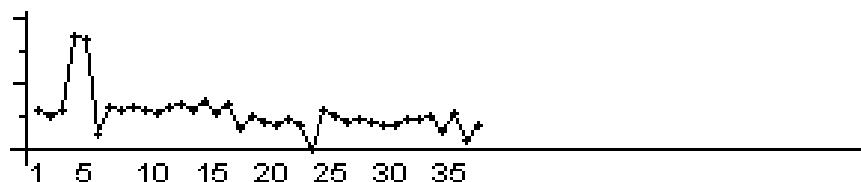
NPN #4 [25 genes] [nearest cluster dist: 22.5600]
Within cluster dist: mn+-sd=11.919+-4.503 CV=0.377
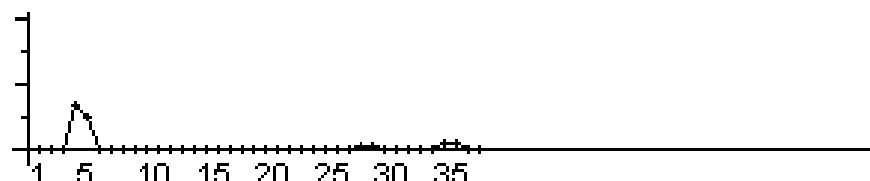
NPN #5 [14 genes] [nearest cluster dist: 0.2357]
Within cluster dist: mn+-sd=1.686+-2.751 CV=1.631
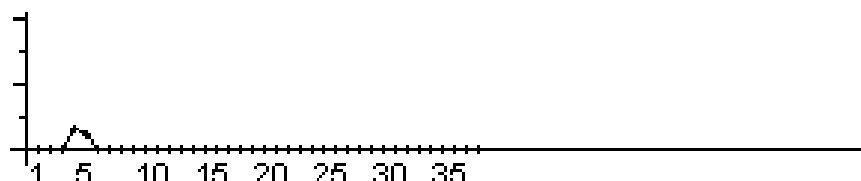
NPN #6 [13 genes] [nearest cluster dist: 0.2357]
Within cluster dist: mn+-sd=0.223+-0.081 CV=0.365

☐ Err   1X   Curve   HP names   Close

Hierarchical Clustering of Expression Data
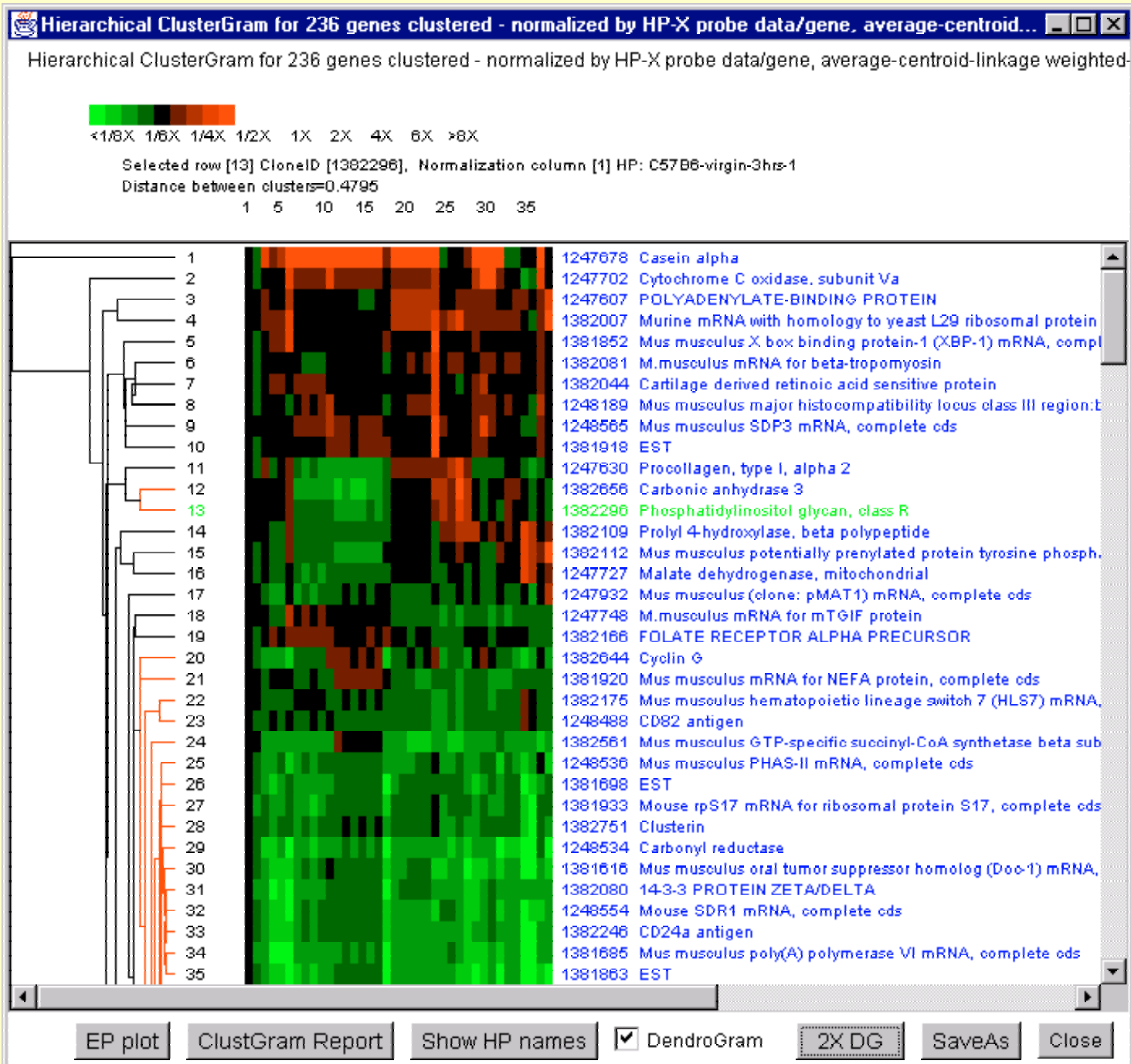
Reports

- Data reported as:
    1. Web-accessible <u>dynamic spreadsheets</u> or
    2. <u>tab-delimited text</u> exportable to Excel

- <u>Gene set</u> reports - linked to UniGene, GenBank, mAdb, LocusLink, etc.

- <u>Array sample</u> reports - linked to histology and model Web pages

- <u>Pop-up Web browser</u> on specific data from dynamic reports or plots

Gene Set Operations Help Manage Data and Search Results

- <u>All gene sets are named</u> with a directory of existing sets

- <u>Set operations</u> (AND, OR, DIFFERENCE) may be used to create new derived sets

- Special sets:
  1. <u>Filtered genes set</u>   holds genes passing the data filter
  2. <u>Edited Gene List</u>   holds results of clustering or editing
  3. <u>Normalization set</u>   may be used as normalization method
  4. <u>User data filter set</u>   may be used as a data filter

- Sets are <u>saved</u> when the session is saved, <u>restored</u> when MAExplorer restarted

Find Gene or Gene Subset by <u>Name Guesser</u>

Gene Set Operations - e.g. 'AND' of Two Sets

Scrollable Dynamic Gene Reports

Scrollable Dynamic Gene Reports

Correlation Report of all Samples

- Sample vs Sample correlation coefficient report for current set of Filtered genes

HP vs. HP correlation coefficients table, Pregnancy 13 days: C57BL/6 vs. stat5a (-,-), 8 probes

HP vs. HP correlation coefficients table, Pregnancy 13 days: C57BL/6 vs. stat5a (-,-), 8 probes

D2 | rSq=0.982, n=405, HP:2(mn+-sd)=(1+-0), HP:3(mn+-sd)=(1+-1)

		C57B6-p13-totalRNA8	C57B6-p13.1	C57B6-p13.2poly-A	Stat5a.--.p13-15min	Stat5a.--.p13-15min2
1	C57B6-p13-totalRNA8	-	rSq=0.715, n=405, HF	rSq=0.729, n=405, HF	rSq=0.953, n=405, HF	rSq=0.958, n=405, HF
2	C57B6-p13.1	-	-	rSq=0.982, n=405, HF	rSq=0.756, n=405, HF	rSq=0.757, n=405, HF
3	C57B6-p13.2poly-A	-	-	-	rSq=0.772, n=405, HF	rSq=0.773, n=405, HF
4	Stat5a.--.p13-15min	-	-	-	-	rSq=0.997, n=405, HF
5	Stat5a.--.p13-15min2	-	-	-	-	-
6	Stat5a.--.p13-1hr2	-	-	-	-	-
7	Stat5a.--.p13-30min	-	-	-	-	-
8	Stat5a.--.p13-30min2	-	-	-	-	-
9	-	-	-	-	-	-

Close

Tab-Delimited Reports Exportable to Excel

Plug-in Extensions to MAExplorer

- <u>Plug-ins</u> will allow investigators to extend capabilities of core MAExplorer program to other analysis methods

- Open Java API, Java examples, donated plug-ins & links to Plugins will be published on our Web site

- <u>Typical plug-ins</u>: normalization, Filters, PCA, clustering, Web-server functional analysis of cluster results, etc.

- <u>Plug-ins</u> will have three types of implementations:
  1. Using 100% Java code
  2. Accessing local programs written in any language
  3. Web-CGI or client-server to specialized genomic DBs

Data Format Conversion for MAExplorer

- <u>Cvt2Mae</u> reads a variety of array data - for both one-of-a kind academic and commercial arrays (eg. Affymetrix, Incyte, etc).

- Users can create an <u>Array Layout</u> description that may be used in subsequent conversions

- Will add generic "MAML" (now "MGED") standard compatibility (MicroArray XML) for data portability

- Cvt2Mae is undergoing *Beta*-testing

Select a Chip Array-Layout

Converted Affymetrix files for Analysis

Affymetrix Data from Cvt2Mae Conversion

Converted Incyte files Ready to Analyze

Cvt2Mae: convert array data to MAExplorer files - Version: 08-21-2001 (Beta)

Enter data for steps 1, 2, and 3. Then press 'Run' to convert your data to MAExplorer format.

1. Select Chipset: Affymetrix - Human

1.1 (opt.) Select Quant. software used: [NONE] pick array data ()

2. Select Input Data Files: Browse file name | Browse folder

U937-Affymetrx-2subclones.txt

2.1 Samples to use '<<file>> sample name'. Remove sample | Rename sample

<<C:\Temp\AffyData\U937-Affymetrx-2subclones.txt>> [DSD-1-Tel10A-1-U95A]
<<C:\Temp\AffyData\U937-Affymetrx-2subclones.txt>> [DSD-1-Tel10B-1-U95A]
<<C:\Temp\AffyData\U937-Affymetrx-2subclones.txt>> [DSD-1-Tel17A-1-U95A]
<<C:\Temp\AffyData\U937-Affymetrx-2subclones.txt>> [DSD-1-Tel17B-1-U95A]

Vendor	Affymetrix
Layout name	Affymetrix - Human
Species	Human
Spots/microarray	12630

3. Select Project Output Folder: Create New project folder

Project output folder: C:\Temp\junk\

MAExplorer startup File: C:\Temp\junk\MAE\Start.mae

4. Edit and Run Run | Edit Layout | Save Layout | Assign GIPO fields | Assign Quant fields | Done | Reset

☐ Expert assign-mode

Status: ===> Finished writing out data files. Press 'Done' to exit

To start MAExplorer, go to project folder & click on Start.mae.

Summary

- MAExplorer is a flexible microarray data-mining tool running on the user's computer

- Uses direct-manipulation, data filtering, built-in graphics, statistics, clustering, gene and sample set operations

- Manages multiple samples, replicates, sets, expression profile lists where the session state may be saved on the disk for later use

- The Cvt2Mae data conversion tools allows use with common chips

- Plug-ins will allow extension with new analytic methods by users

- MAExplorer identified genes in preferentially expressed during lactation, *Nucleic Acids Res.* (2000) **28**:4452

- Freely available for download with documentation on Web site