

# MicroArray Explorer - a Tool for Data Mining of cDNA Microarrays: Overview

<http://www.lecb.ncifcrf.gov/MAExplorer>

PF Lemkin<sup>1</sup>, GC Thornwall<sup>2</sup>, K Walton<sup>3</sup>, L Hennighausen<sup>3</sup>

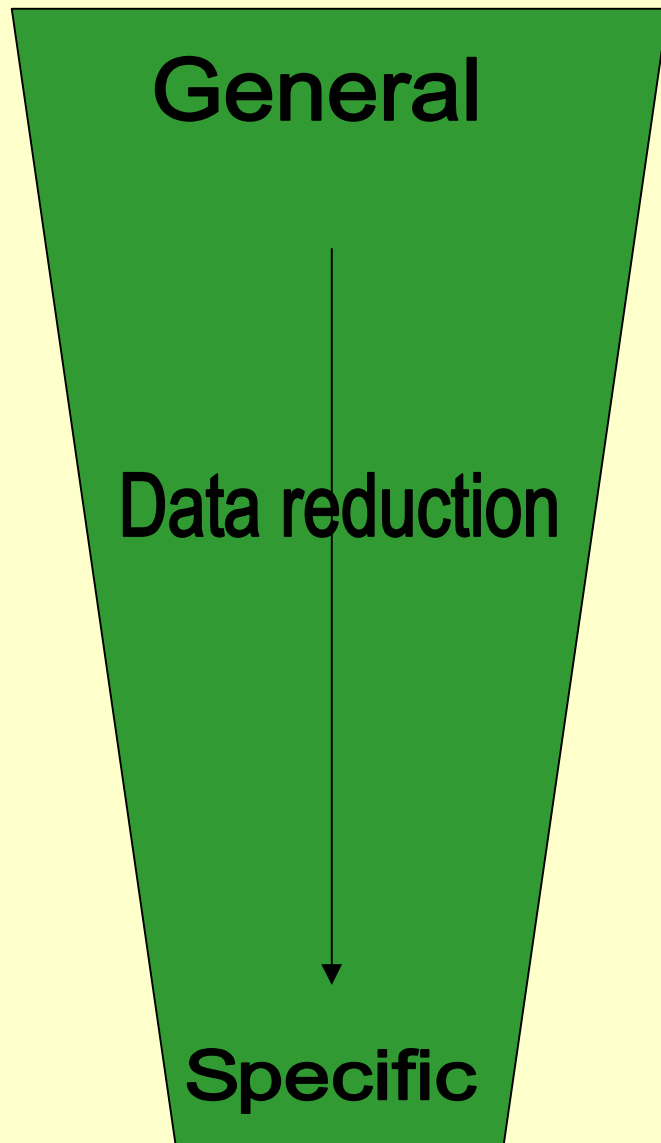
(1) LECB, NCI/FCRDC (2) SAIC/FCRDC (3) LGP, NIDDK

mail: lemkin@ncifcrf.gov

# I. MAExplorer Overview

- MAExplorer is a Java-based data-mining tool for analyzing microarrays
- Java provides real time response required for data-mining
- Runs either as a stand-alone application or Web-browser applet
- Stand-alone installers are available on the Web site for Windows 95/98/NT/2000, MacOS, Solaris, Linux, Unix, other Java compatible systems
- Initially developed for Mammary Genome Anatomy Program for spotted membranes, <http://www-lecb.ncifcrf.gov/mae>
- Extended for different array substrates, geometries,  $^{33}\text{P}$  or Cy3/Cy5 spot-labeling, and scanners using configuration files

# Data Mining - Finding Putative Relevant Patterns



Quantified array spot data for multiple samples

Organize by sample, gene expr, gene sets

Change views: normalization, data filters

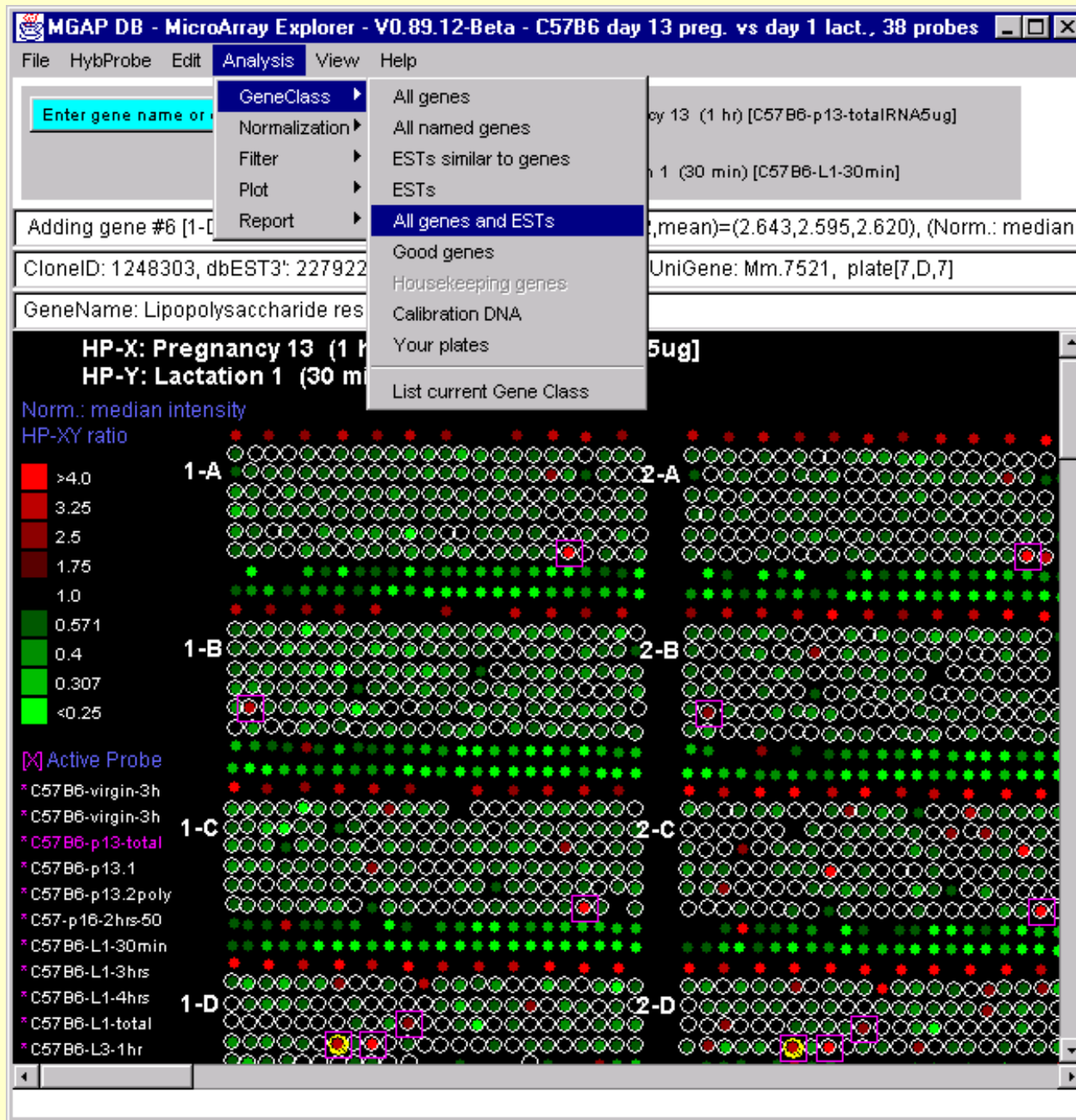
Visualize and query: plots, cluster, reports

Explore external genomic databases

**Results**

# MAExplorer User Interface

- Showing named genes and ESTs from the MGAP database



# MAExplorer Overview - continued 2

- Helps analyze genes and gene family expression patterns for multiple samples
- Samples organization: X-Y paired samples, sets of X-Y replicate samples (X- and Y-sets), ordered expression profile list of samples (E-list)
- Data filters: use gene sets, spot intensity and ratio range, statistics and clustering to drill-down to subsets of genes of interest
- Generates plots: pseudo array image, scatter, histogram, expression profile, clustergram, dendrogram, silhouette plots

# Operations on 2-conditions & N-conditions

Set of HP-X replicate samples

Set of HP-Y replicate samples

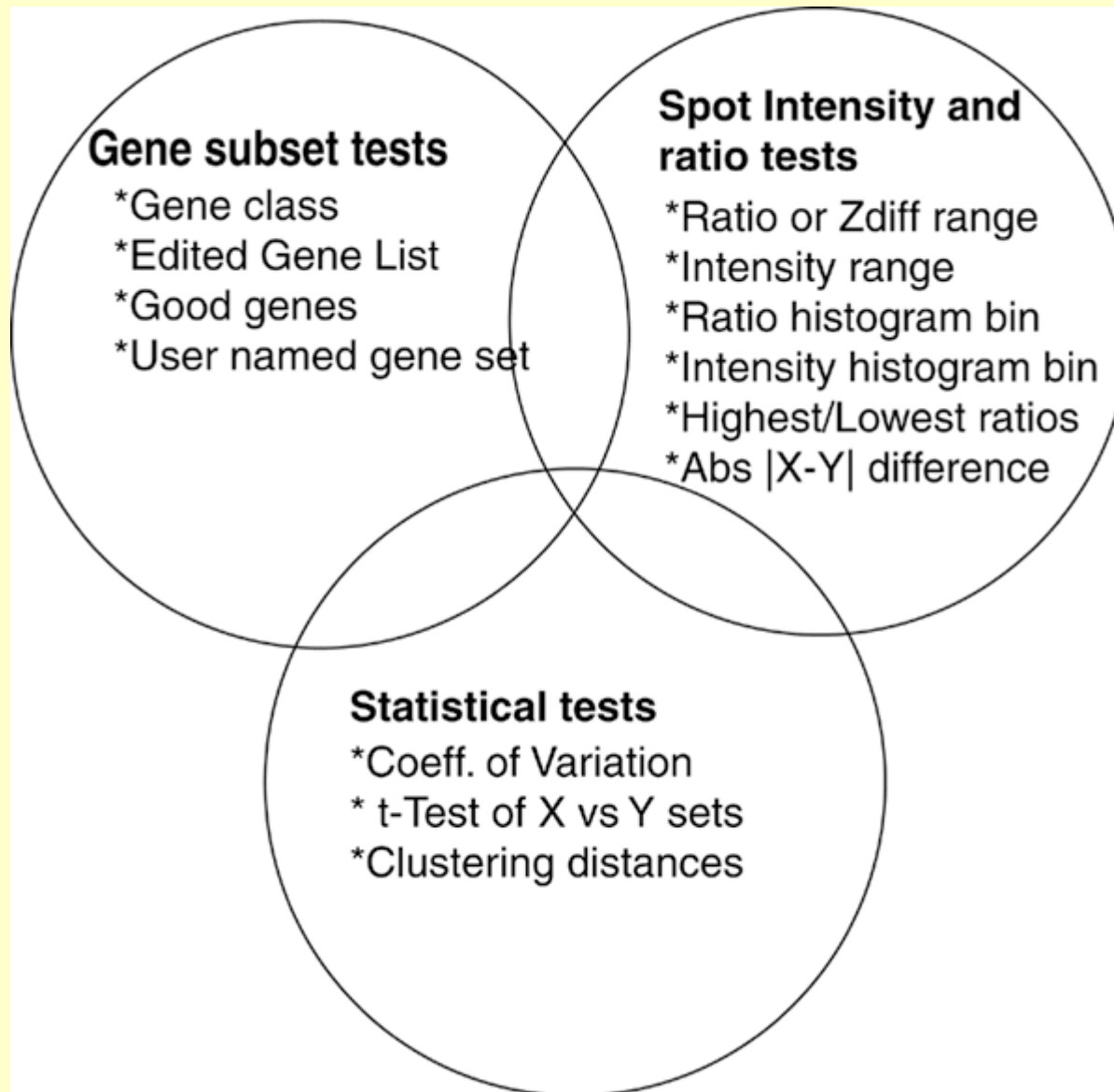
Operations on sets of replicates- e.g. *t*-test, CV

Order list of HP-E samples

Operations on order lists of samples: e.g. clustering, EP plots

# Gene Data Filter is Intersection of Tests

- Current set of genes is **intersection** of gene sets each passing selected filter tests
- Filtered gene subset is used as **pre-filter** for subsequent clustering, plots, and tables
- Changing any filter parameters causes the data filter to be re-computed



# MAExplorer Overview - continued 3

- Cluster methods:
  - a) similar genes
  - b) K-means
  - c) hierarchical clustering
- Reports: Web-accessible spreadsheets, or tab-delimited text exportable to Excel
- Direct manipulation of genes and gene sets, plots and reports, filter parameters, etc.
- Set operations on gene subsets and condition lists (samples) help manage search results
- Integrated Web browser access to public genomic, histology and model Web databases

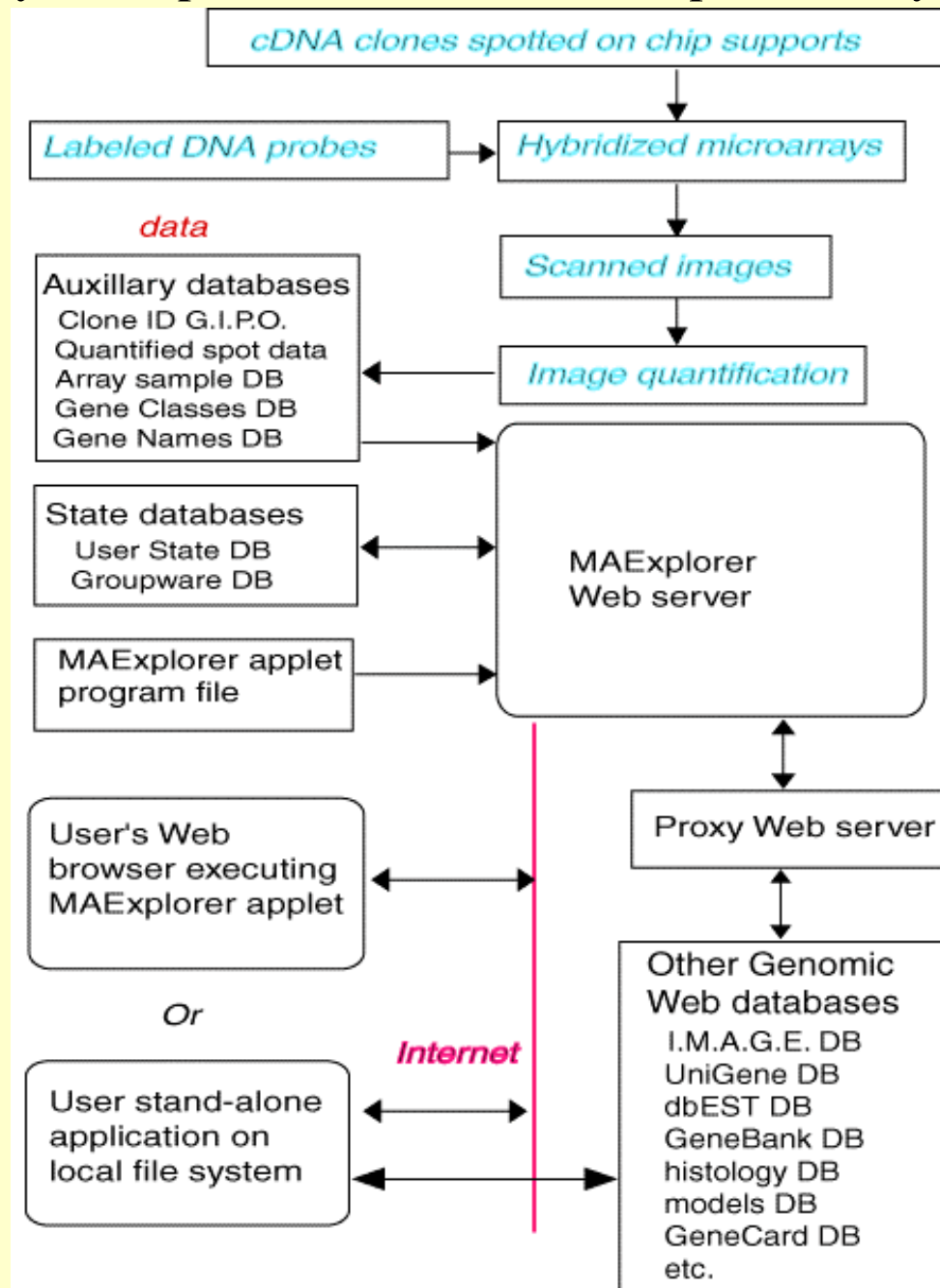


# MAExplorer Overview - continued 4

- Data is read from
  1. local files on a user's computer and/or
  2. downloaded from a Web array database server
- Data is cached on local computer so analysis can continue off-line
- Stand-alone version allows saving intermediate results and state of data-mining session to local disk
- Research groups could use MAExplorer to publish their array data on their own Web servers similar to MGAP MAExplorer server
- On-line documentation (manual, tutorials, etc.) available from Web site

# Relationship of MAExplorer to Internet

(Steps in cyan are performed before MAExplorer analysis.)



# Client-Centric Computations - Advantages/Disadvantages

- **Client-centric approach primarily using Java**
- a) + Java programs run (pretty much) on all operating system platforms as either stand-alone or applets (in browsers)
- b) + handles rapid response required for direct manipulation on new generation of very fast desktop computers
- c) + stand-alone version may be restarted quickly from data cached off-line
- d) + size limitations not a problem with stand-alone Java applications
- e) + Java plug-ins allows prototyping new local and Web DB analysis methods functionality by any group of users
- f) - slow startup for applet version because the program and all data has to be downloaded each time it is run
- g) - difficult to build large stable Web-applets handling very large data sets
- h) - applications must be installed on clients computer where they may be some level of incompatibility

# Server-Centric (CGI or Applet) Computations - Advantages/Disadvantages

- **Server-centric approach using mix of HTML, CGI and Java**
- a) + may have better resources for very large data sets but with dependence on server
- 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
- c) + may be easier to prototype and distribute new functionality using 3rd party software such as RDBMS, S-plus, etc. using centralized CGI or servlets where only one copy is required on the server
- d) - susceptible to Internet traffic bandwidth problems and server-load dependencies
- e) - difficult to get very rapid response required for direct manipulation if all computations done on the server

# Expression Data Used in MAExplorer

- **Database configuration data table** for specific array layout and content \*\*
- **Hybridized array samples table** describing their experimental conditions \*\*
- **Gene-In-Plate-Order table** (print table) listing Clone Ids, gene names, genomic DB Ids, spot and source plate coordinates \*\*
- **Quantified array spot data table** for samples from quantification software such as GenePix<sup>TM</sup>, Molecular Dynamics' ImageQuant<sup>TM</sup>, Research Genetics' Pathways<sup>TM</sup>, etc. \*\*
- **Data is optionally cached** from a microarray Web database server data to the local computer. Future analysis of this data is then independent of the Web database server
- **External Web genomic databases** corresponding to probes and Clone IDs are accessed as needed: I.M.A.G.E, GeneBank, dbEST, UniGene, NCI/CIT mAdb Clone DB, GeneCard, MGAP histology and model DBs, etc.

\*\* Auxiliary data required for MAExplorer is indicated in blue

# MAExplorer Home Page

<http://www.lecb.ncifcrf.gov/MAExplorer>

The screenshot shows a Netscape browser window titled "MicroArray Explorer - MAExplorer - Netscape". The address bar is empty. The browser's menu bar includes "File", "Edit", "View", "Go", "Communicator", and "Help". The toolbar contains "Back", "Forward", "Reload", "Home", "Search", "Netscape", "Print", "Security", "Shop", and "Stop".

The main content area features a logo for "MAE" (MicroArray Explorer) and the title "MAExplorer - MicroArray Explorer". Below this is a large heading: "MicroArray Explorer for Data Mining Gene Expression Patterns".

The text describes the tool: "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."

It lists capabilities: "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."

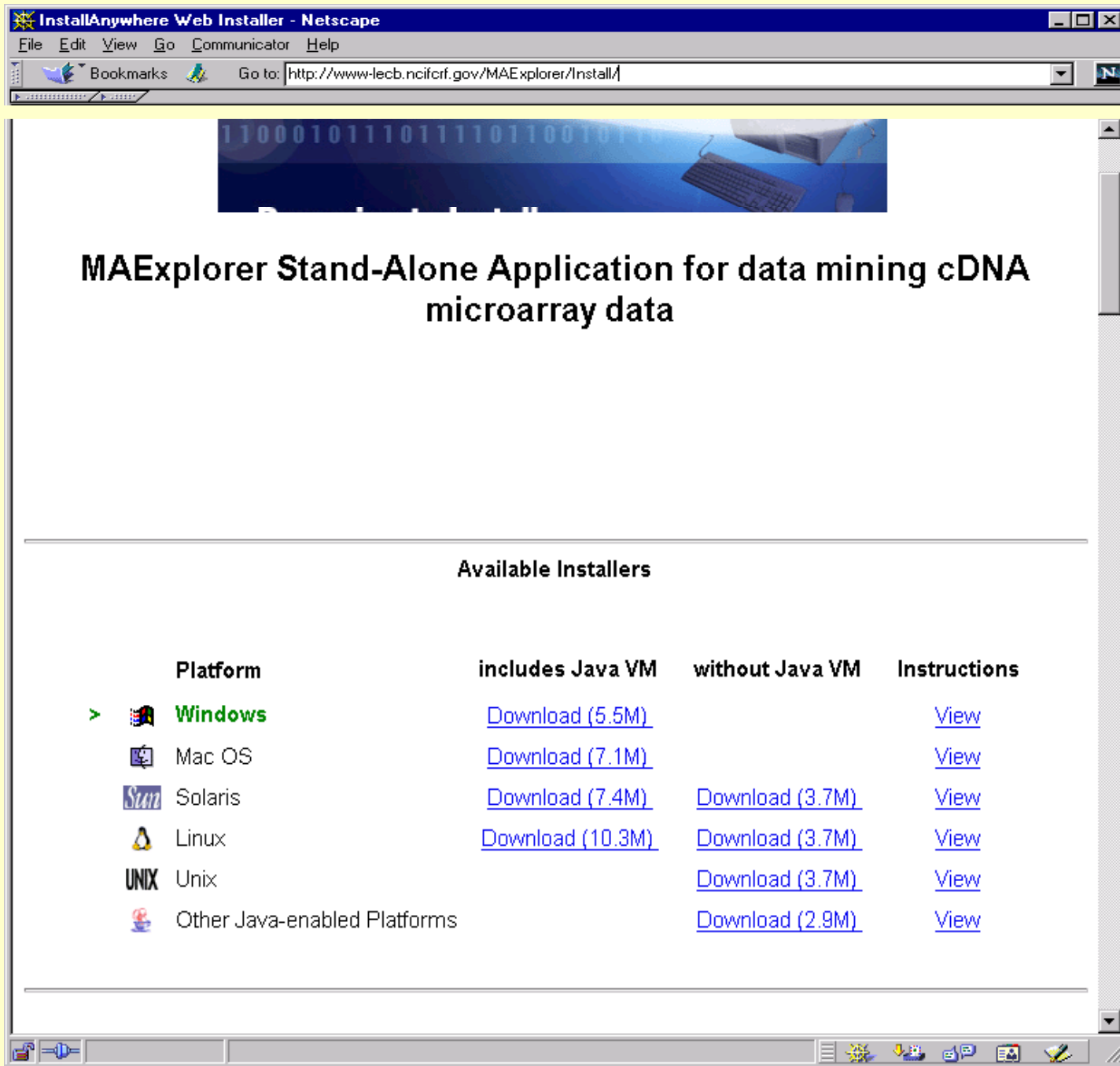
It also describes the data visualization and analysis: "Microarray data may be viewed and directly manipulated in array pseudoimages, scatter plots, histograms, expression profile plots, cluster analyses (similar clones, K-means, hierarchical clusters, etc.), and reports. A key feature is the clone data 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."

Finally, it states: "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."







The left sidebar contains a navigation menu for "MAExplorer" with links for "Introduction" and "Demonstrations". Under "Documentation", there are links for "Manual (on right)", "Manual (new window)", "Manual (>6Mb Zip)", "Manual (entire)", and "Newsletters". Other links include "Short Tutorial", "Advanced tutorial", "Menu summary", "Quick Start", "Glossary", "Index", and "Help Desk". Under "Downloading", there are links for "User's array data", "Stand-alone version", "NEW Revision notes", and "Installer information". A prominent blue "download" button is located at the bottom of the sidebar.

The status bar at the bottom of the browser window shows "Document: Done" and a taskbar with various system icons.

# Download Stand-alone version Web page



The screenshot shows a Netscape browser window titled "InstallAnywhere Web Installer - Netscape". The address bar contains the URL "http://www-lecb.ncicrf.gov/MAExplorer/Install/". The main content area features a banner with binary code and a computer keyboard, followed by the heading "MAExplorer Stand-Alone Application for data mining cDNA microarray data". Below this is a section titled "Available Installers" which contains a table of download links for various operating systems.

Platform	includes Java VM	without Java VM	Instructions
>  <b>Windows</b>	<a href="#">Download (5.5M)</a>		<a href="#">View</a>
 Mac OS	<a href="#">Download (7.1M)</a>		<a href="#">View</a>
 Solaris	<a href="#">Download (7.4M)</a>	<a href="#">Download (3.7M)</a>	<a href="#">View</a>
 Linux	<a href="#">Download (10.3M)</a>	<a href="#">Download (3.7M)</a>	<a href="#">View</a>
 Unix		<a href="#">Download (3.7M)</a>	<a href="#">View</a>
 Other Java-enabled Platforms		<a href="#">Download (2.9M)</a>	<a href="#">View</a>

# Mammary Genome Anatomy Program DB page MAExplorer <http://www.lecb.ncifcrf.gov/mae>

**MAExplorer**

[MGAP Introduction](#)  
[MAExplorer Startup](#)

[Hybridizations](#)

[Startup DBs](#)

[Public databases \(Click Once\)](#)

- [Freq. vs Lact.](#) ( [large font](#) )

- [C57B6 all stages](#)
- [C57B6 Yr. 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.](#)
- [B-Inhib. KO vs C57 Yr+Lact](#)
- [B-Inhib. KO vs C57 Yr.](#)
- [B-Inhib. KO vs C57 Lact.](#)
- [C/ERP-#H.O. Yr+Preg+Lact](#)
- [C/ERP-#H.O. vs C57 Y+P+L](#)
- [C/ERP-#H.O. vs C57 Y+P+L](#)
- [Rehybridization repro \(4\)](#)
- [Tumor models](#)
- [C57 dev vs models, Whorm](#)
- [C57 dev vs models, P.horm](#)
- [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](#)
- [Quick Start](#)

## 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).



# Arrays accessible by MAExplorer

- NCI/CIT/mAdb Cy3/Cy5 data from the NCI/ATC facility (<http://nciarray.nci.nih.gov>). CGI connectivity lets users download sets of hybridized arrays data for use with MAExplorer
- <sup>33</sup>P membranes used in MGAP project (<http://mammary.nih.gov/>)
- <sup>33</sup>P membranes (neuroarray) in collaboration with Mark Vawter (NIDA) & Kevin Becker (NIA)
- Databases have been constructed for other arrays using Excel editing of user data
- Incyte and Affymetrix arrays using pseudo-arrays using the new Cvt2Mae data converter tool (<http://www.lecb.ncifcrf.gov/Cvt2Mae>)
- Other array data may be converted using Cvt2Mae <User-defined> array layouts

## II. Ongoing extensions to MAExplorer

- Adding other analysis and clustering tools to core MAExplorer program
- Directly connecting to other array database servers - but with a secure connection
- Cvt2Mae Java setup tool makes it easier to use MAExplorer on academic or commercial arrays
- Extending to other analysis schemes and existing software such as multi-dimensional scaling, clustering, etc. using user-specific Java Plug-ins
- Java plug-ins will be able to:
  - a) implment new functions with Java code
  - b) access other programs on local computer
  - c) access client-server programs over the Internet or on same computer

# Summary

- MAExplorer is used as a stand-alone application or as applet over the Web
- Accepts different array geometries, spot supports, <sup>33</sup>P or Cy3/Cy5 glass slides, scanner data
- Analyzes multiple probes, X-Y replicate sets, expression profiles, replicate spots
- Provides direct manipulation of scatter-plots, histograms, clustergrams, dendrograms, silhouette plots, spreadsheets
- Filters genes by gene subsets, spot intensities and ratios, and statistical tests
- Set operations on gene subsets and sample condition lists help manage search results
- Uses Web links to genomic, histology and model Web databases
- Generates reports as Web-accessible spreadsheets or exportable to Excel
- Users may save their data-mining session state locally to continue later
- Share user states on a collaborative Web server i.e. “groupware” [Future]
- MAExplorer was used to identify genes in MGAP DB preferentially expressed during lactation. Results were verified using northern blots (NIDDK) (*Nucleic Acid Res.* **28**:(22) 4452-4459).
- On-line documentation (manual, tutorials, etc.) is available on the Web site
- MAExplorer is available at <http://www.lecb.ncifcrf.gov/MAExplorer>
- MAExplorer is undergoing *beta*-testing