User's Manual

Version 3.1.5

Center for Toxicoinformatics
National Center of Toxicological Research
U.S. Food and Drug Administration

ArrayTrack 3.1.5

User's Manual

An Integrated Software System for the Support of Toxicogenomics Research through Managing, Mining, Visualizing, and Interpreting DNA Microarray Gene Expression Data

Center for Toxicoinformatics
National Center for Toxicological Research
U.S. Food and Drug Administration
3900 NCTR Road
Jefferson, Arkansas 72079
U.S.A.
Te: +1-870-543-7142, +1-870-543-7387
Fax: +1-870-543-7662
NCTRBioinformaticsSupport@nctr.fda.gov
Weida.Tong@fda.hhs.gov
http://edkb.fda.gov/webstart/arraytrack/

August 1st, 2005

Notice

ArrayTrack software is constantly evolving with both major and minor features, improvements, and inevitable bug fixes. This manual is for use with ArrayTrack major release version 3.1.5. While we strive for consistency between the manual and version 3.1.5, you might observe some slight differences. For updated information, please check the ArrayTrack web site at http://edkb.fda.gov/webstart/arraytrack/, where, for example, updated manuals for version 3.1.5 will be periodically provided.

TIPS

Many functions in ArrayTrack are accessible from multiple paths, for example, left-side window panels, pull-down menus and right mouse-click options.

1. Right-click on a (set of) selected object(s) under the Database Contents tree to access the applicable TOOL functions.

2. Multiple sets of arrays can be selected by a combinations mouse-click and SHIFT-CTRL keys.

3. Most functions come with default parameter settings. If you do not know a better setting, use the default.

4. All Spreadsheet viewers share similar functions, e.g. Copy/Paste of selected table content.

Table of Contents

Chapter 1 Introduction 8
What Is ArrayTrack? 8
Center for Toxicoinformatics of the NCTR/FDA 8
Toxicoinformatics Integrated System (TIS) 8
ArrayTrack - A Prototype of TIS 10
MicroarrayDB 10
LIB 10
TOOL 11
Availability of ArrayTrack 11
Online Version 11
Local Installation 11
Installing and Running Online Version of ArrayTrack 11
System Requirements 11
For Windows Users 12
For Mac OS X Users 12
For Linux/Unix Users 12
Installing and Running Local Version of ArrayTrack 12
Chapter 2 Working with Database: MicroarrayDB 14
Input Form 14
Overview 14
Experiment Design 15
Hybridization and Data 15
Data Import 17
Data Export: for information about exporting data please refer 23
ArrayType Information 23
Overview 23
Activate ArrayType Information 23
Export ArrayType Information 23
Create New ArrayType 24
Data Sharing and Security Protection 26
Exploring and Viewing Data in MicroarrayDB from Tree View 27
Structure of Database Contents 27
Exploring Database Contents 27
Operating on Experiment: 28
Operating on Hybridization 29
Operating on Array Data 29
Significant Gene List 29
Display Significant Gene List 30
Import Significant Gene List 32
Export Significant Gene List 33
Delete Significant Gene List 33
To create significant gene list 33
VennDiagram 33

Chapter 3 Working with Libraries 37
Gene Library 38
Working with the Table 38
Filter Genes: 38
Sort Table by Column 39
Rearrange Table Columns 39
Searching Functional Information for a List of Genes 39
More Information about a Gene 40
Link to Other Public Databases 41
Sort Genes Based on Their Common Pathway and Proteins 41
Customize Table 41
Export Table Contents to a Local File 41
Copy/Paste selected rows/columns 42
Stop a Search Session 42
Launch Bar Chart for Selected Genes 42
Pathway Library 42
Ingenuity Library 49
Protein Library 51
IPI Library 53
Orthologene Library 53
GOFFA Library 54
Chip Library 60
Display ArrayType Information 60
Compare Different ArrayTypes 61
Summary: 62
Toxicant Library 62
EDKB Library 64
ID Converter 66
Chapter 4 Working with Tools: Quality Control 68
Overview of TOOL 68
Overview of Quality Control 69
Launch of Quality Control 69
From the TOOL Panel 69
From the TOOL Pull-down Menu 69
From Right-click on Selected Arrays in the 69
Contents of Quality Control View 70
Preview of Scatterplot 70
Preview of Rank Intensity Plot 70
Quality Control Parameters 70
Overall Judgment on Array Quality 70
Function Buttons for Quality Control 71
Overview of Quality Filtering 71
Launch of Quality Filtering 72
Contents of Quality Filtering 72
Functions for Quality Filtering 73
Chapter 5 Working with Tools: Normalization 74

Overview 74
Lowess 76
Smoothing Factor 77
Robustness Iterations 77
Delta 77
Total Intensity Normalization 78
Channel Scaling 78
GenePix Mean Log Ratio Normalization 79
Linear and Lowess 80
Warning 80
Chapter 6 Working with Tools: Analysis Tools 81
Overview 81
T-test and ANOVA 81
P-value Plot 87
Clustering 89
PCA 91
Chapter 7 Working with Tools: Visualization 93
Overview 93
Scatter Plot 94
Default Plot 94
Selected Spots 94
Search for Genes in This Viewer 95
Axes Options 95
Show Lowess curve 95
Adjust Lowess parameters 95
Save as image 96
MA Plot 96
Mixed Scatter Plot 97
Rank Intensity Plot 98
Choosing Data Source for Plotting 98
From MicroarrayDB Arrays 98
From Local Data Files 98
Virtual Array Viewer 99
Overview 99
Adjust Brightness of Virtual Array Image 99
Zoom In and Out of Virtual Array Image 99
Filter Spots 100
Pull-Down Functions on Spots 101
Pull-down Options 101
Select and Deselect Current Spot 101
Mark Spots 102
Right-click Accessible Functions 102
Functions Applicable to Marked Spots 102
Actual Array Viewer 102
Bar Chart 103
Overview 103

Query Specification 103
Bar Chart Display 103
Toolbars 104
Table Display 106
Links between Bar Chart and Table 106
Summary 107
Chapter 8 Working with Tools: DMVS 108
Chapter 9 Data Export 109
How to Access Data Export Functions 109
Options for Data Export 109
Export selected datasets as spreadsheet 111
Preview selected datasets as spreadsheet 111
Export original files for selected datasets 112
Export data in a narrow format 112
Export in TIGR MeV format 112
Export Gene ID matches 112
Chapter 10 What to Expect from ArrayTrack 4.0 114
References 115
ArrayTrack Team 115
Acknowledgments 115

Chapter 1 Introduction

What Is ArrayTrack?

ArrayTrack is an integrated software system for managing, mining, visualizing, and interpreting microarray gene expression data. This software has been developed by the Center for Toxicoinformatics at the National Center for Toxicological Research (NCTR) of the U.S. Food and Drug Administration (FDA) and is continuously updated. ArrayTrack is a module of a comprehensive software system described below, the Toxicoinformatics Integrated System (TIS), that is being developed to integrate analysis of genomic, proteomic, metabonomic data and toxicology data.

Center for Toxicoinformatics of the NCTR/FDA

The mapping of the human genome and the determination of corresponding gene functions, pathways, and biological mechanisms are driving the emergence of the new research fields of toxicogenomics and systems toxicology. Many technological advances such as microarrays are enabling this paradigm shift that portends an unprecedented advancement in the methods of understanding the expression of toxicity at the molecular level. At the NCTR/FDA, core facilities for genomic, proteomic, and metabonomic technologies have been established that utilize standardized experimental procedures to support center-wide toxicogenomic research. Collectively, these facilities are continuously generating an unprecedented volume of data.

To effectively meet the challenges of modern toxicological research, the NCTR/FDA established the Center for Toxicoinformatics on June, 2002. Toxicoinformatics is an emerging scientific discipline that integrates approaches from multidisciplinary fields of bioinformatics, chemoinformatics, computational toxicology, informatics technologies, and physiologically-based pharmacokinetic modeling with the objectives of knowledge discovery and the elucidation of mechanisms of toxicity. The primary function of the Center for Toxicoinformatics is to apply and develop toxicoinformatics approaches for omics research and traditional toxicological studies at NCTR and beyond to FDA. More information about the Center for Toxicoinformatics can be found at http://www.fda.gov/nctr/science/centers/toxicoinformatics/index.htm.

Toxicoinformatics Integrated System (TIS)

The NCTR's Center for Toxicoinformatics has been developing a Toxicoinformatics Integrated System (TIS) for the purpose of fully integrating genomic, proteomic, and metabonomic data with data in the public repositories, as well as conventional in vitro and in vivo toxicology data. Figure 1-1 illustrates the TIS architecture organized around three major components: central in-house data archives (DB), a set of libraries with highly relevant information downloaded from public databases (LIB), and analysis and visualization functions (TOOL). The DB component contains a set of relational databases, each storing experimental platform-specific data (e.g. microarray data for genomics, MS data for proteomics, and NMR data for metabonomics) together with annotation information about the experiments and samples. The TOOL component provides the ability to query, visualize, mine, analyze, and correlate diverse data from both local and public resources. The LIB component hosts information from public databases on genes, proteins, pathways, and small chemicals involved in the pathways or toxicological experiments. Through integration of different data types with analysis capabilities, TIS will be able to extract a tailored dataset for data interpretation, hypothesis generation, and hypothesis testing to aid toxicogenomics studies.

Below the TIS component level is the software module level. Software modules are associated with a particular class of data that, in turn, is associated with a particular type of

experimental platform (e.g., microarray, *in vitro* endpoint, *in vivo* endpoint, or protein gel). Each software module for each data type/experimental platform can be constructed independently, and the overall system can be developed in accordance with extant priorities and experiment progress.

TIS will integrate microarray gene expression data, proteomics data, and metabolite profiling data with classic in vivo or in vitro toxicology data. Expression profiles of a suspected toxicant may provide unique signatures that can be readily compared with expression patterns of known toxicants stored in the TIS. In addition, TIS will consolidate data in a manner conducive to development of models to predict toxic endpoints based on chemical structures and/or gene expression profiles. With TIS, expression data from mechanistic-based assays and *in vivo* pathology or clinical observations can be readily compared, possibly leading to development of less expensive and more timely assays for risk assessment. Finally, expression profiling may become an important component for diagnostics in medicine and TIS could be a significant benefit to aid FDA regulators in evaluating new diagnostic tools that are based on "omics" technologies.

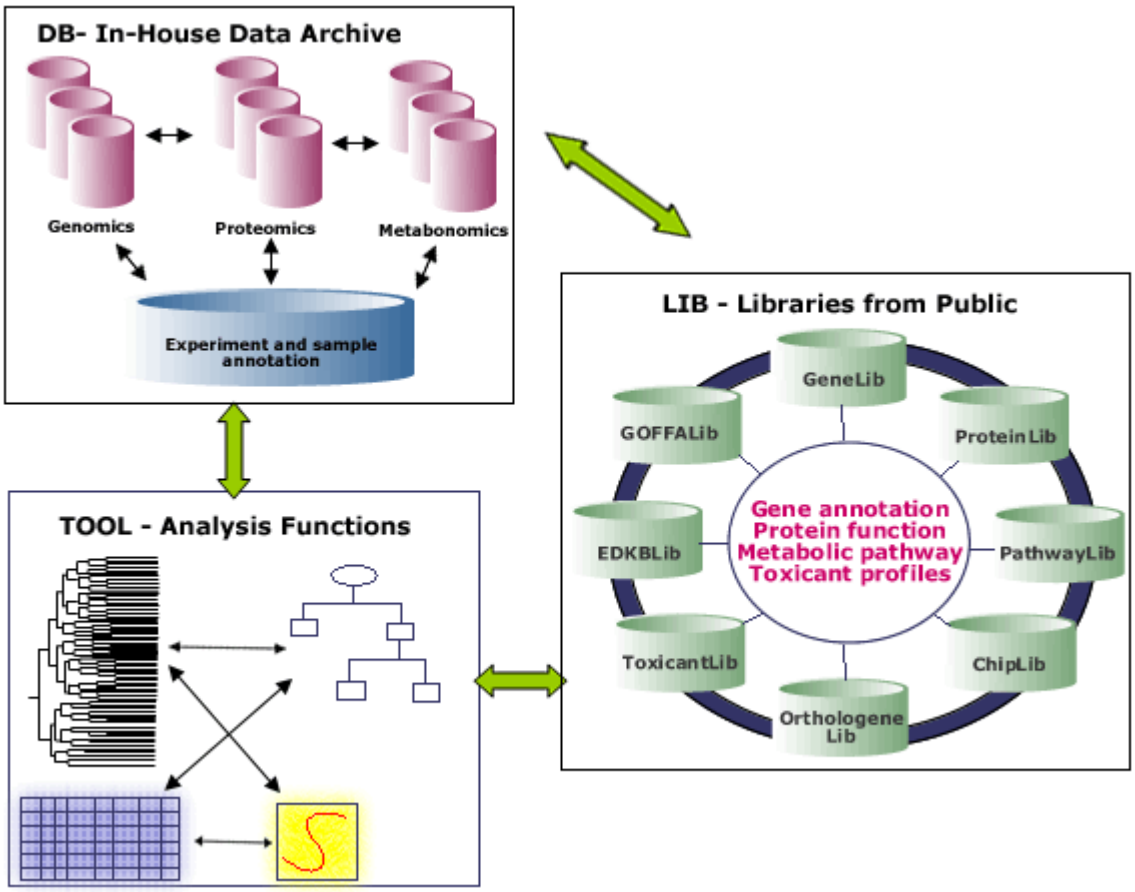


Figure 1-1: An overall system architecture of the Toxicoinformatics Integrated System (TIS).

The system is based on a DB-TOOL-LIB integrated structure. (1) The DB is a central data archive for in-house data storage and management; (2) The TOOL provides data visualization and analysis functions; and (3) The LIB a set of libraries that contain data both from online public databases as well as NCTR in-house databases that together integrate information, for example, on sequence, gene annotation, gene and protein function, pathways, and toxicant profiles and chemical structure.

ArrayTrack - A Prototype of TIS

ArrayTrack is a prototype version of TIS and has been developed to effectively handle and analyze DNA microarray data. ArrayTrack is logically constructed of three linked components: (1) MicroarrayDB, a database that stores microarray experiment information and gene expression data in accordance with MIAME (Minimum Information About a Microarray Experiment); (2) LIB, a set of libraries that mirror critical data in a number of public databases; and (3) TOOL, a set of tools that operate on experimental and public data for data analysis, visualization, and knowledge discovery purposes. An overview of ArrayTrack is shown in Figure 1-2 and more detailed descriptions of the three components follows.

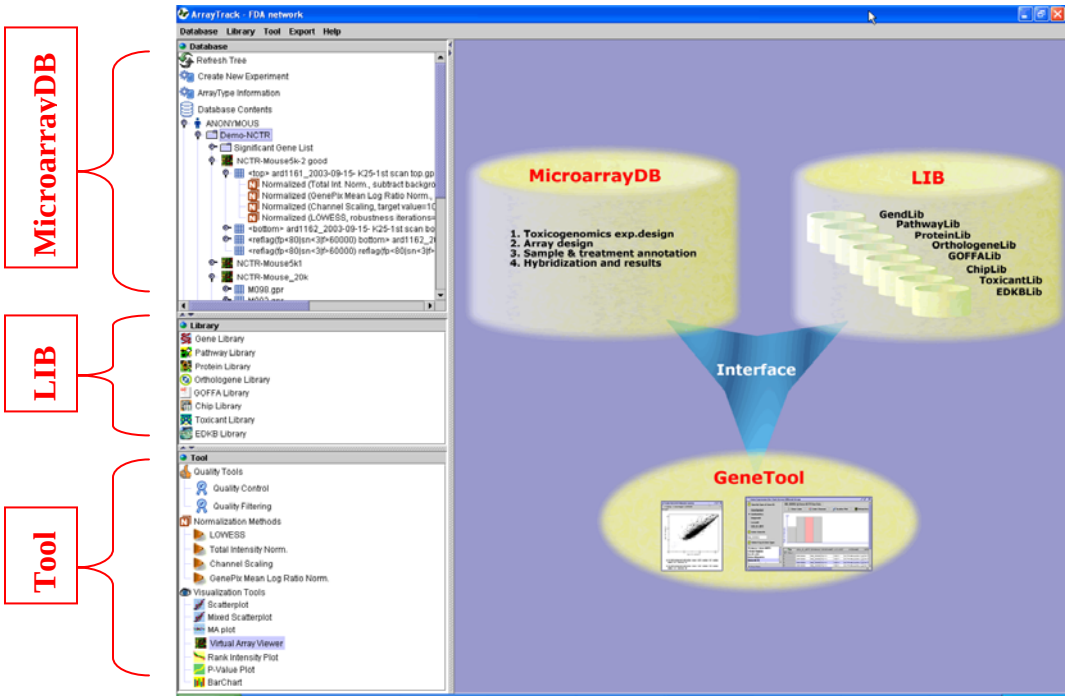


Figure 1-2: ArrayTrack's three main components: MicroarrayDB, LIB, and TOOL.

MicroarrayDB: MicroarrayDB (DB) is part of ArrayTrack's ORACLE-based relational database that stores microarray experimental data. As implemented at NCTR, DB data complies with rigorous criteria to assure a validated microarray repository that will be a rich resource for data mining and experiment comparison. Data may be entered if and only if it meets prescribed standards for completeness, accuracy, normalization as well as conformance to the applicable ontology. Such disciplined data curation requires toxicoinformaticians and biologists work closely together to understand both the structure of the database and the structure of the data to be stored. An NCTR microarray data submission form was developed that specifies the required essential information from both microarray and toxicology perspectives. The format is adheres to the MIAME/Tox (Minimum Information About a Microarray Experiment for Toxicogenomics) guideline for toxicogenomic research.

LIB: LIB is part of ArrayTrack's ORACLE-based relational database that comprises nine libraries: Gene Library, Pathway Library, Protein Library, IPI Library, Orthologene Librrary, GOFFA Library, Chip Library, Toxicant Library, and EDKB Library, that mirror the most essential information in public databases related to genes, proteins, pathways, toxicants, and as well as detailed data for various microarray systems (chips). Library data sources include UniGene, LocusLink, GeneCards, SWISS-PROT, and KEGG from which information on gene annotation, protein function, and metabolic pathways are obtained. Efforts are underway to

incorporate other types of information (e.g. PubMed) from public databases into LIB. The libraries are highly interlinked within ArrayTrack, and are frequently updated to reflect the ever-increasing and refined data within the public repositories. LIB and its use are described in detail in Chapter 3. Scientists not analyzing microarray data will nonetheless find much utility in using ArrayTrack to query public data owing to the nature of the integration of information from disparate data sources.

TOOL: The TOOL section of ArrayTrack provides functions for microarray data visualization, normalization, significance analysis, clustering, classification and, importantly data quality assessment (QA) and control (QC). Several standard or common methods for data normalization, analysis, visualization, and QA/QC of data are available. Novel or essential tools to enhance our capabilities that are tailored to toxicology-specific problems are implemented or in development. A number of data visualization capabilities have already been developed. Additionally, given the ever-increasing commercial and public software packages providing common or specialized data analysis capabilities for microarrays, we have developed interfaces/interface formats for other software packages to conveniently access and analyze data stored in ArrayTrack.

It is important to point out that many of the various functions in ArrayTrack are accessible from either the pull-down menus or the left-side panels, and thus a particular function may be invoked by different sequences of operations.

To illustrate the logical operation of ArrayTrack, the user can select an analysis method from the TOOL, apply the method to selected microarray data stored in the MicroarrayDB, and the analysis results can be directly linked to information in the LIB. The user can also hyperlink from data within ArrayTrack to the corresponding detailed data in the many supported public data repositories. In this way, ArrayTrack can be very helpful to scientists in interpreting the biological meaning of microarray results by convenient access to information on genes, proteins, and pathways, etc.

Availability of ArrayTrack

Currently, ArrayTrack is being distributed free of charge by the NCTR/FDA to the research community.

Online Version: Users within the fda.gov domain can download and run ArrayTrack at http://weblaunch.nctr.fda.gov/jnlp/arraytrack/index.html, and have access to ORACLE-based DB to store microarray data. For the users outside of the fda.gov domain, they can download and run ArrayTrack at http://edkb.fda.gov/webstart/arraytrack/. Users outside the fda.gov domain can access all of the functions of ArrayTrack except for uploading microarray gene expression data since at the time of this writing a decision has been made not to use the online version of ArrayTrack as a public repository for microarray data.

Local Installation: For those who are seeking to have the entire client-server system installed at their local sites for independent use, please contact NCTRBioinformaticsSupport@nctr.fda.gov to request the CD or DVD. In this case, you will need Oracle license to run ArrayTrack locally.

Installing and Running Online Version of ArrayTrack

System Requirements: ArrayTrack is a client-server system. To date, the software has been tested on the Windows operating systems (98/NT/2000/XP), Linux/Unix, and Mac OS X. If you have a problem running ArrayTrack on your system, please contact NCTRBioinformaticsSupport@nctr.fda.gov. The application screens are best viewed at 1024x768 (or higher) resolution. For Windows users, we suggest a Pentium 266 MHz or faster processor with at least 256 MB of physical RAM to support the graphical applications.

For Windows Users: To use the online version of ArrayTrack, go to http://edkb.fda.gov/webstart/arraytrack/ and follow two simple steps: (1) Install Java if it is not already installed on your machine, using the provided link to the Sun Microsystems web site in step one (The program will automatically detect whether Java is already installed on your machine. If so, you can skip this step). (2) Install and run ArrayTrack.

For Mac OS X Users: If you have kept up to date with your updates from Apple, then you should already have Java 1.4(+) available on your machine, in which case you should be able to either (1) just click the link in the second step to start the application, or (2) failing that, save the jnlp file which is the target of the second step's link (http://edkb.fda.gov/webstart/arraytrack/arraytrack_ext.jnlp) to your hard drive, then double click it to run the application.

For Linux/Unix Users: Install the Java Runtime Environment for your platform from http://java.sun.com/getjava/, then execute the javaws command which should be contained within the Java installation directory (exact path to it will vary), like so:

javaws http://edkb.fda.gov/webstart/arraytrack/arraytrack_ext.jnlp # (for users external to FDA), or

javaws http://weblaunch.nctr.fda.gov/jnlp/arraytrack/arraytrack_internal.jnlp # (for users within FDA).

You can also use this link (http://edkb.fda.gov/webstart/arraytrack/) in the future to run ArrayTrack, or let ArrayTrack place icons (Figure 1-4) on your desktop and start menu when prompted. You will notice some delay the first time you run ArrayTrack due to the need to download the entire ArrayTrack application (Figure 1-5). You may also be prompted to update your version of Java before ArrayTrack itself is started. Future uses of the software will only download parts of the application that have been changed, if any, and should start much quicker. For the non-FDA users, due to the FDA firewall, you may experience delays when using this online version of ArrayTrack. Interested users may request a CD or DVD for local installation, which will greatly increase speed.

When you activate the ArrayTrack, you will see the login window (see Figure 1-3). If you don't have an account you can just click cancel button or leave the fields blank then click OK button to login and view the demo data.

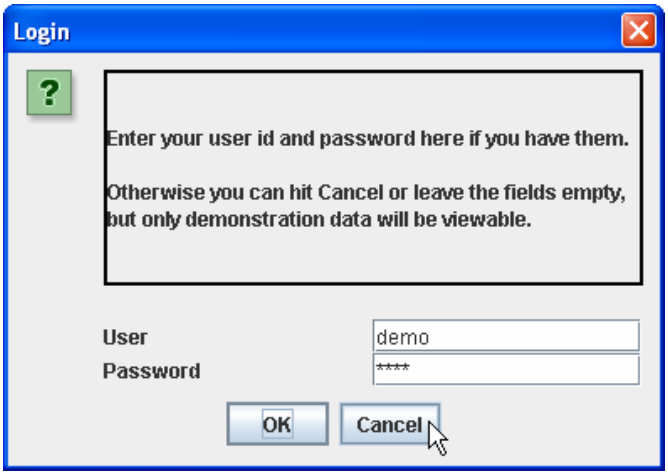


Figure 1-3: Login to ArrayTrack

Installing and Running Local Version of ArrayTrack

If you requested a CD or DVD for local installation of ArrayTrack, you will receive detailed instructions for installation and how to configure your local microarray database(s).

Please note that for local installation, your systems must have ORACLE 9i or higher installed. Updates for the various libraries in LIB are available for download at http://edkb.fda.gov/webstart/arraytrack/.

Enjoy yourself by following the instructions to be discussed in the following chapters!

Figure 1-4: The ArrayTrack icon appears on the desktop of a computer after installation.

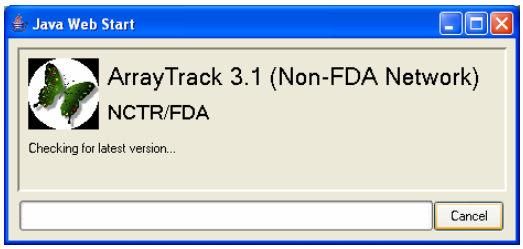


Figure 1-5: Downloading and running the online version of ArrayTrack.

Chapter 2 Working with Database: MicroarrayDB

MicroarrayDB has been designed to store DNA microarray gene expression experimental data together with essential annotation information about an experiment, its protocol and the samples. MicroarrayDB supports data from both one-channel (e.g. filter arrays, Affymetrix GeneChip® arrays, and Amersham CodeLink® arrays) and two-channel (e.g. spotted cDNA or oligo arrays) microarray platforms and adheres to the MIAME guideline for microarray experiments (http://www.mged.org/Workgroups/MIAME/miame.html). A set of tools have been developed for managing, normalizing, visualizing, analyzing, and, importantly, for performing QA/QC of data stored in the database.

Input Form: Inputting Experimental Information and Gene Expression Data including Creating a New Experiment

NOTE: In the publicly accessible online version of ArrayTrack, the data uploading function in Input Form is disabled.

Overview: According to MIAME, an experiment consists of a set of hybridizations. The database structure in ArrayTrack has been arranged in an Experiment➔Hybridization➔Array Data hierarchical, tree-like format (for details, see Figure 2-22).

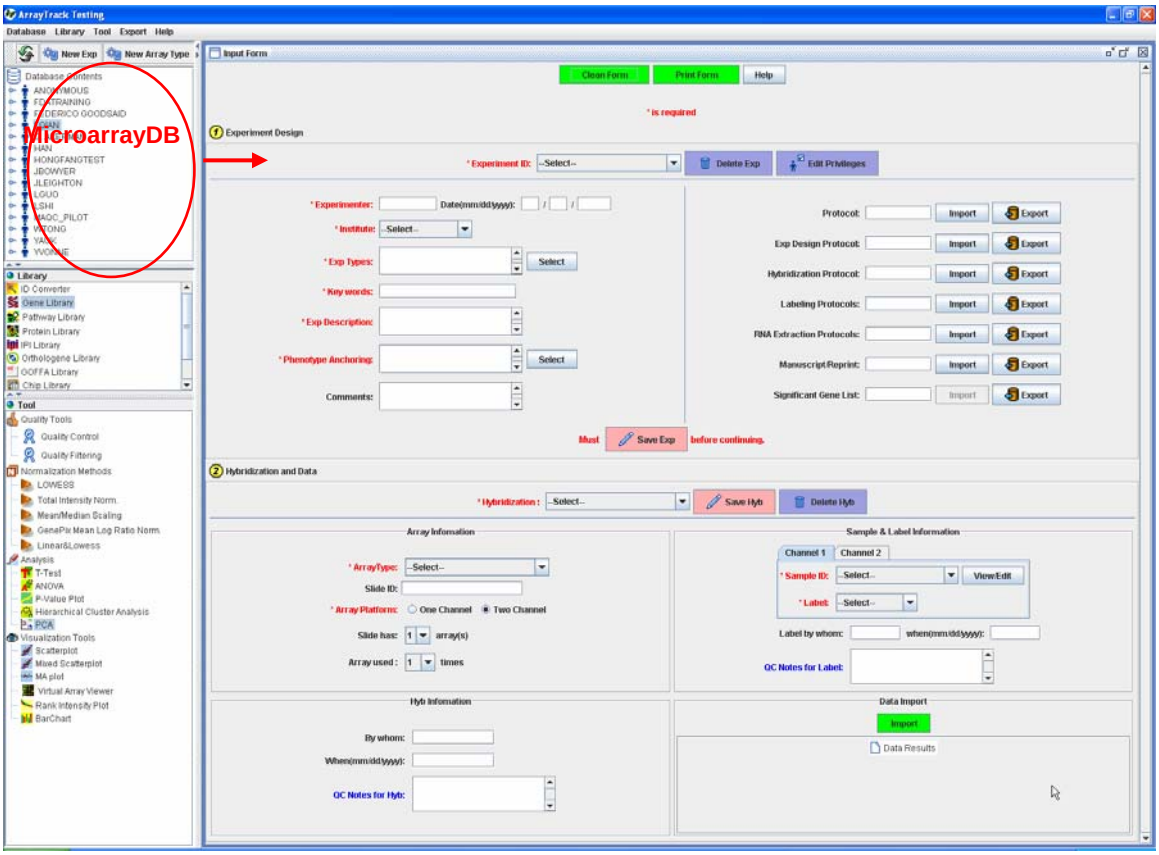


Figure 2-1: Input Form: Data entry form of MicroarrayDB.

Under the Database window panel of functions (Figure 2-1), the user can select the New Exp button to activate Input Form (Figure 2-1) that allows the user to enter all the MIAME-specified information about the experiment. Before any information may be input into Input Form, the user is prompted to specify a unique title for the experiment and experiment group name, as shown in Figure 2-2.

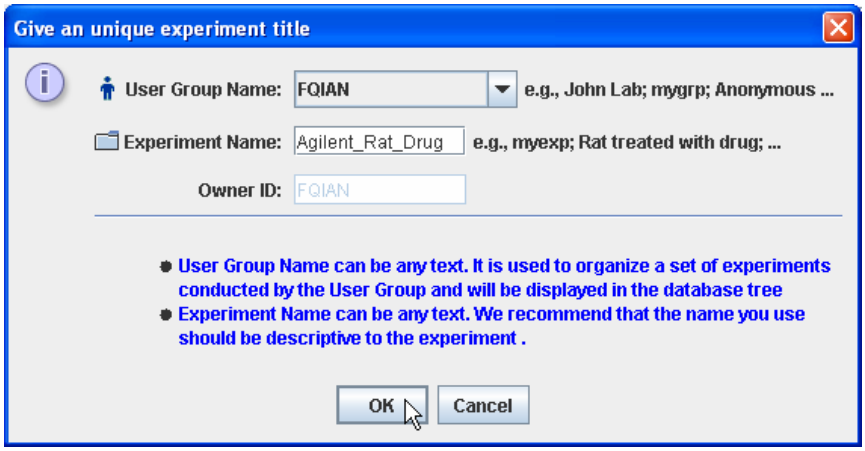


Figure 2-2: The user needs to specify an Experiment ID and Experiment Group Name for the new experiment.

The Owner ID will automatically show the name of the user whoever logged into the Windows system, and is grayed out. After the owner created an experiment, he can click Edit Privileges (see Figure 2-1) to assign different privileges (view or/and write) to the other users; otherwise no body can view/edit the experiment except the owner. The Exp Group Name is arbitrary.

The information items on the Input Form are straightforward and reasonably self-explanatory. The Input Form consists of two sequential sections: (1) Experiment Design, and (2) Hybridization and Data. Just follow the Input Form to enter all required (highlighted in red) and the remaining optional information items.

Experiment Design: The description of an experiment and its associated experimental protocols are input in the Experiment Design section. The information in Experiment Design only needs to be filled in and saved once for an entire experiment, since such experiment design information remains the same for all hybridizations of the same experiment. However, the user first must click Save Exp before moving on to the Hybridization and Data section. An existing Experiment can be deleted by clicking on Delete Exp (see Figure 2-1).

Hybridization and Data: This section is used for input and/or update of data associated with all hybridizations, arrays and samples included in the entire experiment. From a logical perspective, it is best to consider an experiment to consist of a set of hybridizations, each of which has an associated microarray type, sample and label (e.g., florescent dye or radioisotope. The amount and types of data to be entered are determined by many factors of the experiment design. For example, single versus two channel experiment, dye flip experiment and cDNA- versus oligonucleotide-based experiments are major determinates of specifically what data must be entered.

Data describing a hybridization can be input in the lower portion of Input Form that is divided into three areas: 1) Hybridization; 2) Array Information; and 3) Sample and Label Information (i.e., fluorescent dye). The bright green IMPORT button Import in the lower right of the Input Form enables importation from a file of the actual microarray data for each

hybridization. By the end of the data input, every hybridization in an experiment will be linked with its appropriate array information, sample and label information and its experiment data.

Hybridization: The hybridization information is first input in the section of the input form titled HYBRIDIZATION. The user inputs the name of the hybridization associated with a specific sample. The name is arbitrary and determined by the user, and would normally provide some descriptive information. Multiple hybridization names can be input and will then all appear in the pull-down menu. A hybridization can only be saved after the array information and sample labeling information for a sample has been entered. If the user later wants to change the name of the hybridization, this is done by right clicking the hybridization name in the database window and choosing the option "Edit Hyb" name and editing the existing name (see Figure 2-3A).

Array Information: Information for each array is input in the section titled Array Information. There are more than 70 array types predefined within ArrayTrack. If the array type the user is using is one of those predefined in ArrayTrack, it can be selected from the pull-down menu as shown in Figure 2-3B. Instructions for creating a new array type is given further below on Page 23.

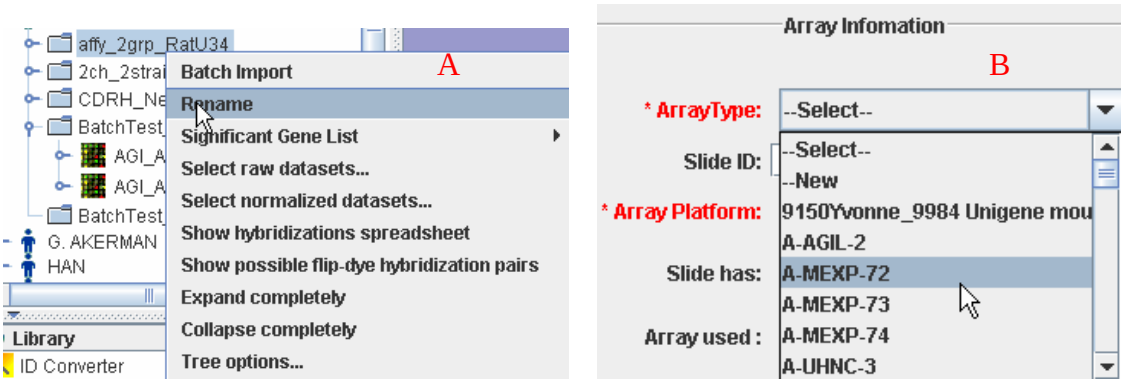


Figure 2-3: Hybridization name editing (A) and pull-down selection of array type (B)

Sample and Label Information: Information for each sample and how it is labeled is input in the section titled "Sample and Label Information". Clicking the View/Edit button (see Figure 2-4B) brings up the Sample Form that is shown in Figure 2-5.

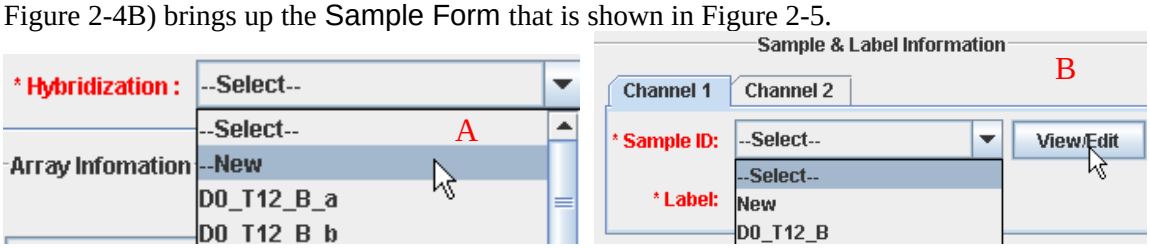


Figure 2-4: Sample ID pull-down selection.

The user must click Save Sample to save sample information to the database. An existing Sample can be deleted by clicking on Delete Sample (see Figure 2-5).

If a sample (e.g. A-2HZN00A) is involved in several hybridization records, the user needs to enter information only once. The remaining hybridization records can be chosen from the Hybridization pull down list (see Figure 2-4A). Similarly, to save time during input, the user can load a similar sample already saved in the database, make minor changes (e.g. SampleID or Sex) to the Sample Form, and then save the new sample with a new name.

Notes: (1) A Sample is associated with only one Experiment and is not shared across different Experiments. However, a Sample can be shared by multiple Hybridizations within the same Experiment (e.g., technical replication). The user can choose to finish inputting all samples altogether before inputting any hybridization information as long as the experiment is saved. When the user inputs hybridization information, all the saved sample IDs will become selectable from the sample ID pull-down list (Figure 2-4B).

The hybridization information is only saved when the user clicks the Save Hyb button, at which point the current section of the form the user is in will be permanently saved into database (an existing Hybridization can be deleted by clicking on Delete Hyb).

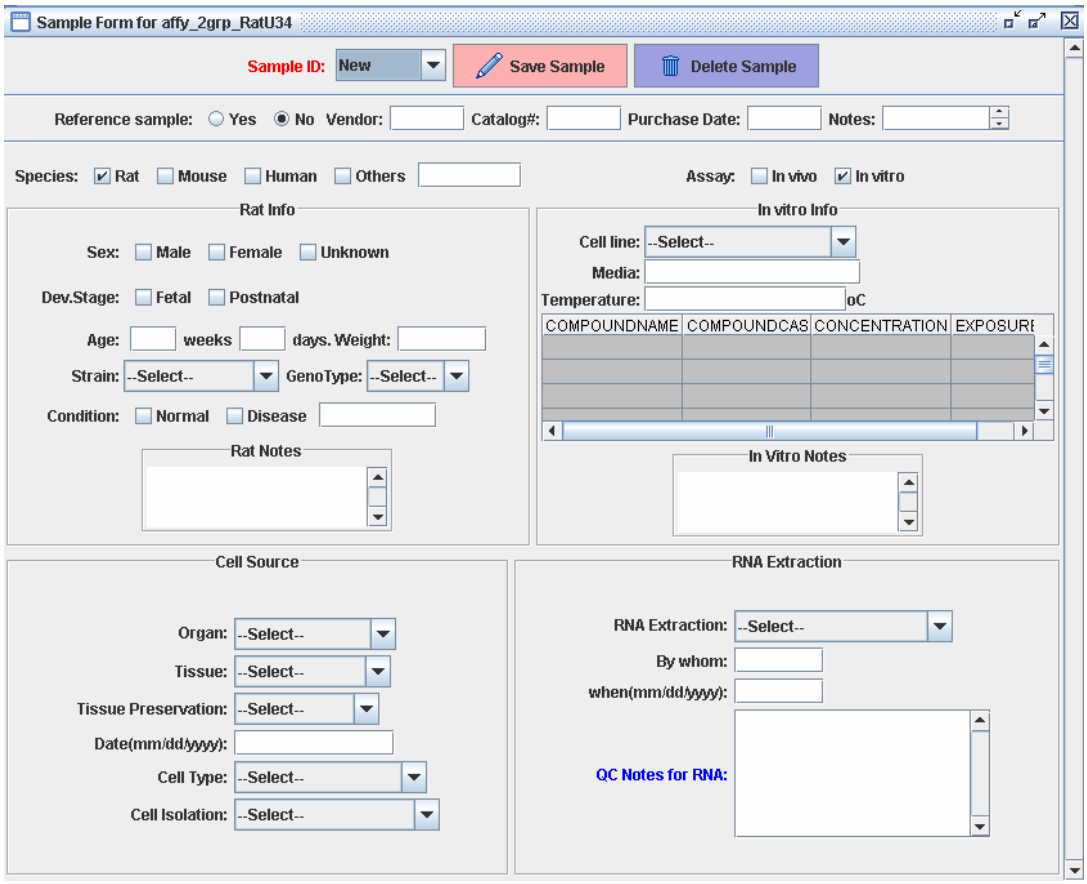


Figure 2-5: Sample information entry form (Sample Form)

Data Import: User can import each dataset one by one or import multiple dataset in one shot using the "Batch import" function. The "Batch import" function is very convenient if you have many hybridizations of the same array type. The two options are addressed as following:

1. Individual import

Microarray gene expression data is loaded into the database by first clicking the Import button in the lower, right panel of the **Input Form**. The user must save the data in each section as the data entry proceeds: The user must save the experiment Save Exp before continuing to Hybridization and Data section, and Data Import is not accessible until the Hybridization has been saved Save Hyb.

Clicking the import button Import brings up the Select Files window that prompts the user to select the Data File, up to two Image Files (optional) and the scanner Settings File

(optional) corresponding to the current hybridization/array. Browse buttons Browse are provided to locate and select the file. After pushing the OK button, the Choose Columns panel pops up that allows the user to map data columns in the Data File to corresponding MicroarrayDB data fields (Figure 2-7). Data columns already mapped will be shown in yellow and their corresponding MicroarrayDB data fields will be marked with ✔.

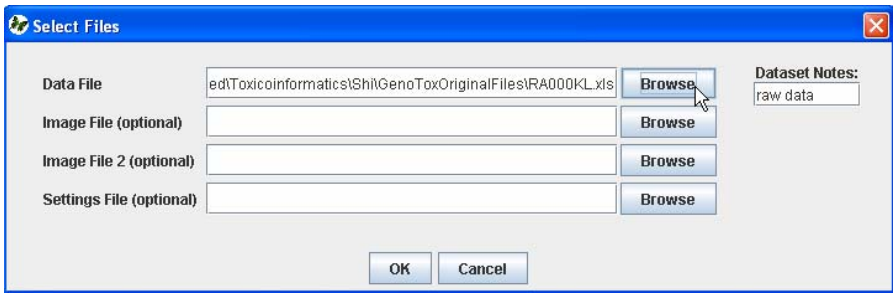


Figure 2-6: Select Files dialog for Data Import.

Tip: According to the data structure shown in Figure 2-3, one hybridization can be associated with multiple arrays that could be a result from scanning the same hybridized slide under different scanning parameters or at different times. The "Data set description" field (Figure 2-6) allows the user to input such descriptive information (e.g. 33% scanning power) that will be shown within the "<>" under the tree structure view.

Tip: The Data Import can be done separately from Input Form. The user can input all hybridization information into the Input Form without importing any data. Data import can be started later from the database tree view by selecting the hybridization (right-click), select "Import data".

The Data File must be in text format (Caution: Very large Excel files could cause "out-of-memory" errors; it is recommended that large Excel files be converted to text format prior to import). For a GenePix GPR Data File, the mapping is automatically determined by ArrayTrack. For other data files other than the GenePix GPR format, the user can easily set the mapping by clicking on a column header and then clicking on its corresponding MicroarrayDB data field (see Figure 2-7).

Figure 2-7: Choose Columns for loading array data file (GenePix GPR file).

After the first successful import from any file format that has recognizable column names at the top, the next time the same file format that is used for the column choices should default to the previous choices, with preference to the most recent import done by the same user.

Clicking on OK at the bottom of Figure 2-7 will save the data for the current hybridization/array to MicroarrayDB. If the Data File does not match the ArrayType definition (e.g. unequal number of spot numbers), ArrayTrack will give a warning and the user should take extreme care to double check data consistency, such as array type or sample label, etc. When the data is successfully loaded into MicroarrayDB, the Database Contents tree will be updated to show the new hybridization/array.

For Affymetrix data, the user can directly import .CEL file. ArrayTrack will automatically convert .CEL file to probe set file. This function only works for FDA users. For non-FDA users they can not import .CEL file because the Bioconductor server which is for the converting is not accessible from out side of FDA firewall. But they can import probe set file (in .txt format) instead of the .CEL file.

The Select Files dialogue (Figure 2-8B) provides 4 types of data file: 1).CEL file, 2).txt file (probe set file, optional), 3) .DAT file (image file), 4) .CHP file. .CEL file and probe set file are the data files, so the user needs to have either of the two formats of the file for data import. If the user has probe set file then the .CEL is optional and vice versa. DAT file and CHP file are image files and optional for data importing. Since the image files are not used in calculations, any format is acceptable.

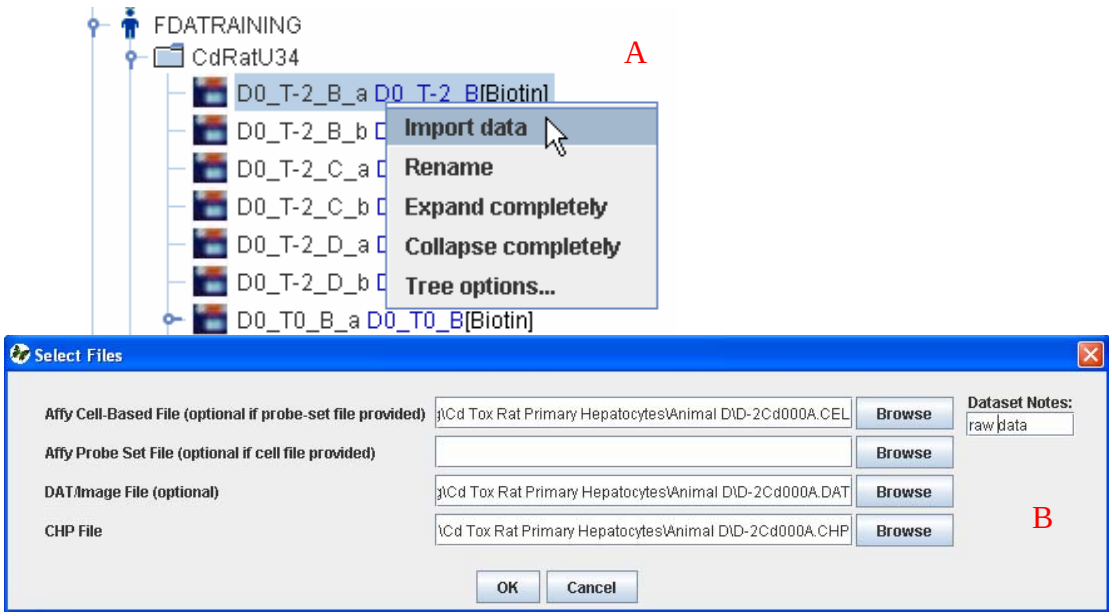


Figure 2-8: Importing Affymetrix data

2. Batch import

If the user needs to import multiple arrays (e.g. 10, or 50 hybridizations) of the same array type, s/he can use the "Batch import" function. Before using the "Batch Import" function, the user needs to prepare a hybridization file (in Excel format) containing all the information about the data (like sample, label, the file names, etc), see Figure 2-9.

The following is an example of batch importing 20 hybridizations.

After the hybridization file is ready, the user can start batch import. There are several ways to activate "Batch Import" function: 1) from "Database" pull-down menu, 2) if experiment

has already been created, right-clicking the experiment name and choose "Batch Import". See Figure 2-10. After choosing "Batch Import", the "Batch Import" window shows up (Figure 2-11).

	A	B	C	D	E	F
1	HybName	Cy3 Sample	Cy5 Sample	Cy3 Organ	Cy5 Organ	FileLocation
2	A1	SUR1	CUR1	Universal	Universal	testA\A1.txt
3	A2	SUR2	CUR2	Universal	Universal	testA\A2.txt
4	A3	SUR3	CUR3	Universal	Universal	testA\A3.txt
5	A4	SUR4	CUR4	Universal	Universal	testA\A4.txt
6	A5	SUR5	CUR5	Universal	Universal	testA\A5.txt
7	B1	SUR6	AB1	Universal	Brain	testB\B1.txt
8	B2	SUR7	AB2	Universal	Brain	testB\B2.txt
9	B3	SUR8	AB3	Universal	Brain	testB\B3.txt
10	B4	SUR9	AB4	Universal	Brain	testB\B4.txt
11	B5	SUR10	AB5	Universal	Brain	testB\B5.txt
12	C1	SUR11	AL1	Universal	Liver	testC\C1.txt
13	C2	SUR12	AL2	Universal	Liver	testC\C2.txt
14	C3	SUR13	AL3	Universal	Liver	testC\C3.txt
15	C4	SUR14	AL4	Universal	Liver	testC\C4.txt
16	C5	SUR15	AL5	Universal	Liver	testC\C5.txt
17	D1	CUR1	AB6	Universal	Brain	testD\D1.txt
18	D2	CUR2	AB7	Universal	Brain	testD\D2.txt
19	D3	CUR3	AB8	Universal	Brain	testD\D3.txt
20	D4	CUR4	AB9	Universal	Brain	testD\D4.txt
21	D5	CUR5	AB10	Universal	Brain	testD\D5.txt

Figure 2-9: Hybridization file for batch import

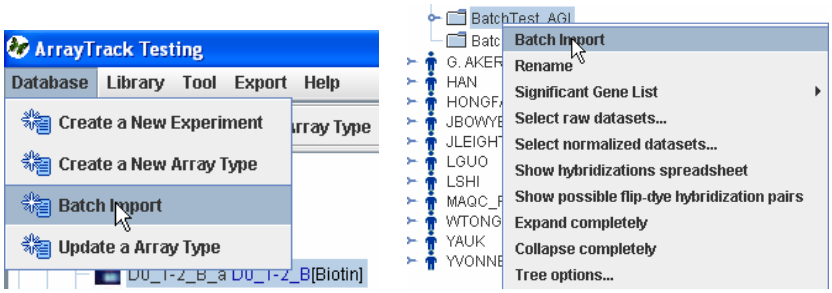


Figure 2-10: Activate "Batch Import" function

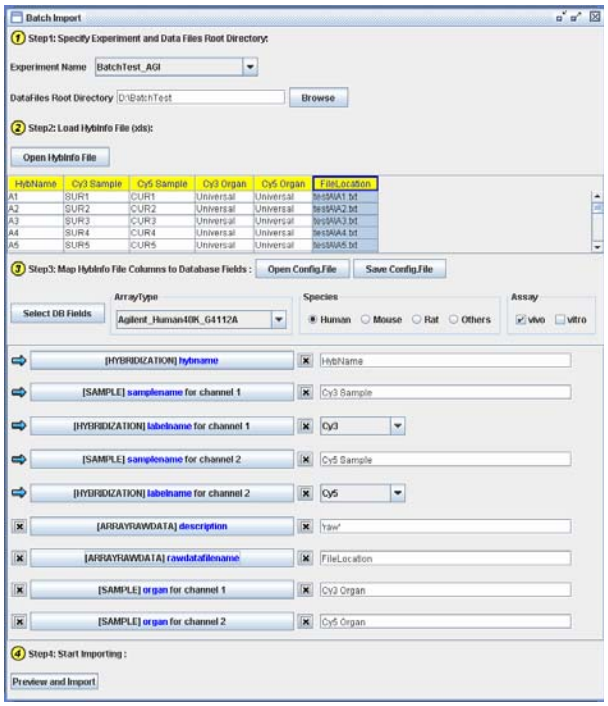


Figure 2-11: Batch import interface

In Figure 2-11, the user can fill the form by following step 1 through step 4.
Step 1: choose the experiment name from the pull-down list. If experiment is not created yet, the user can choose "New" from the pull-down list and type a new experiment name (refer page 14 for creating new experiment). Then click Browse button to locate the file root directory.

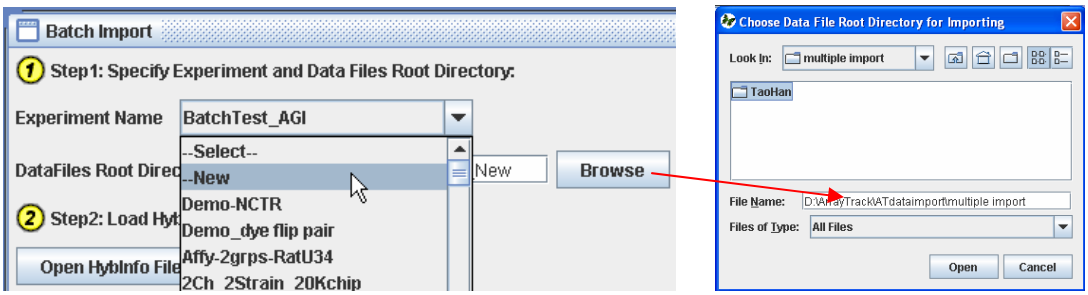


Figure 2-12: specifying experiment name and data file directory

Step 2: Load hybridization file which is the file you created before batch-import. Click Open HybInfo File button and choose the hybridization file. The contents of the file will show up below the Open HybInfo File button (see Figure 2-13).

Step 3: Map hybridization file columns to database fields. First choose array type from the pull-down list, if the list doesn't include the array type you need, then you have to create a new array type (see page 23 - ArrayType Information part for creating new array type). Then choose species (Human, Rat, or Mouse, etc) and assay (In vivo or In vitro). Next map the columns to database fields.

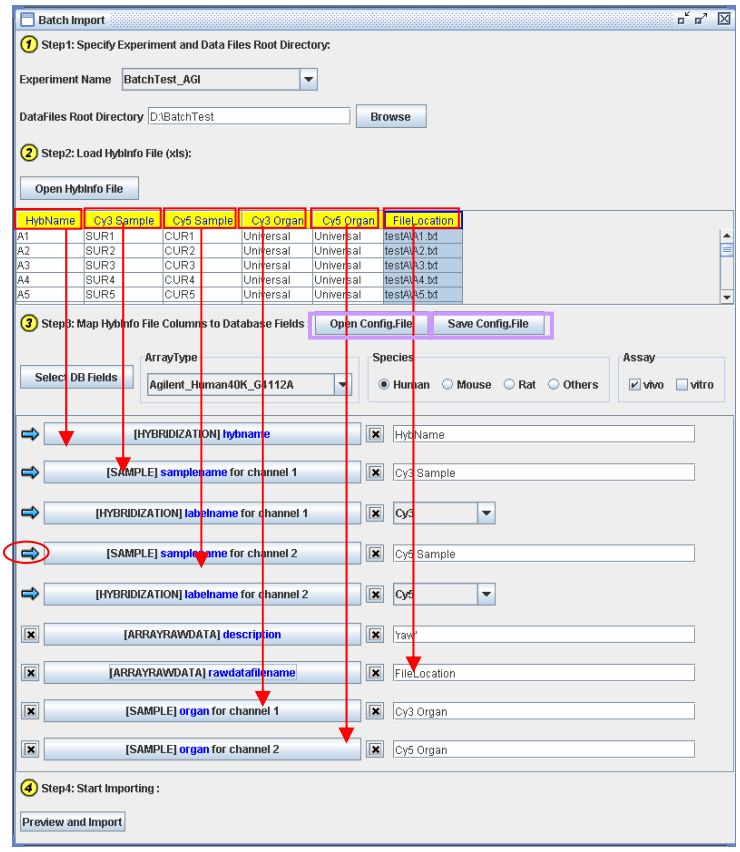


Figure 2-13: mapping the hybridization columns to database fields

In Figure 2-13, the blue arrows at the left side of the field buttons means this field is a required field (the right-side empty box has to be filled), all the others are optional. The user can click the cross sign button at the left side to remove the database button, the button at the right side is used to clear the contents in the text box. To map the column to the database field, click the column title then click the database field button. The column title will show in the text box right to the button. The mapped column will be highlighted in yellow. The user can save this configure after mapping all the columns by clicking the button Save Config.File and open this configure file next time for batch import.

If the user wants to map columns to the other database fields, he can click Select DB Fields to bring up more database fields (Figure 2-14).

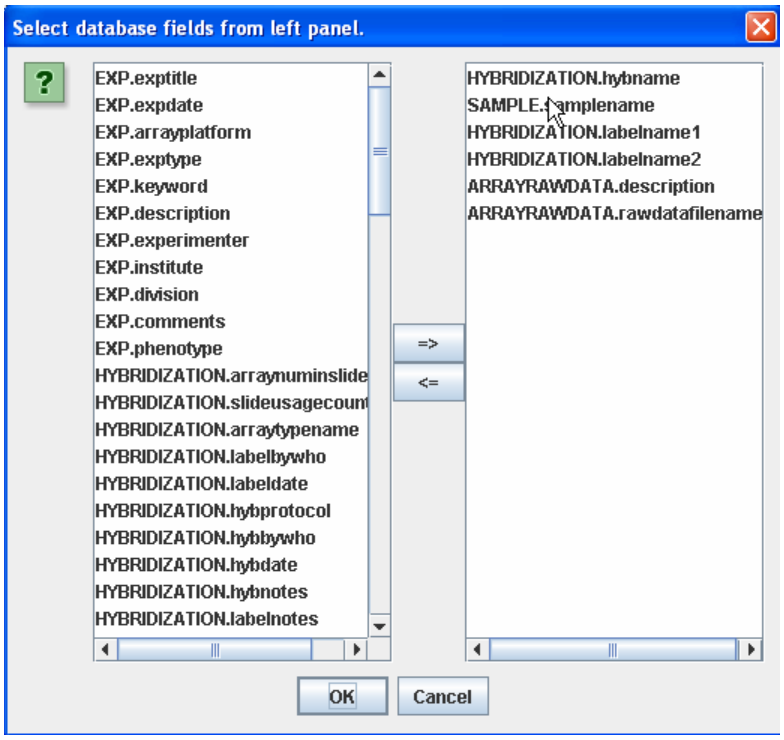
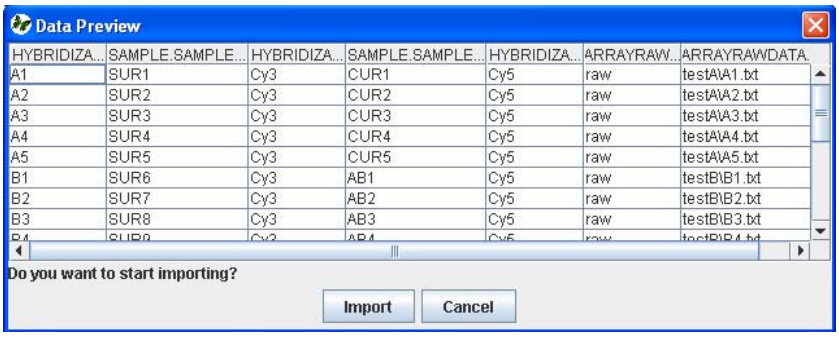


Figure 2-14: more database fields for choosing

In Figure 2-14, the database fields in the right panel are the fields that will be shown in "Batch Import" interface (see Figure 2-11) while the database fields in the left panel are more options available for choosing. The user can move the database fields from left to right side or vice versa. Some of the fields in the right panel are required fields and can not be moved to the left side. The user can select any optional field and bring it to the right side by click the arrow button.

Step 4: After mapping the columns to the database field, the user can click Preview and Import button to preview the data information before importing.



HYBRIDIZA...	SAMPLE.SAMPLE...	HYBRIDIZA...	SAMPLE.SAMPLE...	HYBRIDIZA...	ARRAYRAW...	ARRAYRAWDATA
A1	SUR1	Cy3	CUR1	Cy5	raw	testA\A1.txt
A2	SUR2	Cy3	CUR2	Cy5	raw	testA\A2.txt
A3	SUR3	Cy3	CUR3	Cy5	raw	testA\A3.txt
A4	SUR4	Cy3	CUR4	Cy5	raw	testA\A4.txt
A5	SUR5	Cy3	CUR5	Cy5	raw	testA\A5.txt
B1	SUR6	Cy3	AB1	Cy5	raw	testB\B1.txt
B2	SUR7	Cy3	AB2	Cy5	raw	testB\B2.txt
B3	SUR8	Cy3	AB3	Cy5	raw	testB\B3.txt

Figure 2-15: batch import preview window

In Figure 2-15, the first column lists the hybridization names that will show in ArrayTrack Database panel. The user can click any column title to sort the columns. If the user are not satisfied with the preview, he can click "cancel" button and re-do the Step 1~ 4. After previewing it, click Import button to start importing. If the data are imported successfully the user will see the message "The data are imported without error".

The batch import function is very efficient and error-preventing when importing data in large scale.

Data Export: for information about exporting data please refer Chapter 9 Data Export

ArrayType Information

Overview: One essential item that needs to be input in Input Form is ArrayType. ArrayType defines basic information about the arrays with which an experiment is being conducted. In ArrayTrack 3.1.3 release, more than 70 "standard" array types have been pre-defined, including ~30 array types manufactured by Affymetrix and many other cDNA or oligonucleotides arrays (e.g. from Agilent, MWG, and ClonTech) used in gene expression studies by NCTR scientists or their collaborators. If in your new experiment you are using an array type shown on the pre-defined list, you can simply select it as the array type for your current experiment. Otherwise, you'll need to define your new array type. Fortunately, we have made it relatively easy for the user to define a new array type.

The ArrayType Information for a particular array can be conveniently viewed from the Chip Library (see Figure 3-44 of Chapter 3).

Activate ArrayType Information

By clicking on New Array Type, the user can create new array type (Figure 2-16).

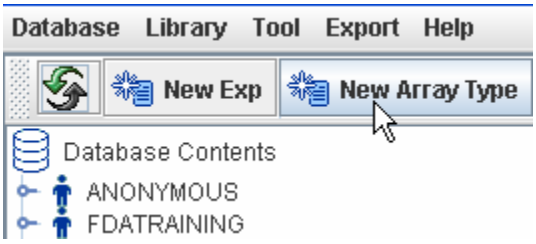


Figure 2-16: Array Type Information - Create new array type.

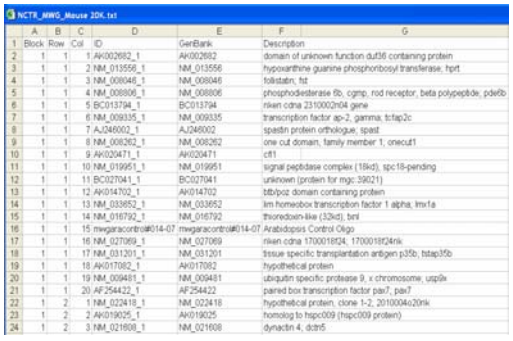
Export ArrayType Information: To export an array type, the user can open the Chip Library, highlight the array type record then click "Export" button. See Chapter 3 Chip Library for detail.

Create New ArrayType: To create a new array type, the user should at first define an ArrayType Information File, which specifies the essential data fields needed for defining the array elements (adherence with the MIAME guideline is recommended). The process is best explained by using examples, as presented below.

Figure 2-17 shows an example of an information file (NCTR_MWG_Mouse 20K.txt) for defining the NCTR_MWG_Mouse 20K array that consists of 48 blocks, and each block consists of 21 rows and 20 columns, resulting in a total of 20,160 spots. Each spot (gene) has its corresponding block#, row#, column#, GenBank accession #, description, etc. This ArrayType Information File can be easily generated from the information about your probes and the way they are printed on the microarray slides.

When the user click New Array Type button

(Figure 2-16), the dialog shown in Figure 2-18 pops up and allows the user to specify the file name for the ArrayType Information File, the way spot location (coordinates) is defined in the file, and the type of array elements (oligo or cDNA). ArrayTrack accepts several types of coordinate definition systems including MIAME's MetaCol/MetaRow/Col/Row notation, Block/Col/Row, Row/Col, and Without coordinates. If no coordinate information is available, the user must specify the number of columns and rows for the array; while this is arbitrary, a square or nearly square arrangement will result in the best graphical displays. In each case, ArrayTrack converts the coordinate information into the MIAME-compliant format: MetaCol/MetaRow/Col/Row. For some cases (e.g. Row/Col and Without coordinates), the MetaCol/MetaRow ratio is automatically set to 1/1, i.e. one MetaCol and one MetaRow.



NCTR_MWG_Mouse 20K.txt

	A	B	C	D	E	F	G
1	Block	Row	Col	ID	GenBank	Description	
2	1	1	1	AK002682_1	AK002682	domain of unknown function duf36 containing protein	
3	1	1	2	NM_013556_1	NM_013556	hypoxanthine-guanine phosphoribosyl transferase; hprt	
4	1	1	3	NM_008046_1	NM_008046	follistatin; fst	
5	1	1	4	NM_008806_1	NM_008806	phosphodiesterase 6b, cgmp, rod receptor, beta polypeptide; pde6b	
6	1	1	5	BC013794_1	BC013794	riken cdna 2310002n04 gene	
7	1	1	6	NM_009335_1	NM_009335	transcription factor ap-2, gamma; tcfap2c	
8	1	1	7	AJ246002_1	AJ246002	spastin protein orthologue; spast	
9	1	1	8	NM_008262_1	NM_008262	one cut domain, family member 1; onecut1	
10	1	1	9	AK020471_1	AK020471	cf11	
11	1	1	10	NM_019951_1	NM_019951	signal peptidase complex (18kd); spc18-pending	
12	1	1	11	BC027041_1	BC027041	unknown (protein for mgc:39021)	
13	1	1	12	AK014702_1	AK014702	btb/poz domain containing protein	
14	1	1	13	NM_033652_1	NM_033652	lim homeobox transcription factor 1 alpha; lmx1a	
15	1	1	14	NM_016792_1	NM_016792	thioredoxin-like (32kd); txl	
16	1	1	15	megaracontrol#014-07	megaracontrol#014-07	Arabidopsis Control Oligo	
17	1	1	16	NM_027069_1	NM_027069	riken cdna 1700018f24; 1700018f24rik	
18	1	1	17	NM_031201_1	NM_031201	tissue specific transplantation antigen p35b; tstap35b	
19	1	1	18	AK017082_1	AK017082	hypothetical protein	
20	1	1	19	NM_009481_1	NM_009481	ubiquitin specific protease 9, x chromosome; usp9x	
21	1	1	20	AF254422_1	AF254422	paired box transcription factor pax7; pax7	
22	1	2	1	NM_022418_1	NM_022418	hypothetical protein, clone 1-2, 2010004o20rik	
23	1	2	2	AK019025_1	AK019025	homolog to hspc009 (hspc009 protein)	
24	1	2	3	NM_021608_1	NM_021608	dynactin 4; dctn5	

Figure 2-17: ArrayType Information File.

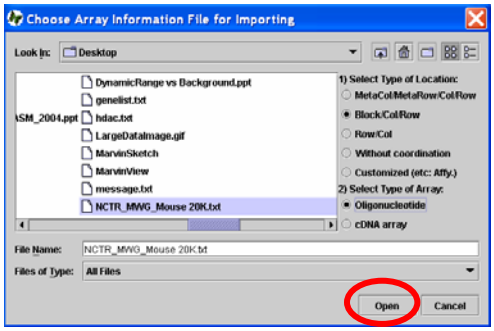


Figure 2-18: Loading ArrayType Information File.

After clicking on Open, the Choose Columns form (Figure 2-19) pops up, which allows the user to enter essential information (e.g. name, platform, manufacturer, layout of array elements, etc.) about the new array type and map data columns in the ArrayType Information File to the corresponding MicroarrayDB data fields by selecting a specific data column header

and clicking on the corresponding MicroarrayDB data field to be mapped. Data columns already mapped will be shown in yellow and their corresponding MicroarrayDB data fields will be marked with a ✔. Clicking on ✖ will undo a previous column assignment. Clicking on OK at the bottom of Figure 2-19 commits data in the ArrayType Information File to the new ArrayType, which becomes available for selection in the ArrayType field of the Input Form (Figure 2-1).

Figure 2-19 gives an example of entering array information in the MetaCol/MetaRow/Col/Row format. In the example there are four Meta columns and three Meta rows, numbered left to right, and top to bottom, respectively. Each Meta column has seven columns and each Meta row has eight rows. Figure 2-19 illustrates assigning data in the ArrayType Information File to MetaCol, MetaRow, Col and Row data fields in ArrayTrack.

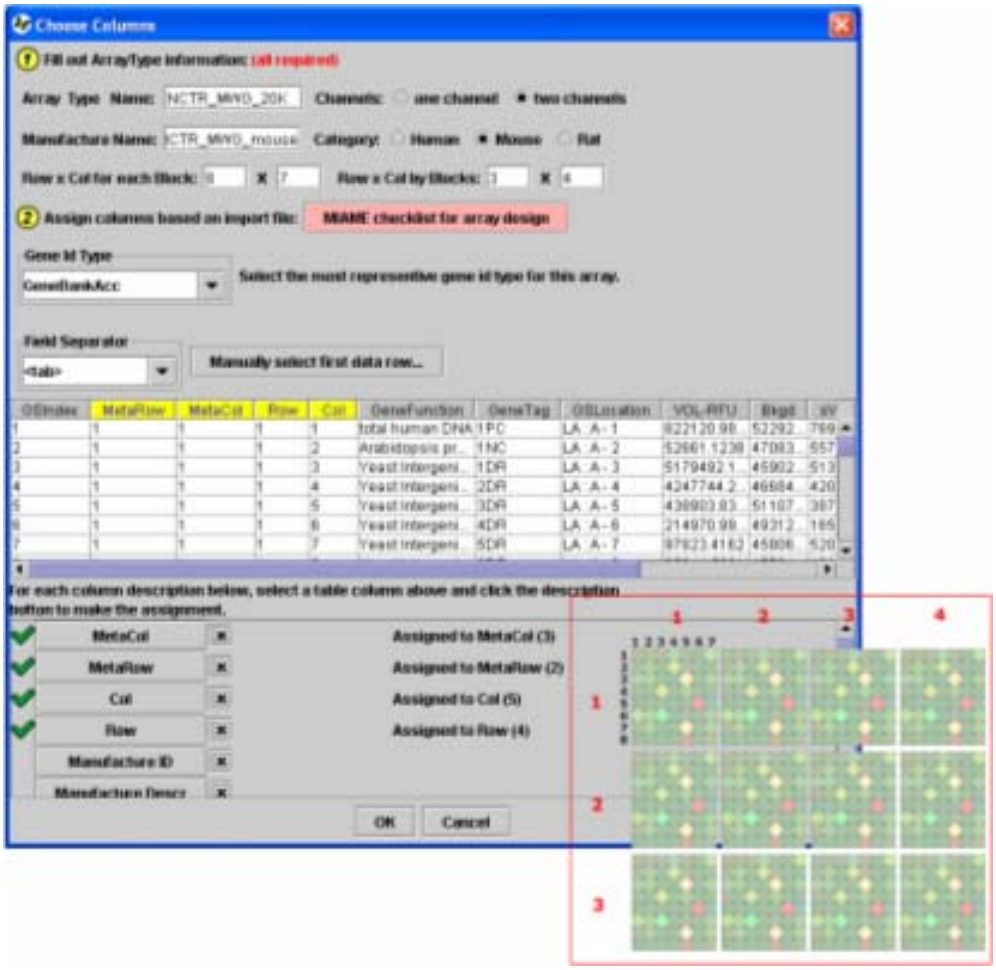


Figure 2-19: Mapping columns in ArrayType Information File to MicroarrayDB data fields and MetaCol/MetaRow/column/row example.

Figure 2-20 gives an example of entering array information in the Block/Column/Row format. In the example, each block has 7 columns and 8 rows and the blocks are arranged in 4 columns by 3 rows and numbered sequentially first left to right then top to bottom. Figure 2-20 illustrates assigning data in the ArrayType Information File to Block, Column, and Row data fields in ArrayTrack.

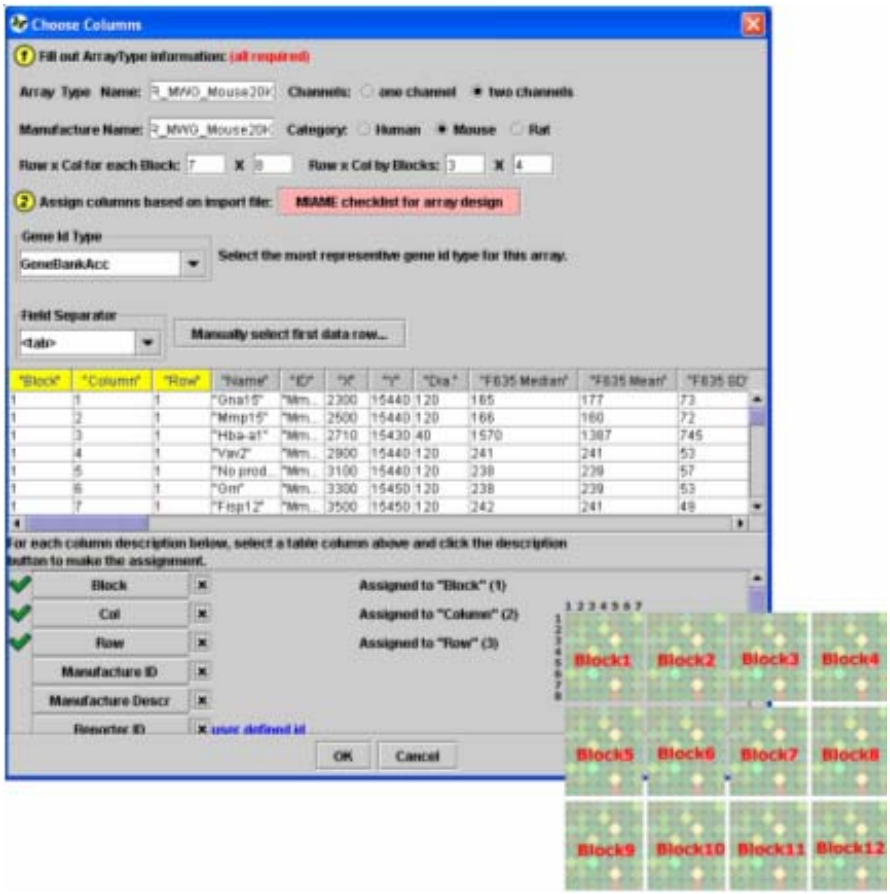


Figure 2-20: Example of Block/Col/Row

Tip: ArrayType information can be conveniently viewed from the Chip Library (see discussion on Chip Library in Chapter 3).

Data Sharing and Security Protection

The owner of the experiment can assign Read/Write privileges to individuals by clicking on Edit Privileges of Input Form (Figure 2-1) to share the experiment data with others (Figure 2-21), e.g. to allow a bioinformatician to access and analyze the data. Data from the experiment is invisible in ArrayTrack to users not granted Read/Write privileges. This security feature allows the owner of an experiment the full control in data sharing while not compromising data security. (Note: ArrayTrack queries the Windows operating system to confirm user login identification)

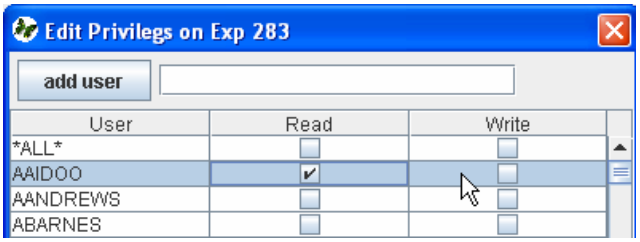


Figure 2-21: Assigning Read/Write privileges to individuals.

To assign privilege to users who are not in the list, the owner can create new user first by typing the user's name (the name that the user uses to login his computer) and then click "add user". The new user name will show in the list, and the owner can assign Read and Write privilege to the new user. See Figure 2-21.

Exploring and Viewing Data in MicroarrayDB from Tree View

Structure of Database Contents: The ArrayTrack database is structured similar to Windows explorer and thus is arranged hierarchically as shown in Figure 2-22. When ArrayTrack is closed, the tree view is saved for the particular user and will be recovered the next time that user restarts ArrayTrack.

The tree structure hierarchy is always arranged in the following nested order: owner, experiment, hybridization, raw data and normalized data. Thus the tree structure appears as follows, with the indicated icons denoting to different types of data:

MicroarrayDB (Database Contents)
- Owner ( )
  - Experiment ( )
    - Hybridization (: two-color array; : filter array; : GeneChip® array)
    - Significant Gene List
      - Raw gene expression array data ( )
        - Normalized gene expression array data (, if available)

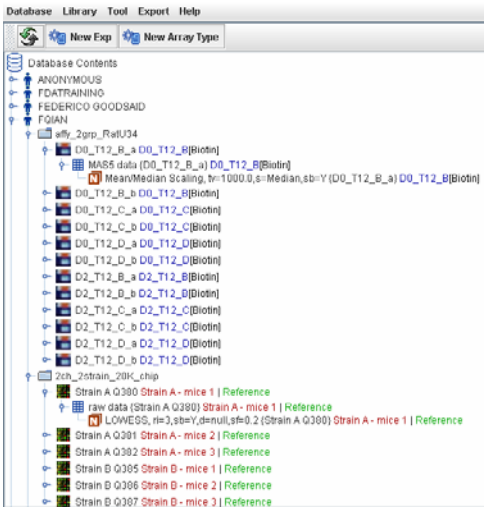


Figure 2-22: Hierarchical Structure of the MicroarrayDB Contents.

Exploring Database Contents: The tree-like Database Contents structure allows the user to select and/or open a particular user ( ), experiment ( ), or hybridization (, , or) by toggling on the lock ( ) unlock ( ) signs. The content structure can also be manipulated by right-click on a particular item (Figure 2-23). For example, by right-clicking on a hybridization, the user can either Expand completely or Collapse completely the content (sub) tree beneath it. A right mouse click makes available various operations that depend on the level of the tree structure, as illustrated in Figure 2-23. Right clicking any level (i.e., experiment, hybridization or dataset) gives a list of options, the last of which is Tree options that when clicked gives the user five ways for displaying (or hiding) more detailed information:

1) Show samples on hybridizations appends sample names to the hybridization name that are colored green or red corresponding to Cy3 and Cy5, respectively;
2) Show hybname on dataset appends the hybridization name to the dataset names.
3) Show samples on datasets appends sample names to the dataset names that are colored green or red corresponding to Cy3 and Cy5, respectively;
4) Show original data filename on raw datasets appends the name of the original raw data file to the beginning of the raw data; and,
5) Show label names appends sample label names (e.g., Cy3 and Cy5) to the all levels of the tree that are colored green or red corresponding to Cy3 and Cy5, respectively.
6) For 1, 2 and 3 above, green and red applies only to two-channel systems using Cy3 and Cy5 labels. For single channel systems, the sample name will be colored blue.

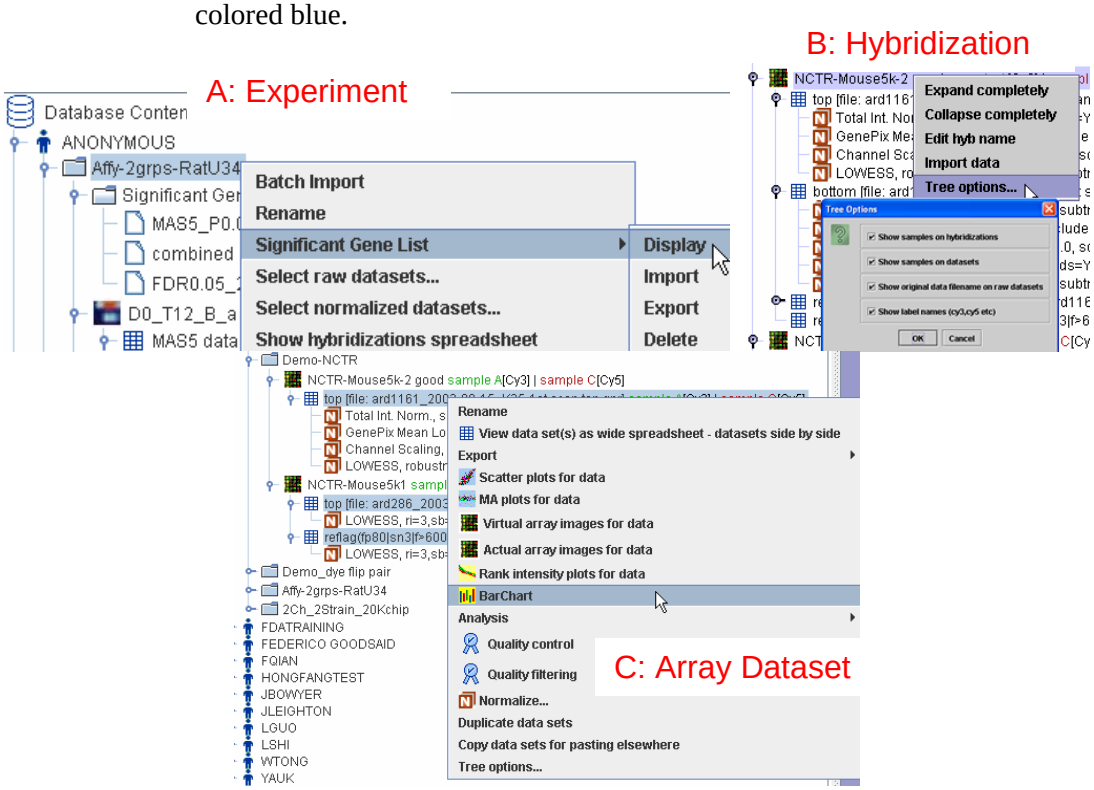


Figure 2-23: Database Contents tree can be expanded or collapsed completely by right-clicking on an item. Depending on the nature of the object that is right-clicked on, a set of applicable functions become accessible. (A) Experiment; (B) Hybridization; and (C) Array Dataset.

Operating on Experiment: Double-clicking on an Experiment will bring up the Input Form (Figure 2-1). Upon right-click on an Experiment, in addition to Expand completely and Collapse completely the tree, the user can choose several other options (Figure 2-23A).

Select raw datasets highlights (i.e., selects) all the raw datasets underneath this Experiment making them collectively available for a subsequent operation (e.g. Normalization).

Select normalized datasets highlights (i.e., selects) all the normalized datasets underneath this Experiment making them collectively available for a subsequent operation (e.g. Data Export).

Show hybridizations spreadsheet collects the information about the all hybridizations contained in the experiment together with the attached toxicological information including all hybridization, sample, treatment and dosing information.

Show possible flip-dye hybridization pairs automatically detects all possible flip-dye hybridization pairs based on the user entered sample and labeling information for all the hybridizations within the Experiment. The term "possible" is used to denote that, for example, a specific sample can be associated with multiple hybridizations that could also be dye-flip paired depending on user input. The flip-dye paring information is listed in a spreadsheet view (Figure 2-24 left). Hybridizations without matched flip-dye pairing will be displayed in a separate spreadsheet (Figure 2-24 right). This function is based on the sample and hybridization information entered in the Input Form (Figure 2-1). Thus, it is important to make sure that information in Input Form is entered correctly. Note that this function only works for an experiment utilizing two dye labels (a so-called two color system).

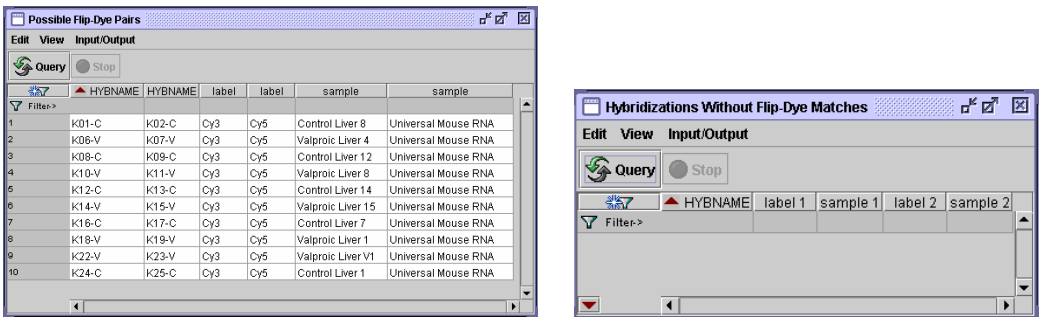


Figure 2-24: Automatic detection and display of Flip-Dye Pairs information for an Experiment. Left: List of matched Flip-Dye pairs; Right: List of Unmatched hybridizations.

Operating on Hybridization: Double-click on a Hybridization will bring up the Input Form (Figure 2-1) with information about the Experiment and the Hybridization shown.

Right-click on a Hybridization allows the user to Rename hybridization and Import data in addition to Expand completely and Collapse completely (Figure 2-23B). As is mentioned in Data Import part of this Chapter, this is another way of loading gene expression array data into MicroarrayDB.

Operating on Array Data: Double-click on an Array Dataset will bring up the Spreadsheet view for this array dataset. Right-click on selected Array Datasets pops up a long list of functions applicable (Figure 2-23C). Instructions for use of these functions are discussed in Chapters 4 through 8.

Significant Gene List

Significant Gene List allows the user to Import, Export, Delete and Display a list of significant genes associated with this Experiment (double-click any experiment will bring up the Significant Gene List showing in a spreadsheet, see Figure 2-25)

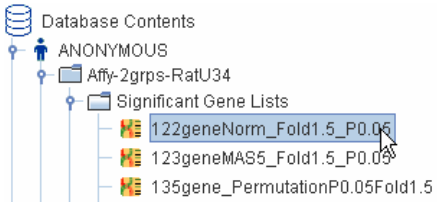


Figure 2-25: Double-click the significant gene lists for display.

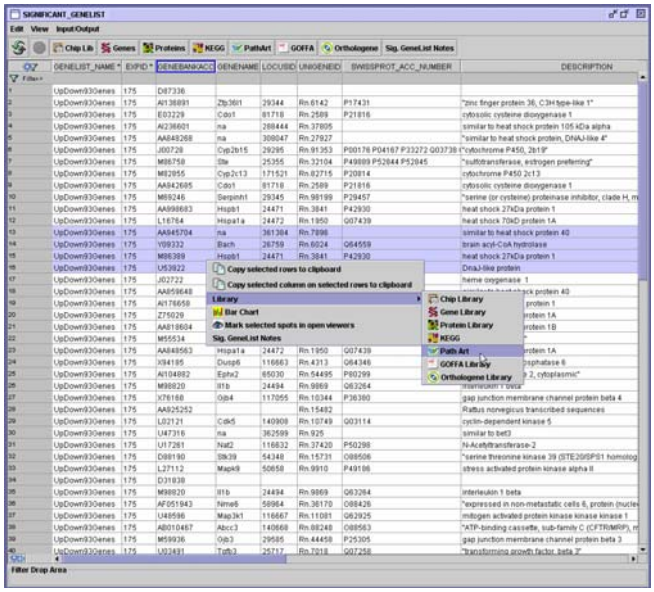


Figure 2-26: Display of Significant Gene List information associated with an Experiment

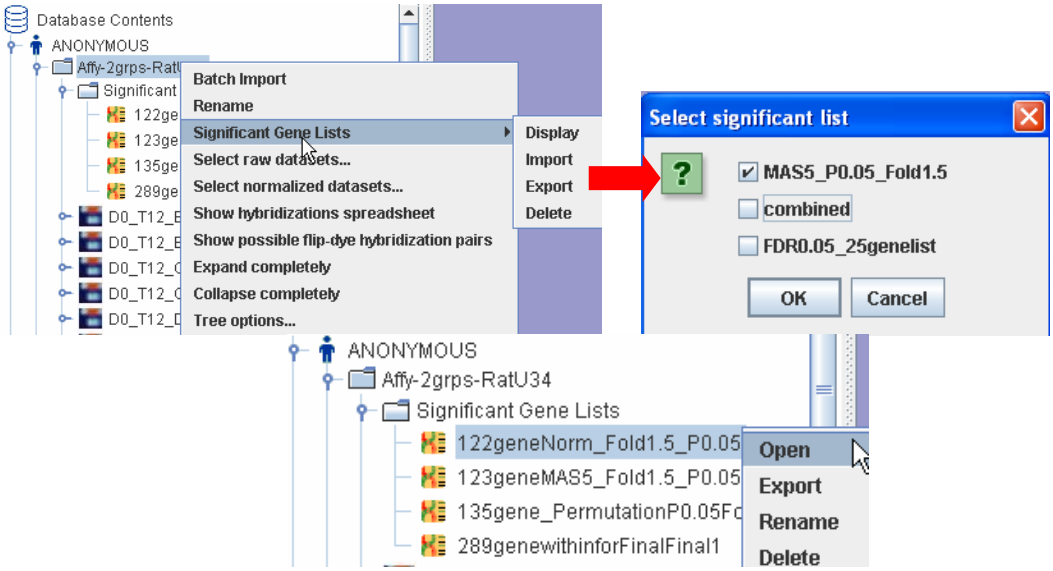


Figure 2-27: select Significant Gene List to display

Display Significant Gene List

Double-clicking the name of a Significant Gene List will bring out the spreadsheet displaying the significant gene list. Another way to bring out significant gene list is right-clicking the experiment-> choosing Significant Gene List-> choosing Display, then a pop-window will show up and list all the significant gene lists under the experiment. The user can choose the significant gene list to display, see Figure 2-27. If the user selects all the significant gene lists to display, then the multiple lists will be shown in one spreadsheet.

The user can open multiple significant gene lists individually and combine the lists to get common gene lists. For example, if the user wants to combine list A and list B. He can open the two lists first separately and then highlight the record in list B, drag to list A, or vice versa, choose the ID's from the two lists for mapping. If the two gene lists are from the same array type, you can select SpotID for matching; if the gene lists are from different array type, then the user

can select LocusID, or GeneName or other ID depending on the column contents in the gene lists. See Figure 2-28 and Figure 2-29.

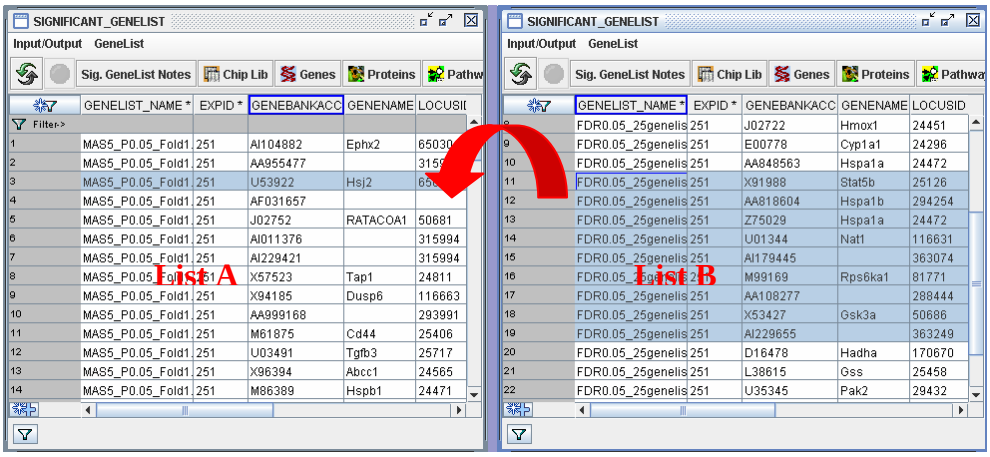


Figure 2-28: combine two lists to get common gene list

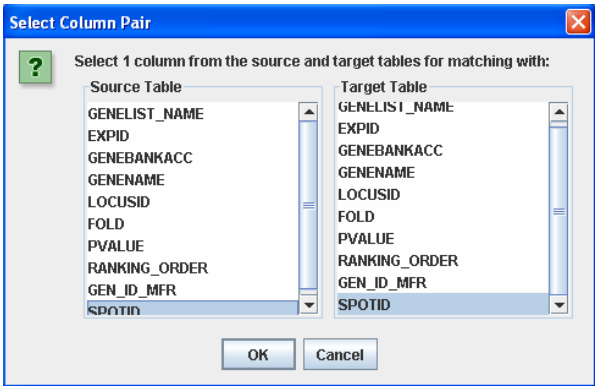


Figure 2-29: choose the column SPOTID for mapping two lists

From the display table, the user can get additional information about the list of genes by accessing other functions including Gene Library, Protein Library, and Pathway Library that appeared on top of the table. The user can also add annotations to the gene lists by clicking Sig. GeneList Notes and then typing in the explanation. Bar Chart (Chapter 6) can be accessed by right-licking on selected gene records.

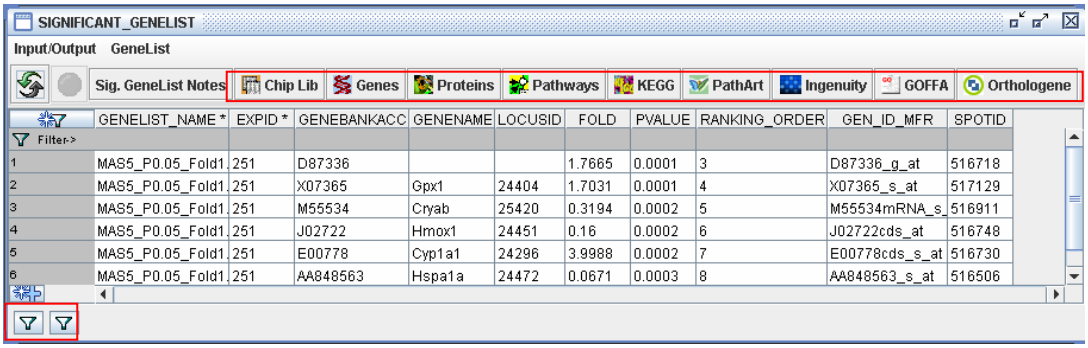


Figure 2-30: The combined gene list result

Figure 2-30 shows the common genes from the two significant gene lists. The number of the filter icons at the left bottom indicates the number of the significant gene lists involved combining. From the combined result the user can access other libraries by clicking the buttons at the top row. Please be aware that if the user click Ingenuity button, the fold change results in the significant gene list will be automatically be converted to $\log_2$Fold or original Fold (when 0< fold <1, it will be – 1/Fold) in Ingenuity analysis results.

Figure 2-31: Fold converting options

The user can continually combine the already combined lists. The lists can be in the same experiment or across different experiments, as long as the array types are the same. This function is very helpful for finding the common genes across different experiment.

Import Significant Gene List:

The text file for Import can be loaded (Figure 2-32) and the data columns in the text file can be mapped to ArrayTrack database fields. The user may have already noticed that the process of importing Significant Gene List is quite similar to that of importing new array type (Figure 2-18).

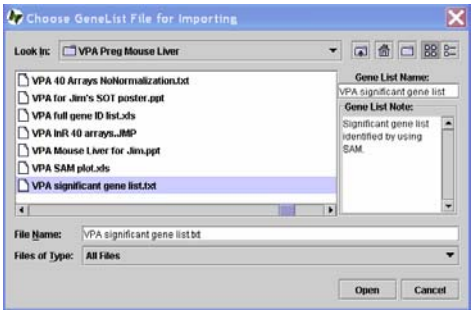


Figure 2-32: Loading a text file with information about Significant Gene List associated with an Experiment.

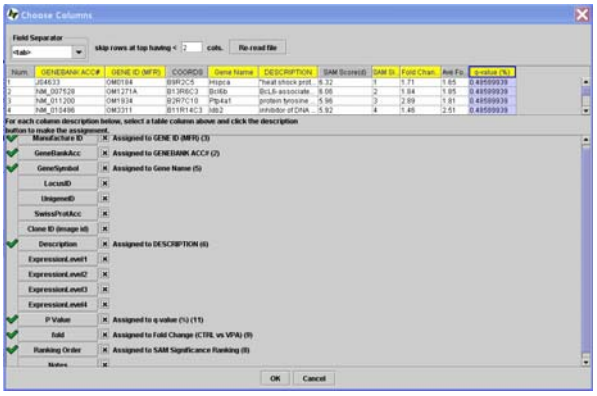


Figure 2-33: Mapping Significant Gene List information to ArrayTrack database fields.

Export Significant Gene List:

Click the significant gene list->export sig. gene list-> save to local drive.

Delete Significant Gene List:

Click the significant gene list->delete.

To create significant gene list:

From T-test/ANOVA result the user can set the filter criteria to get the significant gene list. See Chapter 6.

VennDiagram

The Venn diagram is used to describe the relationships of multiple data set (C_1, C_2, C_3……Cn) which may have some (but not all) elements in common. ArrayTrack can draw the VennDiagram with n<=3. There are 3 ways to activate VennDigram: a) select 2 or 3 significant gene lists and right-click, choose VennDigram then select the ID for the mapping, b) from the Tool panel, click VennDigram icon, c) from the Tool pull-down menu. See Figure 2-34.

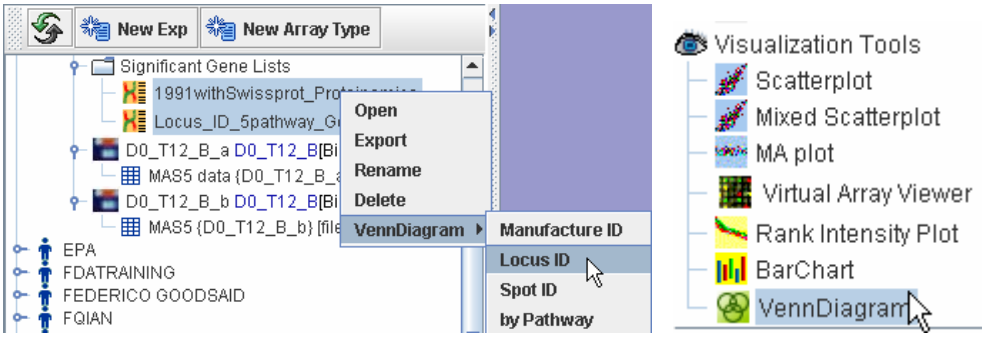


Figure 2-34: get common genes from 2 significant gene lists

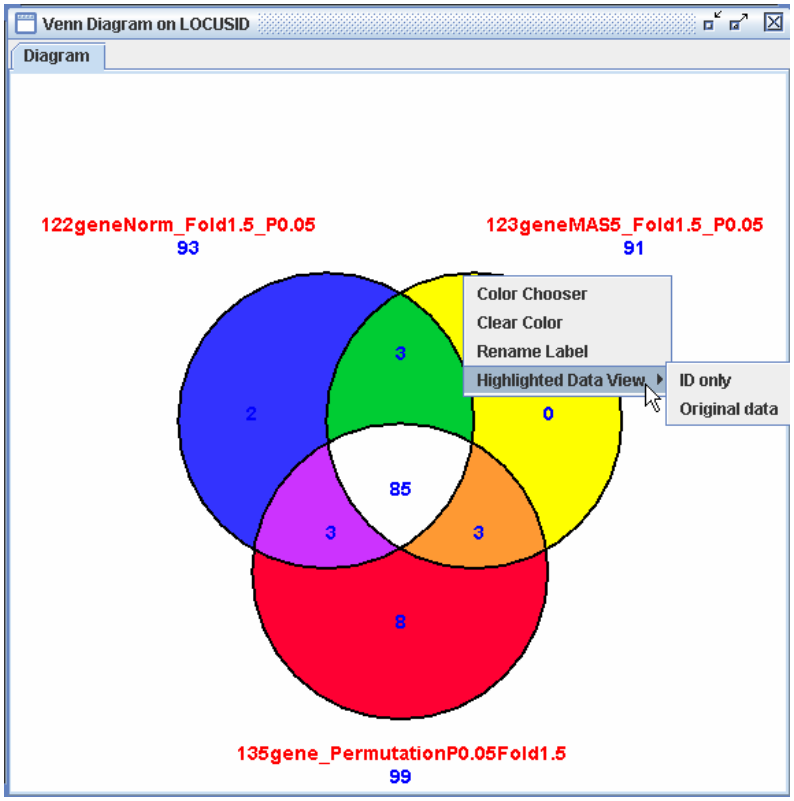


Figure 2-35: The result of the Venn diagram for the common genes

a) If you activate VennDiagram in the first way you will see the Venn diagram (Figure 2-35, an example of the common genes from the 3 gene lists). The red text in the diagram represents the name of the gene list, and blue number represents the number of the unique IDs in the significant gene lists. The blue number in the overlapped part of the circles represents the number of the common genes. The user can label the circles with different colors by right-clicking anywhere in the window and select Color Chooser. Right-clicking -> choosing Highlighted Data View -> selecting Original data will bring up the data tables for all the colored parts and selecting ID only will bring up the table of common ID only. See Figure 2-36.

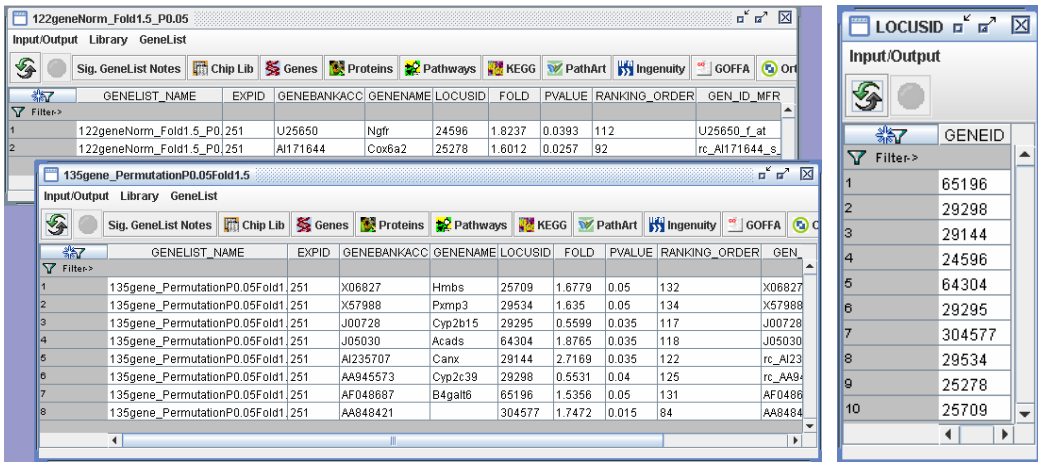


Figure 2-36: The original data view of the common gene list and original data with ID only

b) If you activate VennDiagram (VennDiagram) from the Tool panel, a pop-up window will let you open your gene lists (maximum 3 gene lists) from the local drive. See Figure 2-37. Click open button to load the gene lists and select the common ID column for the three gene lists, then click Draw Venn button. The result of the VennDiagram will show, see Figure 2-38.

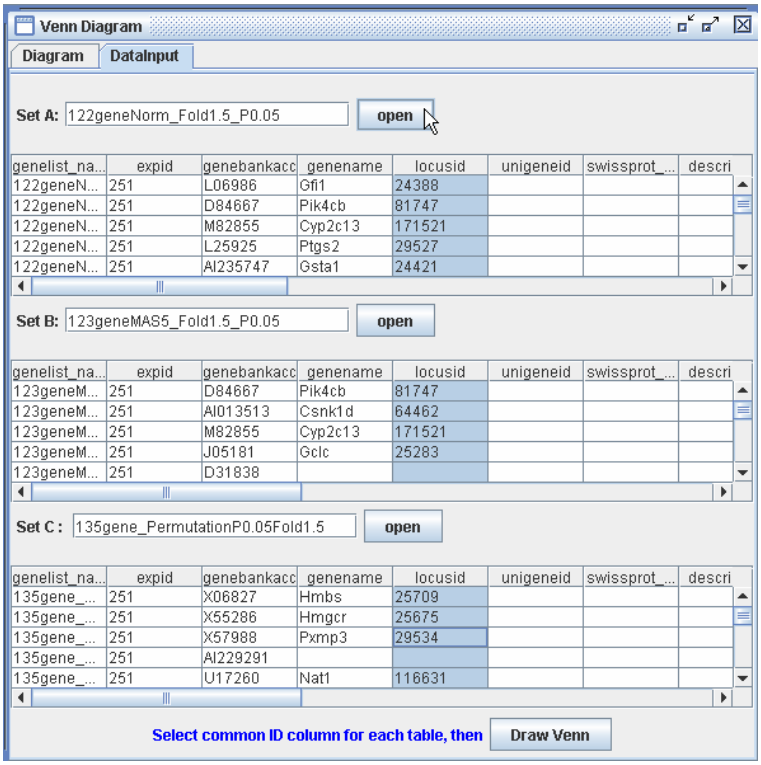


Figure 2-37: open gene lists from local drive

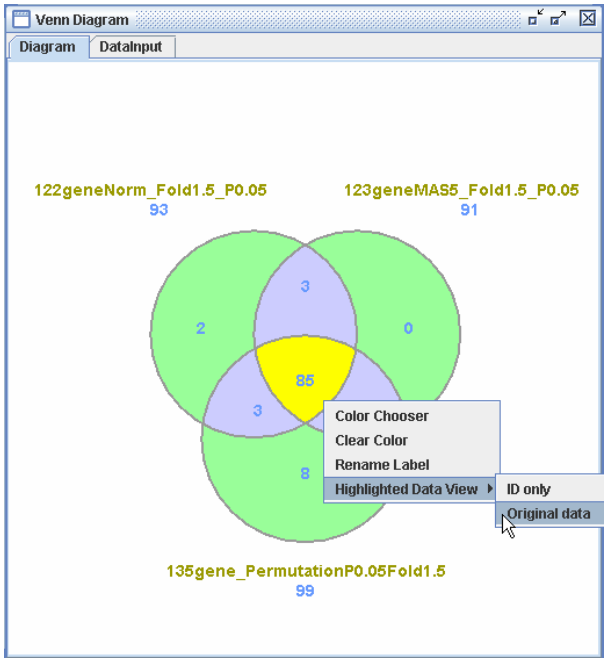


Figure 2-38: VennDiagram for three gene lists

In Figure 2-38, right-clicking any highlighted area in the VennDiagram and choosing HighlightedData View->Original data will bring up the tables (Figure 2-39) under "DataInput" tab showing the three original lists with some highlighted rows corresponding to the highlighted area in the VennDiagram. The user can export those highlighted rows by right-clicking the rows ->choosing Export-> selected rows. The unselected rows also can be exported in the same way. If the user wants to clear all the highlights, s/he can choose "Clear Selection".

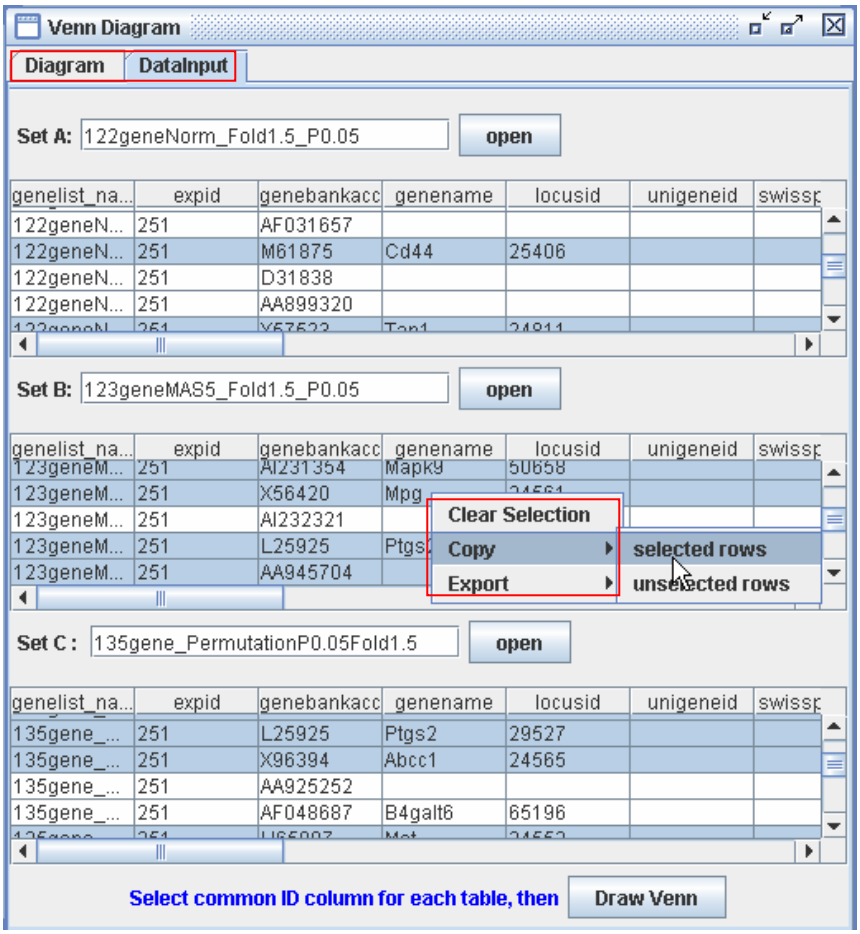


Figure 2-39: Original data view with highlighted rows corresponding to the highlighted area in VennDiagram

c) Activate VennDiagram from the Tool menu.

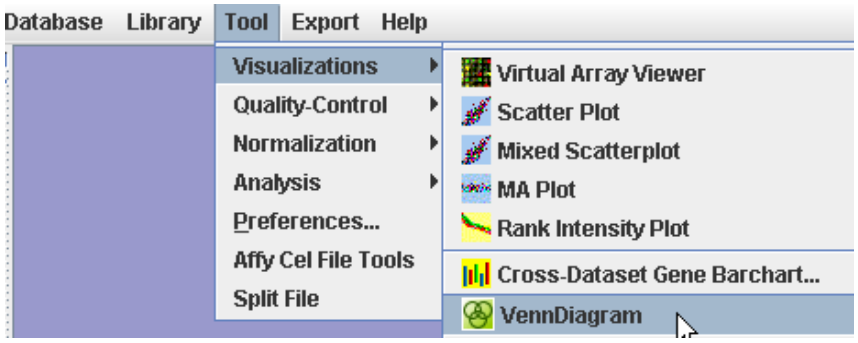


Figure 2-40: Activate Venn Diagram from Tool pull-down menu

Chapter 3 Working with Libraries

The libraries within ArrayTrack provide a powerful capability to augment the analysis of microarray data, and assist in the interpretation of experimental results. Even if the user is not analyzing microarray data *per se*, the libraries provide significant utility for investigating biological data since the salient information and data from public databases on sequence, genes and their function, proteins and their function, conserved orthologous genes and pathways are aggregated and interlinked.

The contents of the online version Libraries in ArrayTrack are updated weekly. If you have a local installation of ArrayTrack in your own server, the Libraries are NOT automatically updated. Instructions for updating your locally installed Libraries are available from our website (http://edkb.fda.gov/webstart/arraytrack/).

The Library panel (Figure 3-1) provides access to the Gene Library, Pathway Library, Protein Library, IPI Library, Orthologene Library, GOFFA Library, Chip Library, Toxicant Library, EDKB Library and ID Converter tool. All libraries can be searched (or filtered) in a number of ways to be explained below. Moreover, results of a query on a particular library can generally be used as the basis for a subsequent search against another library, and so on, allowing the user to drill-down to more detailed and related biological information. For example, the gene library can be searched for a list of genes, the result of which can be used to search for associated protein, pathway and cross-species orthologous gene information in one or more of the other libraries.

The ID Converter tool is used for converting one type of ID to another, e.g. converting Locus ID to Unigene ID or other ID. This tool will be very helpful for library searching.

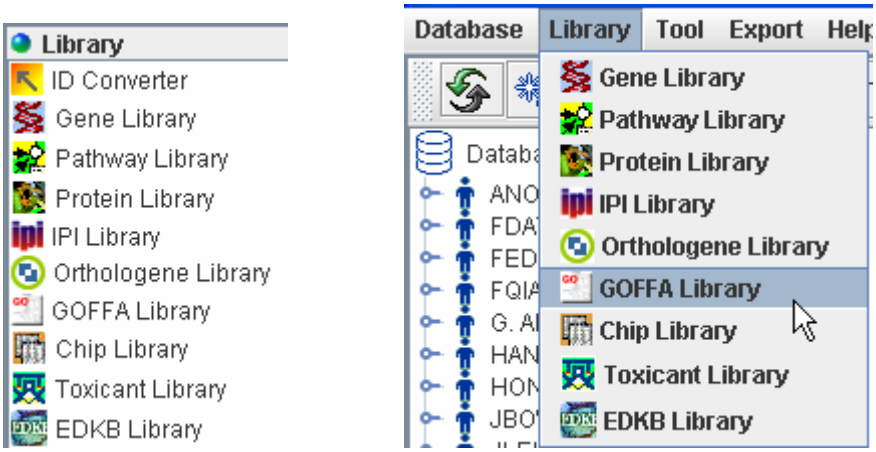


Figure 3-1: Libraries can be accessed from the Library panel or the Library pull-down menu.

ArrayTrack's libraries integrate gene, protein, pathway and other data allowing data interrogation and mining of data across data types. The Gene Library, Pathway Library, Protein Library, IPI Library, Orthologene Library, GOFFA Library and Chip Library are interconnected on the basis of gene, protein or species, whereas the Toxicant Library and EDKB Library are interconnected with genes or pathways based on a chemical's CAS number.

For ease of use, all libraries except the GOFFA present the same user interface, which is split into two panels: 1) the left panel is the search form; and 2) the right panel is for displaying search results (Figure 3-2). Additionally, the user interfaces for these libraries have been to the extent practicable, designed to present identical or similar features and means for executing functions and operations, that is, to have a similar "look and feel". Thus, as the libraries are

individually discussed below, redundancy will be minimized where possible. Therefore, once the user understands the use of the Gene Library, this understanding will mostly apply to the other libraries.

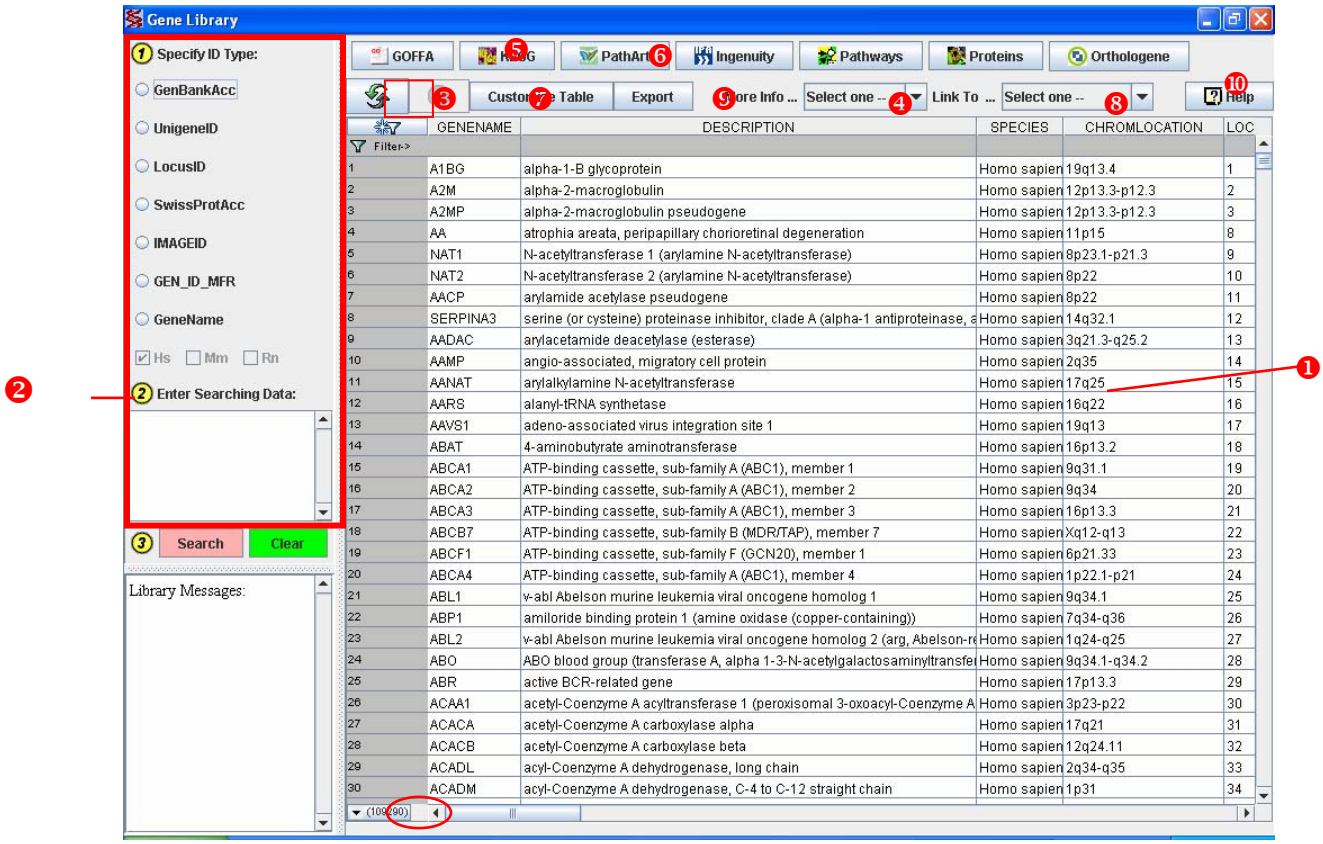


Figure 3-2: Gene Search panel displays the number of records in Gene Library.

Gene Library

Gene Library has human, mouse, and rat data. The user can double-click on Gene Library to bring up the panel. The detailed functionality of Gene Library is summarized as follows:

Working with the Table ❶

The main part of the Gene Library is an Excel-like spreadsheet table. As shown in Figure 3-2, 109,290 records are present in the table (on 5-20-2005). All the gene records in ArrayTrack are searched and the total number of records in Gene Library is displayed; however, only the first few thousands (6,000) are retrieved and displayed in the result table. To get a more complete list of the genes, the user needs to click on the (▼) button shown at the left-bottom of the result table (Figure 3-2).

Filter Genes: A very useful feature about the table is that the user can use a combination of filters to display gene records of interests. The filters can be entered on the line under the individual column names and at the right of the Filter-> sign. After the filters have been entered, clicking the search icon Query (or pressing the Enter key) will display all the records meeting the filter criteria. In the example shown in Figure 3-3, the user entered a histone deacetylase filter under gene DESCRIPTION and a homo filter under SPECIES. As a result, 12 records met these search criteria: 12 human (*Homo sapien*) genes are (related to) histone

deacetylases (HDACs). It is clear from the PATHWAY and ONTOLOGY columns that the biological functions of HDACs are mainly cell cycle-related.

It is important to note that by default all the filters are operating in a logic "AND" manner, *i.e.*, only those records that meet *all* the filter criteria will be displayed.

More sophisticated queries can be entered. For example, the user can add additional Filters by clicking on and the entered search criteria can be combined in complicated logic operations as illustrated by example in Figure 3-3. Newly inserted filtering rows can be deleted by selecting it, right-clicking on them and then choosing "remove this filter row".

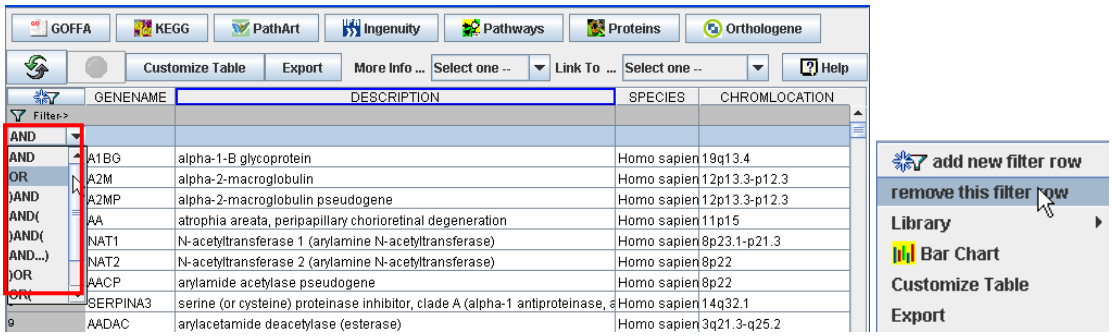
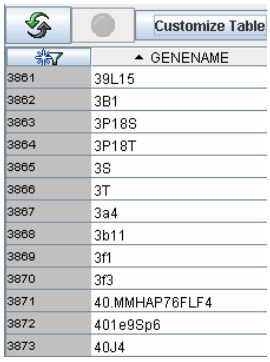
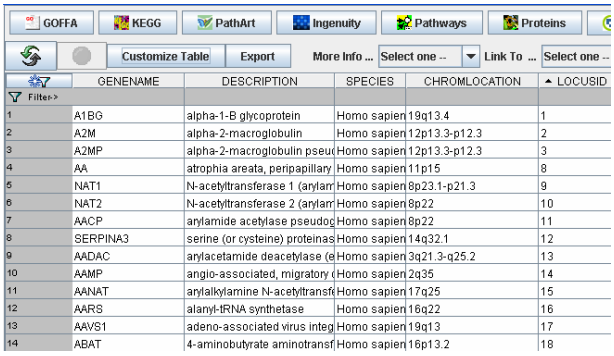


Figure 3-3: Use of a combination of filters to find out interesting genes.

Sort Table by Column: The table can be sorted (or) by toggling on the column header and then pressing Query. Note that the sorting is performed in a way that is consistent with the inherent definition of the data type of the column field. For example, GENENAME is searched by ASCII order (Figure 3-4A), whereas LOCUSID is searched by numerical order (Figure 3-4B). An additional sorting column can be added to a previously sorted table.

A





B

Figure 3-4: Genes can be sorted by columns. (A) Sort by GeneName; (B) Sort by LocusID.

Rearrange Table Columns: The order in which the columns are arranged can be changed by dragging and moving a particular column header to its desired position.

Searching Functional Information for a List of Genes ❷

There are three simple steps to find functional information for a list of genes using the functions in Box ❷ as shown in Figure 3-2:

1. Select a gene IDs (e.g. GenBank accession number, UniGene ID, LocusID, Swiss-Prot accession number, manufacturer's gene ID, and gene symbol) that matches the ID of the gene list.

2. Cut/paste a list of gene IDs of the gene list from e.g. an Excel spreadsheet into the Enter Searching Data box. The gene IDs can be separated by space " ", <tab>, ":", ";", or <Enter>. Checkboxes for Hs, Mm, and Rn are provided for specifying a gene search query (when the ID type is Gene symbol). A **Clear** button is provided for clearing all the entries in the Enter Searching Data box.
3. Click **Search** to find functional information for the gene list in the Enter Searching Data box, which are displayed in a spreadsheet-like table on the right side (Box ❶). Further search can be conducted within the domain of the results from a previous search by checking the within result option next to the **Search** button.

Figure 3-5 displays the results of searching against GeneSymbol for a group of 12 histone deacetylase (HDAC) related genes. The first column is the input gene name that the user types/pastes in the searching section. The rest columns are the matching results for the genes.

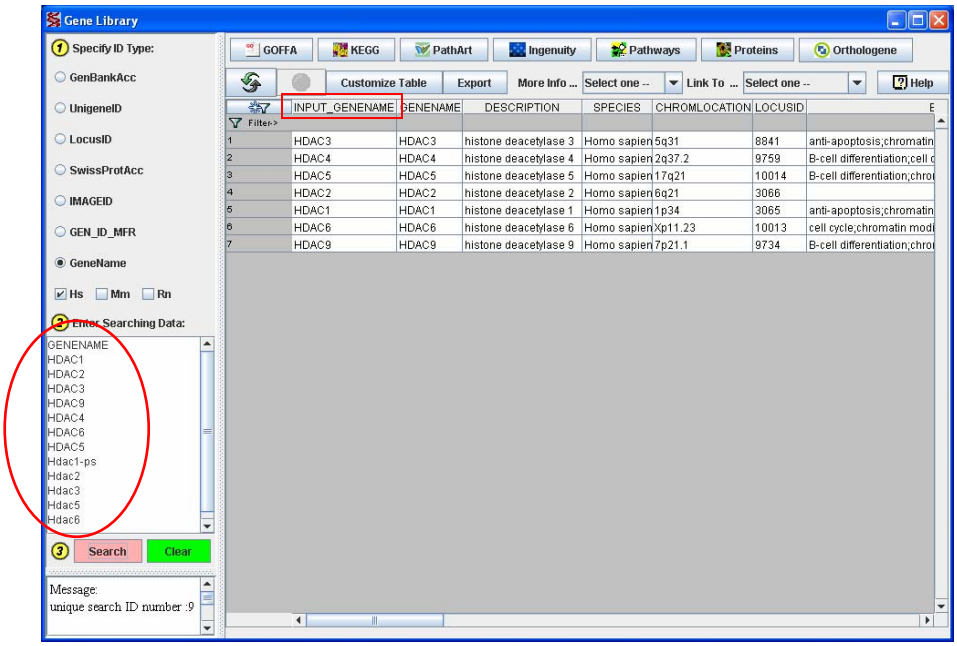


Figure 3-5: Gene Search results by searching against GeneSymbol and Homo sapiens.

More Information about a Gene ❹: On the top of the result table there are several comboboxes: "More info" allows the user to view detailed information for a selected gene record (Figure 3-6), including Synonym, NCBI RefSeq, GenBank Sequences, Gene Ontology, References, Chromosomal Map, Pathway, and Summary. Select the item that you wish to view and click on OK.

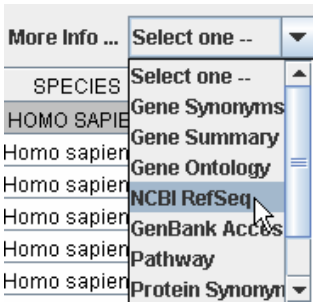


Figure 3-6: More information can be displayed for a selected gene record.

Link to Other Public Databases ❽: "Link to" allows the user to open the official web pages (UniGene, EntrezGene, LocusLink, OMIM, GeneCard, Swiss-Prot, GDB, KEGG, IPI, GeneBank, UniSTS and Homologene) for one selected record (Figure 3-7).

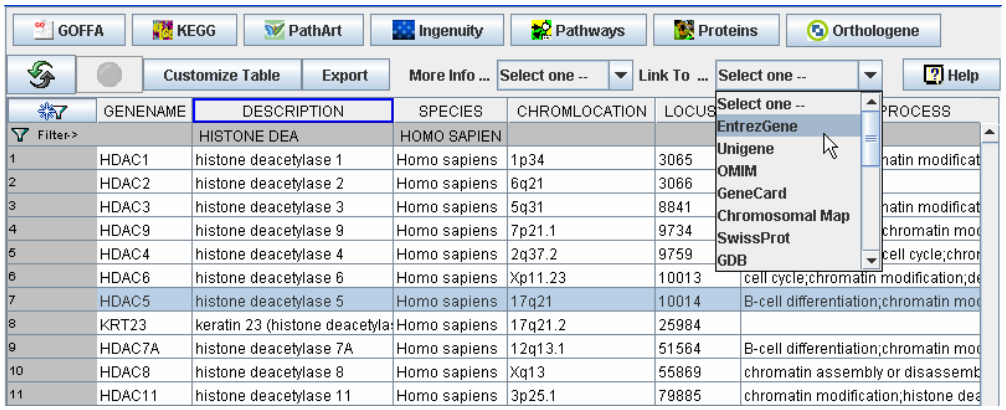


Figure 3-7: Official web sites are linked for a selected gene record.

Sort Genes Based on Their Common Pathway and Proteins (❺ and ❻): The user can highlight any number of genes in the table and click Kegg or PathArt to reorder the highlighted genes based on their shared common pathways. Details are discussed under Pathway Library.

Proteins ... allows the user to explore the protein information for the selected genes. Details are discussed under Protein Library.

Customize Table❼: The table that displays search results can be customized by clicking on Customize Table button in Figure 3-2. The user can select particular fields to be displayed in the table (see Figure 3-8).

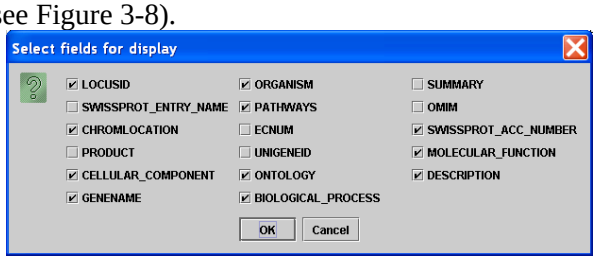


Figure 3-8: Table items that can be chosen for display in the Gene Search results table.

Export Table Contents to a Local File ❾: The contents of a search result table can be exported to a text or Excel file on a local disk by clicking on the Export Table. The dialog box shown in Figure 3-9 allows the user to customize the contents and format of the data to be exported.

Tip: The user can use Copy/Paste to transfer table contents to other applications.

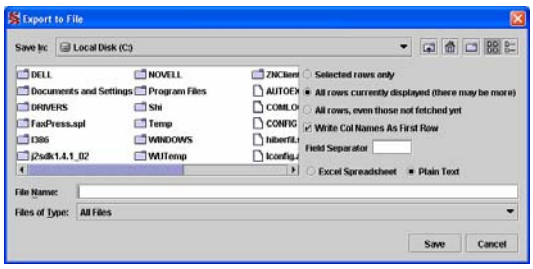


Figure 3-9: Export of search results into a local text or Excel file.

Copy/Paste selected rows/columns: The user can select some of the records, right-click the highlighted rows, choose "Copy selected rows to clipboard" and then pasted at the other place. If the user only want to copy a specific column of the selected rows, he can highlight the rows first, move the mouse to the column to be copied, then right-click and choose "Copey selected columns on selected rows to clipboard". See Figure 3-10.

Stop a Search Session: When a query is being searched, the Query button becomes grayed out and the arrows keep spinning (Query). The user can terminate the current search session by clicking on Stop.

Launch Bar Chart for Selected Genes: Bar Chart across multiple arrays (also see Chapter 6 for more information on the use of the Bar Chart) can be launched by right-click on (a maximum of five) selected genes (Figure 3-10).

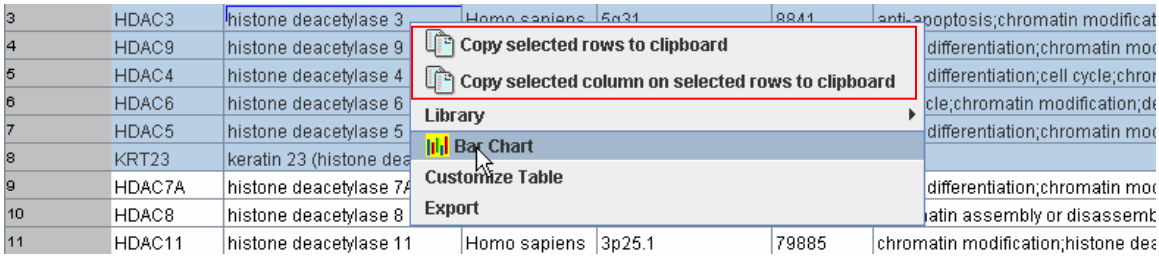


Figure 3-10: Launch Bar Chart for selected gene records.

User can also click Help to get information about the libraries (Figure 3-11).

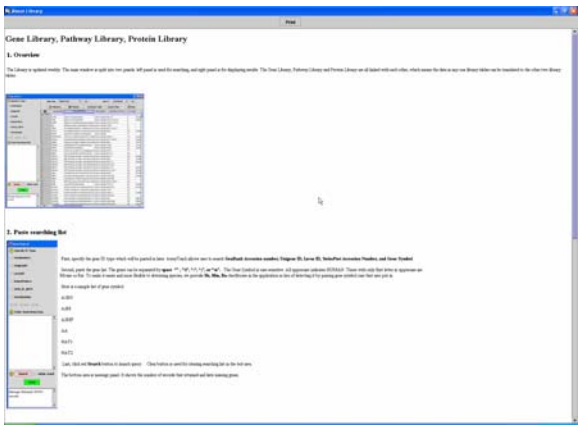


Figure 3-11: Information on the libraries under Help menu

Pathway Library

The Pathway Search panel can be activated by clicking on Pathway Library in the Library panel or Library pull-down menu (Figure 3-1).

Figure 3-12 shows that pathway information is available in ArrayTrack for 31,257genes. This library contains pathway information from KEGG and ParthArt.

The user may have already noticed that the user interface for the Pathway Search panel is the same as that of the Gene Search panel (Figure 3-2). As with the Gene Library, clicking

the **Customize Table** button will display a panel for the user to customize the information that will be displayed (Figure 3-13).

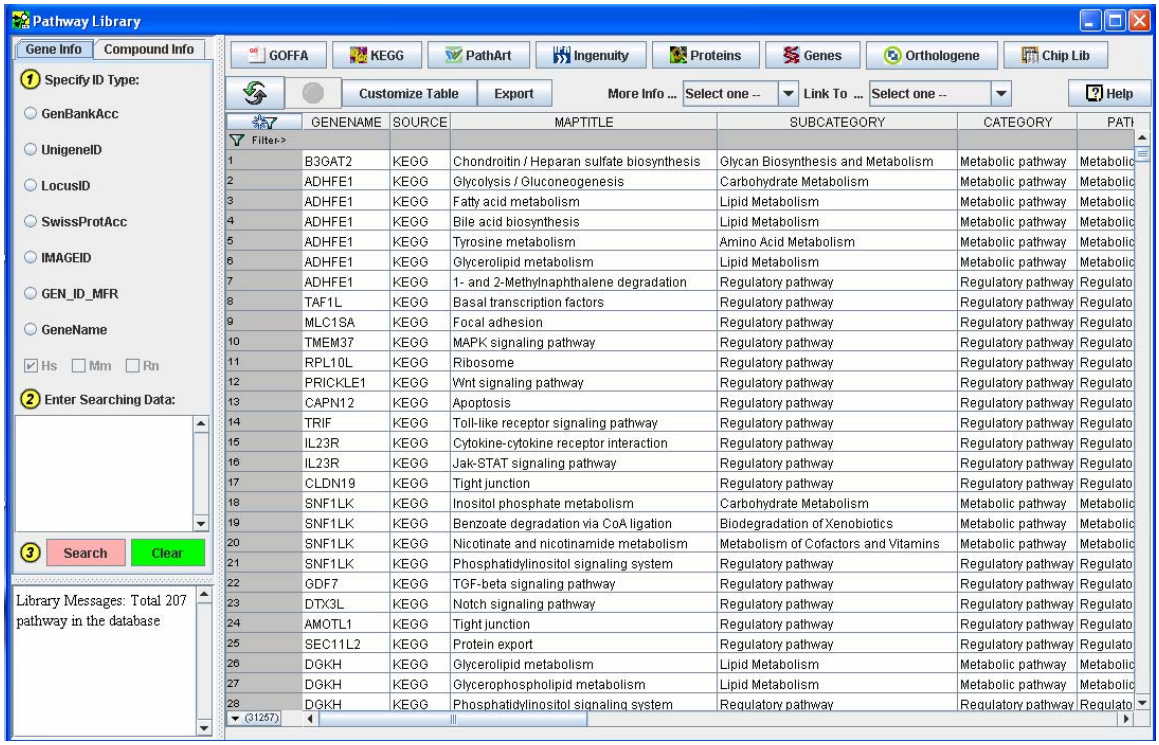


Figure 3-12: Pathway Library window

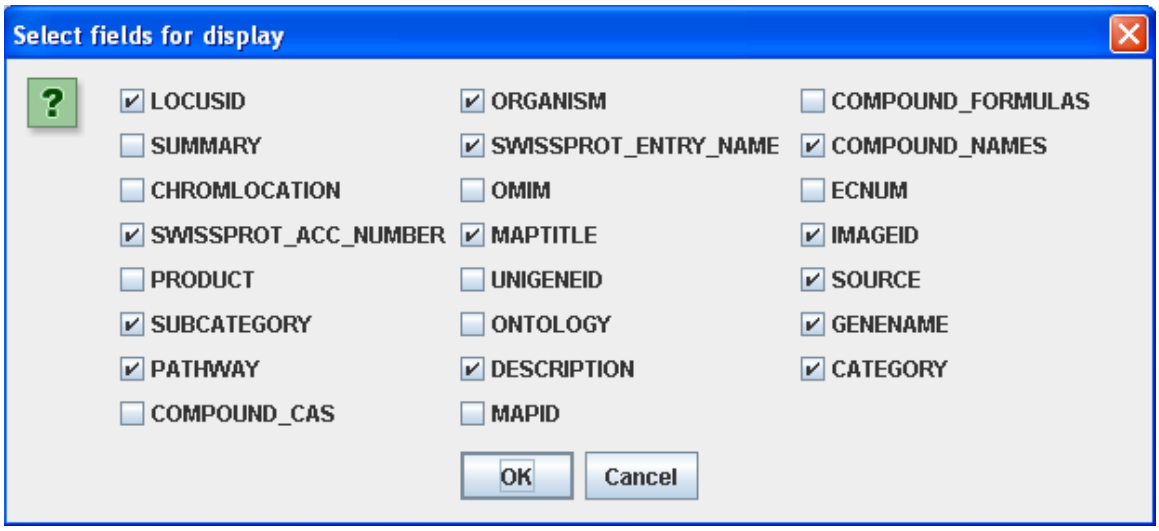


Figure 3-13: Select fields for display

Two types of search can be done in the Pathway Library: 1) search based on gene information; and 2) search based on chemical compound information. The user chooses the type of search by clicking either the Gene Info tab, or the Compound Info tab that are located in the upper left of the panel (Figure 3-12).

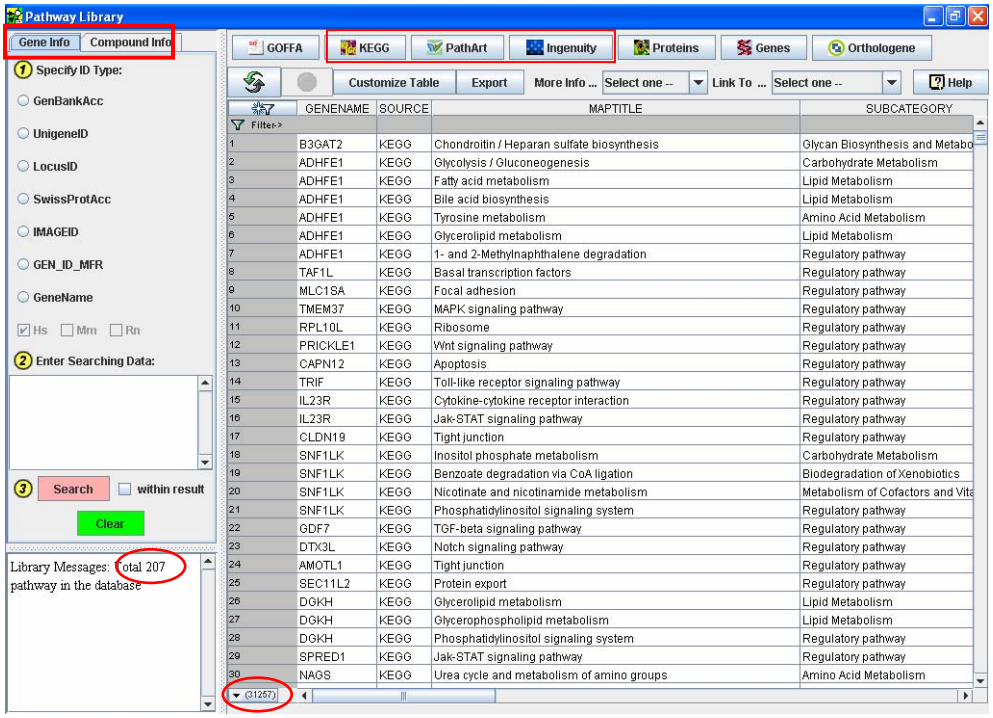


Figure 3-14: Pathway Search panel for gene information.

When searching by Gene Info, the user specifies the type of ID, types or pastes the ID in the window, and then clicks the search button. All the pathways related to the genes are displayed. Also note that a button is provided to clear the contents in the search data window. If searching by GeneSymbol (i.e., gene name), then the user can check the boxes to include Human (Hs), Mouse (Mm) and Rat (Rn) in the search.

When searching by Compound Info, the user can specify compound name ("=" option) or part of a name ("contains" option), and/or the number of carbon, hydrogen or oxygen atoms in the compound's chemical formulae, and/or the compounds CAS Number. CAS numbers can be typed or pasted in the window provided, and can be cleared by clicking the clear button. An example of searching by chemical formula is shown in Figure 3-15.

Note that, for searching for either Gene Info or Compound Info- based search, a subsequent search can be performed that is limited to the genes or compounds associated with the previous search by checking the "within result box".

As for the Gene Library, the user can rearrange the table columns by dragging and moving the column header.

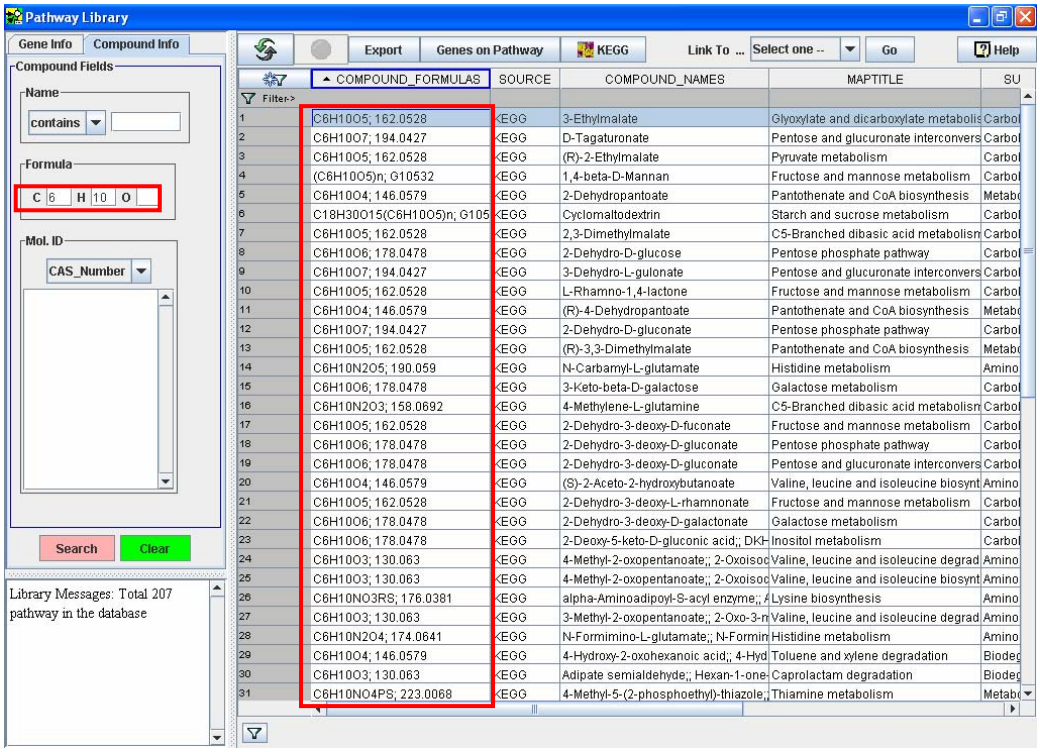


Figure 3-15: Pathway searching panel for compound information

The user can select a group of interesting or relevant genes from the display table and request pathway information from ArrayTrack. After the selecting the genes, the user clicks on either the Kegg or PathArt button; since Kegg and ParthArt are different libraries, probably different search results can be expected. There is another button Ingenuity for pathway that will be addressed in detail later. In Figure 3-16, highlight all the filtered genes after typing "citrate cycle" in the first row, then click Kegg.

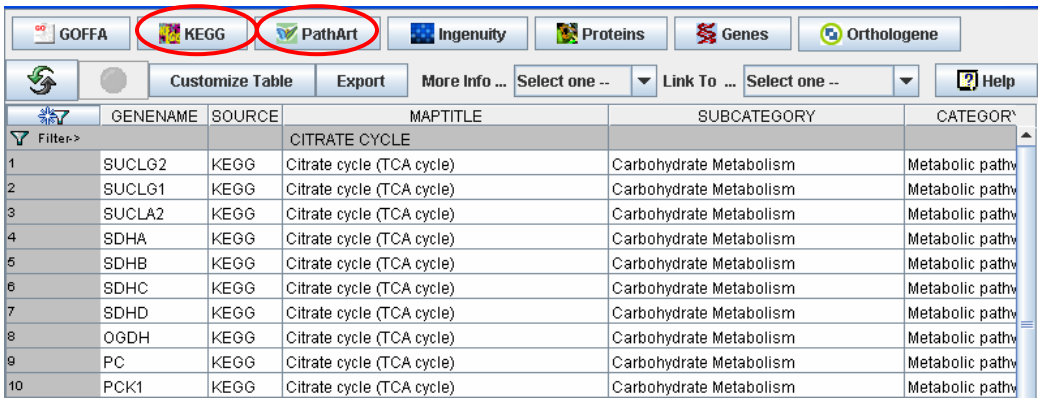


Figure 3-16: type "citrate cycle" in the first row for filtering, then highlight all the filtered results.

After clicking on the Kegg button, the user is prompted to select which species is to be searched (Figure 3-17). Multiple species can be specified for one search.

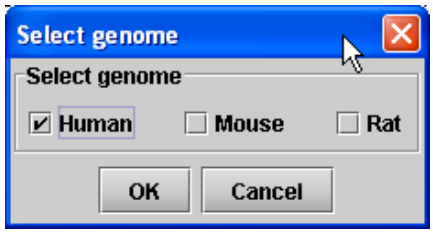


Figure 3-17: Select species for Pathway Search.

All pathway information related to these genes is displayed in a separated table, called Pathway table, as shown in Figure 3-18. At the bottom of the pathway table, a summary on pathway information is provided: *Total submitted genes: 33, 18 genes found, 15 not found, Total 13 pathway maps*, indicating that the user submitted 33 genes to ArrayTrack for pathway information and pathway information is available for 18 genes, which are involved in 13 different pathways. The Pathway table is organized by gene name, Locus ID, pathway map, and the category of the pathway.

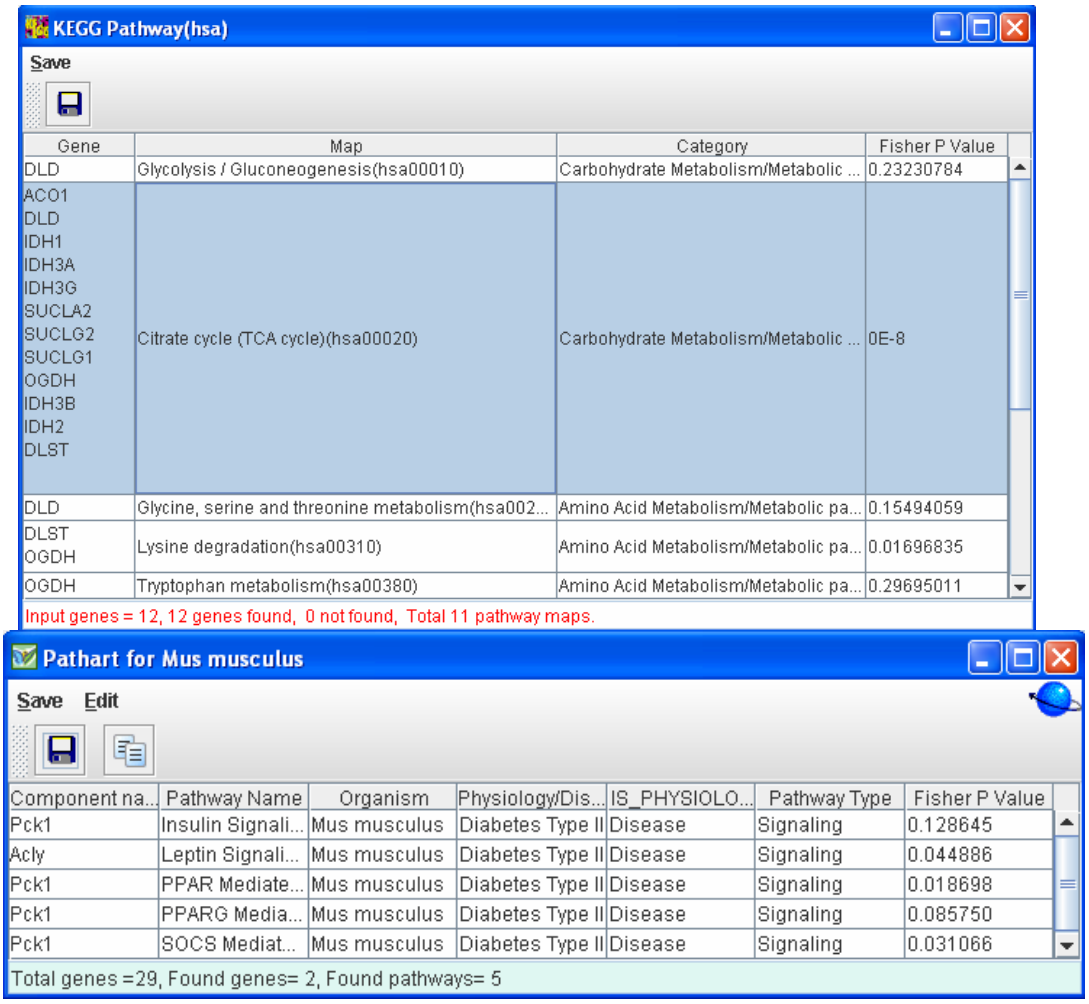


Figure 3-18: Search results from Kegg (A) and PathArt (B).

When you move the mouse cursor over the name of a pathway map or its category, that pathway will be highlighted. In Figure 3-18A, The "Citrate cycle (TCA cycle)(mmu00020)"

pathway is highlighted. By clicking on the pathway that is highlighted, you will be directed to the KEGG/Pathway HTML page (Figure 3-19).

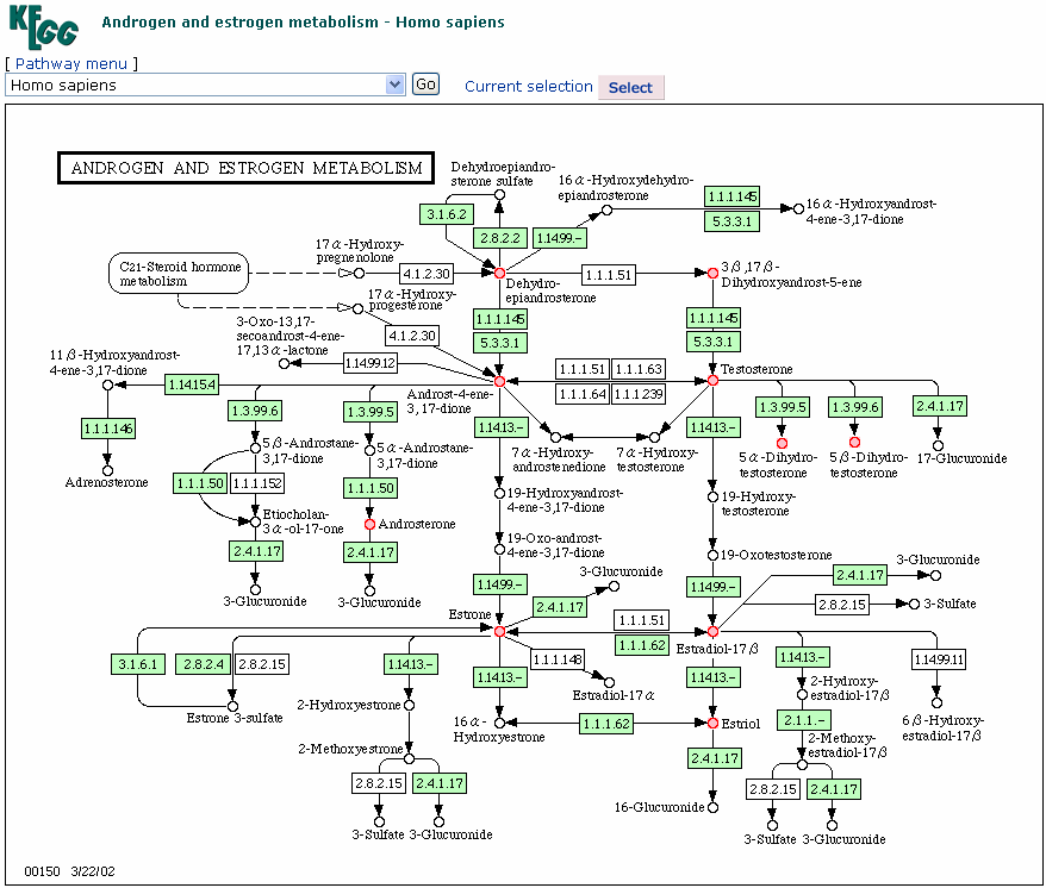


Figure 3-19: A list of the pathways in which the eleven genes are involved is displayed in the format of KEGG/Pathway database search result.

In Figure 3-18B, the user can highlight and then double-click one record, then the Jubilant Biosys's PathArt will show up, see Figure 3-20.

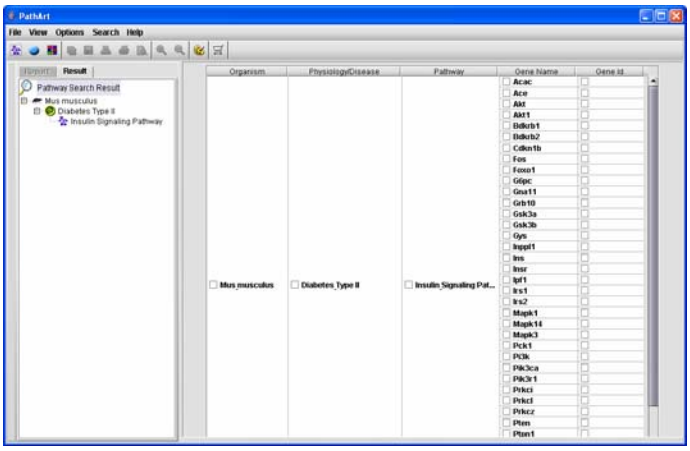


Figure 3-20: Jubilant Biosys's PathArt

Click the pathway name under the "Result" tab in the left panel, the pathway will show up in the right panel. See Figure 3-21.

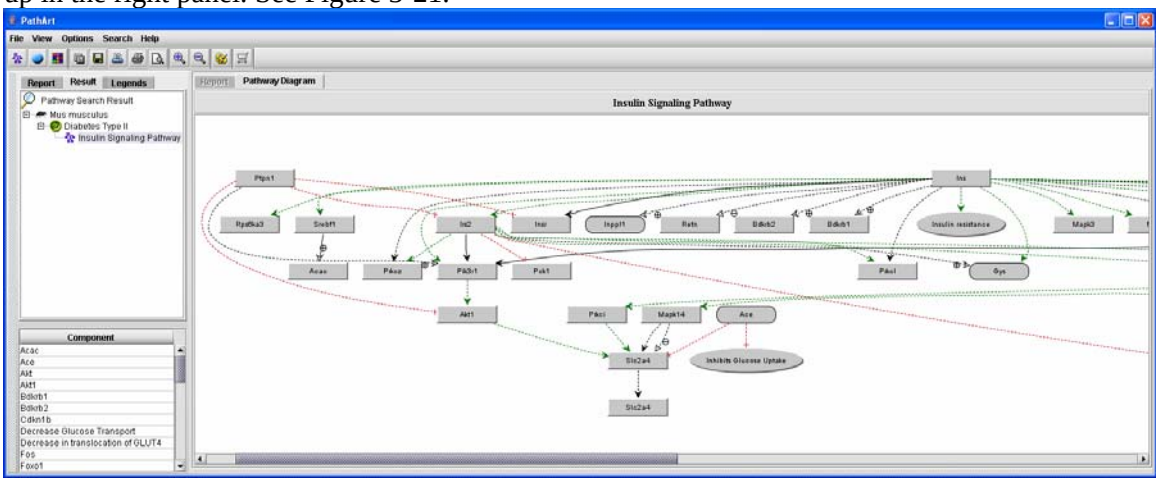


Figure 3-21: Pathway from PathArt

Chemical structures can also be displayed by double-clicking on the compound name. The Pathway table can also be saved in a local file by clicking on , see Figure 3-18.

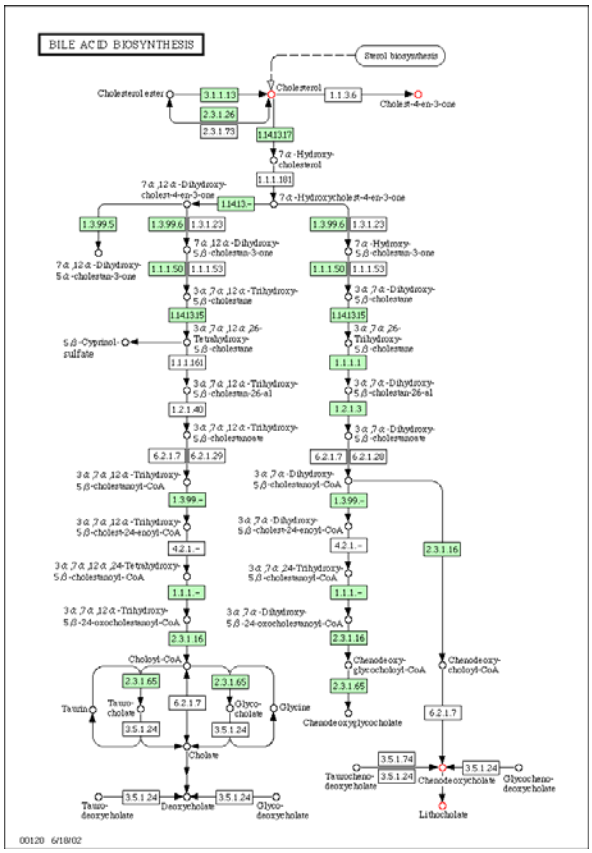


Figure 3-22: KEGG pathway map (hsa00120) on Bile acid biosynthesis. Query compounds are highlighted in red cycles.

Ingenuity Library

Ingenuity delivers systems biology expertise to biologists and bioinformaticians through pathways analysis software, genome-scale computable network databases, and knowledge management services and infrastructure, resulting in increased R&D productivity and faster drugs to market. This manual will not describe how to use Ingenuity software. For more information about Ingenuity, please visit website: http://www.ingenuity.com/. This website provides tutorial and manual for using *Ingenuity* software.

In Figure 3-17, if the user highlight some record and click

button, then the user will go into Ingenuity system to get pathway analysis. The list of genes will be passed to the Ingenuity and directly launch the software. First the user needs to type the name and password to login to the software, see Figure 3-23.

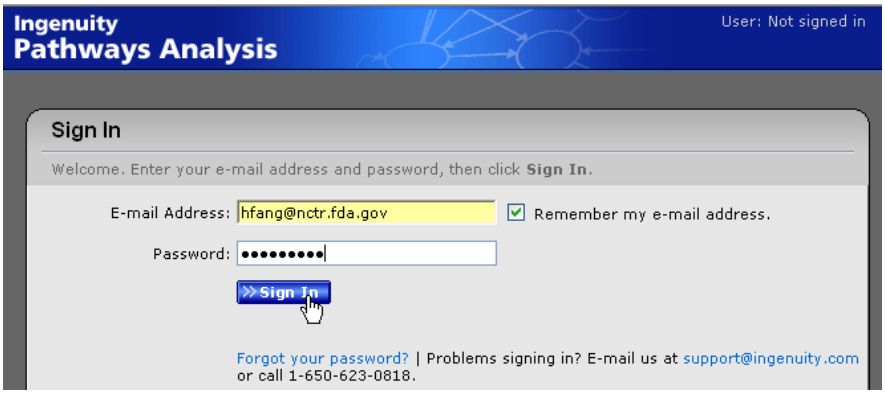


Figure 3-23: login to Ingenuity Pathway software

Once login, the following window will show up. The user can type in the name for the dataset file, then click "Next" button. See Figure 3-24.

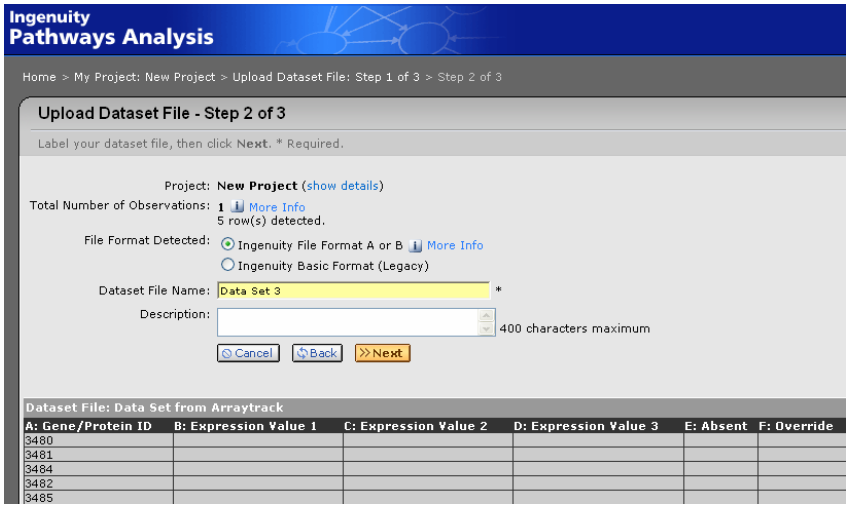


Figure 3-24: Upload dataset file

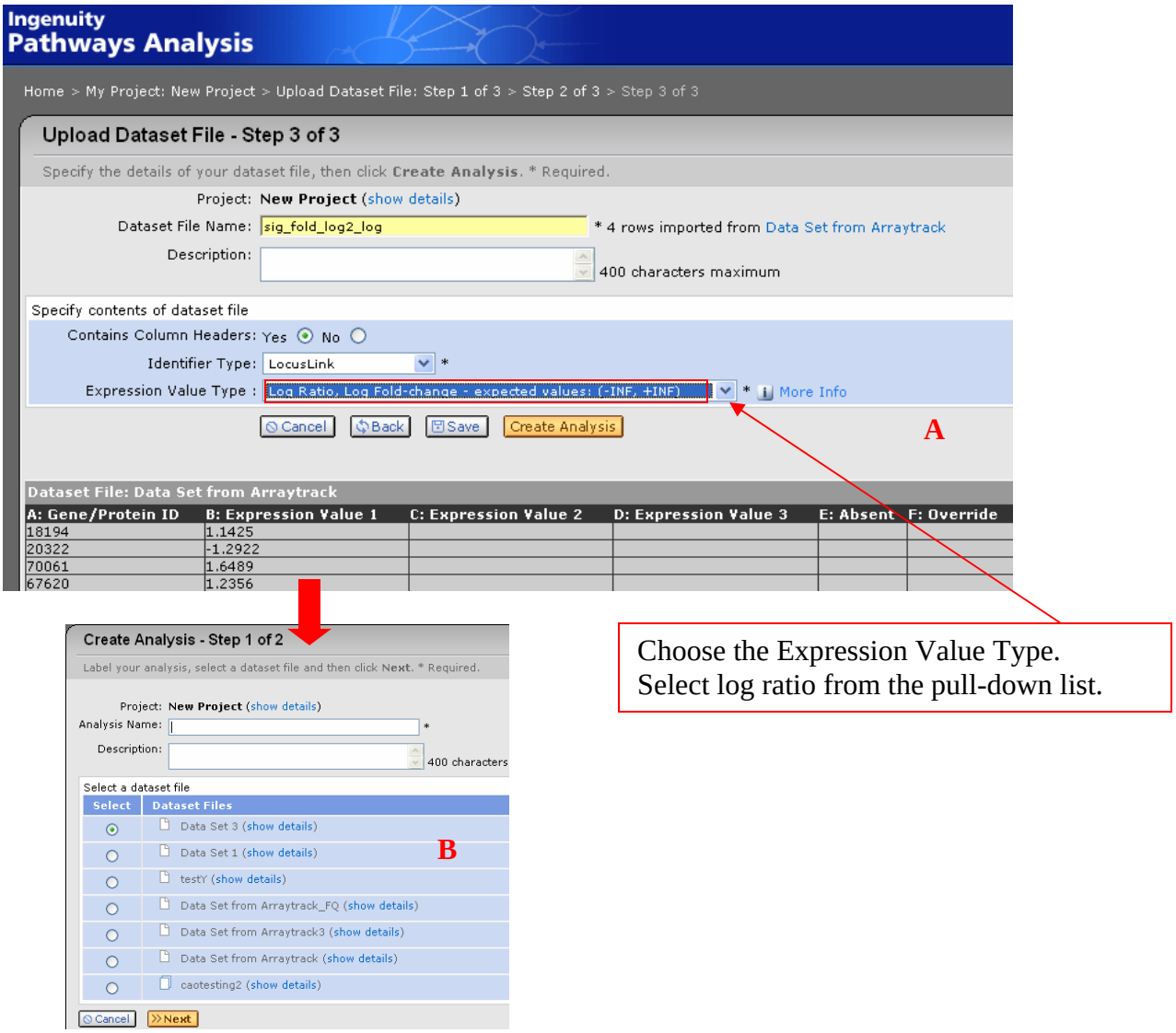


Figure 3-25: State the detail of the dataset before create analysis and give a name for the analysis

In Figure 3-25A, the user needs to specify the identifier type (ArrayTrack uses LocusLink as the identify type) and expression value type (ArrayTrack will automatically convert the fold change to logFold or Fold depending on the selection, so the user should choose Fold or logFold) then click "Create Analysis" button. A new window show up letting the user create a name for the analysis (see Figure 3-25B). Click "next" button and click "Run Analysis" button.

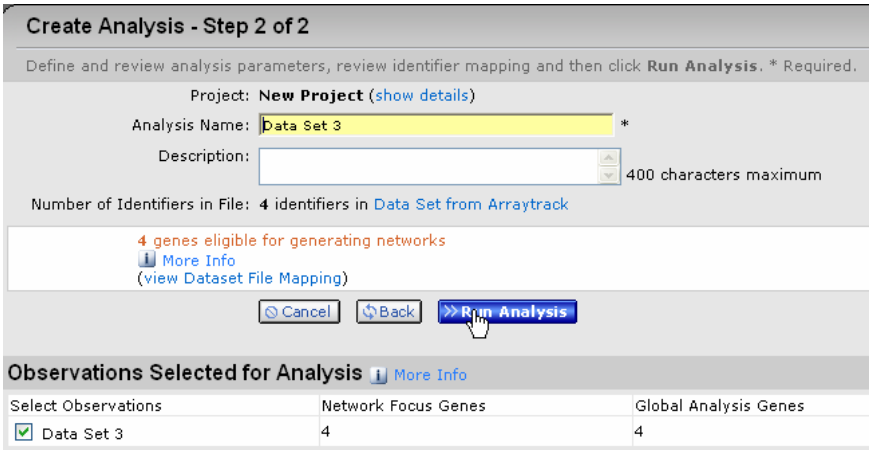


Figure 3-26: Run Analysis

Figure 3-27: Choose the dataset to see the results

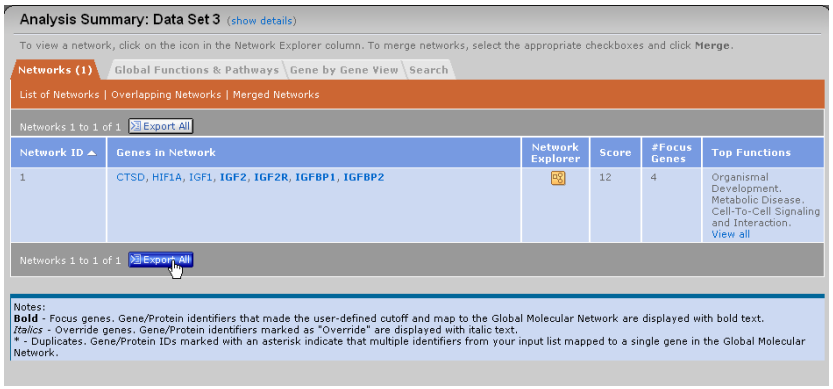


Figure 3-28: The results of the analysis

Protein Library

Similar to accessing the Gene Search and Pathway Search panels, the user can click on Protein Library to bring up the Protein Search panel, which consists of two parts: the left part is the search form, and the right part displays the search results. Figure 3-29 shows that protein information is available for 24,024 records in ArrayTrack.

Similarly, the user can search the Protein Library by pasting a list of gene ID's. The result is represented in a spreadsheet-like table on the right side. The gene ID's that can be searched against include GenBank accession number, UniGene ID, LocusID, Swiss-Prot accession number, manufacturer's gene ID, and gene symbol. Further search can be conducted within the domain of the results from a previous search by checking the within result option next to the **Search** button.

The contents of the table can be customized by selecting data items shown in the Figure 3-30.

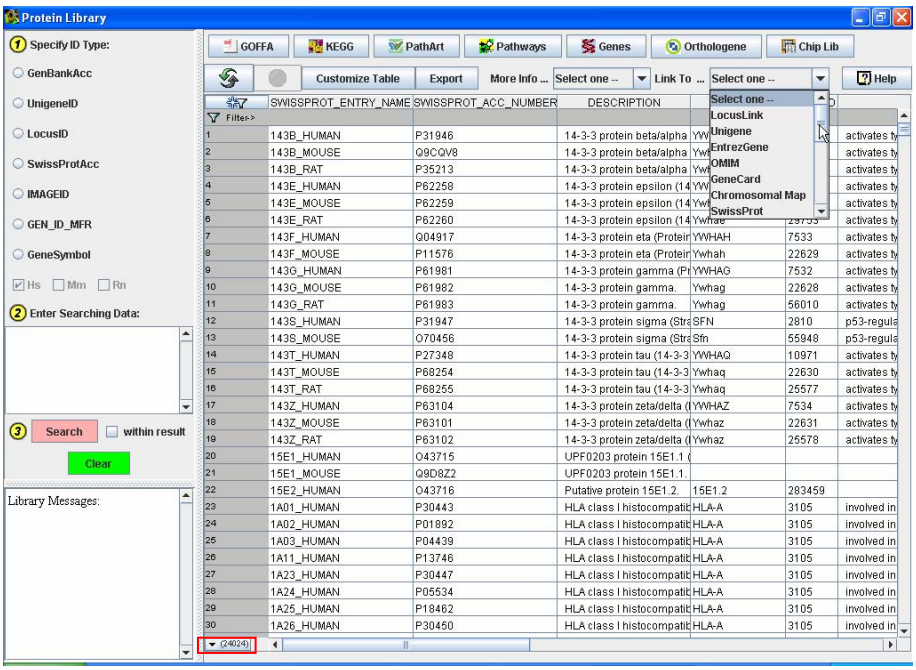


Figure 3-29: Protein Search panel.

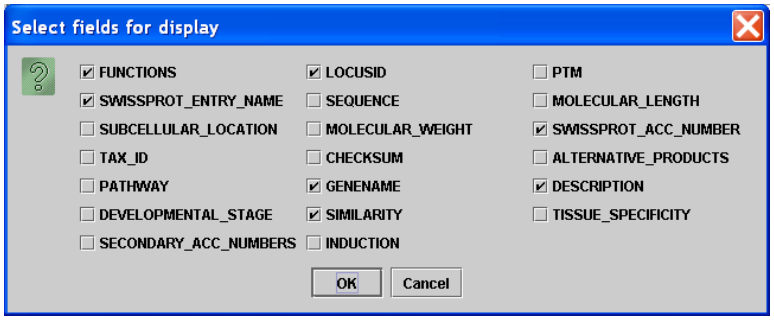


Figure 3-30: Items that can be chosen for display in the Protein Search result table.

Link to Other Public Databases: Just the same as for the Gene Library and Pathway Library, "Link to" provides the user the gateways to other official web pages (Swiss-Prot, PDB, SWISS-2DPAGE, Pfam, PROTSITE, InterProt, SMART, ProDom, UniGene, EntrezGene, LocusLink, OMIM, GeneCard, GDB, Kegg, IPI, UniSTS, Homologene) for obtaining information about selected protein (see Figure 3-29).

IPI Library

IPI Library contains protein information from the International Protein Index (http://www.ebi.ac.uk/IPI/IPIhelp.html). The IPI search panel (Figure 3-31) is similar to that of Gene Search and Protein Search.

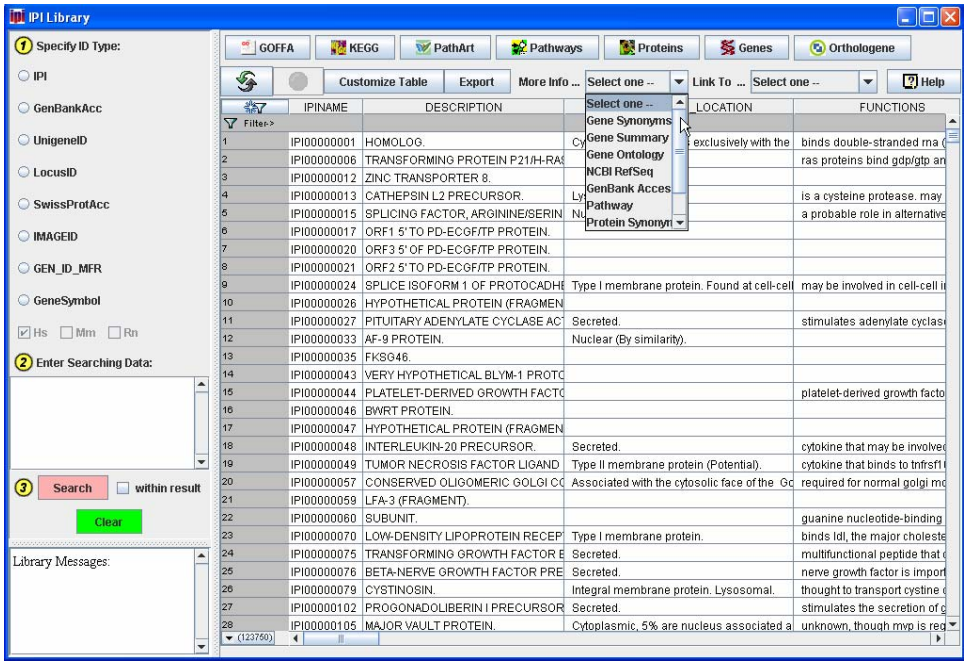


Figure 3-31: IPI Library search panel

Orthologene Library

The Orthologene Library contains genes linked across species that are determined to be orthologous based on analysis of homology using Blast (http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/information3.html). The library currently contains orthologs for human, mouse and rat.

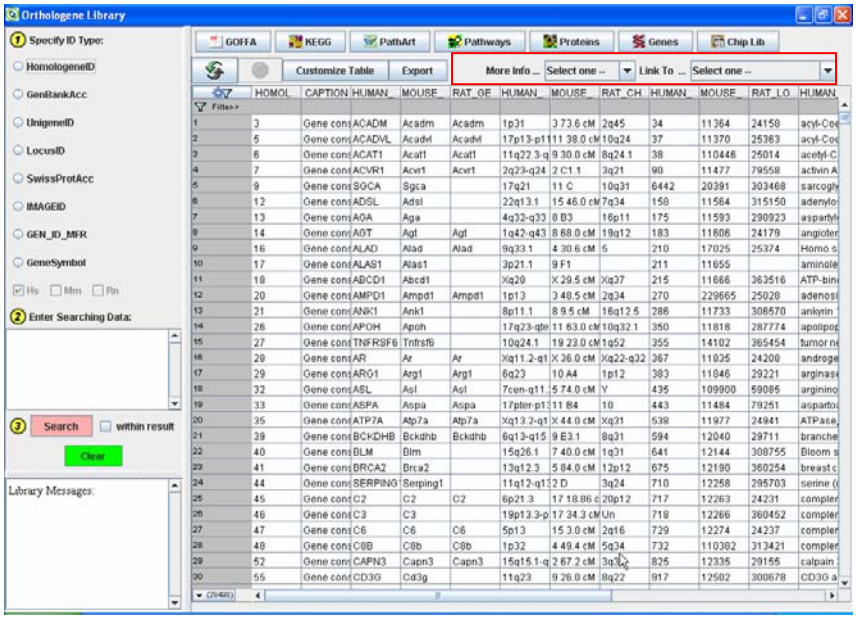


Figure 3-32: Orthologene search panel.

Orthologene Library has the same interface structure as the other libraries, and provides links to other resources as for the other libraries. However, the links to external websites is different, and consists of website databases related to cross-species, chromosome location across species, gene homology and species-specific genome information (see Figure 3-33).

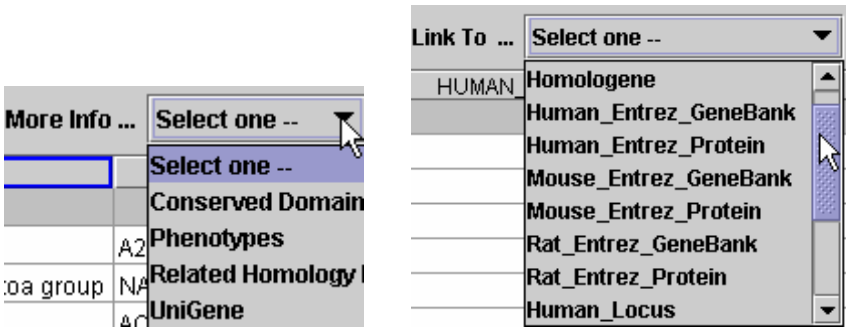


Figure 3-33: More links to other resources

GOFFA Library

The GOFFA Library provides gene ontology information, using the standard vocabulary (terminology) of the Gene Ontology Consortium; the ontology provides standard vocabularies for the description of the molecular function, biological process and cellular component of gene products. These terms are to be used as attributes of gene products by collaborating databases, facilitating uniform queries across them. The controlled vocabularies of terms are structured to allow both attribution and querying to be at different levels of granularity.

The user can access GOFFA library in the same way as the other libraries.

Tips: a convenient way to access GOFFA is through Gene Library. In Gene Library panel, the user can highlight the gene records with interest. Then click GOFFA button at the top, select organism (human, mouse, rat, etc), click OK button. The highlighted gene records will be automatically shown in GOFFA, see Figure 3-34.

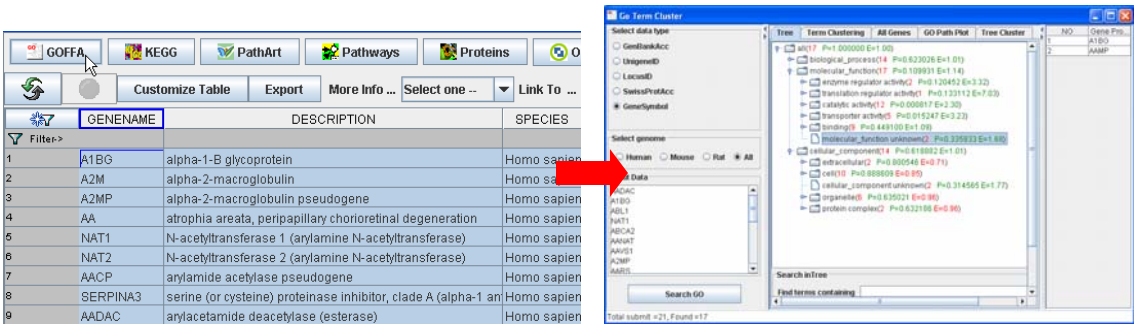


Figure 3-34: Accessing GOFFA via Gene Library

The GOFFA Library displays three vertically parallel panels: 1) the left panel for pasting the name of genes to be searched; 2) after clicking "Search Go", the middle panel will show the associated gene terms displayed in different views. The right panel will show the gene products associated with the terms in the middle panel.

In the middle panel, there are four tabs categorized as: 1) Tree. 2) Term Clustering. 3) All Genes. 4) GO Path Plot.

Under Tree tab: the search results will be shown here in three root groups: a) molecular function; b) biological process; c) cellular component. Each group has more branched sub-groups. A red-colored number is suffixed to the group name of every level tree. That number represents the total number of genes in the search that was found in this ontological category/sub-category. The suffixed green number is Fisher exact test P-value. Following P value is the E (Enrichment Factor, see equation 1.1) value. When you click on a GO term the right panel will show the gene list related to the GO term. See Figure 3-35.

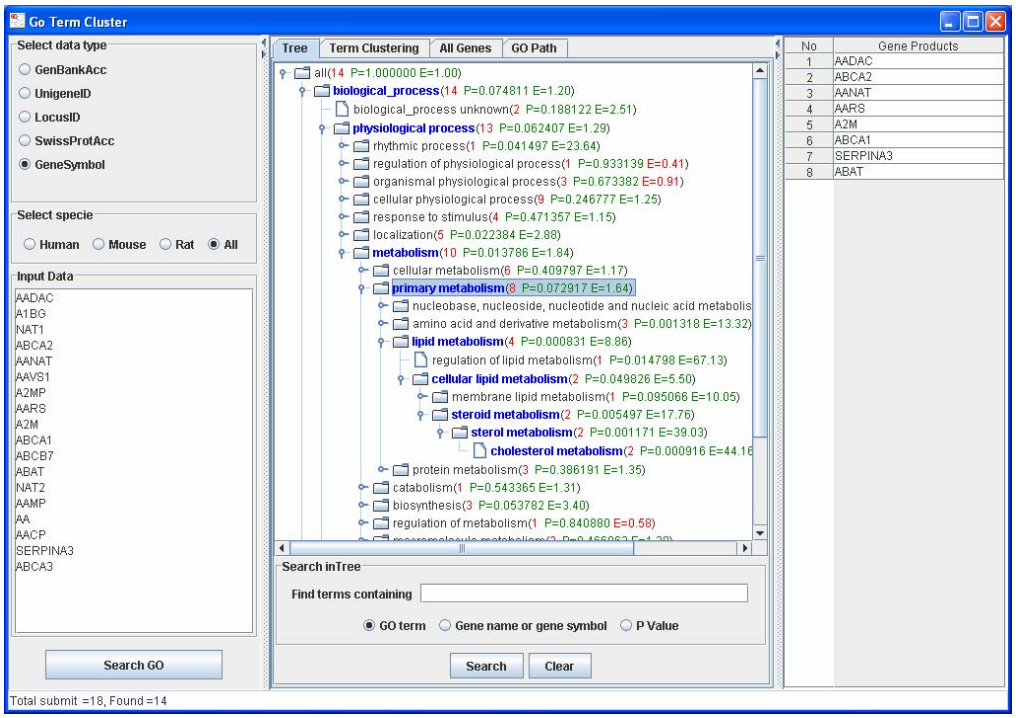


Figure 3-35: GOFFA Library search panel

Enrichment factor =(ni/N)/(gi /G) (1.1)

Where ni is the number of hit genes in term i. N is the number of input genes. gi is the number of gene or protein associated with term i, G is the total number of gene or protein in the database.

At the bottom of the middle panel, the user can enter a term in the "Find terms containing" text box, choose search by "GO term", "Gene name or gene symbol" or "P value", then click search button, the term contained at all levels of the tree will be highlighted in blue color. See Figure 3-35.

Under Term Clustering tab: a spreadsheet with 6 columns titled in No., Term, GO ID (GO accession number), Average Level (the average hierarchical level of the term showing in all the paths), Average Fisher P Value and Gene Hits (the number of the gene products associated with the term). This is an alternative view of the tree structure with P value pre-sorted. Users can sort the table by clicking on the column header. To do multiple column sorting you can click on column header while press ctrl key. Single-click any row will bring up the associated genes shown in the right panel; double-clicking any row will switch back to the tree tab view with interested terms highlighted in blue.

Tree | Term Clustering | All Genes | GO Path

No	Term	GO ID	Average Level	Average P value	Gene Hits
1	arylamine N-a...	GO:0004060	9.00	0.000000	3.00
2	N-acetyltransf...	GO:0008080	8.00	0.000003	3.00
3	N-acyltransfer...	GO:0016410	7.00	0.000004	3.00
4	acetyltransfera...	GO:0016407	7.00	0.000007	3.00
5	ATPase activity	GO:0016887	8.00	0.000056	4.00
6	acyltransferas...	GO:0008415	6.00	0.000099	3.00
7	transferase ac...	GO:0016747	5.00	0.000101	3.00
8	transferase ac...	GO:0016746	4.00	0.000128	3.00
9	ATPase activity...	GO:0042626	6.33	0.000139	3.00
10	hydrolase activ...	GO:0016820	5.00	0.000142	3.00
11	acute-phase r...	GO:0006953	6.40	0.000229	2.00
12	nucleoside-tri...	GO:0017111	7.00	0.000460	4.00
13	pyrophosphat...	GO:0016462	6.00	0.000558	4.00
14	hydrolase activ...	GO:0016818	5.00	0.000571	4.00
15	hydrolase activ...	GO:0016817	4.00	0.000574	4.00

Figure 3-36: Under Term Clustering tab in GOFFA library

Under All Genes tab: This spreadsheet lists all the genes and their associated info such as gene symbol, Average P-value and average hierarchical level. Similarly the users can sort the table by clicking the column header. Double-click on any row will bring up tree view with highlighted gene number in blue on any associated term.

Tree | Term Clustering | All Genes | GO Path

No	Gene Symbol	Average P Value	Average Level
1	AANAT	0.002690	8.89
2	NAT1	0.013394	3.25
3	NAT2	0.013501	3.26
4	ABAT	0.024740	5.94
5	AADAC	0.026330	3.26
6	AARS	0.032951	8.28
7	ABCA1	0.081833	6.38
8	ABCA2	0.087474	5.04
9	A2M	0.089308	6.12
10	ABCB7	0.152133	4.89
11	AAMP	0.154991	3.88
12	ABCA3	0.163110	4.76
13	A1BG	0.165918	2.00
14	SERPINA3	0.245559	4.39

Figure 3-37: Under All genes tab in GOFFA library

Under GO Path tab: under this tab, there are three sub-tabs named biological, molecular function and cellular component. P Path plots provide a numerical figure-of-merit for the statistical significant of paths, for the purpose of comparing the potential significance between paths.

For each of three sub categories, tree paths are ranked in statistical significance based on equation 1.1. The top 10 tree paths with max value are plotted for each category.

$$-\sum \log P_i \qquad (1.2)$$

Where Pi is Fisher exact test (right-tail) probability value for each node.

The user can zoom in/out the plot by right-clicking anywhere in the plot and choosing zoom in/out. The plot can also be saved. Clicking any spot on a colored line or clicking on legends will cause a return to the Tree tab, which displays with clicked path highlighted in blue, see Figure 3-38.

In Figure 3-38, the x-axis is the level of the tree (or called node), y-axis is Log P-value. Moving around the mouse in the plot, the user can see a blue line and P-value above the line. The user can zoom in/out the plot by right-clicking anywhere in the plot and choosing zoom in/out. The plot can also be saved. Clicking any spot on a colored line will cause a return to the Tree tab display with the pathway nodes on that line colored in blue, as shown in Figure 3-38C.

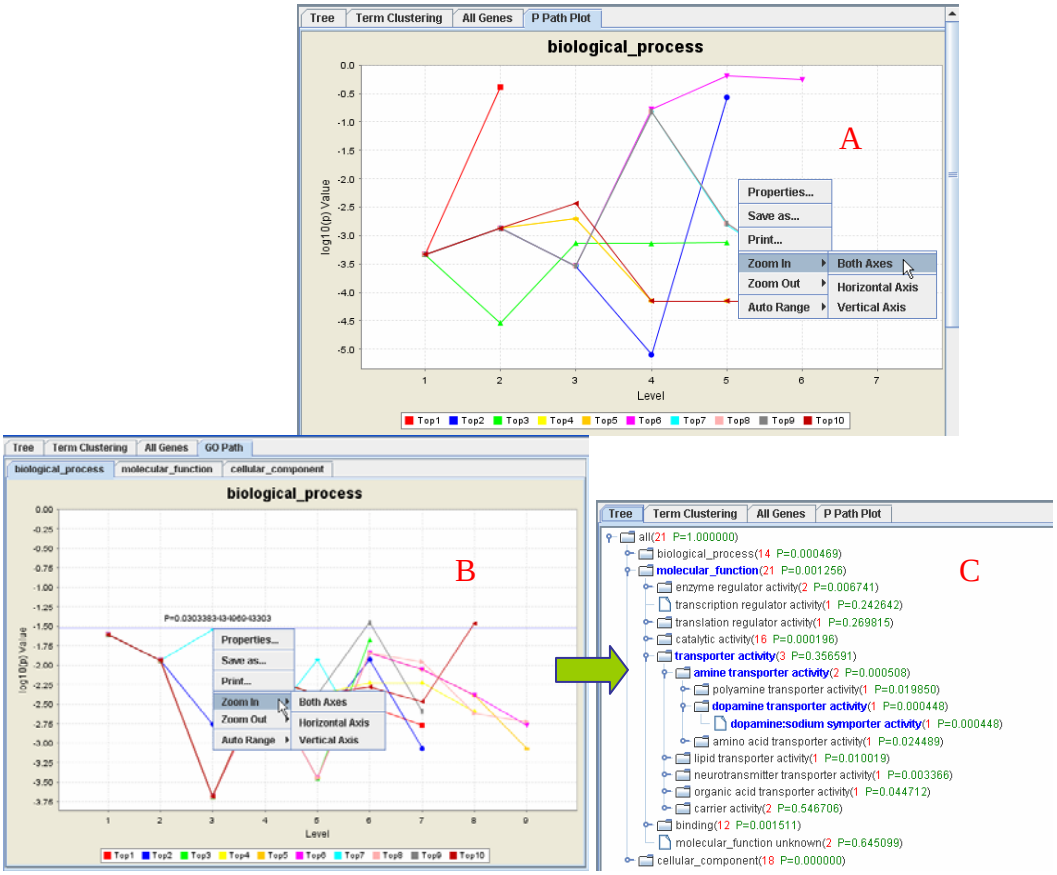


Figure 3-38: Top ten paths with lowest average P value

What information you can get from the GO Path Plot?

GO Path Plot lists the top ten paths ranked in statistical significance. The x-axis is the level of the path (from root to tip, 1~10), while the y-axis is the logP value. The plot supports quick identification of most significant terms through visualization.

In GO, the terms (functions) from the root to the leaves decent from general to specific. While comparing the significance of two terms that has the same p-value, the following questions have to be addressed:

1. If two terms have same P-value, and in the same level of the tree but in different path, which term is more significant? The Path Plot identifies the term that associates with the terms along the path with the smaller p-values significant (Figure 3-39).

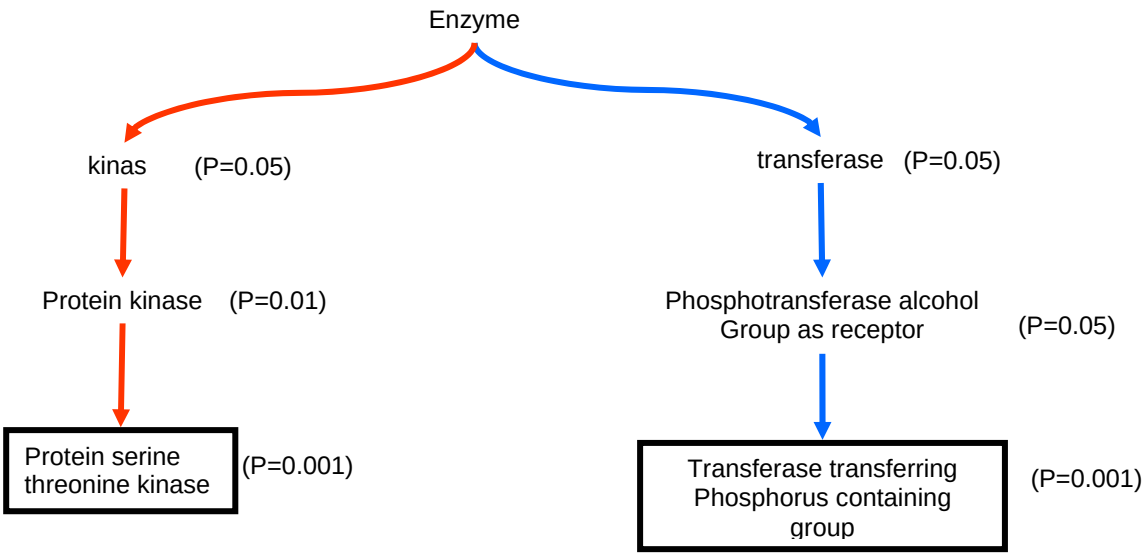


Figure 3-39: "Protein serine threonine kinase" is more significant than "Transferase transferring Phosphorus containing group"

2. If two terms have same P-value, but in the different level of the tree and in different path, which term is more significant? The Path Plot identifies the term in the lower level significant (Figure 3-40).

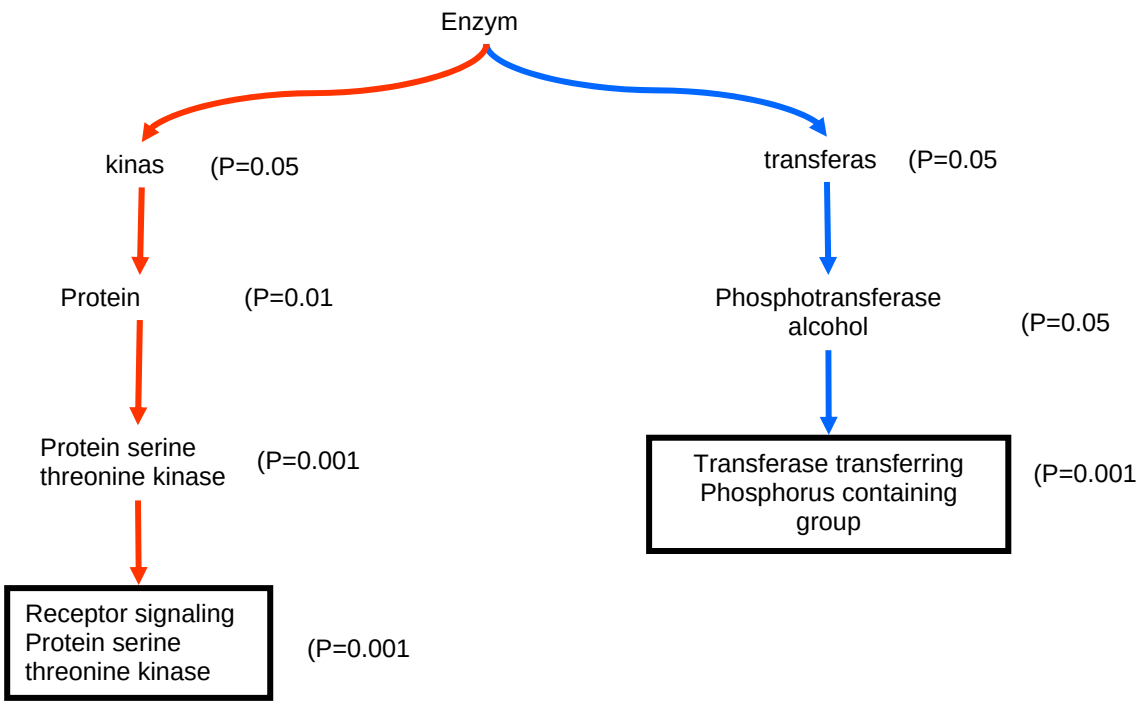


Figure 3-40: "Receptor signaling Protein serine threonine kinase" is more significant than "Transferase transferring phosphorus containing group"

3. From different path you can reach to the same term, which path is more significant? Usually the smaller the average P-value of the path is, the more significant the path is. If the path is longer, then the average P-value will be smaller which means this path is more significant.

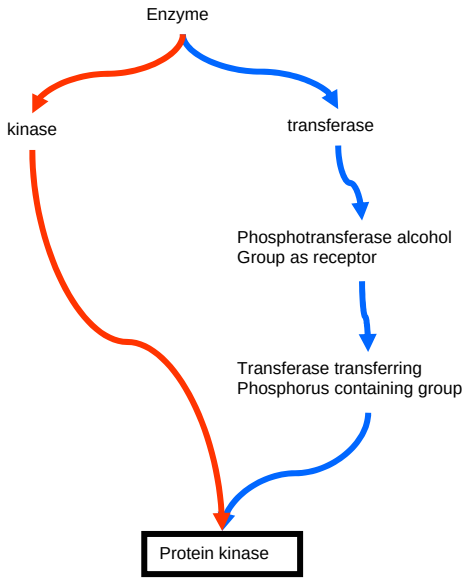


Figure 3-41: The function of the term "Protein kinase" is more close to "Transferase transferring Phosphorus containing group"

4. According to the GO Path Plot trend, how to decide which level you should stop at and how to interpret the trend of GO Path Plot? If two terms have the close p-values, you need to toggle back to the tree view by double click the path and the term containing most genes should be the significant one, see Figure 3-42.

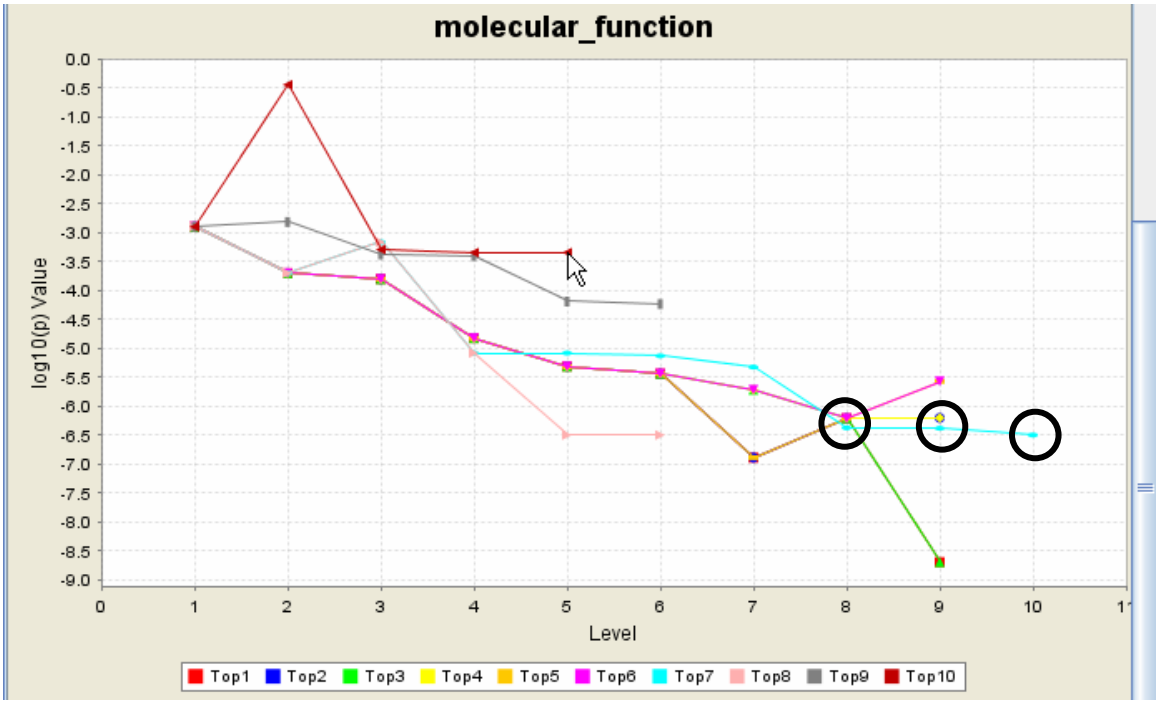


Figure 3-42: Determine which level you should stop at

Chip Library

Display ArrayType Information: The Chip Library hosts information for spots from any microarray array types available within MicroarrayDB. The Chip Search panel (Figure 3-43) is similar to that of Gene Search and Protein Search. By default, all array elements are searched and displayed at the launch of Chip Library.

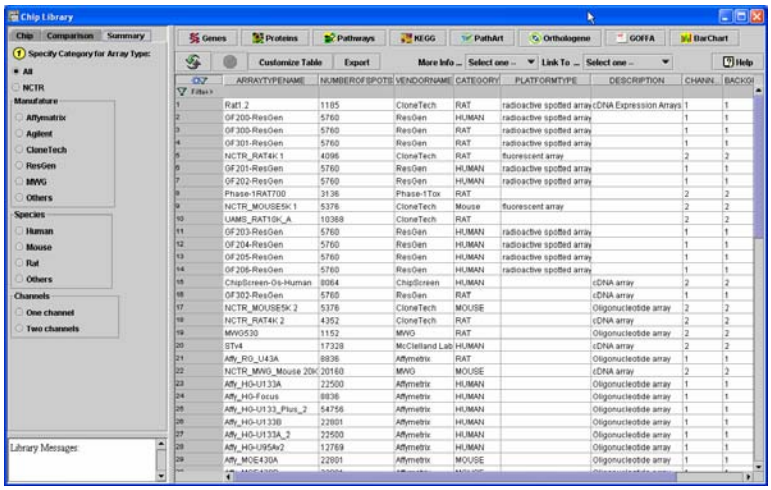


Figure 3-43: Default display of Chip Search panel.

However, the user can also view the arraytype information for a particular array type (e.g. different manufacture, species, channels etc) by specifying it under **Specify Category for Array Type:**. Figure 3-44 displays all the Affymetrix array type. Double-click any one of the chip will bring out a spreadsheet showing the detail of the individual chip.

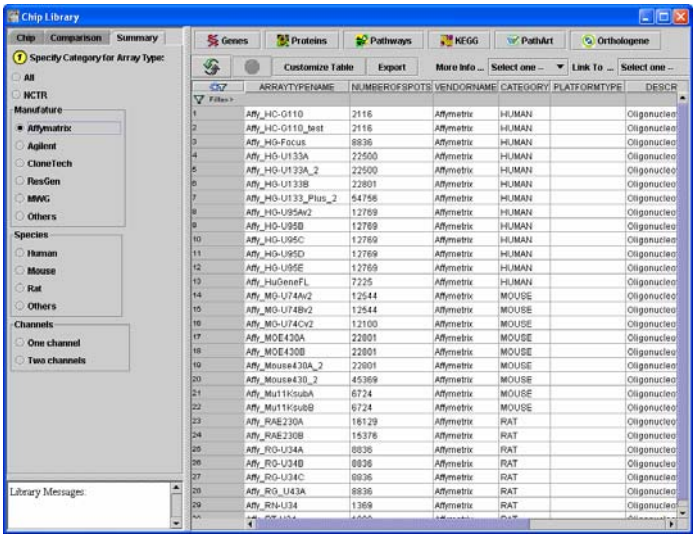


Figure 3-44: Display all the Affymetrix arraytype.

Compare Different ArrayTypes: By clicking on Comparison at the top left of Chip Library (**Chip** **Comparison** **Summary**), the user can query from a selected list of array types for the overlapped (AND) or combined (OR) genes. First choose the array types to be compared, click "search" button, decide the common ID for the two array types. The overlapped gene lists will show in the right panel.

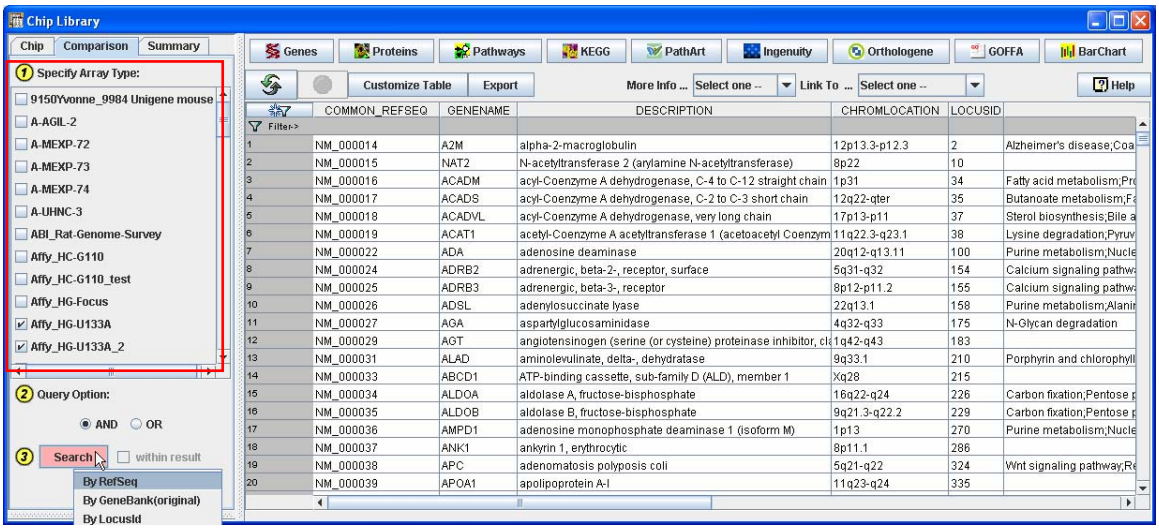


Figure 3-45: Overlapped gene list between Affy_HG_133A array and Affy_HG_133A_2 array.

If the user want to see some specific genes in the viewer, she/he can highlight those genes and right click (Figure 3-46), choose **Mark selected spots in open viewers**. The highlighted genes will be marked in the pre-opened array viewer, see Figure 3-47.

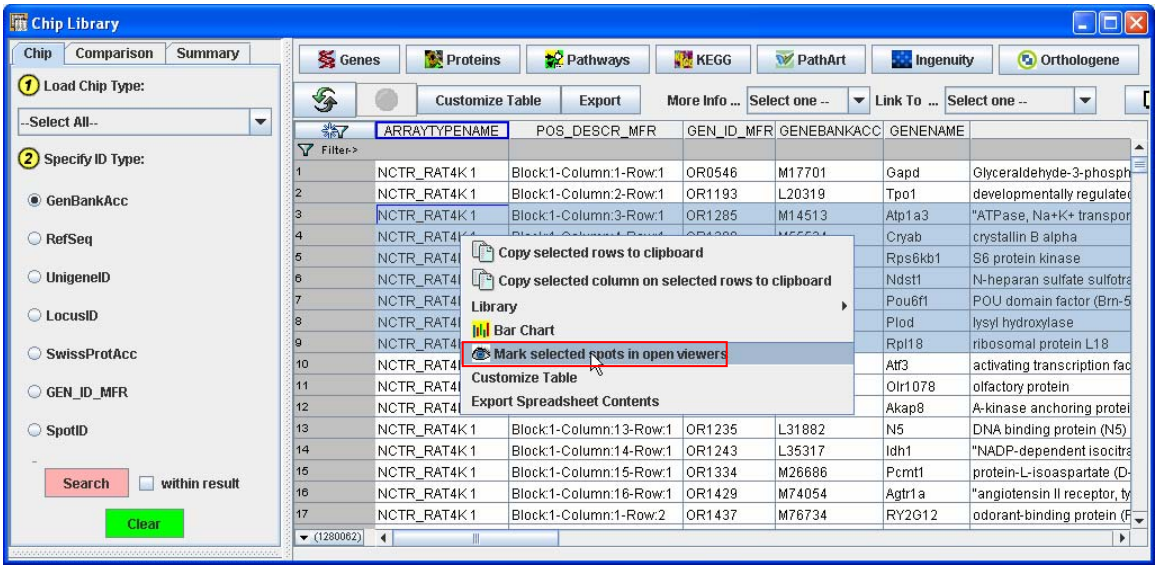


Figure 3-46: Choose specific gene records to be marked in the viewer

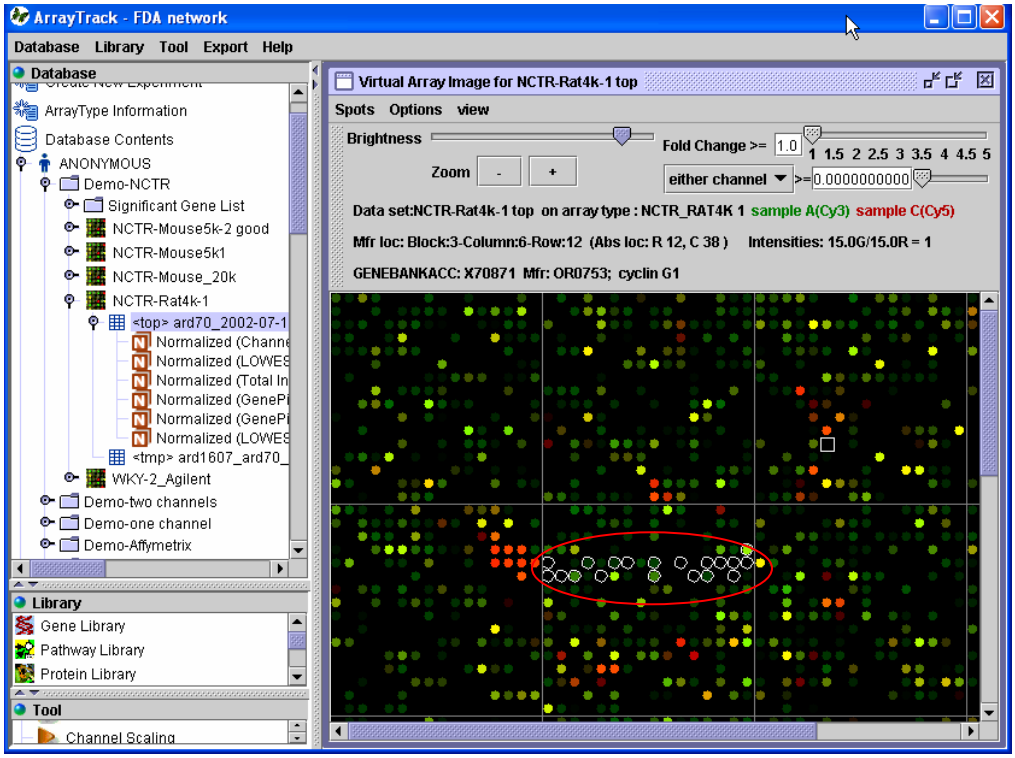


Figure 3-47: the marked genes chosen in the Chip Library

Summary: By clicking on Summary tab at the top left of Chip Library (Chip Comparison Summary), the right panel will show the array types grouped in different categories. For example, if the user wants to see all the Agilent array types, s/he can select "Agilent" in the Manufacture frame, and the right panel will list only Agilent array types.

Toxicant Library

Toxicant Library contains toxicological data and chemical structures (Figure 3-48). The result table shows all compounds in the Toxicant Library, one compound per row. The structure

of a compound is displayed in the left panel when its' name is clicked on. Search can be conducted based on substructure, similarity, (partial) compound name, formula, and compound ID. It is linked to several public databases (Toxnet, Cactus, ChemIDplus, ChemACX, ChemFinder, NCI DTP) on small molecules (see Figure 3-48). Currently, binding affinity data obtained at the NCTR on estrogen receptor and androgen receptor along with information from the CPDB (Cancer Potency DataBase) for the NTP tested chemicals have been made available in Toxicant Library.

By clicking the "Edit" button, the user can edit the chemical structure (e.g. adding functional groups, changing atoms, etc). Structure editing is done similar to with ISIS draw software, see Figure 3-51, using pull down menus, icons and options available by left or right-clicking the chemical structure.

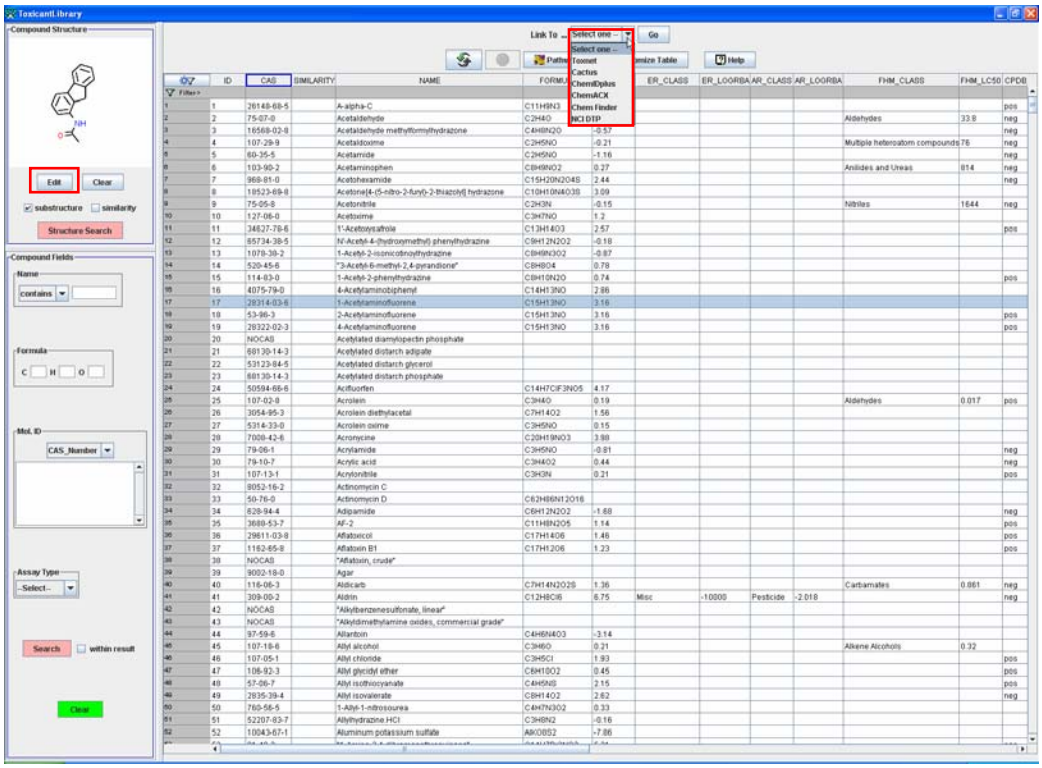
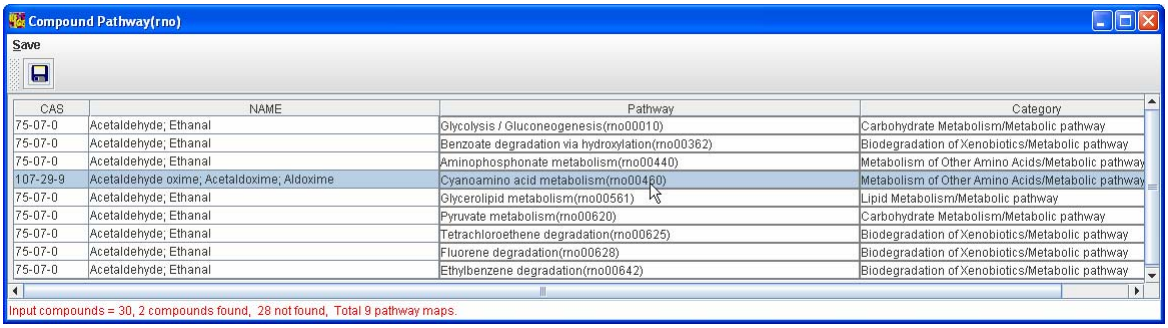


Figure 3-48: Toxicant Library panel.

The user can also highlight a group of chemicals and click

button and will see how many compounds and pathway information have been found, see Figure 3-49.



CAS	NAME	Pathway	Category
75-07-0	Acetaldehyde; Ethanal	Glycolysis / Gluconeogenesis(rno00010)	Carbohydrate Metabolism/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Benzoate degradation via hydroxylation(rno00362)	Biodegradation of Xenobiotics/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Aminophosphonate metabolism(rno00440)	Metabolism of Other Amino Acids/Metabolic pathway
107-29-9	Acetaldehyde oxime; Acetaldoxime; Aldoxime	Cyanoamino acid metabolism(rno00460)	Metabolism of Other Amino Acids/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Glycerolipid metabolism(rno00561)	Lipid Metabolism/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Pyruvate metabolism(rno00620)	Carbohydrate Metabolism/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Tetrachloroethene degradation(rno00625)	Biodegradation of Xenobiotics/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Fluorene degradation(rno00628)	Biodegradation of Xenobiotics/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Ethylbenzene degradation(rno00642)	Biodegradation of Xenobiotics/Metabolic pathway

Figure 3-49: Compound pathway

In Figure 3-49, click any record will bring up the pathway map. See Figure 3-50.

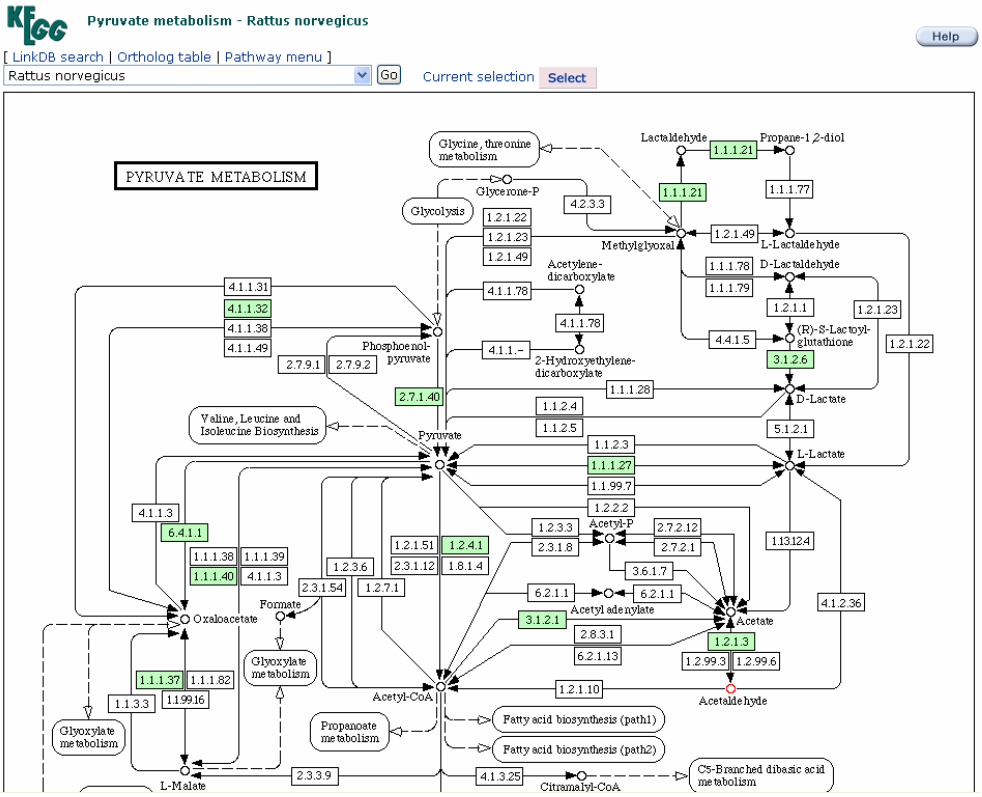


Figure 3-50: Pathway map

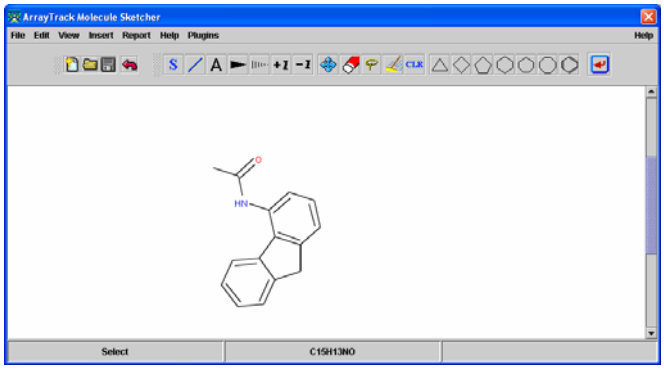


Figure 3-51: Molecular structure editing

EDKB Library

The EDKB Library contains chemical structures and endocrine activity properties of compounds tested in several assays including more than 3200 records of endocrine activity related endpoints such as binding the estrogen and androgen nuclear receptors, uterotropic weight gain, E-screen and combined receptor-reporter gene data (Figure 3-52). All the data are linked to their associated citations. Activities across different assays are scaled relative to estradiol, such that they can be viewed together in a Graphic Activity Profile. The user can link each chemical to the databases such as Chemfind, Chemplus, etc. and search data by assay type (in left panel, specify assay type combo) or directly search any column by typing a key word in the first row of

the spreadsheet and hitting return. The user can also perform chemical structure or chemical similarity search in the upper left panel. In similarity search, the 50 most similar chemicals will be reported in a spreadsheet, one compound per row, with multiple columns listing the activity information. The structure of a compound is displayed when its' name is clicked. Search options are similar to those in the Toxicant Library.

The biological activity for compounds in the EDKB Library is displayed in a graphic way (Figure 3-52). The X-axis is arranged first by assay type then by compound. The Y-axis represents the relative potency (in log10 scale). Each box represents a particular compound and the height of the box reflects the (max – min) of multiple biological data for the same compound in the same assay.

If the user is interested in a particular compound, he can highlight the compound and select Individual Compound from the pull down list of "More Info…", and a new window will pop up showing the detail of the compound information (Figure 3-53).

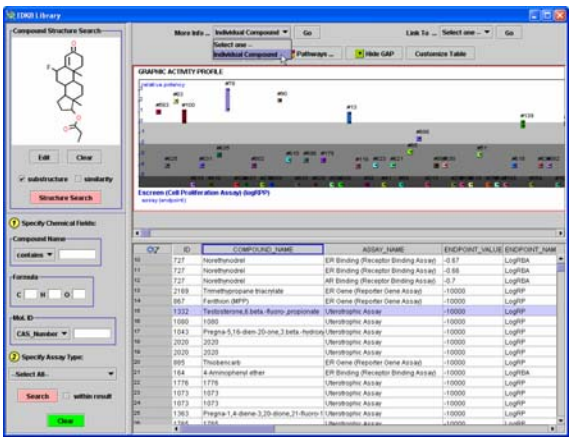


Figure 3-52: EDKB Library panel.

The user can also highlight a group of chemicals and click

button and will get the chemicals pathway information.

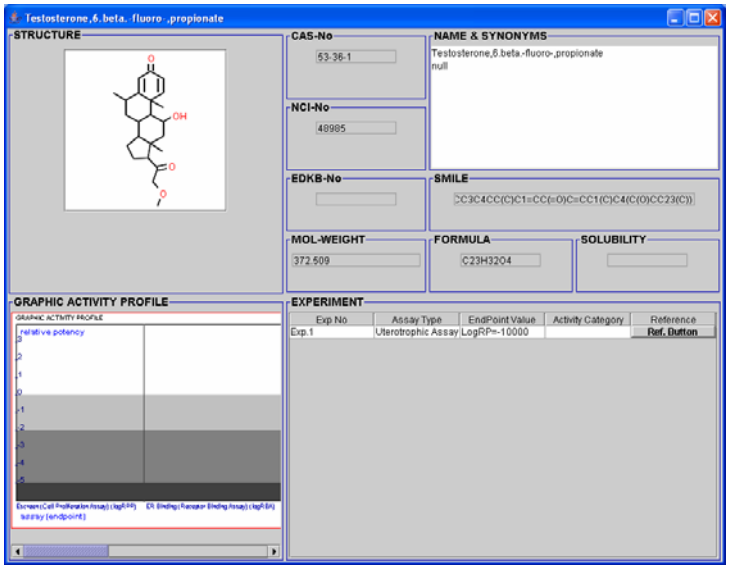


Figure 3-53: Individual compound information

For more information about EDKB (Endocrine Disruptor Knowledge Base), please visit http://edkb.fda.gov/.

ID Converter

ID Converter is a very useful tool for converting one kind of ID to another kind when searching. The user can activate ID Converter by clicking the icon in the Library panel, see Figure 3-54.

Figure 3-54: ID Converter in Library panel

Once the ID Converter is activated, the user can choose the ID type to be converted and select the species (Human, Mouse, Rat). There is a radio button labeled "official name only" under the species option. If this radio button is checked, the output will only display the official name without any synonyms or other unofficial name. The user needs to type/paste the ID in the left panel, then choose the output ID type, click Convert button. The converted results will be shown in the right panel. See Figure 3-55. The user can highlight the searched results and click the library buttons (Gene Library, Chip Library, etc) at the top to go to different library directly.

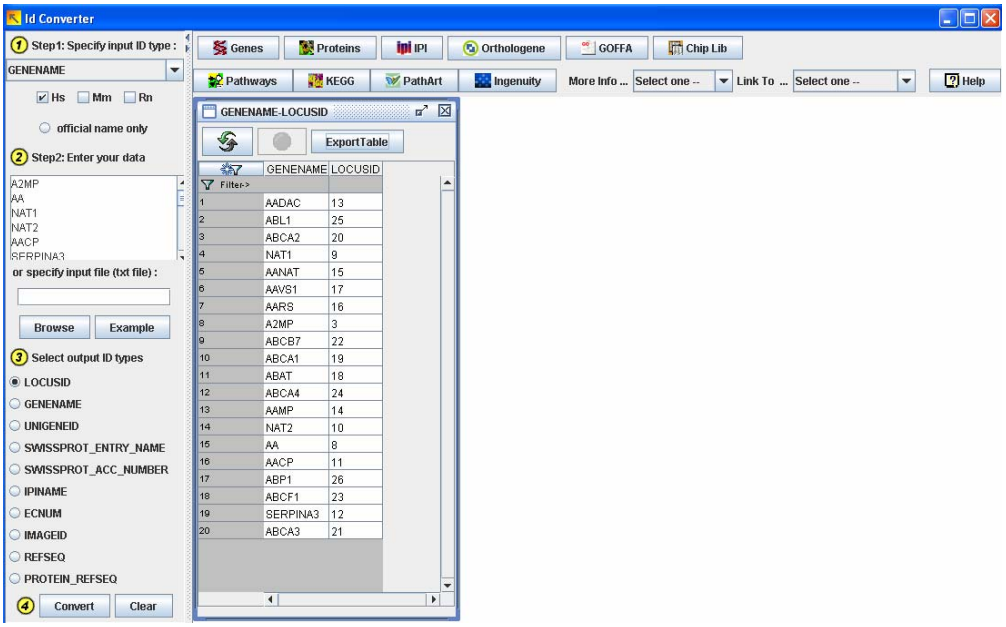


Figure 3-55: ID Converter tool

The ID types we accept are: Gene Bank Accession Number, Gene Name, Locus ID, Swissprot Entry Name, Swissprot Accession Number, Unigene ID, Affy ID, Image ID, IPI name.

The ID types we can give out are: Locus ID, Gene Name, Unigene ID, Swissprot Entry Name, Swissprot Accession Number, IPI name, Enzyme Number, Image ID, Reference sequence number, Protein reference sequence number.

Chapter 4 Working with Tools: Quality Control

Overview of TOOL

The third component of ArrayTrack is TOOL that consists of various functions for normalizing and visualizing microarray data. These functions are accessible within ArrayTrack either from the TOOL panel or Tool pull-down menu (Figure 4-1). The TOOL functionalities are classified into three categories: Quality Tools, Normalization Methods, and Visualization Tools. As has been pointed out earlier, these tools can also be accessed by right-clicking on selected array data sets.

Quality Tools is discussed in this Chapter; Normalization and Visualization are discussed in Chapter 5 and Chapter 6, respectively.

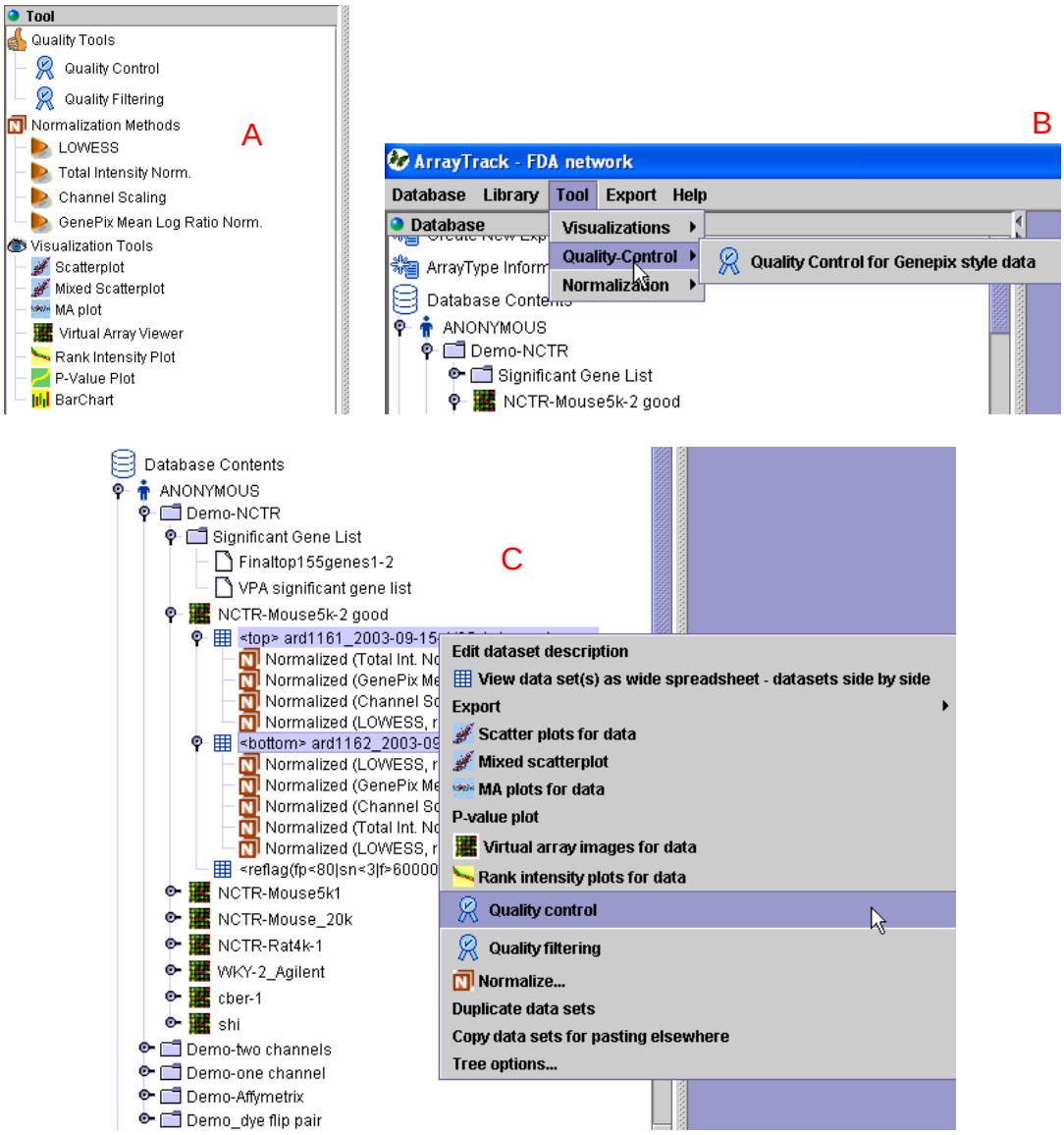


Figure 4-1: Quality Tools can be accessed from (A) TOOL panel; (B) Tool pull-down menu; and (C) Right-click on selected array(s).

Overview of Quality Control

Quality Control provides various visual plots and numerical parameters for measuring the quality of a hybridization (array). Currently, Quality Control is available only for arrays for which the original gene expression file data were input from the GenePix GPR file format (see Figure 2-7 of Chapter 2).

Launch of Quality Control

From the TOOL Panel: When Quality Control is launched from the TOOL panel (Figure 4-1A), the user is asked to select array(s)/hybridization(s) for Quality Control view from a list of hybridizations (Figure 4-2), which by default are sorted first by the Exp ID and then by Hybridization name. A Quality Control panel for each of the selected arrays will be displayed after the OK button is pushed.

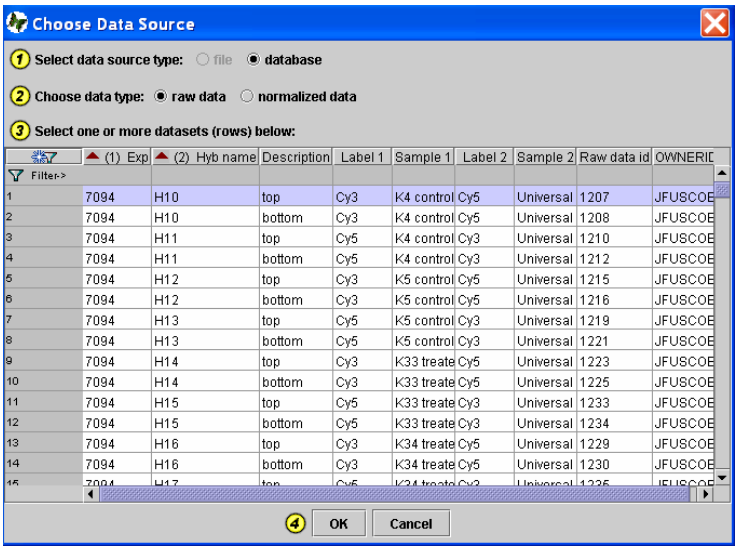


Figure 4-2: Select arrays for Quality Control view.

From the TOOL Pull-down Menu: When Quality Control is launched from the TOOL pull-down menu (Figure 4-1B), a Quality Control view will be opened for each of the currently selected arrays under the MicroarrayDB Contents tree structure (Figure 2-23C). However, if no array is selected under the MicroarrayDB Contents structure and the user tries to launch Quality Control from the pull-down menu, a warning is displayed (Figure 4-3).

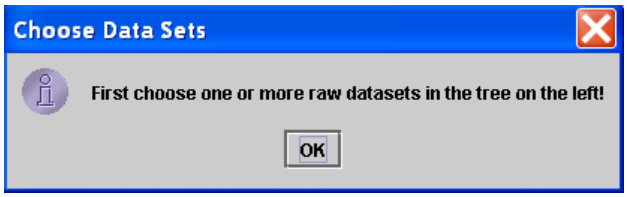


Figure 4-3: The Quality Control panel can not be launched from the pull-down menu until one or more arrays are selected from the MicorarrayDB Contents tree.

From Right-click on Selected Arrays in the MicorarrayDB Contents Tree: This may be the most commonly used way of launching the Quality Control view (Figure 4-1C). A Quality Control panel will be launched for each of the selected arrays.

Contents of Quality Control View

Figure 4-4 and Figure 4-5 are two example views of Quality Control. Each view is consisted of the following main sections:

Preview of Scatterplot: The Scatterplot displays the Cy5 (F635 Median) versus Cy3 (F532 Median) intensities for spots on the array. The user can have the options of background subtraction, showing flagged spots, and switching between Scatterplot and MA Plot. Details about Scatterplot can be found in Chapter 7.

Preview of Rank Intensity Plot: The Rank Intensity Plot is displayed for both Cy5 (red) and Cy3 (green) channels. The user can have the options of background subtraction and adjusting the two channels to a common mean. Details about Rank Intensity Plot can be found in Chapter 7.

Quality Control Parameters: Various quality control parameters are shown in the middle of the panel (Figure 4-4 and Figure 4-5) for both the Cy5 (F635 Median) and Cy3 (F532 Median) channels. A **PASS** or **FAIL** mark is shown for each QC/QA parameter based on the corresponding Threshold value preset. Threshold values can be reset by clicking on Save. Other QC/QA notes on RNA quality/integrity, hybridization, and labeling are also shown if they were entered in the Input Form (Figure 2-1) when loading data to MicroarrayDB.

Overall Judgment on Array Quality: A final judgment (Pass, Fail, Review, or None) can be assigned to each array and save in MicroarrayDB.

The array shown in Figure 4-4 has a much higher signal-to-noise ratio than the one shown in Figure 4-5.

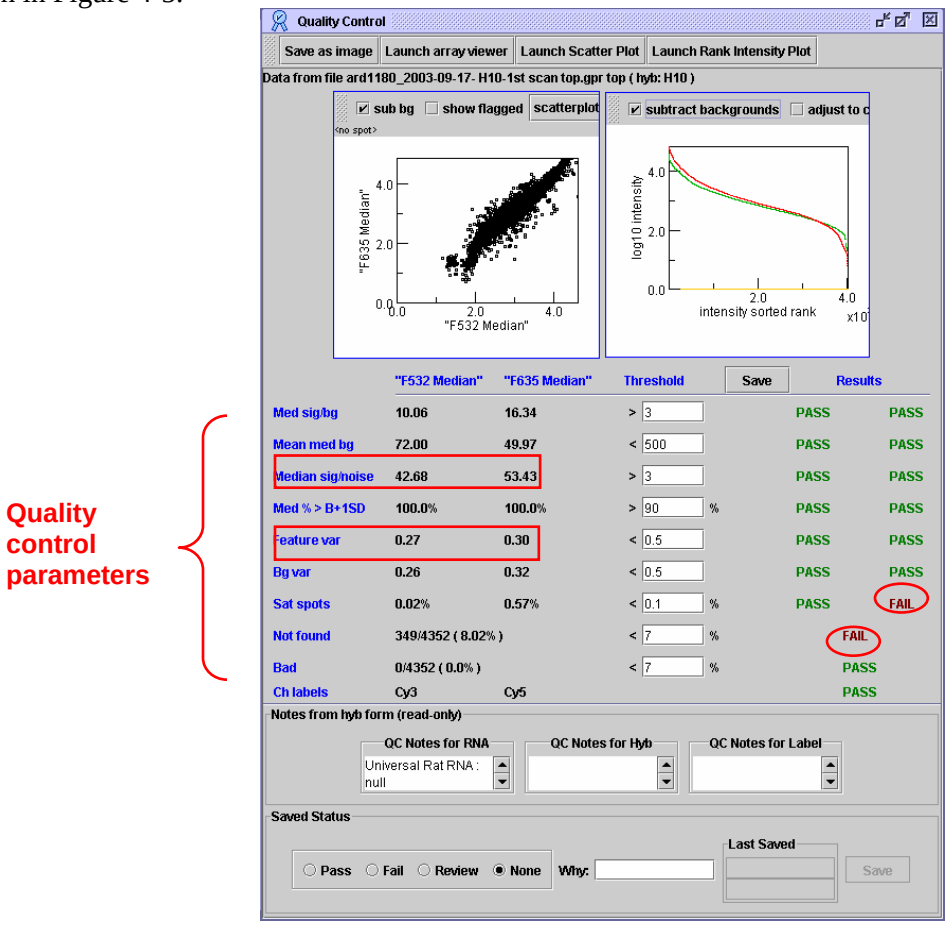


Figure 4-4: Quality Control view and parameters – a successful hybridization with much higher signal-to-noise ratio.

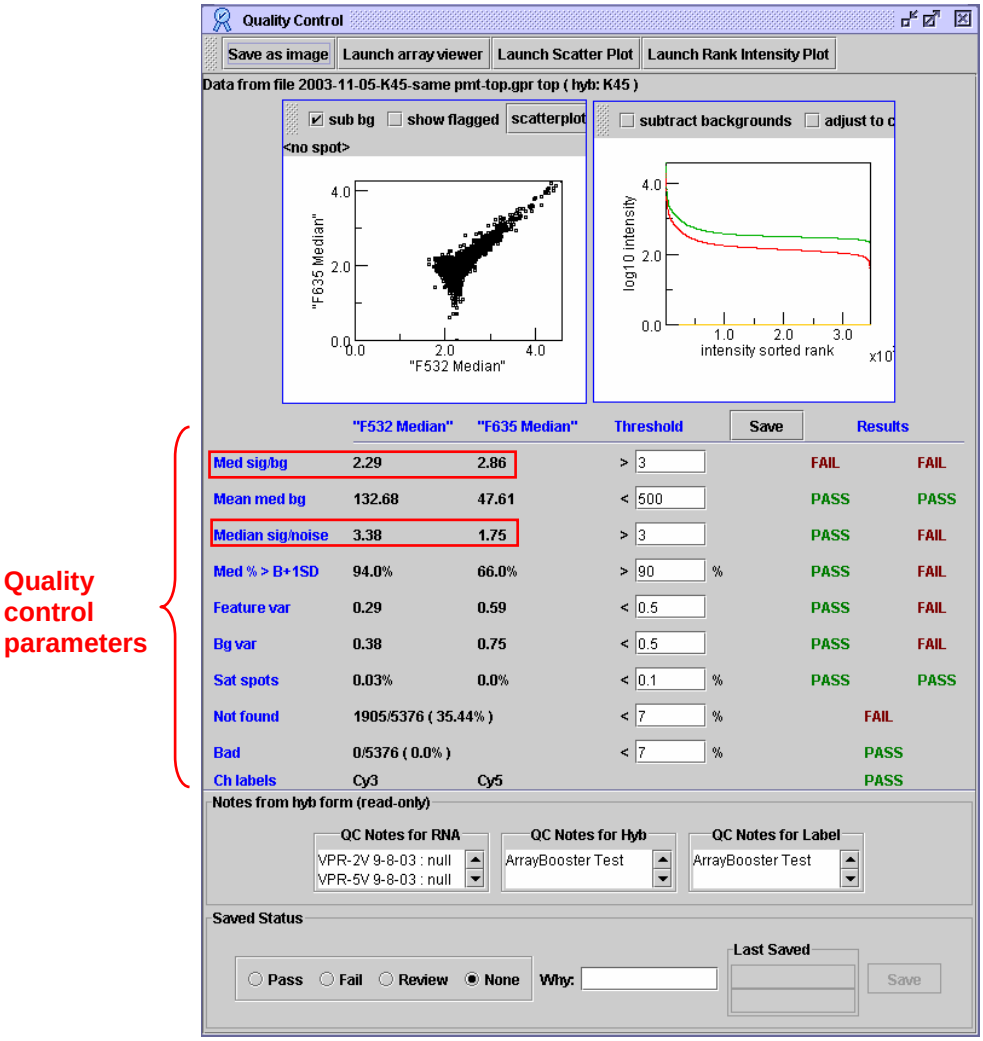


Figure 4-5: Quality Control plots and parameters – a failed hybridization with very low signal-to-noise ratio.

Function Buttons for Quality Control: On the top of the Quality Control view, there are four function buttons (Figure 4-6). Save as image allows the user to save the whole Quality Control view into an image file in JPEG, TIFF, or PNG format. The exact file format is specified by the file name extension of .JPG, .TIF, or .PNG, respectively. The user can Launch array viewer, Launch Scatter Plot, and Launch Rank Intensity Plot from the Quality Control page. Details of these functions are discussed in Chapter 7 on Visualization.

Figure 4-6: Functional options available within the Quality Control view.

Overview of Quality Filtering

Quality Filtering provides the view of filtered spots that meets certain criteria. The users can immediately see how many gene spots are of good quality.

Launch of Quality Filtering

Quality Filtering is launched in a similar way to launch Quality Control.

Contents of Quality Filtering

Figure 4-7 shows the Quality Filtering window that has three colored sections of filtering criteria: 1) Not-Identified (gray). 2) Un-Detected (blue). 3) Saturated (white). When the user enters a value of filtering criteria in the text box in each section, the filtered spots will be marked with the corresponding color in the viewer (Figure 4-7).

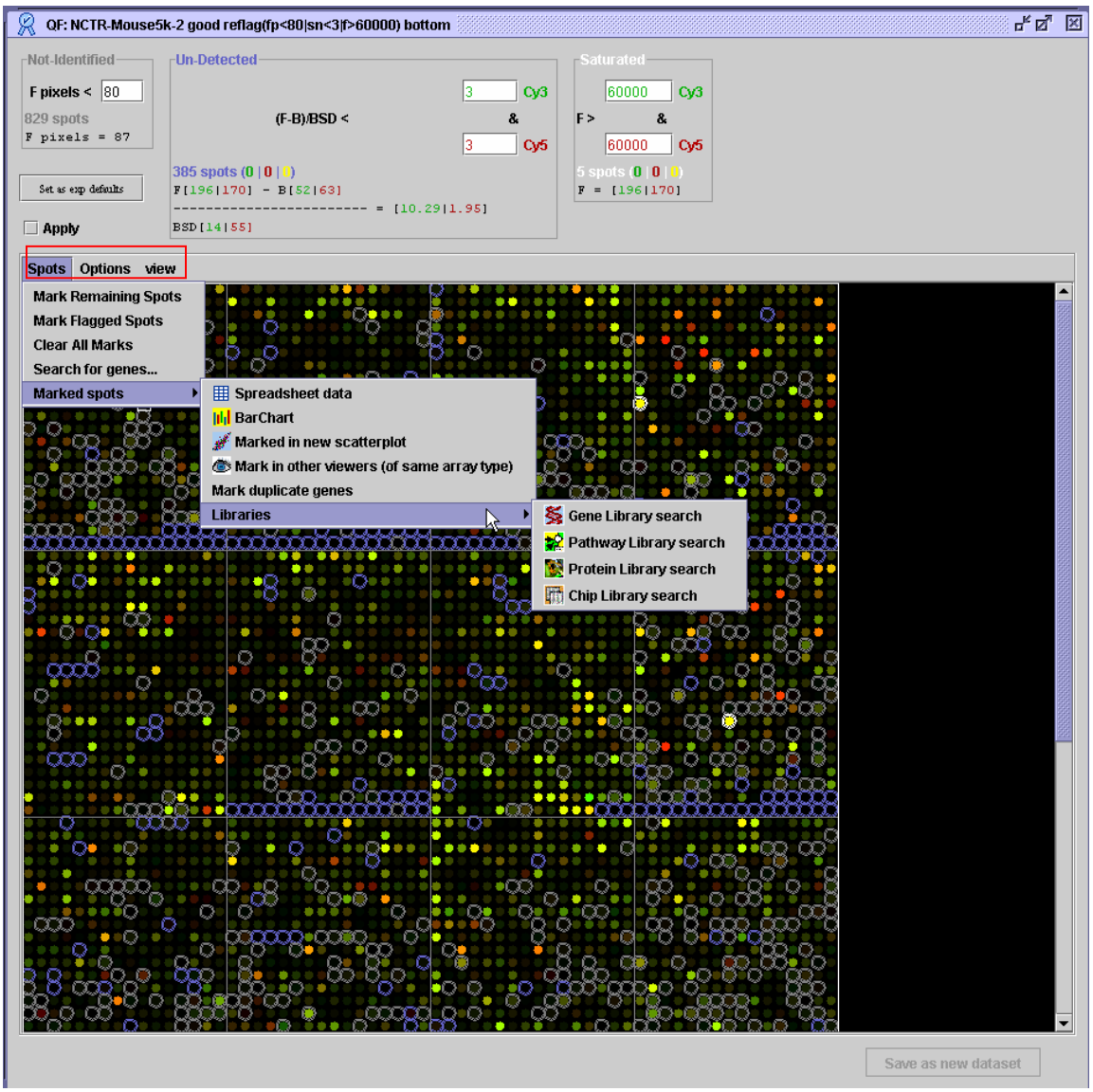


Figure 4-7**:** Functional options available within the **Quality Filtering** view

In Figure 4-7, the user can set the criteria for filtering the spots that s/he can define as "Not-Identified", "Un-Detected", and "Saturated". After typing the numbers in the white boxes for filtering, click "Apply", then the spots that meet the criteria will be marked in the corresponding colors. For example, the spots marked in grey color meet the criteria that the user

set for "Not-Identified"; the spots marked in blue meet the criteria for "Un-Detected" and the spots marked in white are "Saturated" spots.

If the user moves the mouse over any spot, the intensity value for that spot will be shown in the three colored sections.

Functions for Quality Filtering: For marked spots, the user can do further searches in the other libraries by clicking Spots > Marked spots > Libraries, or the user can paste them in spreadsheets and/or export, see Figure 4-7.

Under the View menu, the user can choose the style and color of the spot marking, and toggle the tool bar about the spot information, see Figure 4-8.

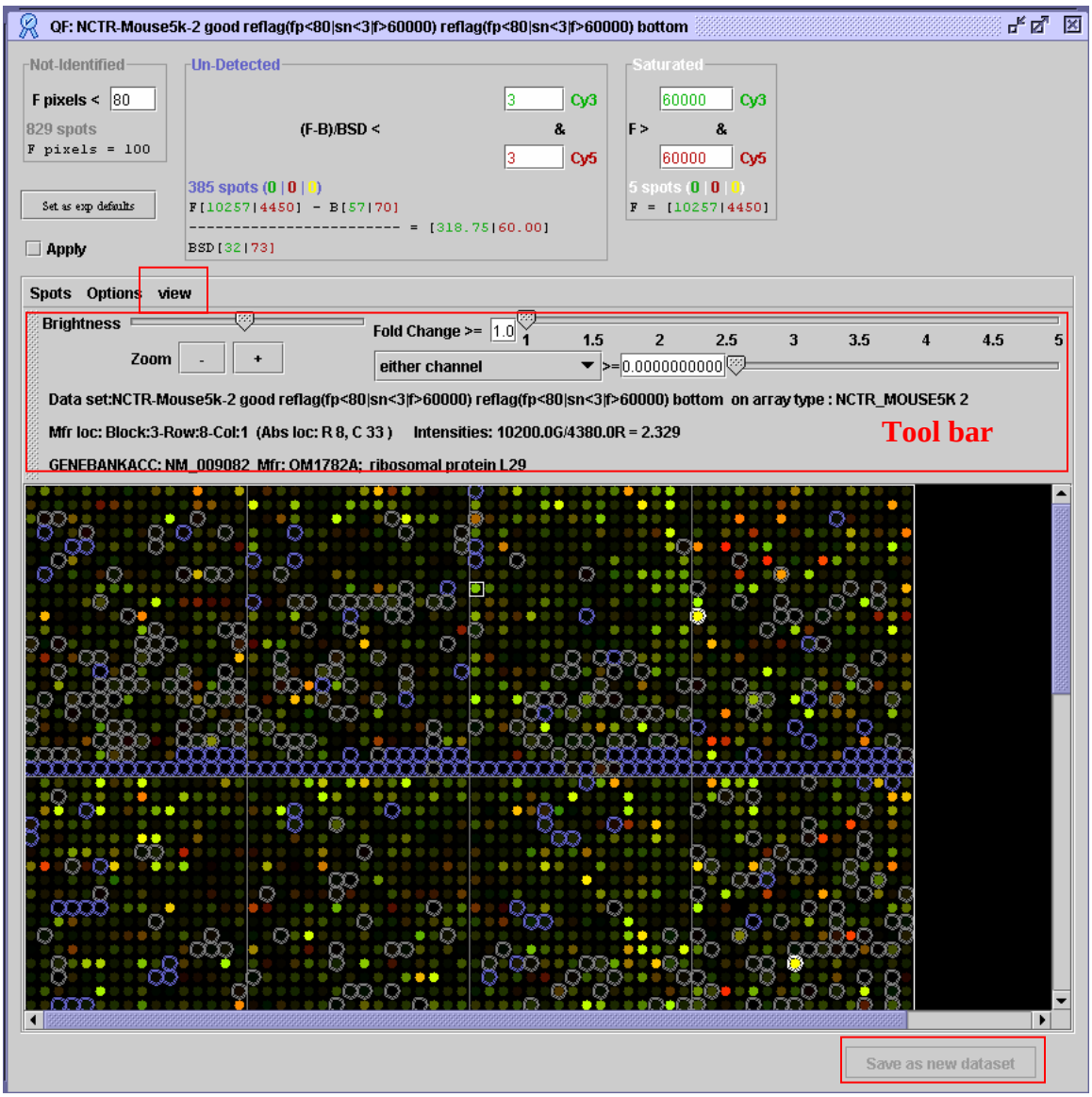


Figure 4-8: Toggle the tool bar about spot information.

The filtered data can be saved as new dataset by clicking the button "Save as new dataset".

Chapter 5 Working with Tools: Normalization

Overview

A microarray experiment is based on the analysis/comparison of multiple arrays (hybridizations). Across-array (hybridization) reproducibility is the most important criterion for judging the quality of a microarray experiment. There are many experimental factors that may render the microarray data inconsistent. Therefore, normalization methods are needed to (partially) correct systematic variations in microarray data introduced by experimental factors such as dye bias, efficiency difference in the cDNA synthesis and labeling reactions, nonlinear optical feature of the detector (scanner), etc.

Experiment normalization methods are used to standardize microarray data so that real (biological) variations in gene expression levels can be differentiated from variations due to the measurement process. Normalization scales microarray data so that you can compare relative gene expression levels. There are four normalization methods in ArrayTrack that can be invoked from the TOOL panel, the Tool pull-down menu, or right-click on selected array(s) (Figure 5-1).

Double-click on Normalization Methods at the TOOL panel hides or shows the contents (normalization methods) underneath it.

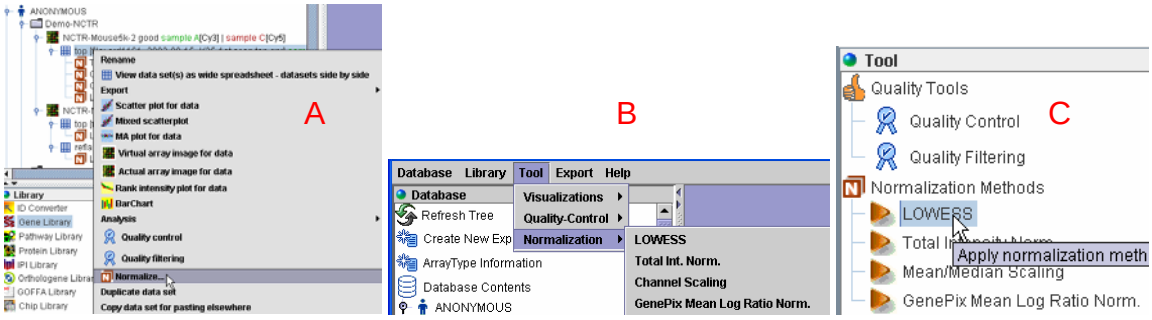
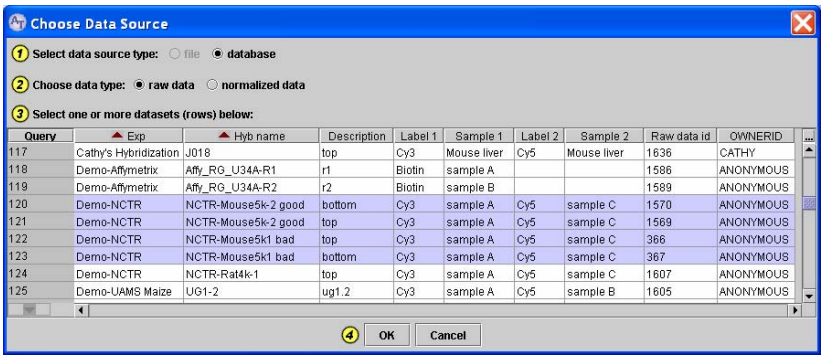


Figure 5-1: Four normalization methods have been implemented in ArrayTrack and can be accessed from (A) Right-click on selected array(s); (B) Tool pull-down menu; and (C) TOOL panel.

By selecting one of the four normalization methods, the user will be prompted with a table of arrays (hybridizations) which s/he has access to. The user can select one or multiple arrays for the normalization method to be applied to by clicking on the OK button (Figure 5-2).



Query	Exp	Hyb name	Description	Label 1	Sample 1	Label 2	Sample 2	Raw data id	OWNERID
117	Cathy's Hybridization	J018	top	Cy3	Mouse liver	Cy5	Mouse liver	1636	CATHY
118	Demo-Affymetrix	Affy_RG_U34A-R1	r1	Biotin	sample A			1586	ANONYMOUS
119	Demo-Affymetrix	Affy_RG_U34A-R2	r2	Biotin	sample B			1589	ANONYMOUS
120	Demo-NCTR	NCTR-Mouse5k-2 good	bottom	Cy3	sample A	Cy5	sample C	1570	ANONYMOUS
121	Demo-NCTR	NCTR-Mouse5k-2 good	top	Cy3	sample A	Cy5	sample C	1569	ANONYMOUS
122	Demo-NCTR	NCTR-Mouse5k1 bad	top	Cy3	sample A	Cy5	sample C	366	ANONYMOUS
123	Demo-NCTR	NCTR-Mouse5k1 bad	bottom	Cy3	sample A	Cy5	sample C	367	ANONYMOUS
124	Demo-NCTR	NCTR-Rat4k-1	top	Cy3	sample A	Cy5	sample C	1607	ANONYMOUS
125	Demo-UAMS Maize	UG1-2	ug1.2	Cy3	sample A	Cy5	sample B	1605	ANONYMOUS

Figure 5-2: Normalization method applies to a set of selected arrays.

In our experience, however, the user is more likely to use the Database Contents tree (see Figure 2-23. for details) to select a set of arrays on which a particular normalization method

is to be applied. For example, by right-clicking on an experiment, the user can quickly select all the raw datasets arrays contained within this experiment for normalization (Figure 5-3). By right-clicking on any of the selected raw arrays, a set of functions including Normalize… can be applied (Figure 5-4) and the user can choose one of the four normalization methods (Figure 5-1C).

If you don't have Write permission to the selected array data, you are still allowed to use the Normalize… function and save your normalized data into MicroarrayDB. However, you cannot change the original data or anyone else's normalized data.

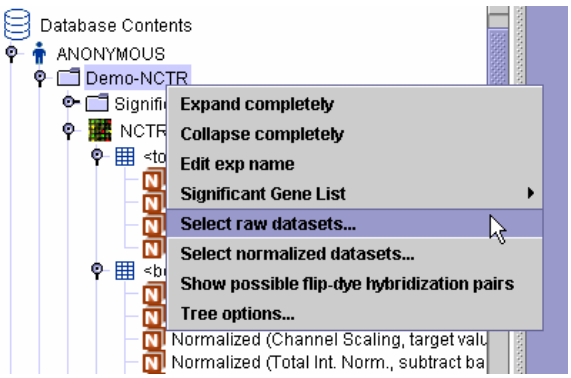


Figure 5-3: All raw data (arrays) under an experiment can be conveniently selected by right-click on the name of the experiment.

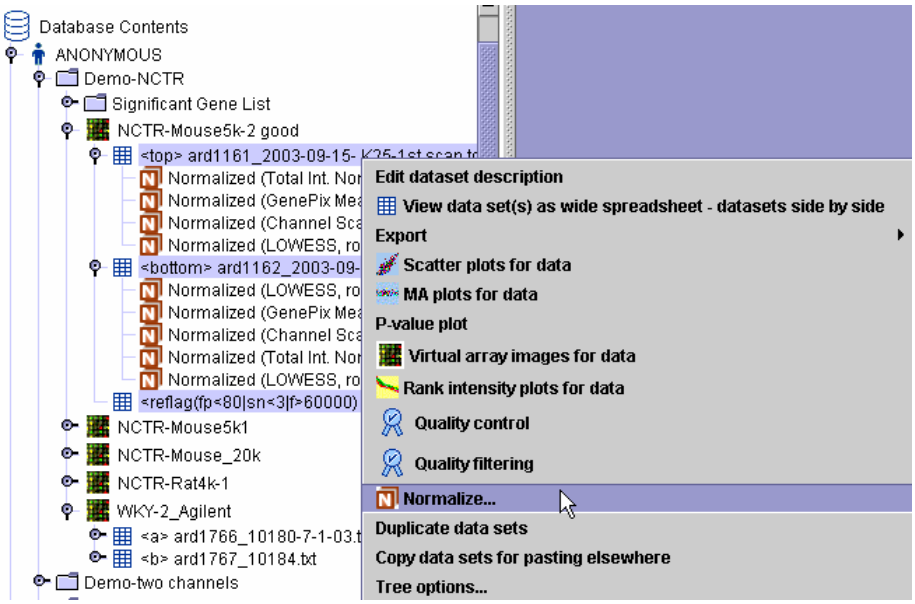


Figure 5-4: Normalization methods can be applied to a set of selected arrays under the Database Contents tree.

Each normalization method adjusts the original intensity values in a way as is defined in the individual normalization method and saves the adjusted (normalized) intensity data in MicroarrayDB.

Lowess

Lowess (Loess) refers to Locally Weighted regression and smoothing scatterplots proposed by W.S. Cleveland (Cleveland, W.S. (1979) "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, Vol. 74, pp. 829-836). Lowess combines the simplicity of linear least-squares regression with the flexibility of nonlinear regression. It does this by fitting simple models to localized subsets of the data to build up a function that describes the deterministic part of the variation in the data, point by point. In fact, one of the chief attractions of this method is that the data analyst is not required to specify a global function of any form to fit a model to the data, only to fit segments of the data.

Lowess has been proposed to normalize microarray data in the hope of correcting intensity-biased ratio measurement as seen in the MA plot (see discussion in Chapter 6). The fundamental assumption for Lowess is that the expression level for most of the genes in the two samples is unchanged. Lowess only applies to data from two-color platforms. Lowess is the default method of normalization in ArrayTrack (Figure 5-5).

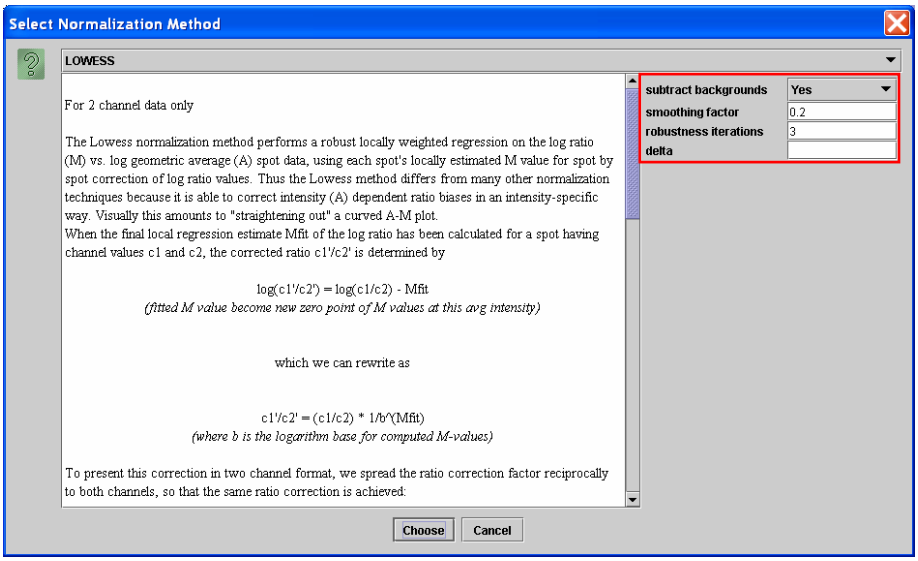


Figure 5-5: Parameter settings for Lowess normalization.

The Lowess normalization method performs a robust locally weighted regression on the log ratio (M) versus log geometric average (A) spot data, using each spot's locally estimated M value for spot by spot correction of log ratio values. Thus, the Lowess method differs from many other normalization techniques because it is able to correct intensity (A) dependent ratio biases in an intensity-specific way. Visually, this amounts to "straightening out" a curved MA plot. When the final local regression estimate Mfit of the log ratio has been calculated for a spot having channel values c1 and c2, the corrected ratio c1'/c2' is determined by

$$\log(\frac{c1'}{c2'}) = \log(\frac{c1}{c2}) - Mfit$$

(fitted M value become new zero point of M value at this average intensity)

which can be rewritten as

$$\frac{c1'}{c2'} = (\frac{c1}{c2})\frac{1}{b^{Mfit}}$$

(where b is the logarithm base for computed M-values)

To present this correction in two-channel format, we spread the ratio correction factor reciprocally to both channels, so that the same ratio correction is achieved:

$$c1' = c1\sqrt{\frac{1}{b^{Mfit}}}$$

$$c2' = c2\sqrt{\frac{1}{b^{Mfit}}}$$

There are three parameters that need to be set for a Lowess (Figure 5-5).

Smoothing Factor: The smoothing factor parameter determines the number of data points having nearby (or equal) A values around the A value for a spot that are included in the spot's local regression estimate, expressed as a fraction 0 < f <= 1 of the total number of spots in the dataset. Thus, for a smoothing factor of 0.2, roughly 20% of the spots with the closest (or equal) A values to a given spot will be included for the regression estimate at that spot (however, see the proviso below about equal runs of A values). Within this window, the spots with the closest A values to the spot to be estimated are given the most weight, with the weight falling to zero near the edge of the smoothing window. However, one proviso applies here: the regression window's edges will never lie in the middle of a run of equal A values. A range of equal A values at the edge of the local regression window may cause the number of points to include in the regression to differ somewhat from that determined by the smoothing factor alone. If the right window edge would end within a range of equal A values based on the smoothing factor, then the window is extended to the right to include the full run of equal A values. On the other hand, the left window edge never lies within a range of equal A values because of the way the algorithm moves the window through ascending A values when choosing new points to estimate: When the algorithm estimates an M value at one point, that fitted M value is automatically set for all points with equal A values, and the next window will begin past all of these equal A values. This behavior delivers an important property of Lowess normalization which is that any two spots with the same A value will have the same fitted M value.

Robustness Iterations: The Lowess regression algorithm is "robust", meaning that it resists giving undue influence to outlying data points. Once a fitting of M values for the A values has been obtained, the regressions can be repeated but this time penalizing points with outlying M values via robustness weights based on residuals relative to the latest estimated fit. Spots with large residuals relative to the fitted M values will be given relatively less weight in the subsequent regressions. Note that these robustness weights are separate from the weights applied based on the A values' distance to the center of the regression window, which are always in effect even in the initial regression. The robustness iterations parameter determines the number of times after the first that the regressions will be done at each spot, for the purpose of reweighting the outlying data points into relative insignificance in the regressions. Thus, to give outlying data points the same weights as any other points (i.e. weights based on their A values alone), 0 can be chosen for robustness iterations.

Delta: The delta parameter is provided to speed up Lowess calculations on large data sets. Setting delta > 0 will let the algorithm skip over A values that are closer than delta to an A value that already has an estimated M value, using linear interpolation between the estimates. Leaving delta unset (leaving the field empty) will result in delta defaulting to 0.01*the range of A values, which is usually a good compromise between speed and accuracy. Setting delta to 0 is the most accurate but slowest setting; it means that the regression is done for every A value in the data.

The default setting for smoothing factor, robustness iteration, and delta is 0.2, 3, and empty (0.01*the range of A values) as shown in Figure 5-5.

Figure 5-6 shows the effect of Lowess on a dataset that showed intensity-based ratio bias. Lowess effectively removed intensity-based bias in the ratio values.

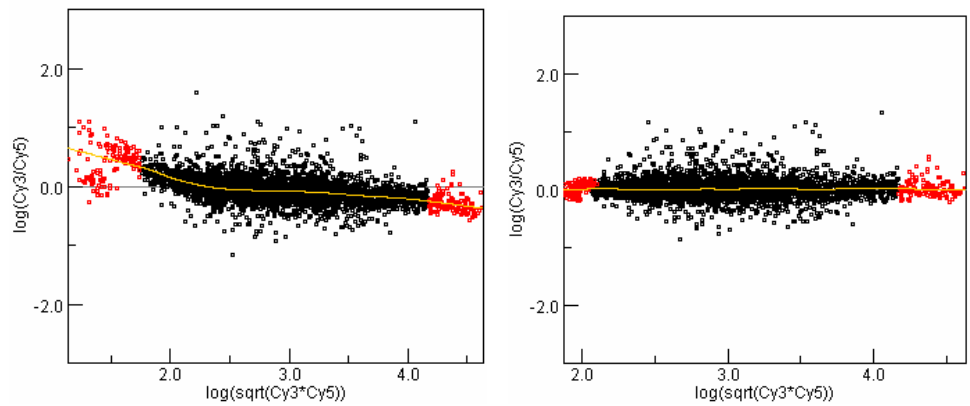


Figure 5-6: Systematic, intensity-based ratio bias (Left) is corrected by Lowess (Right). The yellow line shows the Lowess fitted values, Mfit.

Total Intensity Normalization

Total Intensity Norm. only applies to two-color platforms and tries to "balance" the total intensity of the two channels (samples) in three steps:

1) Compute the sum of each channel's intensities, optionally subtracting backgrounds;
2) Let r be the ratio of these sums, i.e., r = (sum ch1 vals)/(sum ch2 vals);
3) Scale factor for first channel is $\frac{1}{\sqrt{r}}$ and $\sqrt{r}$ for channel 2.

Normalized data are permanently saved in MicroarrayDB and have the following properties:

1) Ratio of intensity sums for the two channels computed for the normalized data should be 1.0;
2) Each spot is adjusted so that the ratio of channel values is $1/r$ times it's un-normalized value.

The effect of Total Intensity Norm is the same as mean intensity normalization.

Channel Scaling

Channel Scaling applies to both one-channel and two-channel platforms and adjusts each channel's intensity values according each Median, Mean, and user-specified Target Value (Figure 5-7). It is accomplished by multiplying each intensity value by T/m, where m is the mean or median of the channel data and T is the target mean/median value option (default is 1,000). Channels are scaled separately in the case of two-channel data. Normalized channel data will have a mean/median matching the target value option.

The effect of Channel Scaling is the same as median intensity normalization.

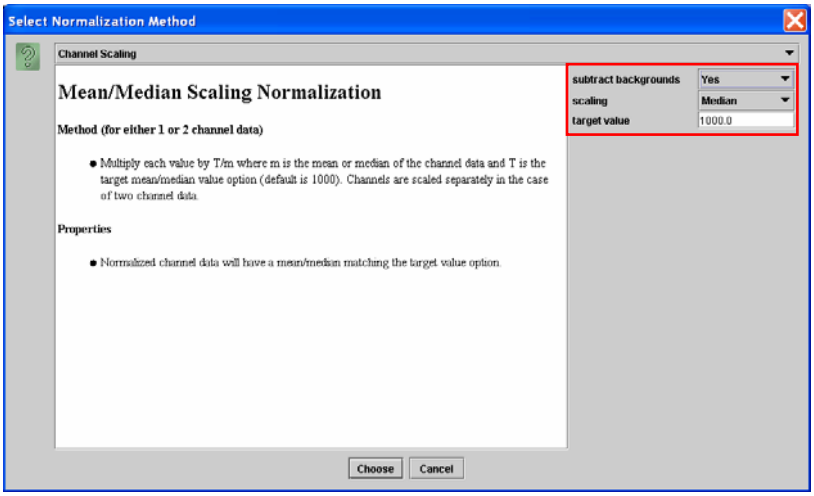


Figure 5-7: Channel Scaling normalization.

GenePix Mean Log Ratio Normalization

GenePix Mean Log Ratio Norm. only applies to two-channel platforms (Figure 5-8). The following steps are involved:

1) Compute channel ratios after respective background subtraction if specified (it doesn't matter which channel is the numerator and which is the denominator);
2) If the exclude ratio limit parameter M has been specified non-zero, then spots are ignored whose ratios don't lie between $1/M$ and M; the default value for M is 10;
3) Take log of remaining ratios (base doesn't matter, will cancel out);
4) Apply anti-log to the average of these log ratios to get r;
5) Scale factor for numerator channel is $\frac{1}{\sqrt{r}}$ and $\sqrt{r}$ for denominator channel (applied after background subtraction, if specified).

After normalization, the average of the log of channel ratios is 0, corresponding to r = 1.0. Each spot value is adjusted such that the ratio of channels is $1/r$ times it's un-normalized value.

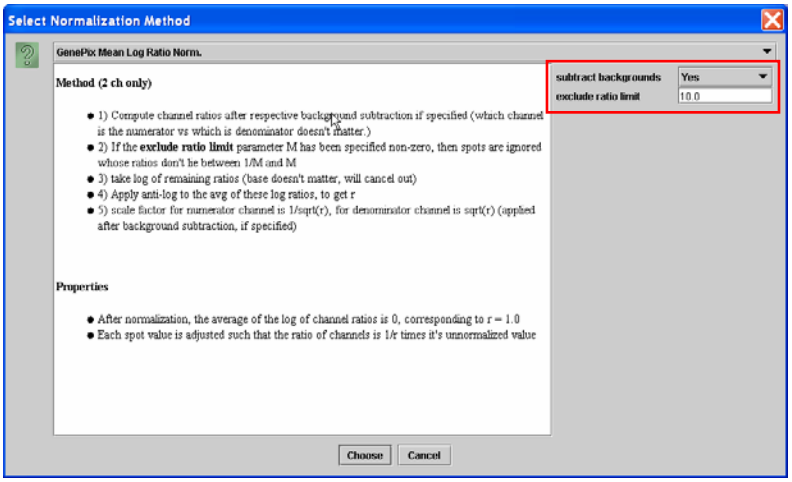


Figure 5-8: GenePix Mean Log Ratio Normalization

Linear and Lowess

This normalization method is the combination of Linear and Lowess. First, values for each channel are multiplied by T/m where m is the mean or median of the channel data and T is the target mean/median value option (default is 1000). Then a Lowess normalization is performed on the resulting scaled channel data. See Figure 5-9. At the right side of the window, there are three pull-down lists letting you have the choice for background subtracting, scaling and flagging the spot. Be aware that the default for Scaling is Geometric Mean. The user can select Median or Mean for different approach.

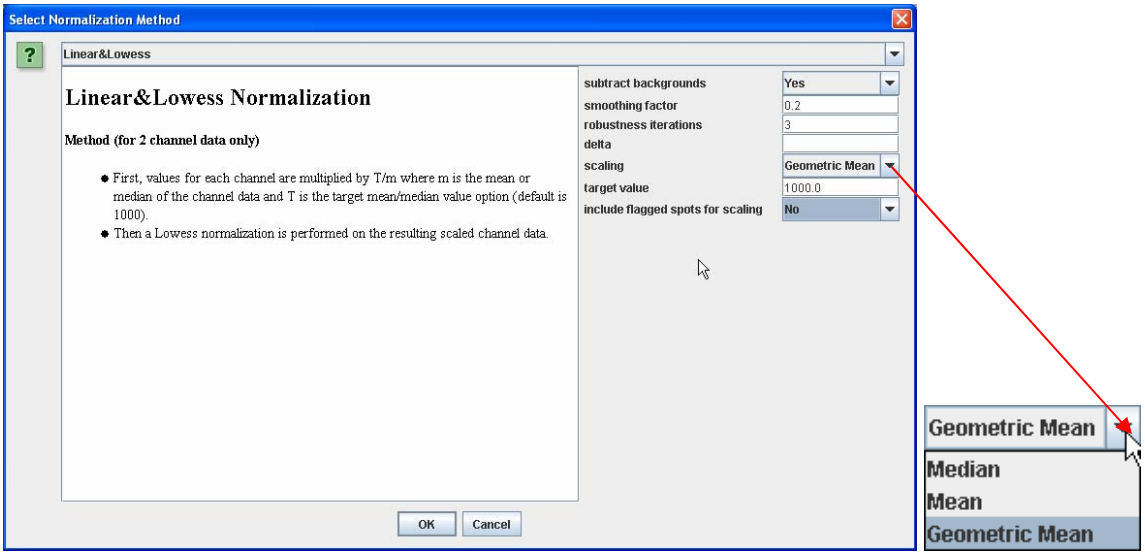


Figure 5-9: Linear and Lowess Normalization

Warning

Although different normalization methods can be applied to the same raw data, the user should apply the same normalization method to all the arrays within the same experiment for a meaningful microarray data analysis and comparison. The choice of a particular normalization method is solely the responsibility of the user. It is also advised that data normalized within ArrayTrack may need to be further "pre-processed" (e.g. centering around mean zero and variance one) before being systematically compared and analyzed by other data analysis software, depending on the particular analysis methods.

Chapter 6 Working with Tools: Analysis Tools

Overview

The analysis tools can be accessed within the "Tool" window (Figure 6-1A), or from tool menu (Figure 6-1B), or from database with data selected and right-clicking (Figure 6-1C). Analysis tools are provided to perform four mathematical/statistical operations: 1) T-test, 2)ANOVA, 3) P-value, 4) Clustering, 5) Principal Components Analysis (PCA).

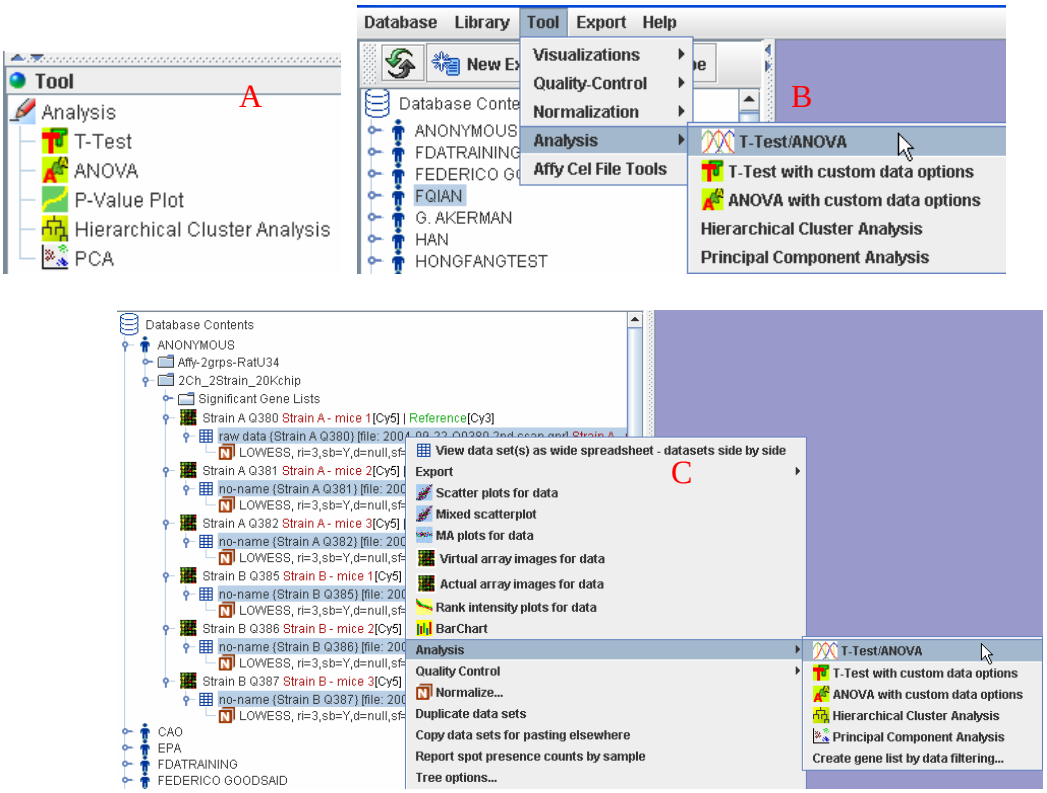


Figure 6-1: Access analysis tools

Before proceeding to explain the use of each analysis tool, please note that the various analysis tools are interlinked. For example, after applying the T-test to two sets of data, ArrayTrack will enable additional analysis tools to be applied to the selected data (more about this will be explained further below).

T-test and ANOVA

The T-test is used to compare two groups of data. It tells us if the variation between two groups is "significant". ANOVA (Analysis of Variance) is used to compare multiple groups. The users might ask if they can just do T-tests for all the pairs of groups. Multiple T-tests are not the answer because when the number of groups grows, the number of needed pair comparisons grows quickly. For example, if there are 7 groups of data, there will be 6+5+4+3+2+1 = 21 pairs. So the comparison will be too complicated. ANOVA puts all the groups of data into one test and gives us one P for the null hypothesis. There are three ways to activate T-test analysis and ANOVA:

1) Clicking the T-test icon under "Analysis Tool" in the Tools panel will pop up a window (Figure 6-2A) that is used for choosing the dataset to do the T-test analysis. The user can click "Browse" button to choose the data file which must be combined in advance into one data file containing multiple data set. If the user click "Gene ID's" button, a window will pop up and allow the user choose different ID types that will be shown in the T-test result, see Figure 6-2B.

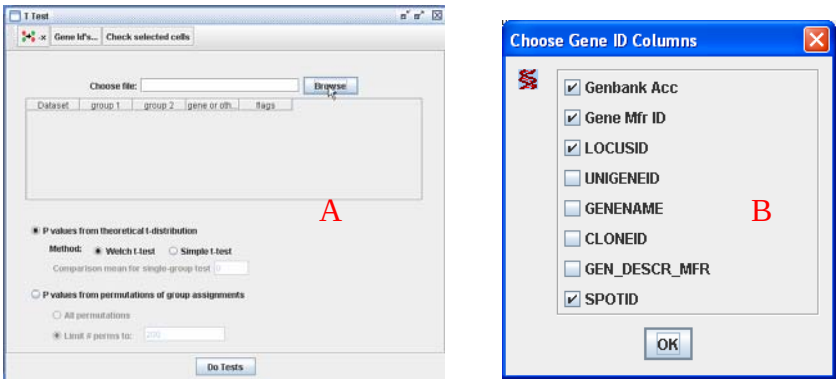


Figure 6-2: Choosing the data after activating T-test through the Tools panel

2) First choose the dataset from database panel as shown in (Figure 6-1C), and then right click and choose Analysis -> "T-test …". Accessing in this way, the user doesn't need to combine the data files.
3) Choose datasets from database panel and then click Tool pull-down menu ->choose "Analysis" -> "T-test …".

As shown in Figure 6-1B &C, there are three options for choosing T-test/ANOVA: 1) "T-test/ANOVA" let the selected data be exported directly to the T-test or ANOVA; and 2) "T-test with custom data options" provides the user the options to select other part of the data to be analyzed by T-test; 3) "ANOVA with custom data options" provides the ability to select additional data options as illustrated in Figure 6-3 and further explained below.

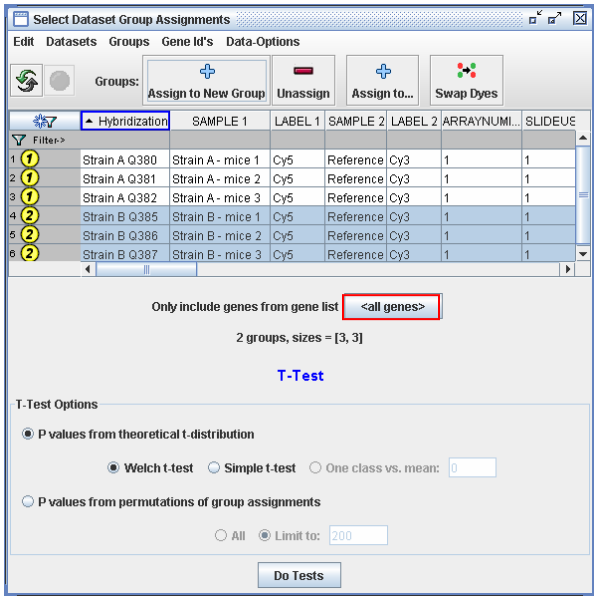


Figure 6-3: Assign dataset to groups

In Figure 6-3, the user can assign the dataset to different groups by highlighting the datasets first and then clicking "Assign to new group" button. The assigned datasets will be marked with a yellow-colored number. The dataset can also be assigned to an existing group by clicking the button Assign to... and typing the group number. The user can unassign the group by clicking the "Unassign" button. If there are two assigned groups then a T-test will be run, if there are three or more groups then an ANOVA test will be run. Before running the T-test, the user can choose to run T-test on all genes or on selected gene list by clicking the button <all genes>. A window will pop up letting the user choose the gene list for the T-test.

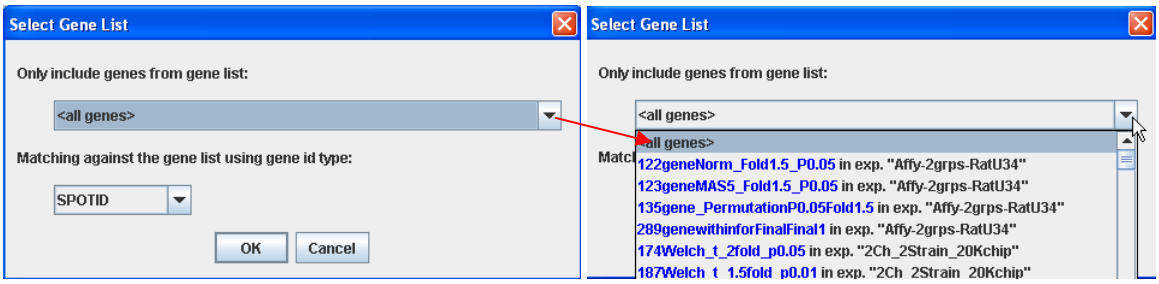


Figure 6-4: Select gene list for T-test analysis

For the T-test analysis, there are two options – 1) P values from theoretical t-distribution, 2) P values from permutations of group assignments. Under option 1 the user can choose "Welch t-test" and "Simple t-test". Under option 2 the user can set the criteria for the permutation T-test.

The following is the formula of permutation T-test:

$$P_{permuatation} = \frac{Numberof\,(P_k < P_0)}{NumberofPermutation}$$

where $${}_nP_k \equiv \frac{n!}{(n-k)!}$$

The user can choose to calculate $P_{permutation}$ using any number that less than or equal to *Numberof Permutation*. For example, if n=12, k=6, then $P_k = 924$. So the user can decide to calculate $P_{permutation}$ using NumberofPermutation = 924 or limit the number of permutation to any number that is less than 924. See Figure 6-5.

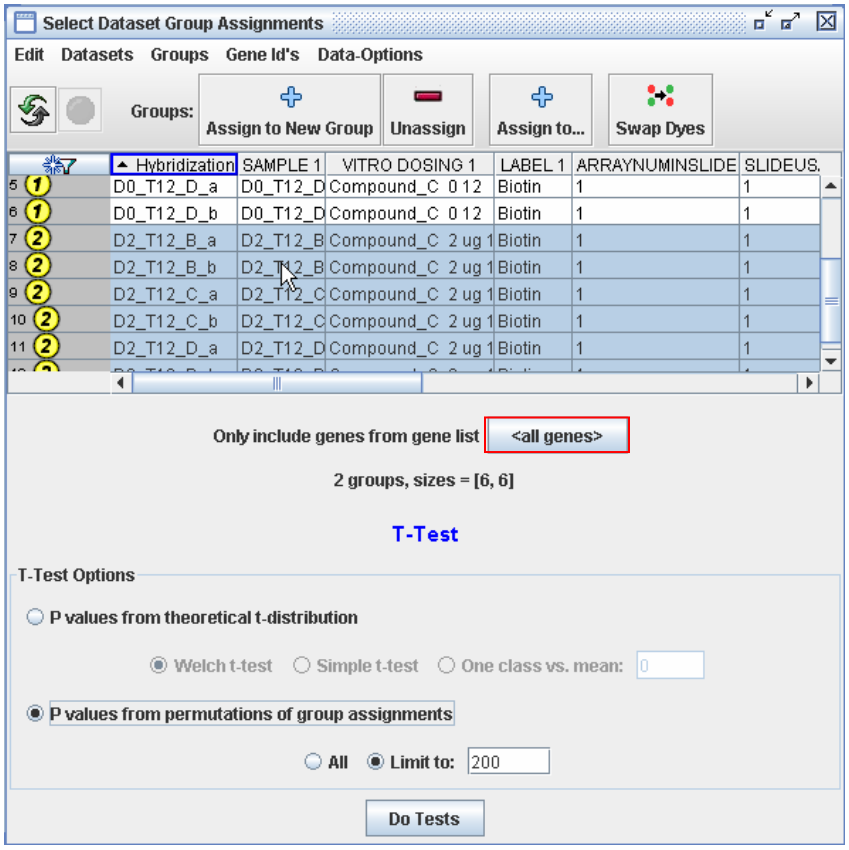


Figure 6-5: Set the criteria of the T-test

Click "Do tests" button, the results of T-test/ANOVA will be shown, see Figure 6-7.

The "Export Options" window is used to select the forms in which data will be exported prior to the T-test. For example, in Figure 6-3, only log intensity data will be exported in log base 2, and flag values will be included as integers. By clicking the "Dataset Naming" tab, the user can choose what parts of the data element names are to be included in the export data table, including the hybridization name, raw dataset description and normalization description. After option selection, clicking OK results in the "T Test" window appearing; in this window the user clicks the boxes to assign each data element to either the first or the second group, as shown in Figure 6-6A. The "Gene Id's" button opens another window (see Figure 6-6B) that allows choosing the ID types that will be shown in the data export results.

Note that multiple data can be selected for export, for example, by selecting both "include log intensities" and "include intensities", both types of the data will be included in the data to be exported. The data options available to be selected in the "Export Options" table depend upon the form of the data that was selected prior to initiating the T-test from the Analysis Tools icon. For example, if the initial data was log data, then intensity data cannot be selected.

Clicking the "Gene Id's…" button will pop up the window that allows choosing different types of id's to be shown in the analysis results, as shown in Figure 6-6B.

When the "Do Tests" button is clicked, the T-test is performed and the results are displayed in a new window titled "T-test Results", and shown in Figure 6-7. The bottom of the results window contains additional functions that can be applied to the results, such as filtering out results above a specified p-value, etc. Buttons are also provided at the bottom of the "T-test Results" window that allows additional operations on the filtered results, such as volcano plot,

HCA, PCA, etc. The T-test results window also allows for launching searches of the gene library, pathway library and protein library, as shown in Figure 6-7 and further explained below.

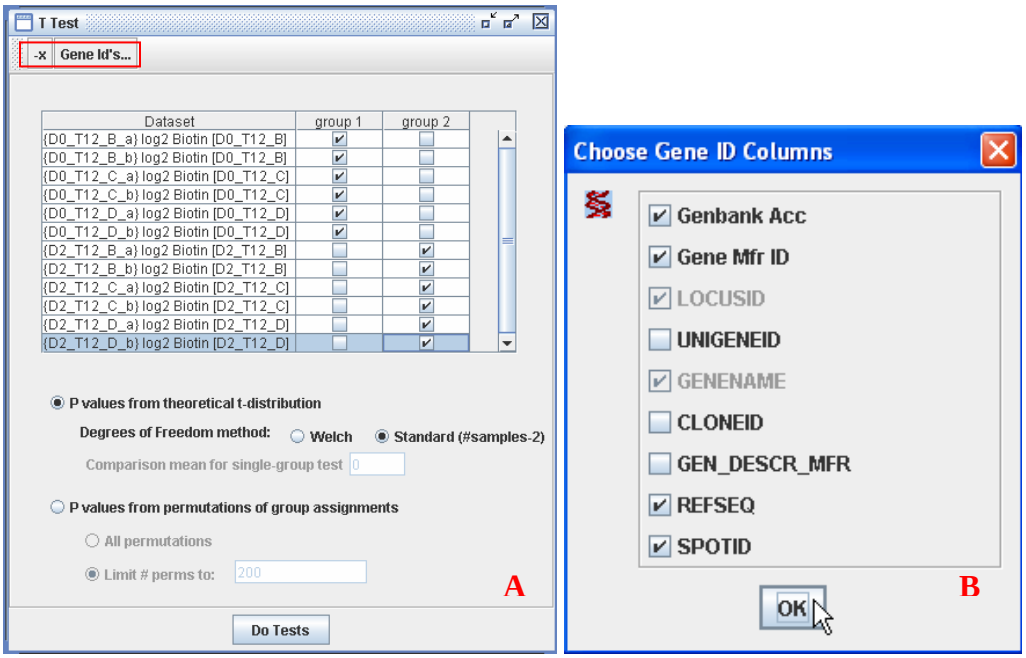


Figure 6-6: assign data in two groups (A) and choose Gene ID types (B)

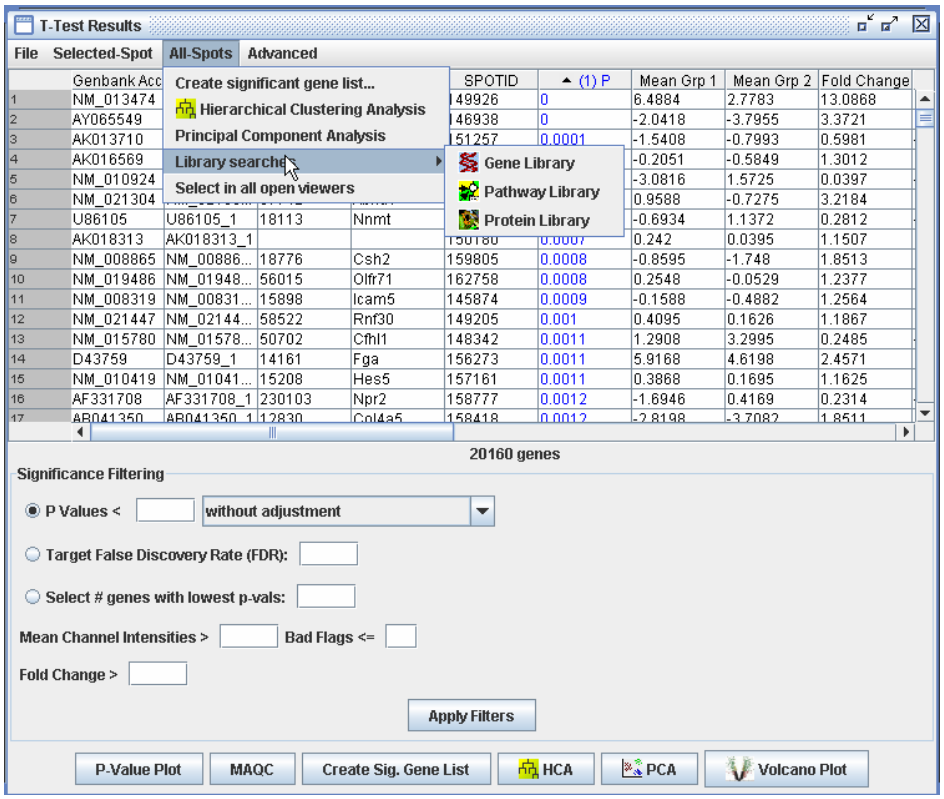


Figure 6-7: T-test results window

In Figure 6-7, the users can get the value of the T-test result like mean of group 1, mean group 2 and mean difference (group 2 - group 1), etc. Once the T-test results are obtained, ArrayTrack provides several types of filters and interactive graphics tools to aid the user in evaluating the data, generally, and in choosing a significant set of genes, particularly. As shown in Figure 6-7, the user can perform significance filtering. A P-value cutoff can be placed in the text box, with an option for Bonferroni correction in the adjacent drop-down menu. Alternative, a text box is available for specifying a certain number of genes with the lowest P-values. Text boxes are also provides for removing spots below the specified mean channel intensity and below a specified minimum fold-change. The filters can be applied in parallel.

Highlight one gene and choose "Selected Spot" pull-down menu -> Create Bar Chart… will launch the Gene Expression Bar Chart.

Clicking the

button causes a volcano plot of the selected export data to be produced and displayed in a new window. Notice that in the example volcano plot of Figure 6-8, the plot is partitioned into six areas by two vertical (x-axis representing the fold-change scale) and one horizontal (y-axis representing the p-value scale) dashed lines. The volcano plot is intended as a graphical tool to select a list of significant genes based on some combination of p-value/fold-change criteria, or to examine the effects of p-value and/or fold change cutoff values on the significant gene list. Usually, the genes appearing in the upper left and upper right areas, areas A and C in Figure 6-8 will comprise the significant gene list, that is, the spots denoted by red in Figure 6-8.

The volcano plot produced by ArrayTrack has a number of features and interactive capabilities providing particular utility for selection of a significant gene list, as summarized below:

1) The mouse cursor can be used on one of the vertical dashed lines to drag the lines either further apart or closer together, increasing or decreasing, respectively, the fold-change encompassed between the vertical lines. The corresponding absolute fold change is displayed to the right of the right side vertical dashed line, and can be seen to change as the vertical position along the x-axis is changed. Alternatively, a fold change difference can be typed into the text box above the plot, causing the vertical lines to be adjusted to that value.

2) Similarly, the mouse cursor can be used to move the horizontal dashed line up or down to change a hypothetical P-value cutoff, with spots above the horizontal line being below the p-value of the intersection of the y-axis. Alternatively, the p-value cutoff can also be typed into the text box above the plot, which will cause the horizontal line to be adjusted to that value.

3) Floating the cursor over a spot will cause information about the spot to be displayed above the plot. The upper line gives numerical values for fold-change, p-value and average channel intensity (average intensity of the spot across all channels and all microarrays). The second line gives the identification information for the spot that was selected prior to the T-test, as shown in Figure 6-6. The third line displays the number of significant genes corresponding to: 1) both the fold-change and P-value cutoff (areas A and C of Figure 6-8); 2) the P-value cutoff alone (Areas A, B and C of Figure 6-8); and, 3) the fold-change cutoff alone (areas A, C, D and F in Figure 6-7); the number of non-significant genes is also displayed (area E of Figure 6-8).

4) The adjustable P-value and fold change lines divide the plot into several color-coded regions (see the keys at the right of the plot) that correspond to regions of significance or non-significance that depend on cutoff values. There are three display options chosen by drop-down menu: 1) color by region; 2) color by mean channel intensity (Red to blue); and color by mean channel intensity (gray scale), as shown in Figure 6-8.

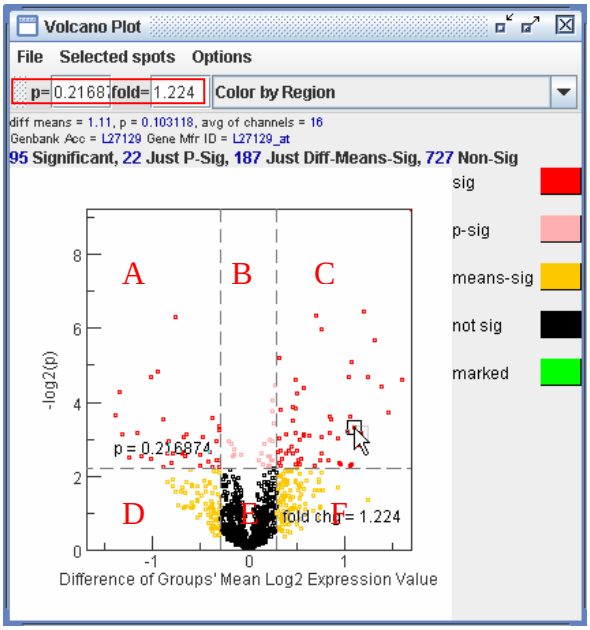


Figure 6-8: Volcano Plot

6) The pull-down menu labeled "Selected Spots" provides the user the ability to: a) create a significant gene lists that corresponds to the desired P-value and fold-change combination; b) perform cluster analysis on the selected significant genes; c) perform search of the ArrayTrack gene, protein and pathways libraries for the selected significant genes; d) mark the selected significant genes in other ArrayTrack viewer windows that might be open (e.g., image viewer).

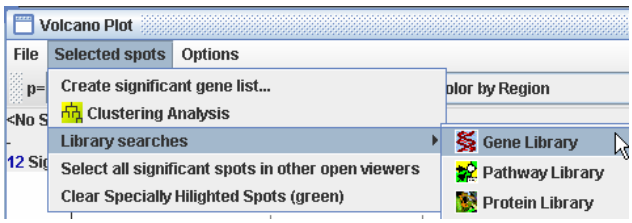


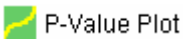
Figure 6-9: Selected spots pull-down menu

P-value Plot

A p-value is associated with a test statistic. It is "the probability, if the test statistic really were distributed as it would be under the null hypothesis, of observing a test statistic [as extreme as, or more extreme than] the one actually observed". The smaller the P value, the more strongly the test confirms the null hypothesis (Econterms).

The user can get P-value plot from T-test result window by clicking the

button (see Figure 6-7).

Also P-value Plot tool can be activated by clicking 

under the Tool/Analysis section. See Figure 6-1A. A pop up window with title "Choose Data Source" allows the user to choose the data file (Figure 6-10). The user needs to assign the columns to the corresponding buttons (e.g. clicking column "P" and then click "P-value" button), then click OK to get P-value plot.

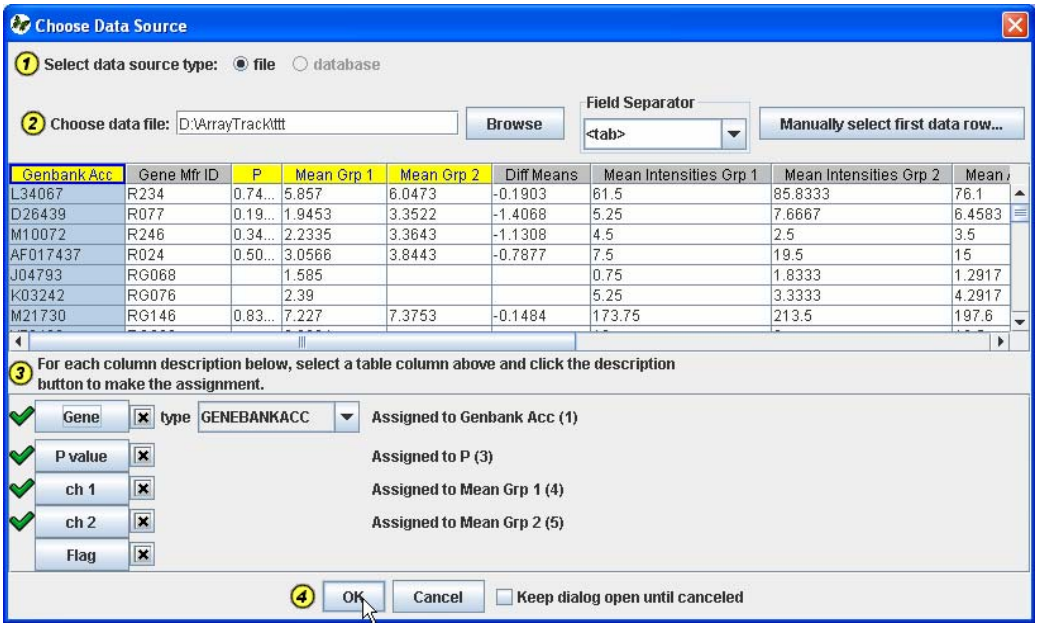


Figure 6-10: Choose data file do get p-value plot

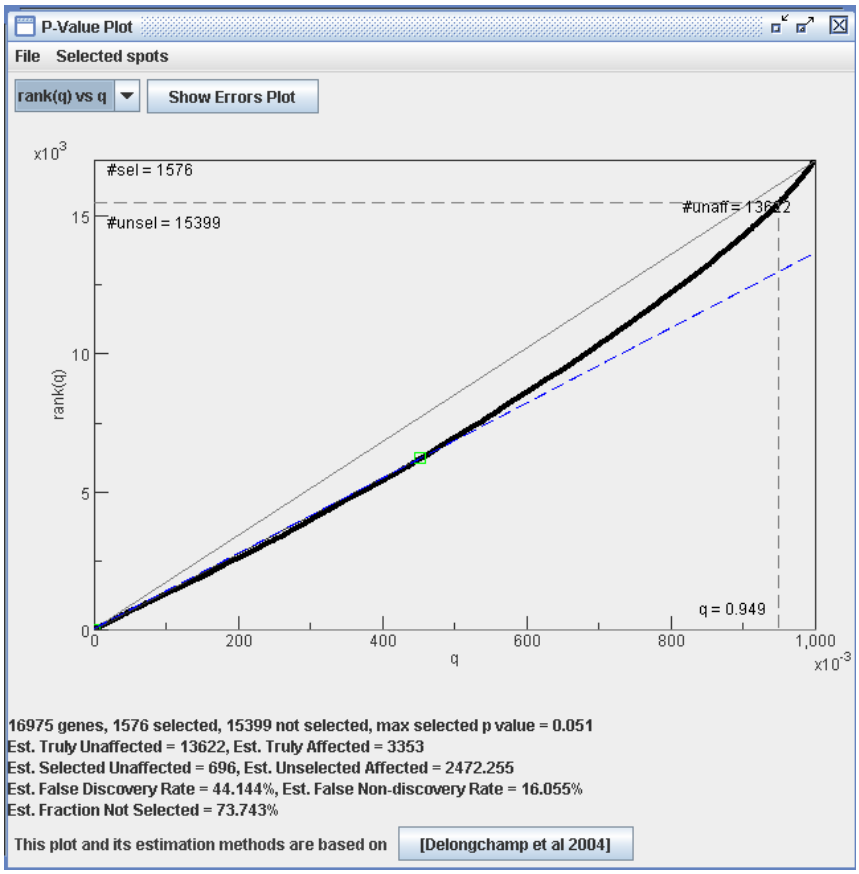


Figure 6-11: P-value plot

In P-value plot window, the y-axis is rank and x-axis is p or q (q= 1-p, user can choose rank(q) vs q or p vs rank(p)). By dragging the gray dashed line, the user can change the p(or q) value and consequently the selected spots(#sel) and unselected spots(#unsel).

The blue dashed line can indicate the number of truly affected and truly unaffected genes.

In P-value plot window, the user can also access volcano plot, clustering analysis and other libraries. See Figure 6-12.

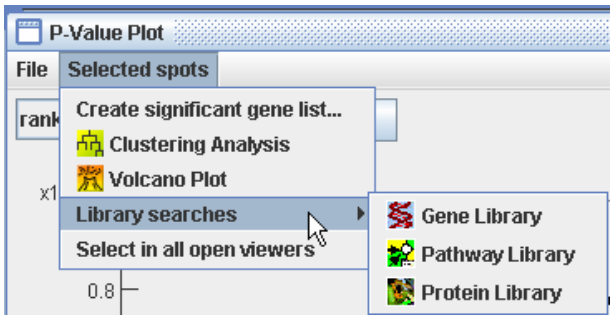


Figure 6-12: Access other functions from p-value plot

Clustering

Hierarchical Clustering Analysis (HCA) can be activated by 1) clicking the HCA icon under the Analysis tool, or 2) choosing the dataset in database panel, right clicking and then choosing "Analysis-> Hierarchical Component Analysis", see Figure 6-1.

1) If HCA is activated in the first way, a pop up window will ask the user to select a file to do analysis. The data file must be text file format (.txt) (see Figure 6-13A). The text file will be shown after the user click "Open" button (see Figure 6-13B). Then the user can do data analysis by clicking "Analysis" in the menu bar or clicking HCA icon . The result is shown in Figure 6-15.

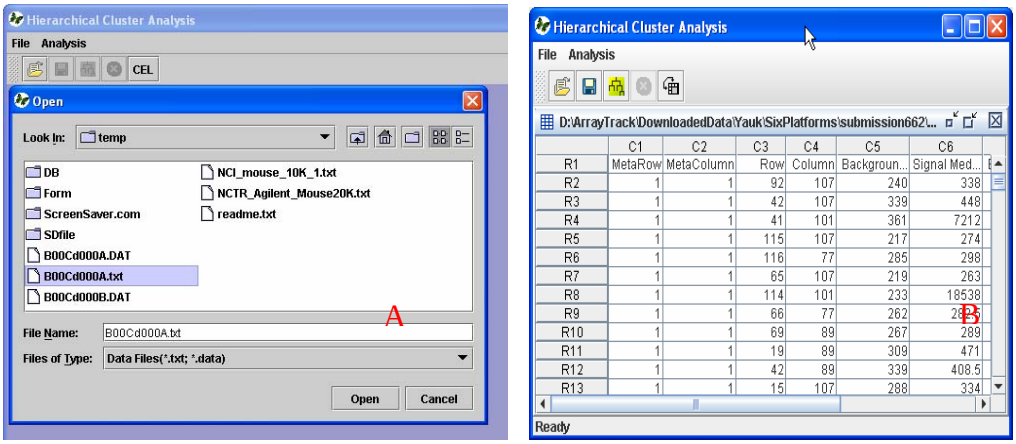


Figure 6-13: Select data file to do data analysis

2) If HCA is activated in the second way, the "Export options" window pops up (Figure 6-14). Clicking OK will bring out the HCA plot, see Figure 6-15.

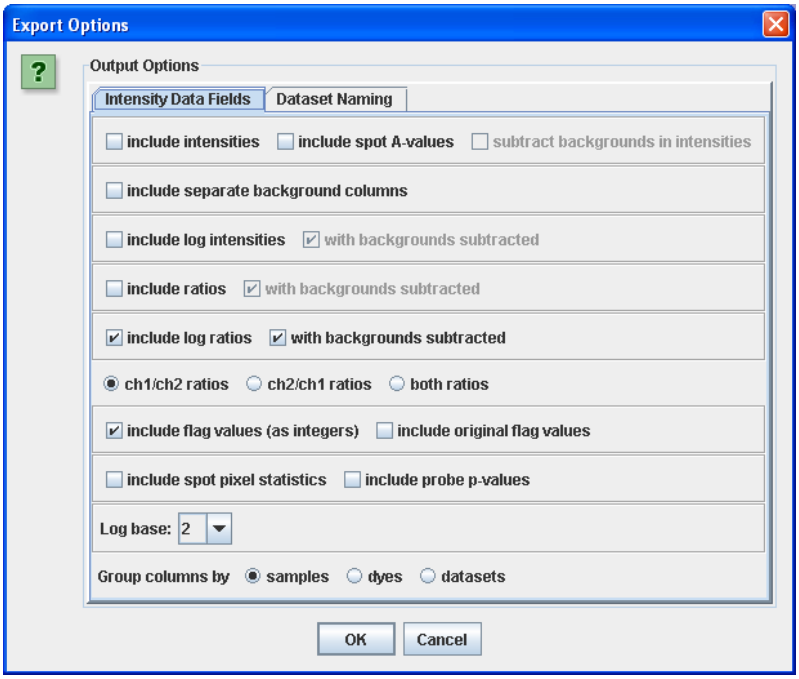


Figure 6-14: Export Option window

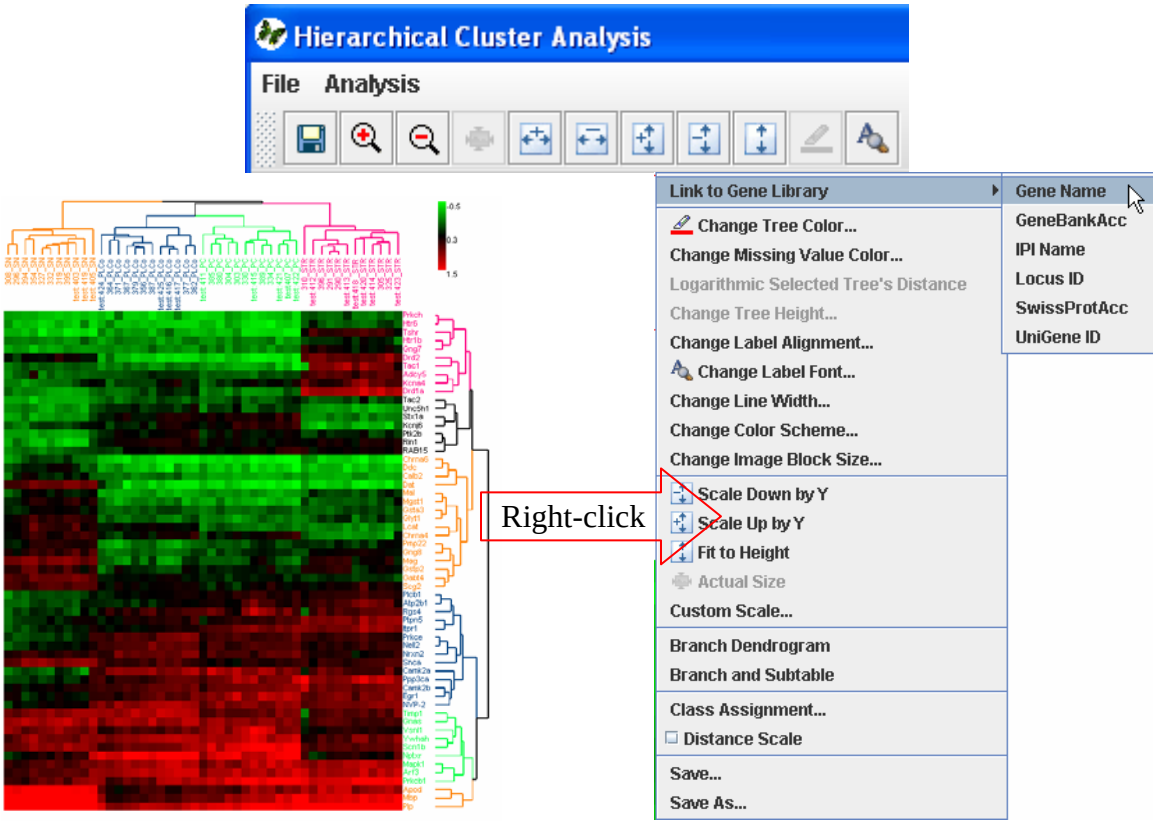


Figure 6-15: HCA analysis result

The user can zoom in/out the plot by clicking or icon at the top of the window. The user can also change the font and the color of the label for each branch of the tree cluster by right-clicking the branch and choosing the right options. From the HCA plot there is a link to Gene Library according to the available IDs. The user can save the HCA image to the local drive. See Figure 6-15.

PCA

Principal Component Analysis is a classical statistical method and is a way of identifying the data patterns and highlighting the data's similarity and differences. The main use of PCA is to reduce the dimentionality of a data set while retaining as much information as possible. In PCA plot, the first principal component is the combination of variables that describe the greatest amount of variation. The second principal component defines the next largest amount of variation and is independent to the first principal component, and so on.

Similar to the other analysis tools, PCA can be activated by 1) clicking the icon PCA under the Tool/Analysis section or 2) choosing the normalized dataset in database panel, right clicking and then choosing "Analysis-> Principal Component Analysis.

1) If the user activates PCA in the first way, the following window pops up. The user can open a file from the local drive or import data from database to do PCA analysis. See Figure 6-16.

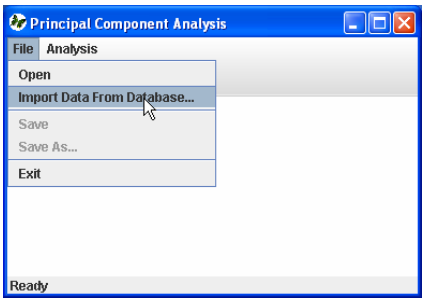


Figure 6-16: Open a file or import data from database for PCA

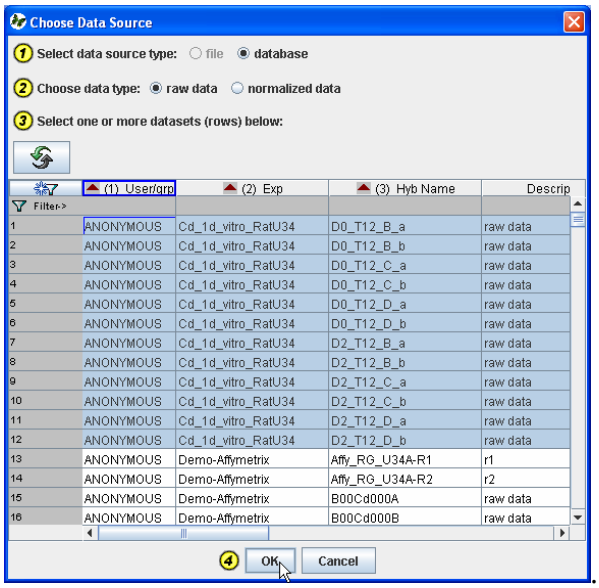


Figure 6-17: Choose data from the database to do PCA

In Figure 6-17, multiple hybridization data can be selected. Clicking OK will bring up the "Export option" window (Figure 6-14). Click OK button, then the PCA plot will show up (see Figure 6-18A).

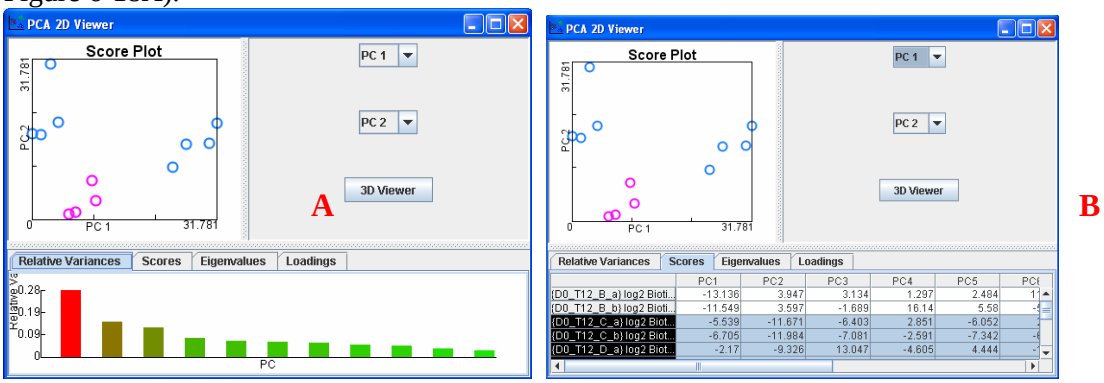


Figure 6-18: PCA plot

2) If the user activates PCA in the second way, the "Export option" window pops up, click OK then PCA results will show up.

In Figure 6-18A, if the user circle any spots in the plot, the "Relative Variance" view will be switched to the "Score" tab view, with the corresponding records highlighted (Figure 6-18B). The user can also click EigenValue and Loadings tab to see the Eigen value and loadings value for each principal component.

The user can also view the PCA plot in three dimensions by clicking the 3D view button, see Figure 6-19. The spot color can be changed by selecting it and then right-clicking-> choose new color. The spot shape can be changed to cube or sphere.

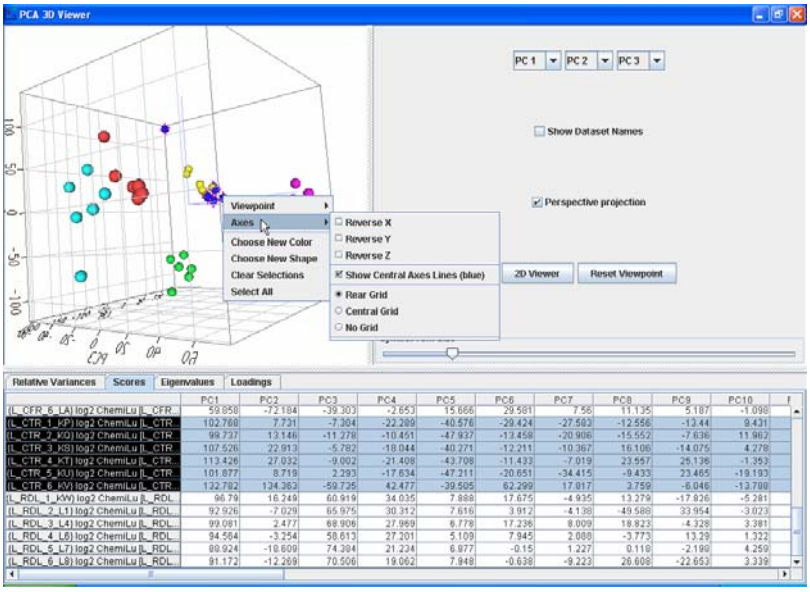


Figure 6-19: PCA three dimension plot

For the three-dimension PCA plot, the user can do the following movements:

1) Zoom in/out: hold shift key while drag the plot to bigger/smaller view.
2) Move the plot without rotation: hold ctrl key while drag the plot to the desired position.
3) Rotate: just drag the plot.
4) Reset: click the "Reset Viewpoint" button will bring the plot to the original position.

Chapter 7 Working with Tools: Visualization

Overview

Visualization is an important step in analyzing microarray data and can be used to identify abnormalities within the data. For example, when two (replicate) arrays are compared to each other, the user can gain an understanding on reproducibility of the experiment. It is highly recommended that the user maximize the use of the visualization tools made available within ArrayTrack before doing massive (and time-consuming) data mining and statistical analysis. Any suspicious arrays should be dealt with care when biological and statistical conclusions are withdrawn. Don't forget: "*Garbage-in-garbage-out*."

The following visualization tools have been implemented within ArrayTrack: Scatter Plot, MA Plot, Mixed Scatter Plot, Virtual Array Viewer, Rank Intensity Plot, P-Value Plot and Cross-Dataset Gene Bar chart.

Most of these functions can be accessed in three different ways (Figure 7-1): (A) from the TOOL panel; (B) from the Tool pull-down menu; and (C) by right-clicking on one of the selected arrays.

Double-click on Visualization Tools at the TOOL panel hides or shows the contents (visualization tools) underneath it.

Many of the visualization tools are interconnected to each other, thus allowing the user to gain a more in-depth view of the data from different perspectives, as will be seen from the detailed discussions in the following sections.

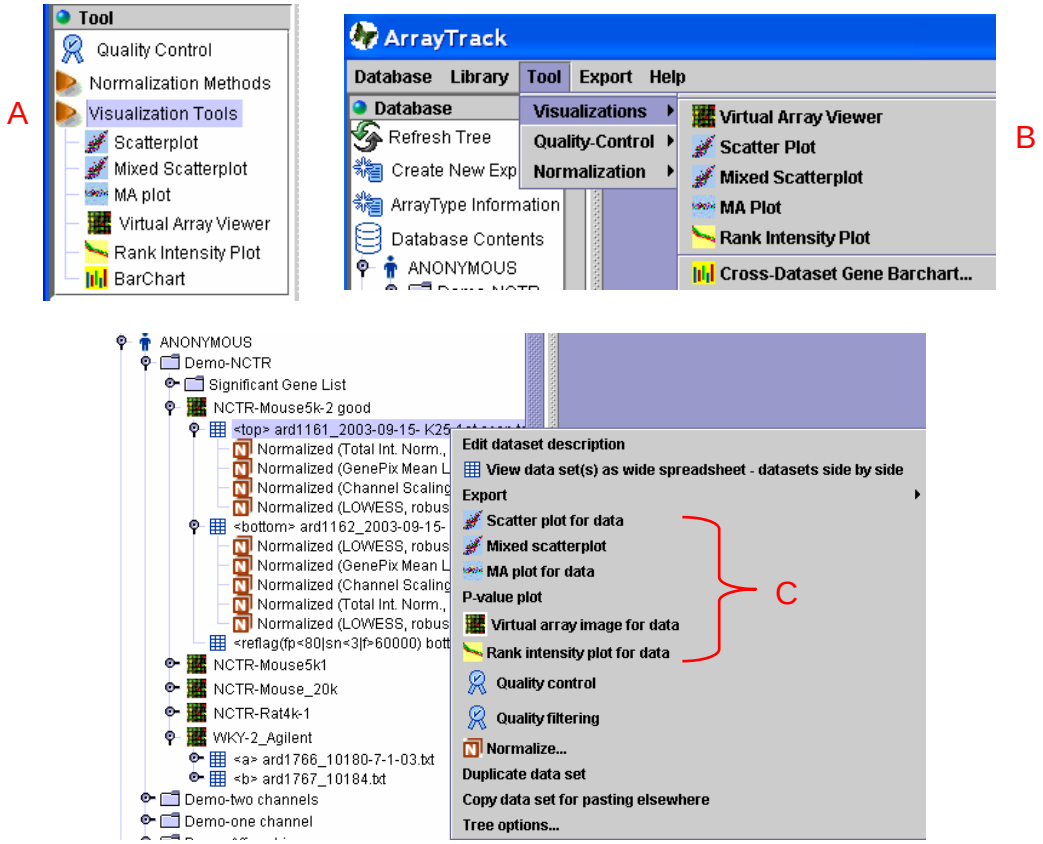


Figure 7-1: Three ways of accessing many Visualization Tools implemented in ArrayTrack: (A) TOOL panel; (B) Tool pull-down menu; and (C) Right-clicking on one of the selected arrays.

Scatter Plot

Default Plot: By default, Scatter plot for data plots the fluorescence intensity data of the Cy3 channel versus those of the Cy5 channel for the same array. An error message will be displayed if it is trying to be applied to one-channel data. Figure 7-2 shows the scatter plot for hybridization NCTR_Mouse_20K. The user can toggle data points that are flagged out e.g. by the Axon GenePix Pro software (Flagged data points are generally those spot features that do not show reliable fluorescence intensity signal; shown as grey cross symbols in Figure 7-2). The user can also choose to plot the background-subtracted intensity data. At the bottom of the plot, statistics for the two channels are displayed. When the mouse moves over the spots, the identity and intensity values of that spot are displayed. "Interesting" spots can be selected by click-and-circle using the mouse, and the selected spots are colored in red.

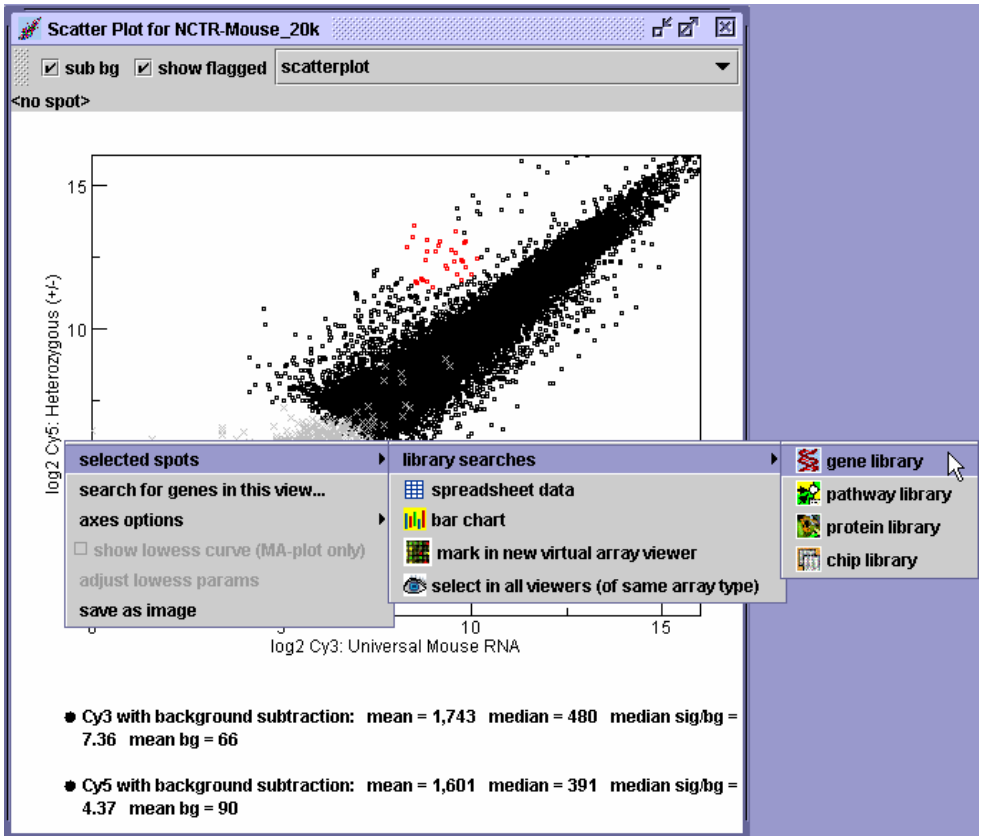


Figure 7-2: Scatter Plot showing the fluorescence intensities for the two hybridization channels.

Right-click on the plot area will pop up a list of actions that can be taken on the scatter plot (Figure 7-2). All the actions are self-explaining and are discussed as follows.

Selected Spots: For selected spots, five actions can be applied:

1) Launch library search against Gene Library, Pathway Library, Protein Library or Chip Library using the GenBank accession numbers of selected spots as queries. See Chapter 3 for details on library search.
2) View gene expression data for the selected spots in a spreadsheet form. For details, see Chapter 9: Data Export.
3) Launch cross-dataset gene Bar Chart for selected spots (for a maximum of five spots). For details, see the section on Bar Chart in this Chapter.
4) Launch a new Virtual Array Viewer with the selected spots marked. For details, see the section on Virtual Array Viewer in this Chapter.

5) Select (mark) the same set of selected spots in all other viewers of the same array type.

Search for Genes in This Viewer: Allows the user to enter/paste the GenBank accession numbers of a list of interested genes to find their location in the scatter plot (Figure 7-3A). Spots already selected can be kept by choosing "Retain current selections". Optionally, the search can be conducted on all other viewers of the same array type. After clicking on Search, genes found on this scatter plot are marked in red (Figure 7-3C); a ✓ or ✗ mark is shown before each GenBank accession number to indicate the presence or absence, respectively, of that individual gene on the scatter plot (Figure 7-3B).

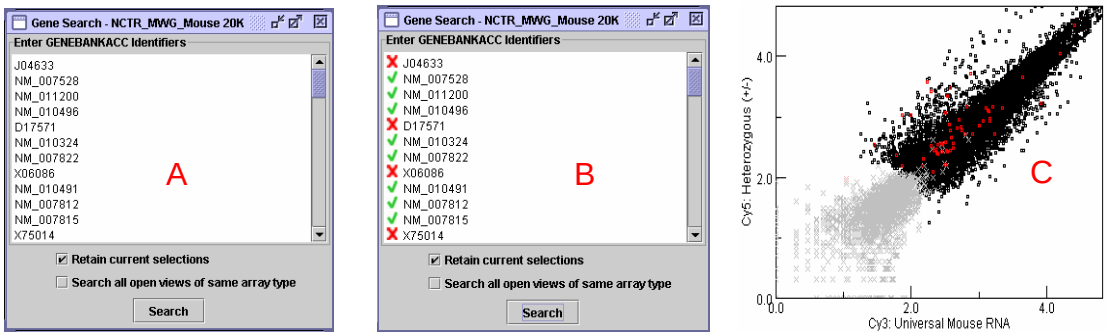


Figure 7-3: Locate a list of interested genes (by GenBank accession numbers) on the scatter plot.

Axes Options

1) Set ranges… sets the range of the plot in X and Y axis.
2) Reverse axes options exchanges the assignment of the X and Y variables (Figure 7-4).
3) Force origin (0, 0) to be visible allows the origin (0, 0) of the Scatter Plot to be visible. Otherwise, ArrayTrack automatically sets the limits of the X and Y axes based on the range of the X and Y values.
4) Constrain MA plot's M axis to [-3.0, 3.0] (MA plot only): If this option is toggled on, the limits of the Y axis (i.e. M or log fold change) is set to be [-3, 3]. Otherwise, ArrayTrack automatically sets the limits based on the range of the M values.
5) Log base: allows the axes to be switched to log base 2, log base 10 or log base e.

Figure 7-4: Axes Options for scatter plot.

Show Lowess curve (MA plot only): The Lowess normalization curve is displayed in yellow (Figure 7-5). For details about Lowess, see Chapter 5 on Normalization Methods.

Adjust Lowess parameters (MA plot only): The Lowess Parameters panel (Figure 7-6) allows the user to adjust the three parameters. As expected, when the Smoothing Factor was adjusted from its default value of 0.2 (e.g. 20% of all data points) down to 0.01 (1%), the

Lowess curve showed more variation (Figure 7-5B), resulting form the fitting of a much smaller portion of the neighboring data points. The user has the option of setting these values as the default values for the remaining session.

Save as image: The graphics can be saved as an image file in JPEG, TIFF, or PNG format. The exact file format is specified by the file name extension of .JPG, .TIF, or .PNG, respectively.

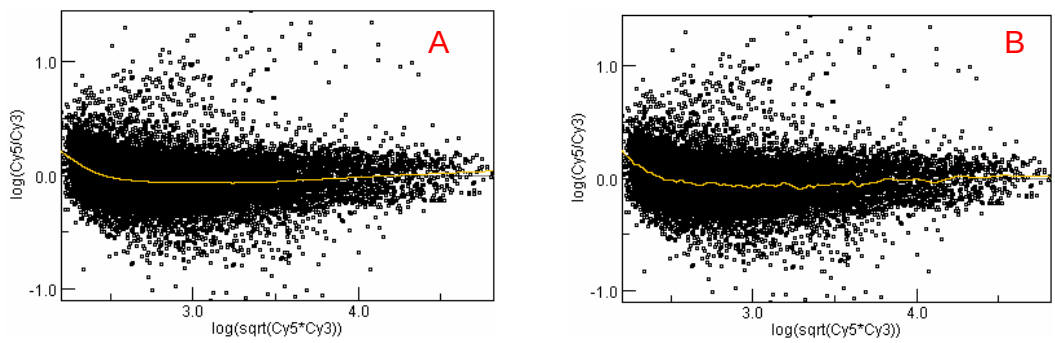


Figure 7-5: Lowess curve is displayed in yellow. A: Default Lowess parameter settings (smoothing factor = 0.2); B: Smoothing facto = 0.01.

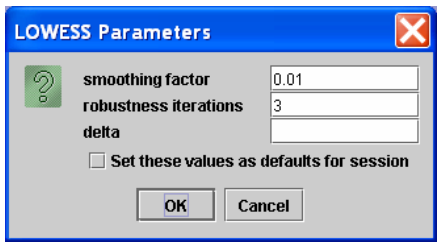


Figure 7-6: Adjusting Lowess parameter settings for the current MA plot.

MA Plot

MA Plot only applies to two-channel data (Figure 7-7). It is a special form of Scatter Plot in which the X axis is the log geometric average (**a**ddition) of the intensity values of the two channels and the Y axis is the log fold change (**m**inus of log intensities):

$$X = \log\sqrt{Cy5 * Cy3} = \frac{1}{2}(\log Cy5 + \log Cy3)$$

$$Y = \log(\frac{Cy5}{Cy3}) = \log Cy5 - \log Cy3$$

The MA Plot is also called *RI* plot, where *R* refers to log ***R***atio and *I* refers to log average ***I***ntensity. All the functions described in Scatter Plot apply to MA Plot. In fact, Scatter Plot and MA Plot are interchangeable (Figure 7-7) and applicable only to two-channel microarray data.

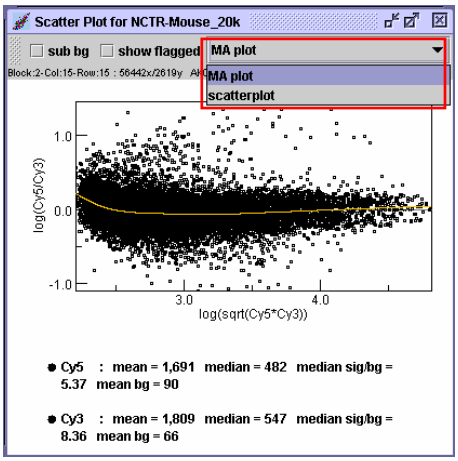


Figure 7-7: MA Plot and Scatter Plot are interchangeable.

Mixed Scatter Plot

Mixed Scatter Plot applies to both two-channel and one-channel data. It allows the user to compare two arrays in one plot in addition to the options of plotting different information items for the same array (if only one array is selected before launching this function). If more than two arrays are selected, Mixed Scatter Plot will not be accessible by right-click on the selected arrays.

After selecting two arrays and launching Mixed Scatter Plot, a plot displaying the Cy3 channel intensities of the two arrays is shown (Figure 7-8). Options such as background subtraction, flagged spots, statistics panel, and log transformation can be selected. There are several data items that can be used as variables for the X and Y axes (Figure 7-8 Right). Combination of such choices makes it possible for displaying various kinds of scatter plots including MA Plot and the regular Scatter Plot. Similar to Scatter Plot and MA Plot, right-click on the plot area will allow access to many functions applicable to the Mixed Scatter Plot.

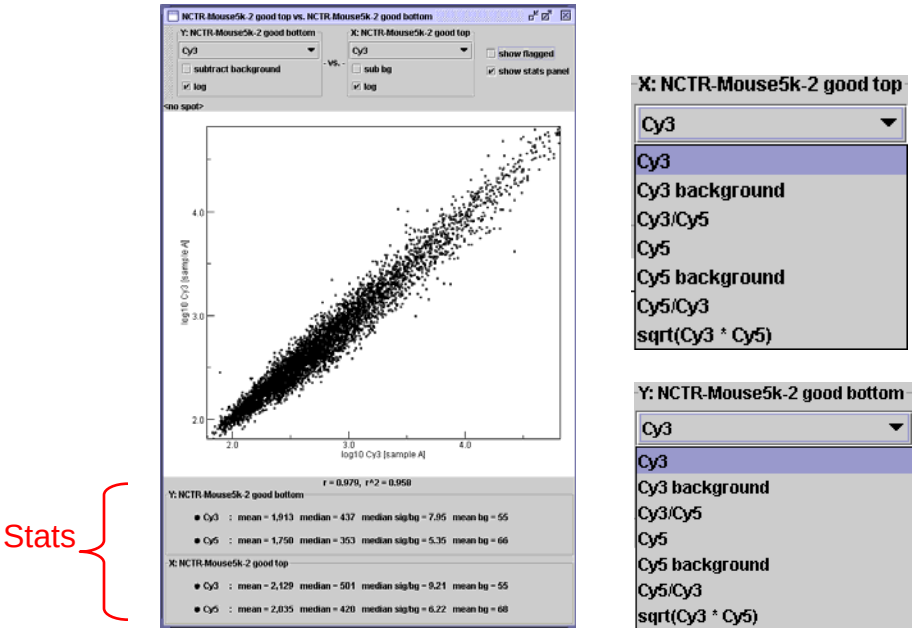


Figure 7-8: Mixed Scatter Plot (Left) and options available for plotting (Right).

Rank Intensity Plot

Rank Intensity Plot applies to both two-channel and one-channel data. It provides a convenient way of visualizing the distribution of intensity data across all the spots on the array: a closer Rank Intensity Plot indicates a closer distribution of the data items (e.g. Cy3 vs. Cy5 intensities from the same array). In the Rank Intensity Plots shown in Figure 7-9, the X axis represents the intensity sorted rank (from low to high) and the Y axis is the corresponding log10 intensity.

Rank Intensity Plot is shown in the Quality Control panel (Chapter 4).

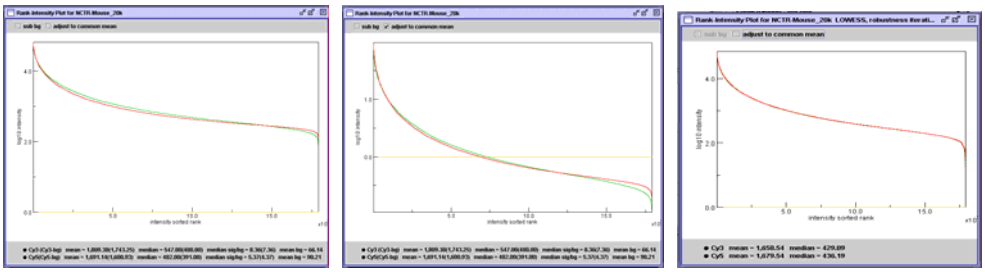
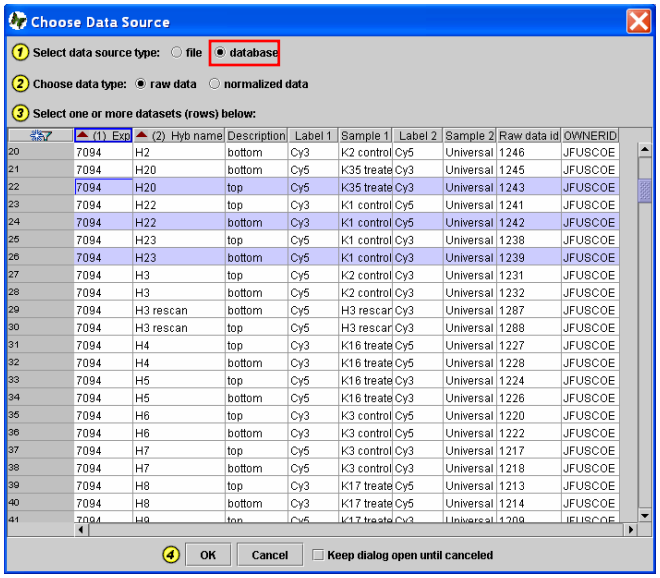


Figure 7-9: Rank Intensity Plot. In the middle plot, mean intensities for Cy3 and Cy5 channels are adjusted to equal mean of zero. In the right plot, Lowess normalized data are plotted.

Choosing Data Source for Plotting

From MicroarrayDB Arrays: When Scatter Plot, Mixed Scatter Plot, MA Plot, and Rank Intensity Plot are activated from the TOOL panel or the pull-down menu, a list of arrays stored in the MicroarrayDB is displayed in a spreadsheet from (Figure 7-10). The default for data source type is MicroarrayDB database and the data type to be plotted is raw data. Multiple arrays can be selected for plotting.

From Local Data Files: When the user choose file as the data source type (Figure 7-10), a local disk file can be input for plotting (Scatter Plot, Mixed Scatter Plot, MA Plot, and Rank Intensity Plot), as shown in Figure 7-11. The table columns in the local file can be mapped to the required data fields (e.g. the X and Y coordinates) for plotting.



	(1) Exp	(2) Hyb name	Description	Label 1	Sample 1	Label 2	Sample 2	Raw data id	OWNERID
20	7094	H2	bottom	Cy3	K2 control	Cy5	Universal	1246	JFUSCOE
21	7094	H20	bottom	Cy5	K35 treate	Cy3	Universal	1245	JFUSCOE
22	7094	H20	top	Cy5	K35 treate	Cy3	Universal	1243	JFUSCOE
23	7094	H22	top	Cy3	K1 control	Cy5	Universal	1241	JFUSCOE
24	7094	H22	bottom	Cy3	K1 control	Cy5	Universal	1242	JFUSCOE
25	7094	H23	top	Cy5	K1 control	Cy3	Universal	1238	JFUSCOE
26	7094	H23	bottom	Cy5	K1 control	Cy3	Universal	1239	JFUSCOE
27	7094	H3	top	Cy5	K2 control	Cy3	Universal	1231	JFUSCOE
28	7094	H3	bottom	Cy5	K2 control	Cy3	Universal	1232	JFUSCOE
29	7094	H3 rescan	bottom	Cy5	H3 rescan	Cy3	Universal	1287	JFUSCOE
30	7094	H3 rescan	top	Cy5	H3 rescan	Cy3	Universal	1288	JFUSCOE
31	7094	H4	top	Cy3	K16 treate	Cy5	Universal	1227	JFUSCOE
32	7094	H4	bottom	Cy3	K16 treate	Cy5	Universal	1228	JFUSCOE
33	7094	H5	top	Cy5	K16 treate	Cy3	Universal	1224	JFUSCOE
34	7094	H5	bottom	Cy5	K16 treate	Cy3	Universal	1226	JFUSCOE
35	7094	H6	top	Cy3	K3 control	Cy5	Universal	1220	JFUSCOE
36	7094	H6	bottom	Cy3	K3 control	Cy5	Universal	1222	JFUSCOE
37	7094	H7	top	Cy5	K3 control	Cy3	Universal	1217	JFUSCOE
38	7094	H7	bottom	Cy5	K3 control	Cy3	Universal	1218	JFUSCOE
39	7094	H8	top	Cy3	K17 treate	Cy5	Universal	1213	JFUSCOE
40	7094	H8	bottom	Cy3	K17 treate	Cy5	Universal	1214	JFUSCOE
41	7094	H9	top	Cy5	K17 treate	Cy3	Universal	1209	JFUSCOE

Figure 7-10: List of arrays in MicroarrayDB. Selected arrays will be plotted after click on OK.

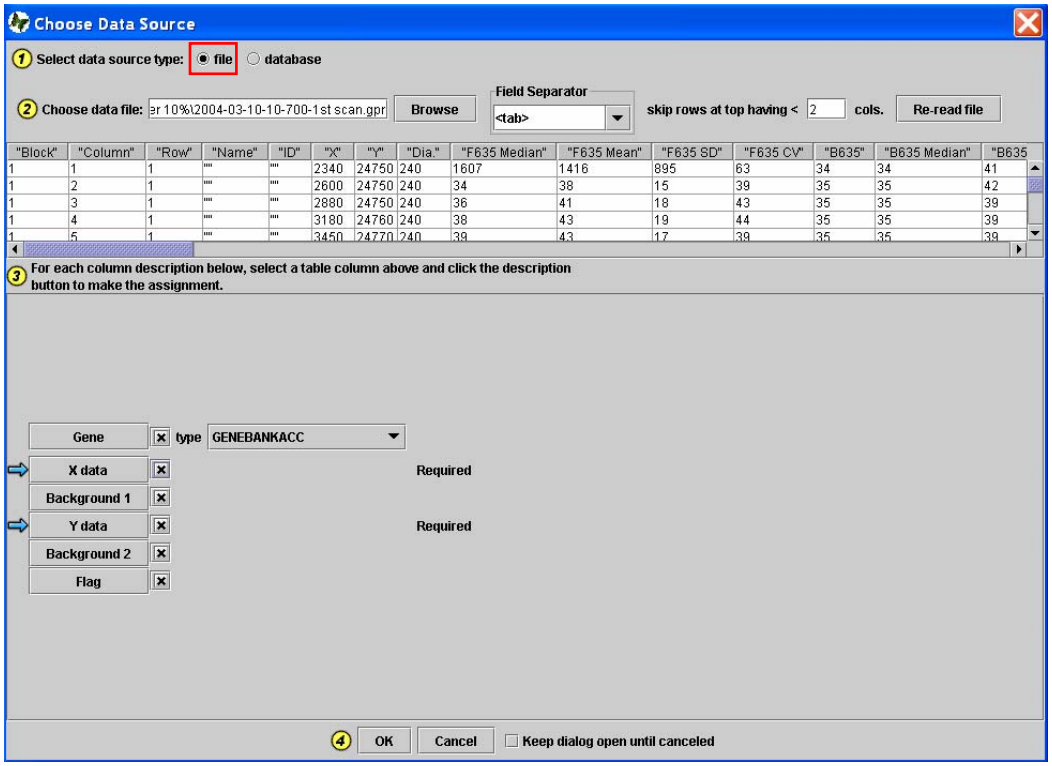
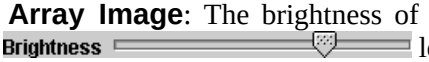


Figure 7-11: Inputting local data file for plotting.

Virtual Array Viewer

Overview: The Virtual Array Viewer (Figure 7-12) displays gene expression data derived from the original array image in a pseudo image format. It applies to both two-channel and one-channel microarray data. In Virtual Array Viewer, the arrangement of spots is exactly the same as in the original image (constructed from the Block/Row/Column description about array elements in the ArrayType Information File). The brightness represents the (average) intensity of the spot, and the color indicates the ratio (for two-channel system). The difference between Virtual Array Viewer and real array image is that some information (e.g. background and spot morphology) is lost in Virtual Array Viewer, but Virtual Array Viewer provides a faster and more convenient way of inspecting the quality and browsing the contents of an array, and looking for significant spots and information about their genes.

Some rudimentary information about the data set and the spots and their genes is displayed just above the array image and changes as the mouse moves over the spots (a white, squared box surrounds the current spot). This includes the data set name and sample descriptions for the two channels, the position of the spot under the cursor both in the manufacturer's coordinate system and in a regular row/column coordinate system, the Cy3 and Cy5 intensities of the spot and their ratio, the GenBank accession number for the gene on the spot, and the manufacturer's description of the gene on the spot. For more information about genes on a group of spots, the array image allows the user to mark spots and then display more in-depth gene information about the marked spots.

Adjust Brightness of Virtual Array Image: The brightness of the image can be changed by adjusting the Brightness bar 

located at the top-left of Figure 7-12.

Zoom In and Out of Virtual Array Image: This is done by clicking on the + or – sign

(Figure 7-12).

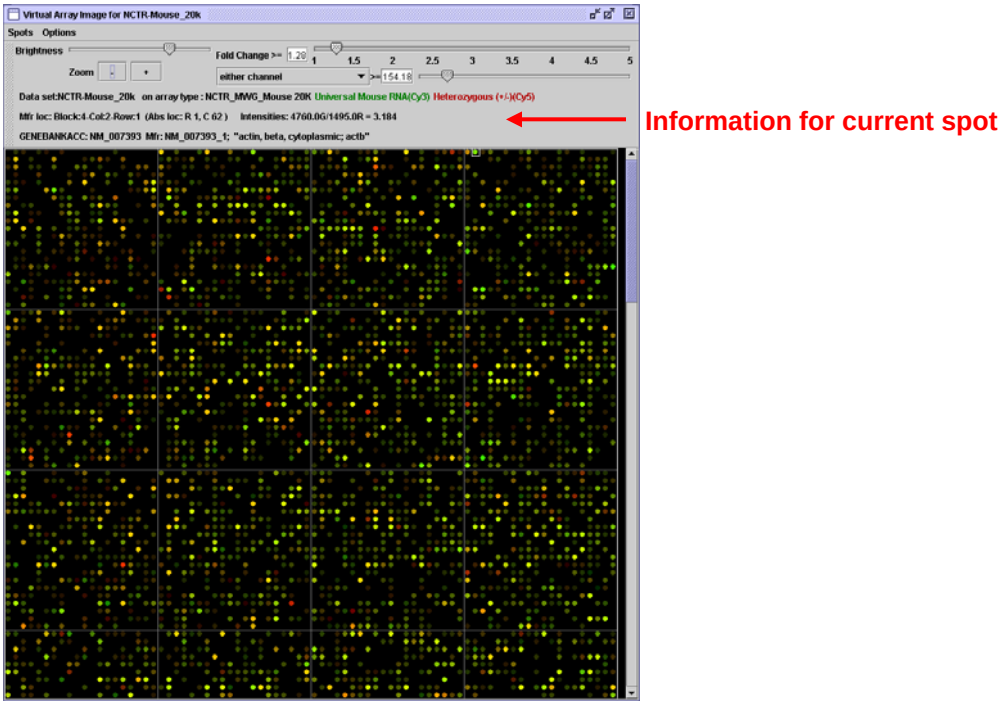


Figure 7-12: Virtual Array Viewer showing a pseudo microarray image reconstructed from fluorescence intensities of a two-channel array. The location of the current gene (*actb*) on the microarray slide is shown on top of the figure.

Filter Spots: There are two slider controls for filtering out unwanted spots(Figure 7-13).

Fold Change slider eliminates spots whose symmetric fold change is less than the chosen number (the precise value is displayed to the left of the slider). Here the symmetric fold change means the maximum of i1/i2 and i2/i1, where i1 and i2 are the intensities for the two channels. For example, positioning the slider at the value 1.3 means only spots will be shown such that one intensity is at least 1.3 times of the other, without regard to which channel is greater. Note: For one channel data, the fold change slider is not adjustable as no ratio data is available for one such array.

Intensity threshold filter can be used to eliminate spots whose intensities do not fulfill the criteria. There are three options for the intensity threshold filter: Either channel only displays those spots for which at least one of the two channel intensities is greater than the threshold value; Channel 1 and Channel 2 filter out spots for which the intensity for the Channel 1 and Channel 2, respectively, is below the threshold.

This is useful since spots with both intensities small might have intensity ratios that are too influenced by noise in the data to be useful. Note that both intensities must fall below the threshold for a spot to be eliminated. This choice was made since a spot with one intensity being small and the other large will have a symmetric ratio that may not be very accurate but is in any case certain to be large and may therefore still be worthy of consideration.

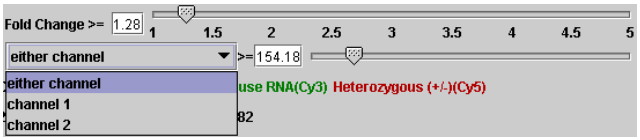


Figure 7-13: Fold Change and Intensity slide controls for filtering out spots for display.

Pull-Down Functions on Spots: There are five pull-down functions under Spots (Figure 7-14):

Mark Remaining Spots marks all spots that passed all filtering criteria and are displayed.

Mark Flagged Spots marks all the spots that are flagged.

Clear All Marks makes all spots as unmarked.

Search for genes opens a panel for the user to paste a list of GenBank accession numbers and mark the corresponding spots. The process is similar to that shown in Figure 7-3 for the Scatter Plot.

Marked Spots opens another list of functions applicable to the currently marked spots (see below on Functions Applicable to Marked Spots).

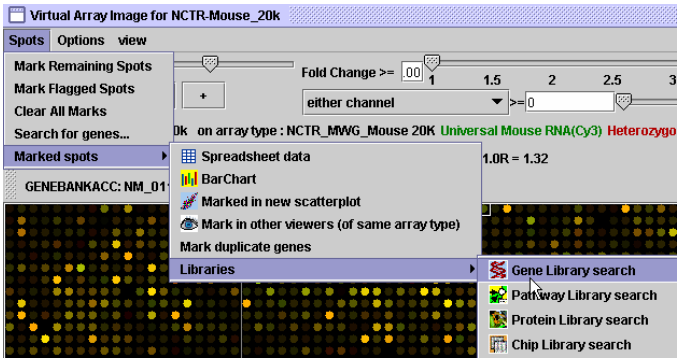


Figure 7-14: Pull-down functions for Spots and Marked Spots.

Pull-down Options: Several other options including flag and background handling are available (Figure 7-15A). The color of the virtual array image can also be set in four different ways (Figure 7-15C). By default, the Red/Green (Cy5/Cy3) ratio is used; however, the user can choose to display either channel intensity in a grey-scale, or to swap the default Red/Green color assignment. The user can also choose the style of marking – circle, crosshair or thick circle (Figure 7-15B) and the color of marking.

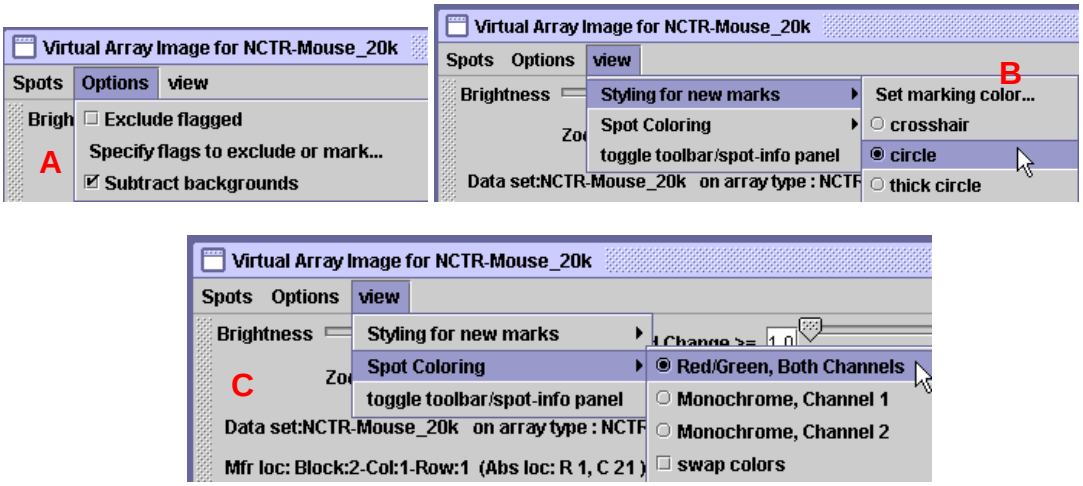


Figure 7-15: Options allow the setting of spot color by channel intensities.

Select and Deselect Current Spot: As the mouse cursor moves over the Virtual Array Viewer, the current spot is covered with a white, squared box (Figure 7-12) and can be selected/deselected by clicking on it.

Tip: If you want to perform any actions on the current spot, you must select (mark) it first.

Mark Spots: There are three ways the user can mark spots, which can be used in combination with each other.

First, the user can always mark a spot directly by clicking on it.

More usually, the user will mark multiple spots simultaneously by filtering out unwanted spots first using the slider controls, and then by either right-clicking on the Virtual Array Viewer or from the pull-down Spots menu to Mark Remaining Spots (Figure 7-14 and Figure 7-16).

The user can also use Search for genes to input the GenBank accession numbers and mark a set of interested genes. The interface is exactly the same as shown in Figure 7-3 for selecting a set of genes in the Scatter Plot. Gene marks can be cleared out by Clear Marks. And the Virtual Array Viewer can be saved as an image file by choosing Save as Image, just like in the Scatter Plot.

Right-click Accessible Functions: Right-click on the Virtual Array Viewer pops up a list of functions (Figure 7-16) that are the same as those accessible from the Spots pull-down menu (Figure 7-14).

Functions Applicable to Marked Spots: They are accessible from either right-click on the Virtual Array Viewer or from the Spots pull-down menu.

1) Spreadsheet Data: View gene expression data for the selected spots in a spreadsheet form. For details, see Chapter 8: Data Export.
2) Bar Chart: Launch cross-dataset gene Bar Chart for selected spots (for a maximum of five spots). For details, see the section on Bar Chart in this Chapter.
3) Mark in New Scatter Plot: Launch a new Scatter Plot and mark the selected spots. For details, see the section on Scatter Plot in this Chapter.
4) Mark in other Viewers: Select (mark) the same set of selected spots in all other viewers of the same array type.
5) Mark Duplicate Genes: If the current gene is spotted at multiple locations on the array, they will all be marked)
6) Libraries: Launch library search against Gene Library, Pathway Library, Protein Library or Chip Library using the GenBank accession numbers of selected spots as queries. See Chapter 3 for details on library search.

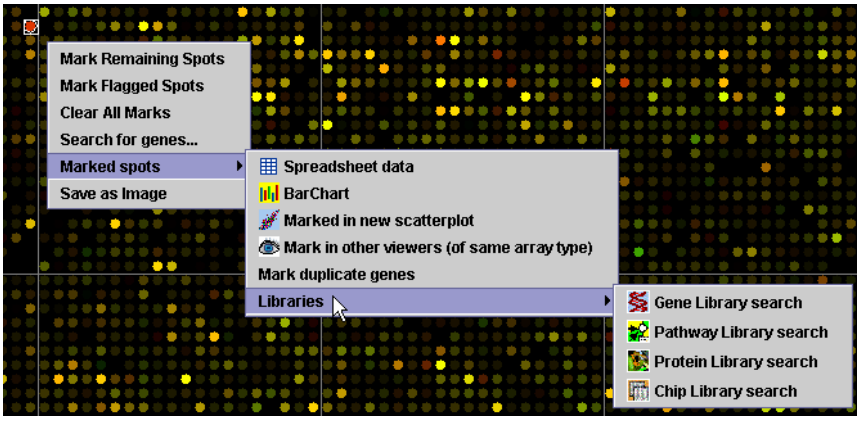


Figure 7-16: Right-click functions and those applicable to Marked Spots.

Actual Array Viewer

Actual Array Viewer (Figure 7-17) displays the original image from the scanner and loaded into ArrayTrack from the Data Import part (Figure 2-1) of Input Form.

Note: Actual Array Viewer has not been made available to the general user of ArrayTrack 3.1.

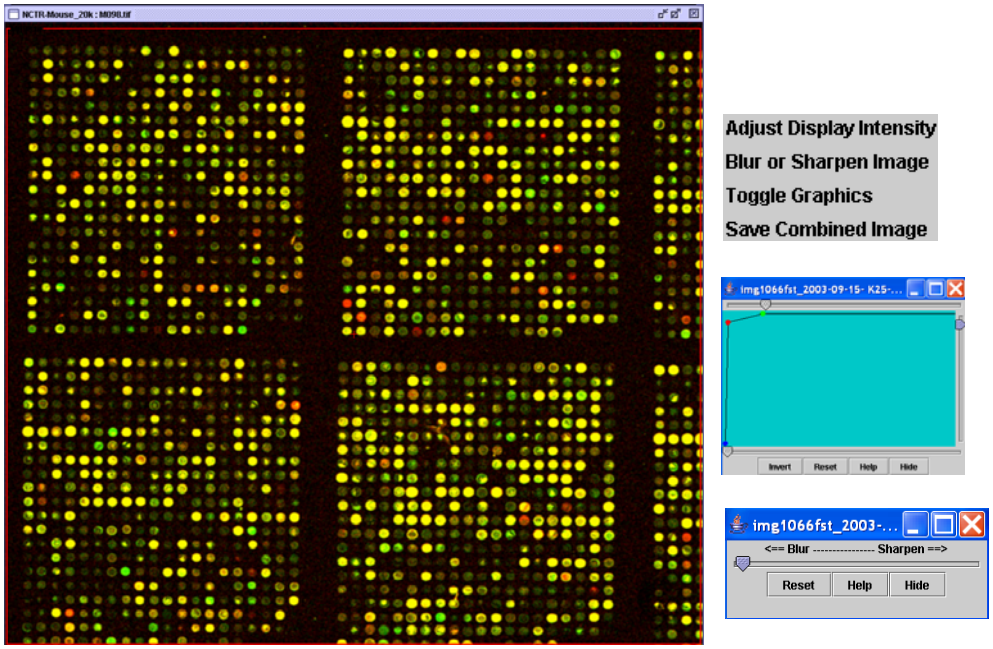


Figure 7-17: Actual Array Viewer displays the original microarray image.

Only a small part of the whole microarray image is shown in this figure. A list of operations for color adjustment pops up after right-click on the Actual Array Viewer.

Bar Chart

Overview: Bar Chart (Figure 7-18) displays expression data for a single gene across multiple arrays within the same experiment or across different experiments. It gives the user an overview of the differential expression levels of this gene across different samples. Bar Chart can be invoked from the TOOL/Visualization panel or pull-down menu, as well as from the Scatter Plot or Virtual Array Viewer on selected/current spots (see detailed discussions on these topics in this Chapter). When Bar Chart is invoked from the TOOL/Visualization panel or pull-down menu, the Gene Expression Bar Chart across Different Arrays panel pops up (Figure 7-18).

As mentioned earlier, the Bar Chart function can also be accessed from Gene Library, Protein Library, Pathway Library, Chip Library, IPI Library, Orthologene Library, T-test/ANOVA result table, Volcano plot, Cluster tree, database tree and the Significant Gene List table.

Query Specification: First, specify the type of gene ID (gene symbol, GenBank accession number, UniGene ID, Locus ID, and Manufacturer's gene ID). Second, enter a gene ID to be searched. Third, select one or more experiments for which expression data for the query gene is to be displayed. Fourth, select the type of data to be displayed (Raw Data vs. Normalized Data). If Normalized Data is to be displayed, further select the normalization method. Finally, click on Draw Barchart to show the Bar Chart.

Bar Chart Display: The X axis displays all the arrays within the selected experiment and the Y axis represents either ratio (for two-color platform) or intensity (for one-channel platform).

There are two tabs above the bar chart, titled: 1) gene name@ experiment name, 2) Standard Deviation (if the bars that representing the hybridizations have been assigned to several groups).

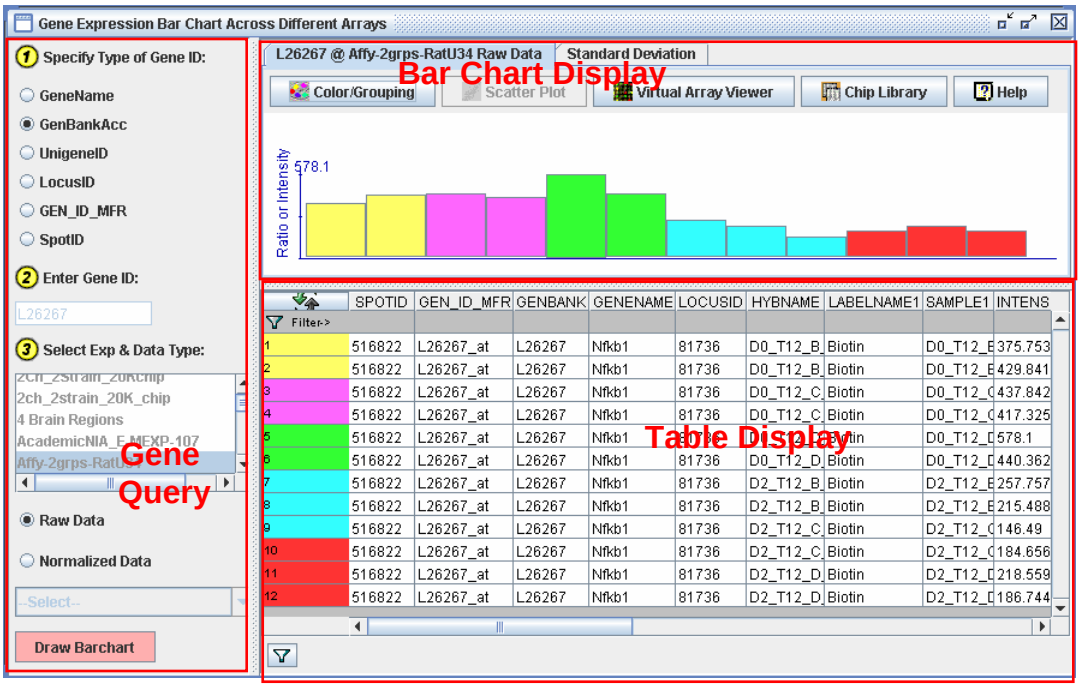


Figure 7-18: Arrangement of Bar Chart panel.

Toolbars: Under the first tab, user can see seven tool buttons (Figure 7-19):

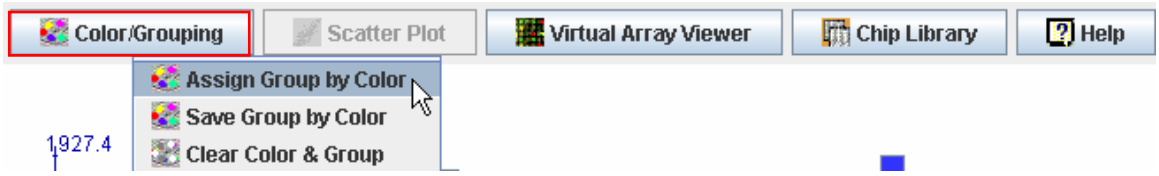


Figure 7-19: Toolbars for Bar Chart

Clear Color clears the current color assignment of the bars and re-displays them in the default color of light gray.

Color Chooser pops up a color chooser dialog box (Figure 7-21) from which the user can select a color to be applied to the following selections of bars (see Table Display and Bar Chart Display) until it is re-assigned. It is very useful for the user to group arrays within an experiment by using different colors.

Save Color saves the color combinations the user chose for different groups. Once the color info is saved, the Standard Deviation bar chart will be drawn and each bar will be marked in the corresponding color that is assigned for the group (see Figure 7-22). Also the records will be marked in the corresponding colors (see Figure 7-20). The way how the arrays are grouped also applies to all the other genes for that specific experiment. Every time the user re-groups the bars that representing multiple hybridizations, the Standard Deviation chart will be re-drawn. The grouping information is visible only for the user who did the grouping.

Scatter Plot launches Scatter Plot viewer for all the arrays currently being selected and color-coded the current gene in red for which the Bar Chart is displayed.

Virtual Array Viewer can be launched in which the gene is marked.
Chip Library can be launched and information about the current gene will be displayed.
Online help is available by clicking on Help.

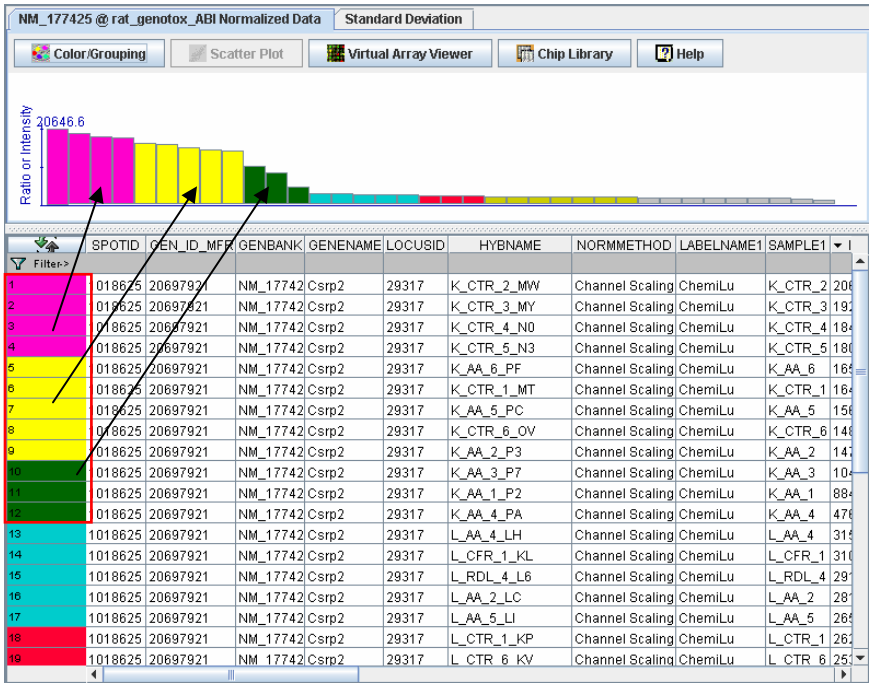


Figure 7-20: After saving the color group each record is marked in the corresponding color

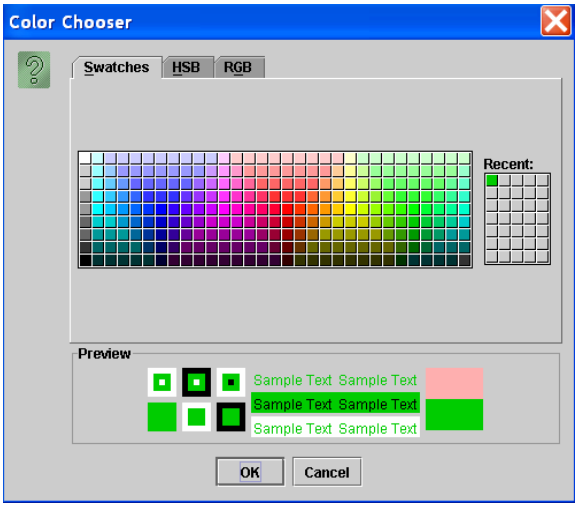


Figure 7-21: Color Chooser for assigning the chosen color as the current color.

If the user click the second tab titled "Standard Deviation", he will see the standard deviation bar chart in the corresponding colors that were chosen for different groups. Each colored bar represents the average intensity value for the corresponding group, while the height of the T-shaped line above the bar is the standard deviation value for the group. The y axis is the intensity. If the user put the mouse over each colored bar, the SD value will show up. If the user clicks a bar, the corresponding records will be highlighted in the spreadsheet below the bar chart.

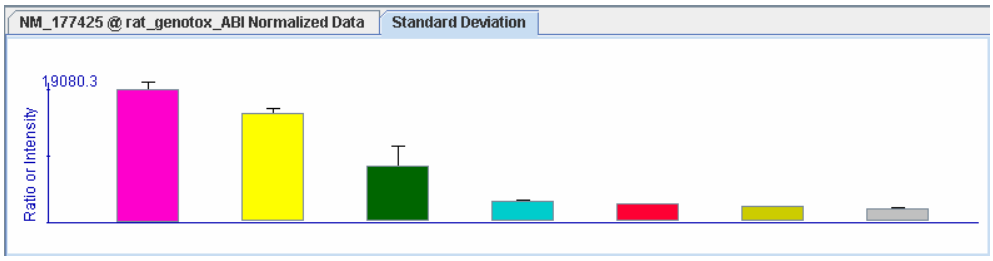


Figure 7-22: Standard Deviation chart

The standard deviation is calculated according to the following formula:

$$S_{N-1} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Table Display: The same results are displayed in a Table format below the Bar Chart. Like the Gene Library table, this Table can be queried. If the user assigned a color before querying, the bars corresponding to the arrays that pass the filtering query will be highlighted. The user can repeat this process to assign a different color to another group of arrays (e.g. samples from the treatment group). It is helpful to group arrays using different colors.

Alternatively, the user can first sort the table by the value of a particular column (e.g. the SAMPLE1 column in Figure 7-23 to separate control samples from treatment samples). Click-on and drag-down on table rows will change the color of the corresponding bars to the currently assigned color. In the example shown in Figure 7-23, control animals are colored in green; whereas valproic acid treated animal samples are colored in red. It is obvious that there is an increase of expression for this gene (NM_007812, Cyp2a5, "cytochrome P450, family 2, subfamily a, polypeptide 5") after valproic acid treatment.

Links between Bar Chart and Table: The Bar Chart and Table are linked together; clicking on a bar will highlight the corresponding row in the Table (and switch the color of the bar between the default color and currently assigned color). By clicking, holding down, and moving the mouse across the Bar Chart, multiple rows in the Table will be highlighted and the color of the selected bars will be assigned to the current color.

Similarly, clicking on a record (row) in the Table will blink the associated bar in the Bar Chart. Multiple rows can be selected by clicking and dragging-down the rows, and the color of their corresponding bars will be assigned as the currently assigned color.

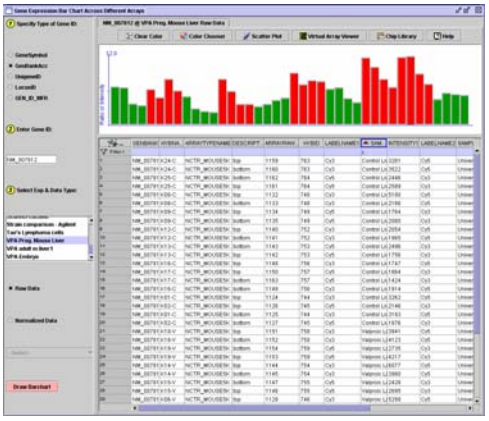


Figure 7-23: Bar Chart displays gene expression data for a gene across multiple arrays. Control samples are colored in green and chemical treated samples are colored in red.

Summary

As we have demonstrated in this Chapter, many visualization tools have been made available for examining microarray data. These functions are highly interconnected. Figure 7-24 is a screenshot of ArrayTrack in which many connected plots and tables are displayed.

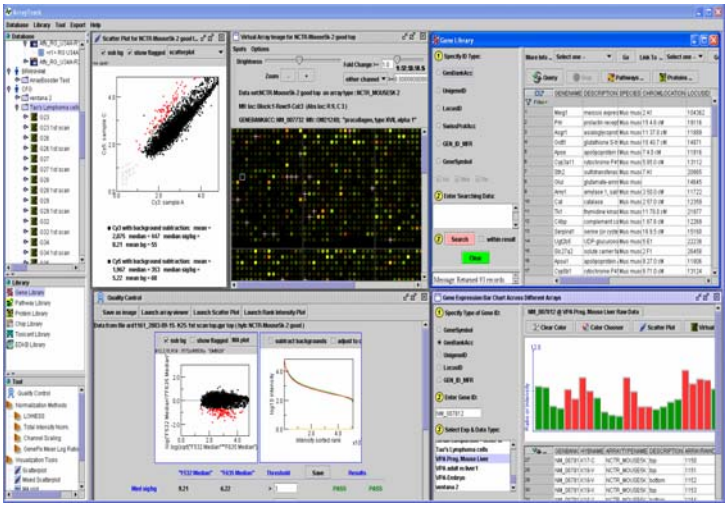


Figure 7-24: A screenshot showing multiple visualization functionalities within ArrayTrack.

Chapter 8 Working with Tools: DMVS

DMVS (Data mining and Visualization System) is a comprehensive software package for the analysis of high-dimensional data including DNA microarray gene expression data. DMVS was originally developed by Zhenqiang Su (zhqsu@chipscreen.com) of Chipscreen Biosciences, Ltd. (http://www.chipscreen.com). Arrangements have been made between NCTR and Chipscreen that allow the convenient data transfer between ArrayTrack and DMVS to achieve integration (Figure 8-1).

Under the NCTR-Chipscreen agreement, NCTR is authorized to distribute DMVS along with ArrayTrack to not-for-profit organizations. For users from the private sectors, please contact us for details on gaining access to DMVS.

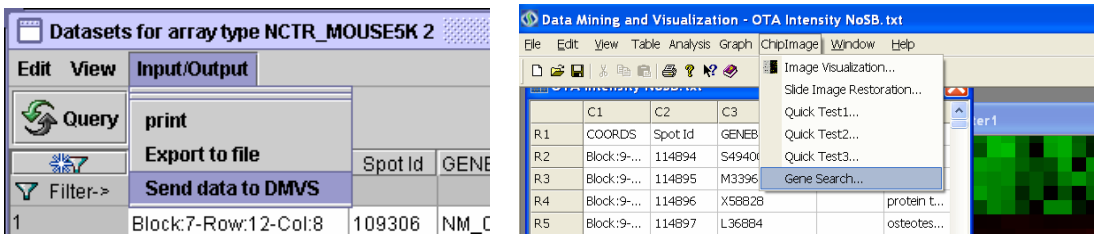


Figure 8-1: Data transfer between ArrayTrack and DMVS.

DMVS Version 2.0 includes the following functions that complement with ArrayTrack: principal component analysis (PCA), hierarchical clustering analysis (HCA), K-means clustering, self-organizing maps (SOM), support vector machine (SVM), partial least-squares regression (PLS), other functions for data analysis, visualization, and chemical structure/similarity handling (Figure 8-2).

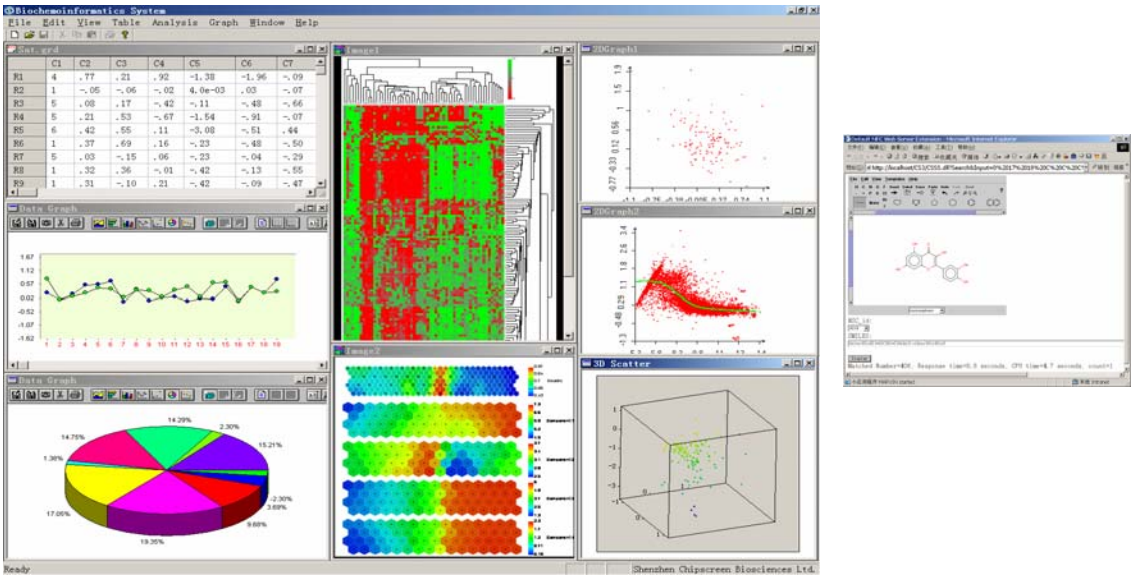


Figure 8-2: Additional TOOL functions available in ArrayTrack/DMVS.

Chapter 9 Data Export

How to Access Data Export Functions

Experimental data stored in MicroarrayDB can be conveniently exported to local data files or ArrayTrack spreadsheets for further analysis purposes. Data export can be accessed from the Export pull-down menu (Figure 9-1), or from right-clicking on selected arrays of the same arrays type (Figure 9-2), where data export options are shown on the top of the list of functions applicable to the selected arrays.

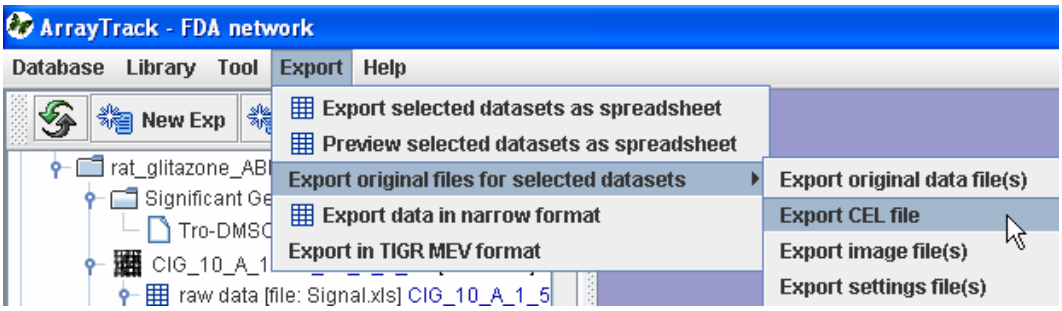


Figure 9-1: Accessing Data Export function from pull-down menu.

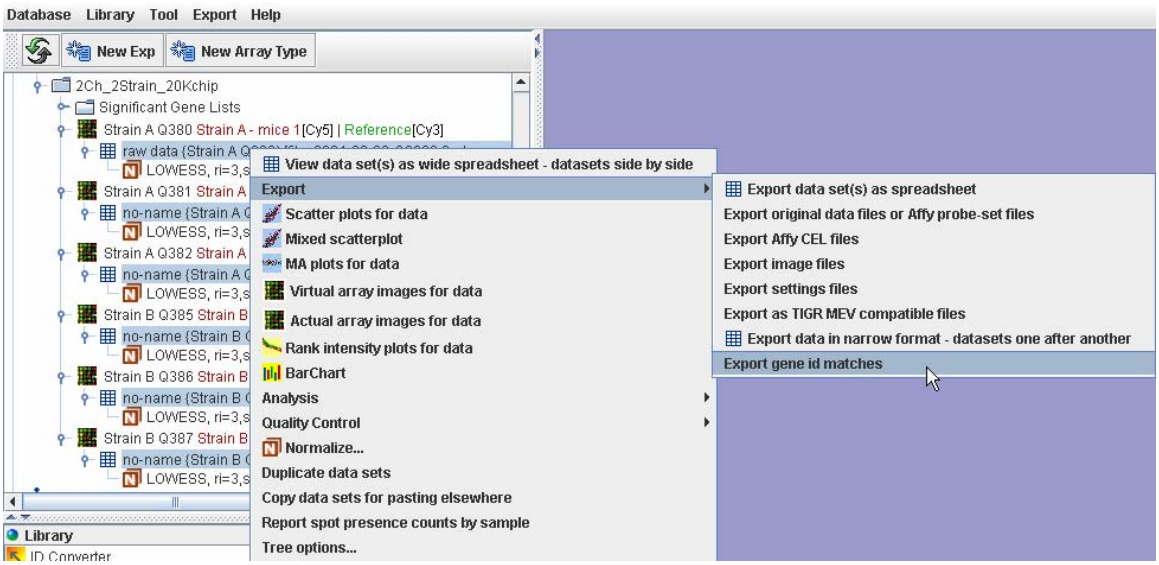


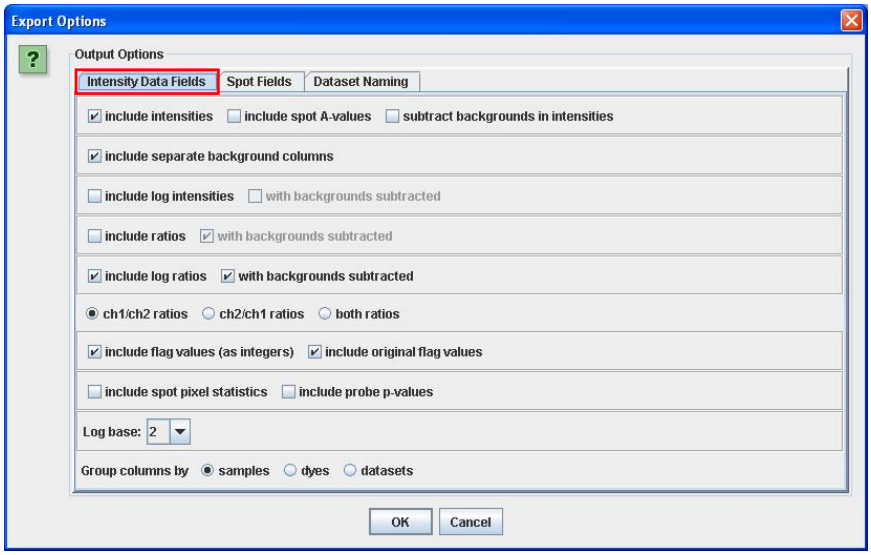
Figure 9-2: Accessing Data Export from right-clicking on selected arrays.

Options for Data Export

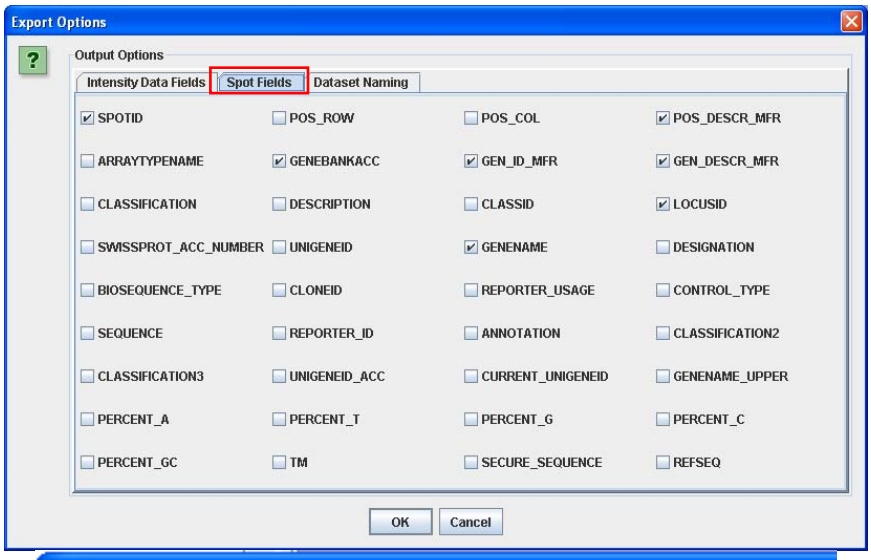
When exporting any data for spreadsheet view, the user will see the Export Options dialog box (Figure 9-3) which allows the user to define exactly what data items and formats to be exported from a variety of combinations by checking appropriate boxes.

These options are grouped into two categories: Intensity Data Fields and Spot Fields (i.e. annotation information about spots that is to be exported along with intensity data columns). These options are self-explaining and will not be discussed here. Options set here will remain effective for Data Export until re-set.

A



B



C

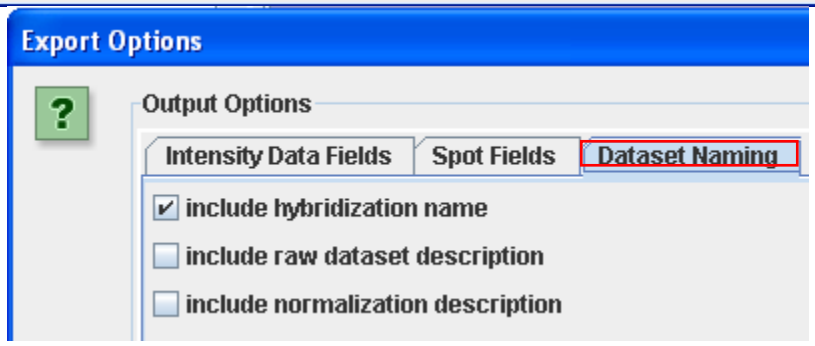


Figure 9-3: Setting Export Options before Data Export. A: Intensity Data Fields options; B: Spot Fields options; C: Dataset Naming options

Export selected datasets as spreadsheet

Data for selected arrays are to be saved as a local text file in which each row represents a spot (gene) and each column is a particular data item from an array. If multiple arrays are selected, additional columns are added in the spreadsheet. (This is the so-called "wide" format, in contrast to the "narrow" format – see below). The user will need to specify the folder and name where this file is to be saved. However, if the user has not selected any arrays before trying to use this option, a warning message will be displayed (Figure 9-4).

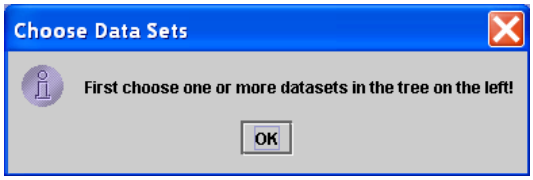


Figure 9-4: One or more arrays need to be selected before Export selected datasets as spreadsheet.

Preview selected datasets as spreadsheet

Instead of saving data as a local text file, Preview displays the data for selected arrays in a spreadsheet within ArrayTrack. If the user has not selected any arrays before trying to use this option, a warning message is displayed (Figure 9-4). Figure 9-5 is a sample spreadsheet view. Note that the column names encode all the information needed to unambiguously identify the identity of that column.

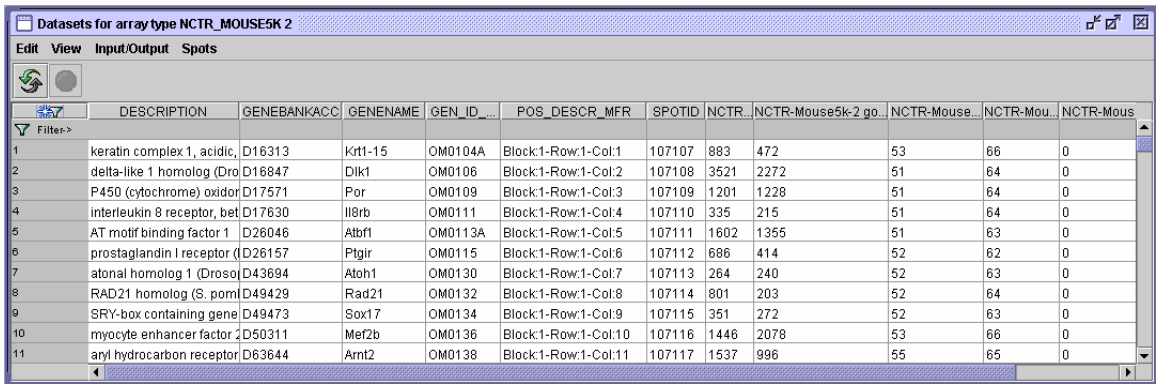


Figure 9-5: Preview of selected datasets as spreadsheet view.

Many functions are available from the Edit, View, Input/Output and Spots pull-down menus for the Spreadsheet view table (Figure 9-6). Under the Edit pull-down menu, the user can (1) Copy selected rows (along with column headers); (2) Copy selected column on selected rows (along with column headers), resulting in the copying of values from multiple table cells from the same selected column; (3) Add new filter row allows additional filter rows to be added on top of the table so that more flexible filtering/querying can be conducted (for more details, see discussion on Gene Library in Chapter 3). Alternatively, new filter rows can be added by clicking on. Furthermore, filter rows can be added or removed by clicking on the existing filter rows and typing or deleting the filter parameter.

Under the View menu, the user can Hide selected column; Unhide column (if any column is hidden); use Dittos Mode (") to display data items having the same value with its previous row; Limit number of records returned initially (current default is 10,000); Count total results (slower queries) fetch all record at once to get the record number; and Resize

cols to fit data reset the width of individual columns so that the columns can be displayed in a clean format, e.g. all column headers are fully displayed.

The Hide, Unhide, and Sort order functions are also accessible by right-click on a column header.

Under the Input/Out menu, the user can Print the current Spreadsheet, Export to file.

Under the Spots menu, the user can mark selected spots in a pre-opened viewer.

Any ongoing query session can be stopped by clicking on .

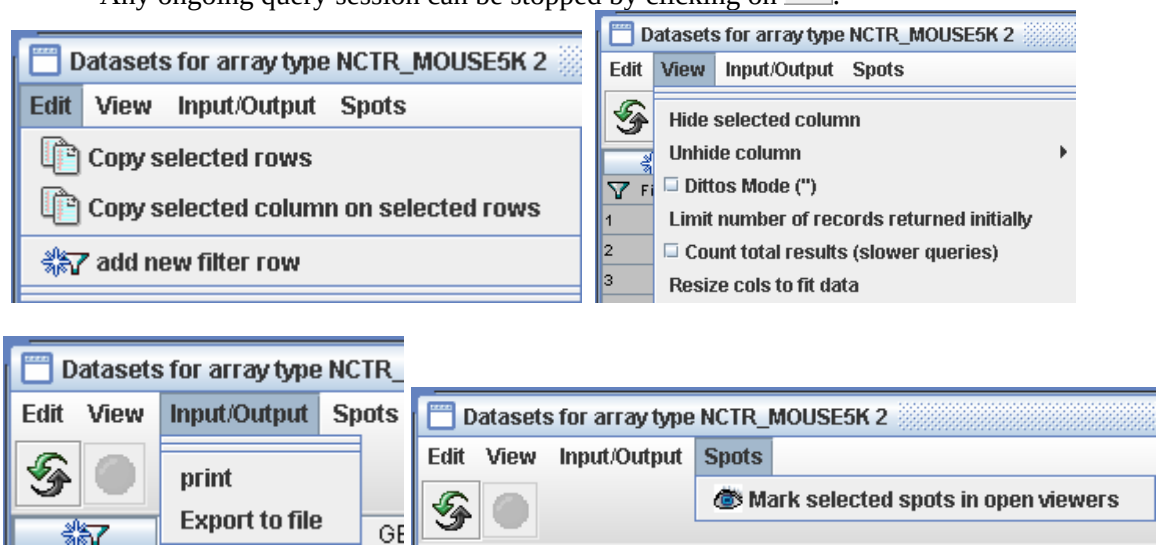


Figure 9-6: Edit, View, Input/Output and Spots pull-down menus for the Spreadsheet view table.

Export original files for selected datasets

Original data files (e.g. GPR files from Axon's GenePix software), CELL file, image file(s), and setting files loaded in MicroarrayDB can be exported in their original formats (Figure 9-1).

Export data in a narrow format

In the narrow format, each row represents a spot (gene) with columns corresponding to the selected fields to be exported. If multiple arrays are selected, additional rows are added in the spreadsheet. This narrow format has been designed for statisticians using SAS.

Export in TIGR MeV format

Exporting in TIGR MeV format (http://www.tigr.org/software/tm4/mev.html) is also supported. MeV is TIGR (The Institute for Genomic Research) Multi-experiment Viewer software.

Export Gene ID matches

If the user choose multiple data set (same or different array type), he can export the data with common genes by right-clicking the data set and choose Export ->Export gene id matches. See Figure 9-7.

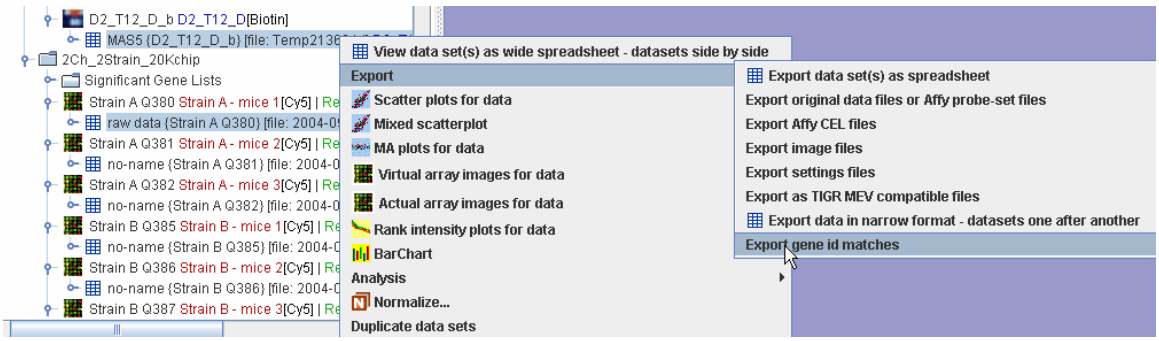


Figure 9-7: Export data with gene ID matches

Chapter 10 What to Expect from ArrayTrack 4.0

ArrayTrack is under continuous development at the Center of Toxicoinformatics of the NCTR/FDA. ArrayTrack contains three integrated components, database, libraries and tools. Retrospectively, the development of ArrayTrack is divided into three stages:

1. The version 1 series (ArrayTrack v1.*) is focused on the database component
2. The version 2 series (ArrayTrack v2.*) is focused on the library component.
3. The version 3 series (ArrayTrack v3.*) is focused on the tool component.

Although there will be many small changes along the way as ArrayTrack development proceeds, the following majors changes are anticipated in the forthcoming releases:

- We will have several minor releases of version 3 in 2005 that include additional analysis tools, such as SAM, F-test and so forth.

- The next major release of ArrayTrack version 4 is scheduled for next year (around April 2006); it will provide all essential functionalities for data analysis and interpretation. In addition, a more flexible data exchange mechanism will be available, by which users can easily exchange data between ArrayTrack and other public microarray data repositories (e.g., GEO, ArrayExpress, CEBS).

- Currently, we only allow the FDA data to be stored in ArrayTrack for an obvious reason, lack of resources. However, we will start to permit some non-FDA users to deposit their data into ArrayTrack in accordance with available resources.

- For the local installation version, ArrayTrack has to be run in the ORACLE environment. Towards the end of 2006 (version 5), we plan the release of an ArrayTrack version that will utilize freeware databases (e.g. Postgres or MySQL) as a backend instead of ORACLE.

References

1. Tong W, Cao X, Harris S, Sun H, Fang H, Fuscoe J, Harris A, Hong H, Xie Q, Perkins R, Shi L and Casciano D (2003) ArrayTrack--supporting toxicogenomic research at the U.S. Food and Drug Administration National Center for Toxicological Research. *Environ Health Perspect* **111**:1819-26.
2. Tong W, Harris S, Cao X, Fang H, Shi L, Sun H, Fuscoe J, Harris A, Hong H, Xie Q, Perkins R, Casciano D (2004) Development of Public Toxicoinformatics Software for Microarray Data Management and Analysis. *Mutation Research* (in press)
3. Tong W, Hong H, Fang H, Xie Q and Perkins R (2003) Decision forest: combining the predictions of multiple independent decision tree models. *J Chem Inf Comput Sci* **43**:525-31.

ArrayTrack Team

The following people (in alphabetical order) have made direct contributions to the development of ArrayTrack:

Cao, Xiaoxi (Megan)
Fang, Hong
Harris, Stephen C.
Hong, Huixiao
Perkins, Roger
Qian, Feng
Shi, Leming
Sun, Hongmei
Tong, Weida
Wu, Jie
Xie, Qian

Acknowledgments

Many colleagues at the FDA's National Center for Toxicological Research have been supportive during the development of ArrayTrack. The Center for Toxicoinformatics is particularly appreciative of the full support of Dr. Dan Casciano, Director of NCTR/FDA and Dr. Ralph Kodell, Director of the Division of Biometry and Risk Assessment, NCTR/FDA.

Many early users of ArrayTrack around the world have provided values suggestions on the update of ArrayTrack. We appreciate their continuing interests in and support of ArrayTrack.

The ArrayTrack team is looking forward to your input for the next release. Thank you!

Contact Us

For further information on ArrayTrack and/or the Center of Toxicoinformatics of the NCTR/FDA, please contact:

Weida Tong, Ph.D.
Director, Center for Toxicoinformatics
National Center for Toxicological Research
Food and Drug Administration
3900 NCTR Road
Jefferson, Arkansas 72079
U.S.A.

Tel: +1-870-543-7142, +1-870-543-7387
Fax: +1-870-543-7662
E-mail: NCTRArrayTrackSupport@NCTR.FDA.GOV
Weida.Tong@fda.hhs.gov, Leming.Shi@fda.hhs.gov
http://edkb.fda.gov/webstart/arraytrack/