

**IMPORTANT!**

With the release of ArrayTrack 3.5.0 this manual is slightly out-of-date. It will be updated shortly, but until then please see the QuickStart Manual

<http://edkb.fda.gov/webstart/arraytrack/Tutorials/AT3.5QuickManual.pdf>

and the tutorials

<http://www.fda.gov/ScienceResearch/BioinformaticsTools/Arraytrack/ucm082498.htm>



# User's Manual

Version 3.4.0



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

# ArrayTrack 3.4

## 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 (NCTR)  
U.S. Food and Drug Administration (FDA)  
3900 NCTR Road  
Jefferson, Arkansas 72079  
U.S.A.  
Tel: +1-870-543-7142  
+1-870-543-7538  
Fax: +1-870-543-7662  
Support Team: [NCTRBioinformaticsSupport@nctr.fda.gov](mailto:NCTRBioinformaticsSupport@nctr.fda.gov)

Home page:

<http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/index.htm>

## Notice

ArrayTrack software is constantly evolving in terms of features, improvements, and inevitable bug fixes. This manual is for use with ArrayTrack major release version 3.4.0. While we strive for consistency between the manual and version 3.4.0, you might observe some slight differences. For updated information, please check the ArrayTrack web site at:

<http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/index.htm>

## Important 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 combination of 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. Each column can be sorted by clicking the column title.
5. Tutorials are available from our web site:

<http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/tutorials.htm>

## Disclaimer

ArrayTrack is the software developed by National Center for Toxicological Research (NCTR) of FDA. All the rights are reserved. The on-line version of ArrayTrack can be accessed via

<http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/index.htm> and is free of charge to use. To install ArrayTrack locally on your machine, you can send your request to us and we will send ArrayTrack CD to you for free. Though ArrayTrack is free of charge, the source code is not open currently. All the rights are reserved.

## Table of Contents

Chapter 1 Overview .....	7
1 The best way to use this manual - Read this chapter first.....	7
1.1 One minute on ArrayTrack .....	7
1.2 Everything before, during and after .....	8
1.2.1 How to get support.....	9
1.2.2 How to get the updated information on ArrayTrack.....	9
1.2.3 ArrayTrack history.....	9
1.2.4 Releases and upgrades .....	10
1.2.5 Frequently asked questions .....	10
1.3 What is new in this version? .....	10
Chapter 2 Working with the Database: MicroarrayDB .....	12
2.1.1 Overview.....	12
2.1.2 Create experiment .....	12
2.1.3 Delete Experiment .....	13
2.2 Data Import .....	14
2.2.1 Batch import regular data.....	14
2.2.2 Batch import SimpleTox data .....	18
2.2.3 View SimpleTox data .....	20
2.2.4 Delete SimpleTox data.....	22
2.2.5 Update Hybridization information.....	23
2.3 Create Array Type.....	24
2.3.1 Overview .....	24
2.3.2 Activate Array Type Information .....	25
2.4 Data Sharing and Security Protection.....	33
2.5 Exploring and Viewing Data in MicroarrayDB from Tree View .....	34
Chapter 3 Gene List .....	37
3.1 Overview.....	37
3.2 Create Gene List .....	37
3.3 Display Gene List .....	38
3.4 Import Gene List.....	42
3.5 Export Gene List.....	43
3.6 Delete Gene List .....	43
Chapter 4 Working with Libraries .....	44
4.1 Overview.....	44
4.2 Gene Library .....	45
4.3 Pathway Library.....	50
4.5 Protein Library .....	56
4.6 IPI Library.....	57
4.7 Orthologene Library.....	57
4.8 GOFFA Library .....	58
4.9 Chip Library.....	63
4.10 Toxicant Library .....	66
4.11 EDKB Library.....	68

4.12 ID Converter .....	70
Chapter 5 Working with Tools: Quality Control .....	72
5.1 Overview of TOOL.....	72
5.2 Overview of Quality Control .....	73
5.3 Launch of Quality Control.....	73
5.4 Contents of Quality Control View .....	74
5.5 Overview of Quality Filtering.....	76
5.5.1 Launch of Quality Filtering .....	77
5.5.2 Contents of Quality Filtering .....	77
Chapter 6 Working with Tools: Normalization .....	79
6.1 Overview.....	79
6.2 Lowess .....	80
6.3 Total Intensity Normalization .....	83
6.4 Mean/Median Scaling .....	83
6.5 GenePix Mean Log Ratio Normalization .....	84
6.6 Linear and Lowess .....	84
6.7 Quantile Normalization.....	85
6.8 Reference Average Comparison Normalization .....	86
Warning.....	87
Chapter 7 Working with Tools: Analysis Tools .....	88
7.1 Overview.....	88
7.2 T-test and ANOVA .....	88
7.3 P-value Plot.....	95
7.4 Clustering.....	97
7.5 PCA.....	99
7.6. Correlation Matrix .....	101
7.7. SAM.....	103
7.8 K-Means.....	117
Chapter 8 Working with Tools: Visualization .....	121
8.1 Overview.....	121
8.2 Scatter Plot.....	122
8.3 MA Plot.....	124
8.4 Mixed Scatter Plot.....	125
8.5 Rank Intensity Plot.....	126
8.6 Choosing Data Source for Plotting .....	126
8.7 Virtual Array Viewer .....	127
8.8 Actual Array Viewer.....	130
8.9 Bar Chart.....	131
8.10 VennDiagram.....	135
8.10.1 Draw VennDiagram by common ID.....	135
8.10.2 Draw VennDiagram by KEGG/PATHART pathway.....	142
8.10.3 Draw VennDiagram by GeneOntology .....	145
Summary .....	145
Chapter 9 Working with other Tools .....	146
9.1 Join table (file) .....	146
9.2 Split table (files).....	147

9.3 Get Unique ID.....	148
9.4 Convert CEL file to probe set files .....	149
Chapter 10 Data Export .....	150
10.1 How to Access Data Export Functions .....	150
10.2 Options for Data Export.....	150
10.3 Export selected datasets as spreadsheet .....	151
10.4 Export original data files or Affy probe-set files.....	152
10.5 Export Affy CEL files.....	153
10.6 Export cross-platform data.....	153
10.7 Export image files and settings files .....	154
10.8 Export data in a narrow format .....	154
10.9 Export to JMP/Genomics .....	154
10.10 Export to DrugMatrix .....	155
Center for Toxicoinformatics of the NCTR/FDA.....	156
Toxicoinformatics Integrated System (TIS) .....	156
References.....	158
ArrayTrack Team.....	159
Acknowledgments.....	159

## Chapter 1 Overview

### 1 The best way to use this manual - *Read this chapter first*

#### 1.1 One minute on ArrayTrack

##### 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.

The system is based on a DB-TOOL-LIB integrated structure. (1) The DB is a central data archive for data storage and management; (2) LIB, a set of libraries that contain data both from online databases as well as NCTR in-house 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 1-1 and more detailed descriptions of the three components follows.

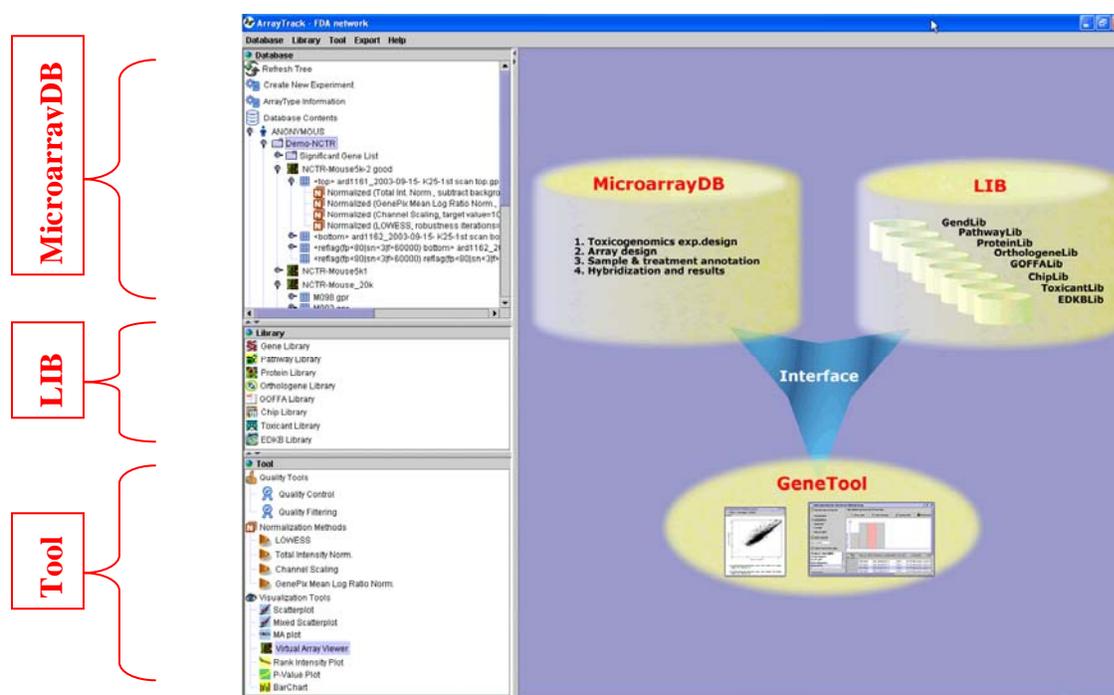


Figure 1-1: ArrayTrack's three main components: MicroarrayDB, LIB, and TOOL.

**MicroarrayDB:** MicroarrayDB (DB) is part of ArrayTrack's ORACLE-based relational database that stores microarray experimental data, experiment along with parameters, clinical and non-clinical data.

**LIB:** LIB is part of ArrayTrack's ORACLE-based relational database that mirrors the most essential information in public databases related to genes, proteins, pathways, toxicants, and as well as detailed data for various microarray systems (chips). Most of the data are from NCBI, others include SWISS-Prot, KEGG, IPI. All this information is not only cross-linked, but most importantly integrated to

provide a better visualization and presentation. The libraries are 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 4. 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 and clustering. 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 are 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.

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 saved in database and 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.

## 1.2 Everything before, during and after

### 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 (FDA intranet) 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.

ArrayTrack Quick manual can be downloaded from web site:

<http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/index.htm>

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

### Running Online Version of ArrayTrack

You can also use this link (<http://edkb.fda.gov/webstart/arraytrack/>) to run ArrayTrack, or let ArrayTrack place icons (Figure 1-3) 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. 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 for local installation, which will greatly increase speed.

When you activate the ArrayTrack from outside of FDA firewall, you will see the login window (see Figure 1-2). If you don't have an account you can just click the cancel button or leave the fields blank then click OK button to login and view the demo data. Non-FDA users without an account can not import

their own data into ArrayTrack. We will create account for users if they send request to us ([NCTRBioinformaticsSupport@fda.hhs.gov](mailto:NCTRBioinformaticsSupport@fda.hhs.gov)).

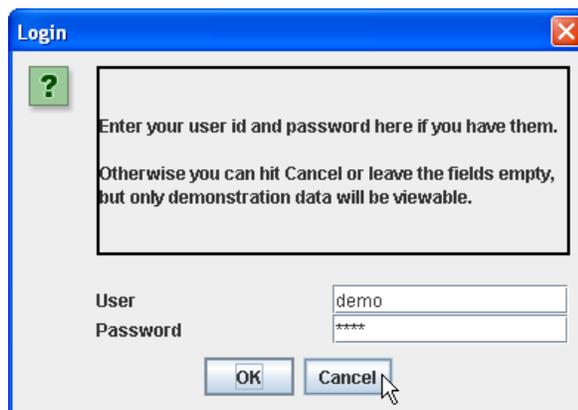


Figure 1-2: Login to ArrayTrack

If you are FDA users, you can activate ArrayTrack within FDA firewall (<http://weblaunch.nctr.fda.gov/jnlp/arraytrack/index.html>).

### Running Local Version of ArrayTrack

If you request CDs 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 10g or higher installed.

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



Figure 1-3: The ArrayTrack icon appears on the desktop of a computer after installation.

### 1.2.1 How to get support

Send your questions or request to: [NCTRBioinformaticsSupport@fda.hhs.gov](mailto:NCTRBioinformaticsSupport@fda.hhs.gov)

### 1.2.2 How to get the updated information on ArrayTrack

The user can download the user's manual from our website:

<http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/index.htm>

The quick help manual give the user a quick overview while the help manual have the detail description about the software.

### 1.2.3 ArrayTrack history

- AT version 1 (2001)
  - Filter array; data management tool;
- AT version 2 (2002): in-house microarray core facility
  - Customized two color arrays; data management, analysis and interpretation;

- Open to public (late of 2003)
- AT version 3.1 (2004): VGDS
  - Affymetrix; analysis capability enhanced;
- AT version 3.2 (2005): MAQC (Microarray Quality Control project)
  - Tested on 7 commercial platforms (Affymetrix, Agilent one- and two-color arrays, ABI, CodeLink, Illumina ...)
  - Integrated with other software (IPA, MetaCore, DrugMatrix, CEBS, SAS/JMP ...)
- AT version 3.3-3.4 (2006 – present)
  - CDISC(Clinical Data Interchange Standards Consortium)/SEND(Standard for Exchange of Nonclinical Data) standard
  - VGDS (Voluntary Genomics Data Submission)→ VXDS

### 1.2.4 Releases and upgrades

Version	Releasing Date
3.1.0 CD	3/1/2005
3.1.4 CD	6/14/2005
3.2 CD	3/8/2006
3.3 CD	1/16/2007
Patch 3.3	2/1/2007
3.4 CD	6/30/2008
Web 3.4	1/30/2008
Patch 3.4	8/30/2008

### 1.2.5 Frequently asked questions

The frequently asked questions are listed at our website:

<http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/arraytrackfaq.htm>

This webpage covers questions such as how to access ArrayTrack, ArrayTrack's system architecture, ArrayTrack's security levels, R server/Bioconductor, etc. This is a quick way to get answer for those common questions.

### 1.3 What is new in this version?

- The ArrayTrack data batch import wizard provides a new function called SimpleTox Format. The SimpleTox input format makes data submission convenient and intuitive, particularly for the toxicogenomics' study data. SimpleTox utilizes controlled vocabulary from MIAME and SEND (Standard for Exchange of Nonclinical Data, <http://www.fda.gov/oc/datacouncil/cdisc.html>) to permit study data (e.g., animal data and toxicological endpoints) to be input together with corresponding array data. The study data is managed in the Study Domain for view and query and, ultimately, is linked to array data for phenotypically anchored analysis.
- The popular SAM-Test tool is available. ArrayTrack's SAM is in R version 2.5.1. There are variations of test types, includes various analysis types like one-class, two-class unpaired/paired, one-class time course, two-class unpaired/paired time course, multiclass, survival, etc.
- Several other new data analysis tools, such as K-mean clustering, and two-way ANOVA are available.
- Common pathways and/or GO terms shared by up to three lists of analytes can be determined. The analytes can be from the same experimental platform, such as genes from DNA microarrays. Alternatively, analytes can come from different technologies such as genomics, proteomics and metabolomics, thus expanding ArrayTrack's utility for systems biology.

- Genes involved in a pathway or GO term can be readily saved for subsequent data analysis and modeling. This function is particularly useful if further analysis will focus on biological phenomena known to be associated with these genes, as would be the case when evaluating the biological plausibility of a particular molecular fingerprint.

## Chapter 2 Working with the 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 two-channel (e.g. Agilent, 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.

### 2.1.1 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.

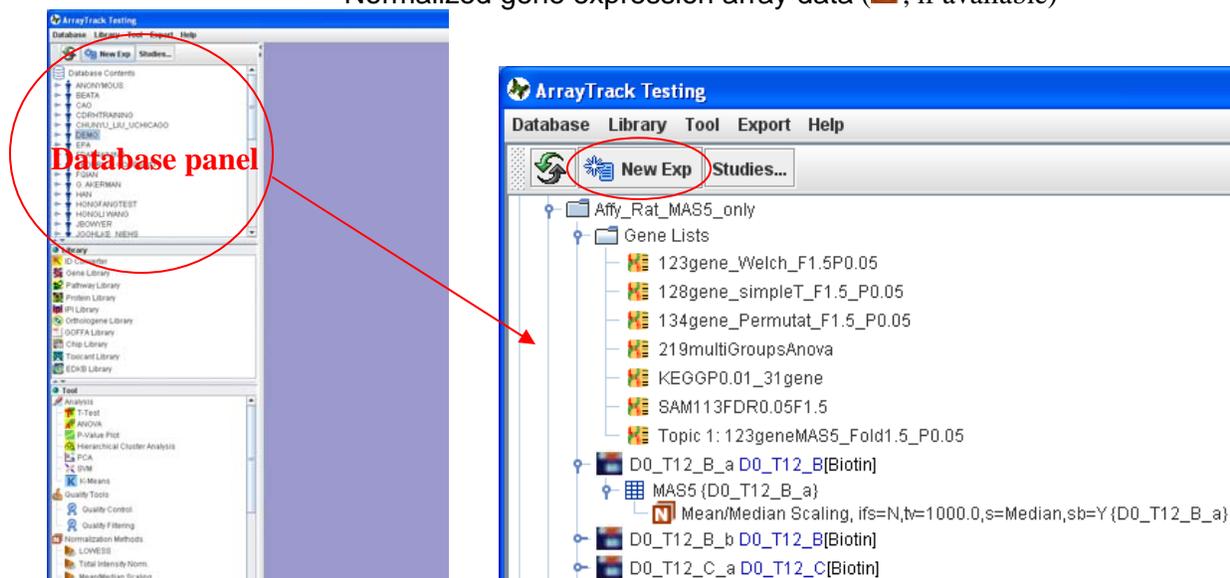


Figure 2-1: Data structure in MicroarrayDB.

### 2.1.2 Create experiment

Under the Database window panel of functions (Figure 2-1), the user can select the  **New Exp** button to create a new experiment. The user is prompted to specify a unique title for the experiment and user group name (owner), as shown in Figure 2-2.

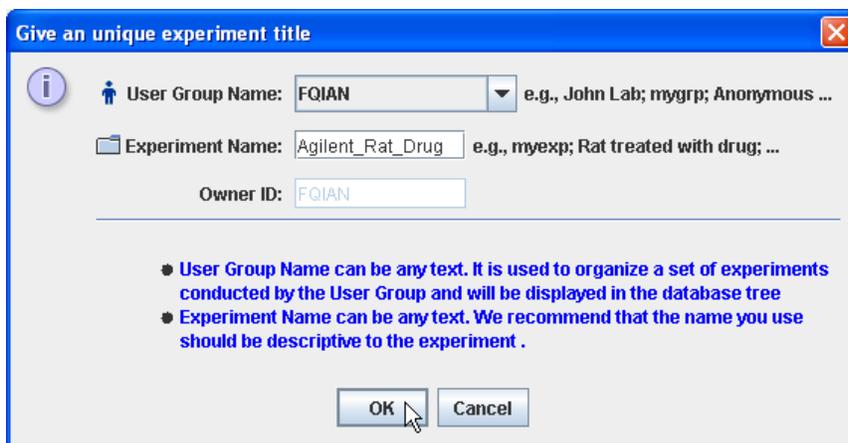


Figure 2-2: The user needs to specify an Experiment ID and Experiment Group Name for the new experiment.

The Owner ID will automatically show the name of the user who is logged into the Windows system, and is grayed out. After the owner created an experiment, the new experiment will be shown under the user group name in the database tree on the left panel.

### 2.1.3 Delete Experiment

To delete an experiment, double-click the experiment name in the database panel to bring out the Input Form. An existing Experiment can be deleted by clicking on Delete Exp.

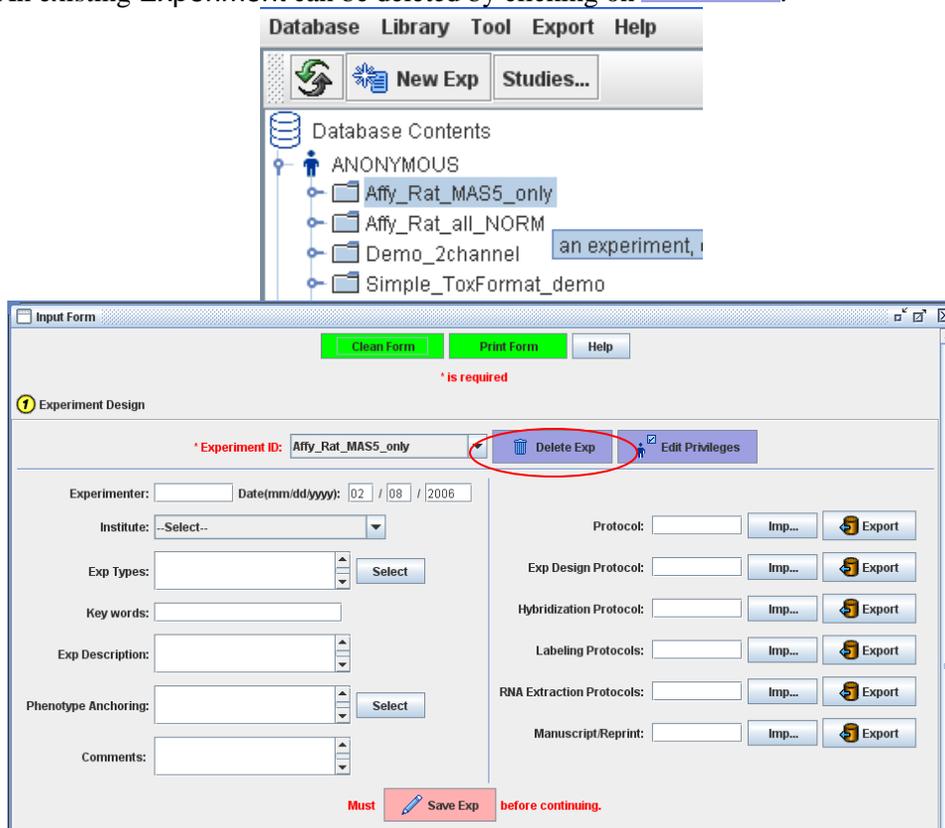


Figure 2-3: Input Form – delete experiment

## 2.2 Data Import

If the user needs to import multiple arrays (e.g. 2, 10, or 50 hybridizations) of the same array type, s/he can use the “Batch import” function. There are two formats for batch import: 1) regular batch import, 2) SimpleTox format. SimpleTox format is designed for animal toxicology study data submission, with combined and extracted main fields from SEND (Standard for Exchange of Nonclinical Data). However, it is not limited to animal study. The data can be customized for clinical data storage. These two types of batch imported will be addressed later.

### Hybridization file

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), ArrayTrack provides some example files under the Help pull-down menu, see Figure 2-4. In hybridization file, the hybridization name must be unique. The column for file name lists the data files that associated with each hybridization. The sample name can be duplicated. The hybridization file must have these three columns at least. The other columns are optional depending how much info the user wants to put in.

	A	B	C	D	E	F
1	HybName	Cy3 Sample	Cy5 Sample	Cy3 Organ	Cy5 Organ	FileLocation
2	A1	SUR1	CUR1	Universal	Universal	testAA1.txt
3	A2	SUR2	CUR2	Universal	Universal	testAA2.txt
4	A3	SUR3	CUR3	Universal	Universal	testAA3.txt
5	A4	SUR4	CUR4	Universal	Universal	testAA4.txt
6	A5	SUR5	CUR5	Universal	Universal	testAA5.txt
7	B1	SUR6	AB1	Universal	Brain	testBB1.txt
8	B2	SUR7	AB2	Universal	Brain	testBB2.txt
9	B3	SUR8	AB3	Universal	Brain	testBB3.txt
10	B4	SUR9	AB4	Universal	Brain	testBB4.txt
11	B5	SUR10	AB5	Universal	Brain	testBB5.txt
12	C1	SUR11	AL1	Universal	Liver	testCC1.txt
13	C2	SUR12	AL2	Universal	Liver	testCC2.txt
14	C3	SUR13	AL3	Universal	Liver	testCC3.txt
15	C4	SUR14	AL4	Universal	Liver	testCC4.txt
16	C5	SUR15	AL5	Universal	Liver	testCC5.txt
17	D1	CUR1	AB6	Universal	Brain	testDD1.txt
18	D2	CUR2	AB7	Universal	Brain	testDD2.txt
19	D3	CUR3	AB8	Universal	Brain	testDD3.txt
20	D4	CUR4	AB9	Universal	Brain	testDD4.txt
21	D5	CUR5	AB10	Universal	Brain	testDD5.txt

Figure 2-4: Hybridization file for regular batch import and example files

### 2.2.1 Batch import regular data

The following is an example of batch importing 20 hybridizations (2 channel data).

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 2-5. After choosing “Batch Import”, the “Batch Import” window shows up (Figure 2-6).

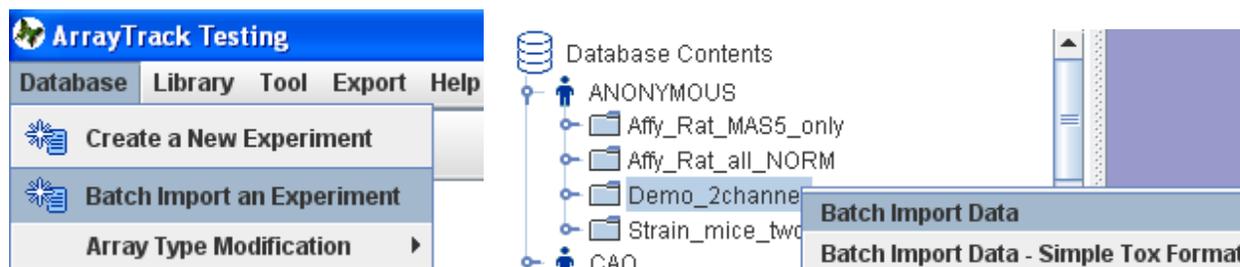


Figure 2-5: Activate “Batch Import” function

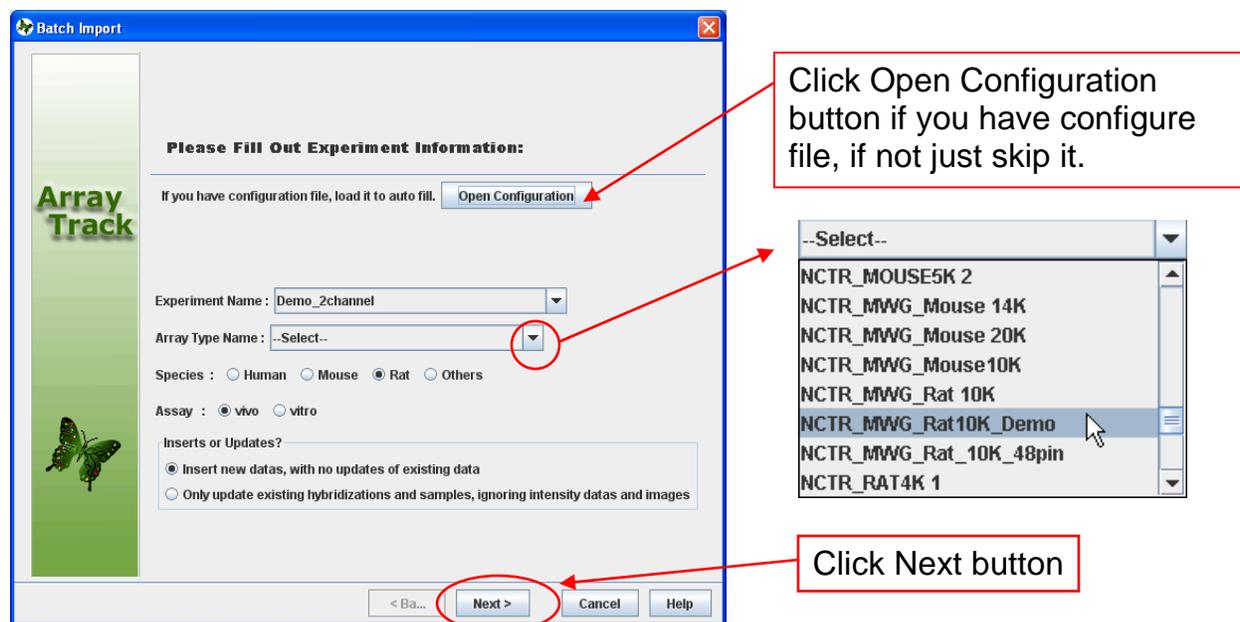


Figure 2-6: Batch import interface

In Figure 2-6, the user can select array type name, choose species, and assay.

**Step 1:** Right now the user can ignore the button “Open Configuration” which will be explained later. Make sure the experiment name is correct. Select the right array type from the pull-down list. If you can’t find the array type from the list, see page 23 about array type information. Choose the right specie, if the specie is other than human, mouse or rat then select “Others”. Select the right assay.

There is an option at the button for inserting new data or updating. The default is “Insert new data, with no updates of existing data” which applies to most cases. The second option is for updating hybridization information without changing data intensity value. For example, after finishing batch import, the user may realize that some sample information is missing. In such a situation, the user can add a sample information column in the hybridization file, and then do the batch-import again using the updating option. Users can add multiple columns but they must make sure that the hybridization name column and file name column are not changed otherwise ArrayTrack can not match the updates to the existing hybridizations.

Using this option the user can edit information related to sample (e.g. sample name or label) without impact on data intensity value.

Click “Next” button.

**Step 2:** map hybridization info file to database. Click the first “Browse” button to locate the data directory then click the second “Browse” button to open the hybridization file.

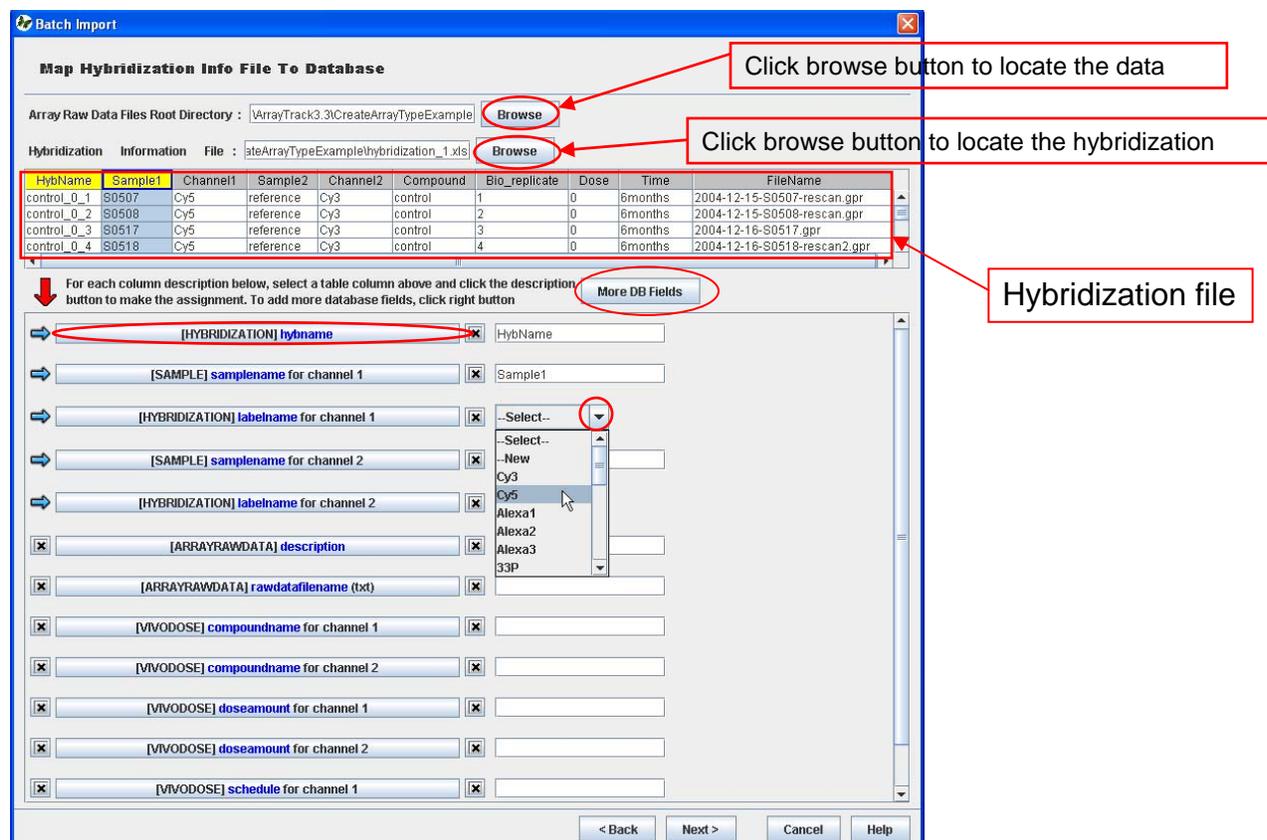


Figure 2-7: mapping the hybridization columns to database fields

**Step 3: Map the columns to database fields.**

In Figure 2-7, 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  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.

If the user wants to map columns to the other database fields which are not showing in the default window, he can click “More DB Fields” button to bring up more database fields (Figure 2-8).

In Figure 2-8, the database fields in the right panel are the fields that will be shown in “Batch Import” interface (see Figure 2-6) 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.

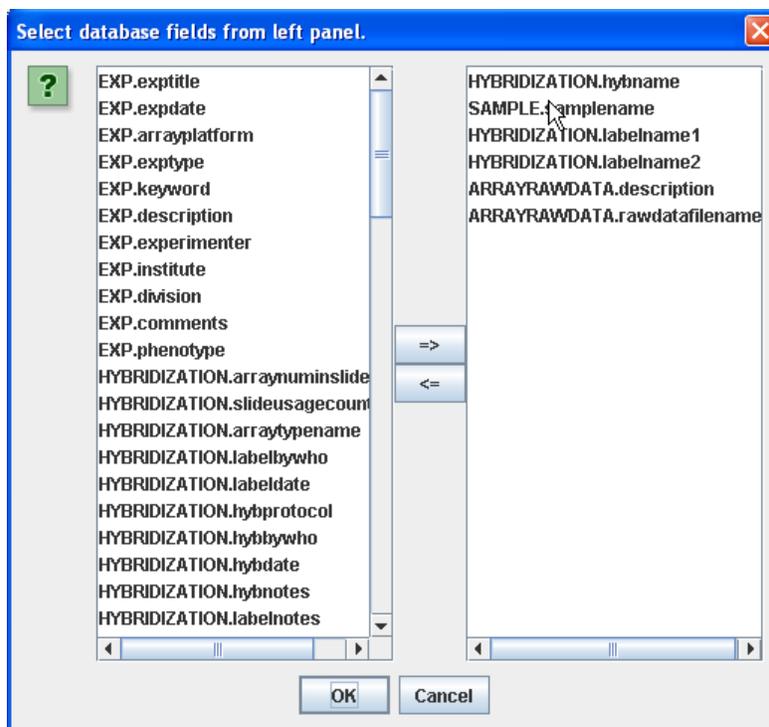


Figure 2-8: more database fields for choosing

**Step 4:** After mapping the columns to the database field, the user can click “Next” button to preview the data information before importing.

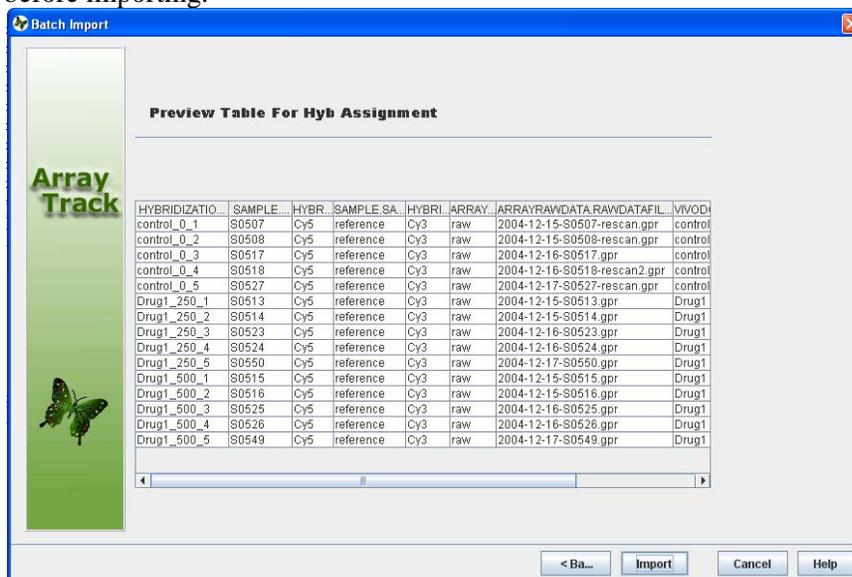


Figure 2-9: batch import preview window

In Figure 2-9, 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 “Back” button and re-do the Step 1~ 4. After previewing it, click **Import** button to start importing. User will be asked to save the configuration file. We suggest saving the configuration file. If the batch import failed this time, user can load the configuration file that saved all the previous choices and

mapping to perform batch import again. The configuration file makes sure that the user does not need to repeat mapping steps. 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 prevents errors when importing data in large scale.

When batch importing Affymetrix data, if the user wants to import .CEL file and probe set file or image file, he needs to specify the data location. It is suggested that the .CEL file and probe set file are located in same directory. The procedure of batch import Affymetrix data is similar to two channel data import.

### 2.2.2 Batch import SimpleTox data

SimpleTox batch import is for storing animal toxicological study data. In Figure 2-10, right-click the experiment, choose “Batch Import Data – Simple Tox Format”

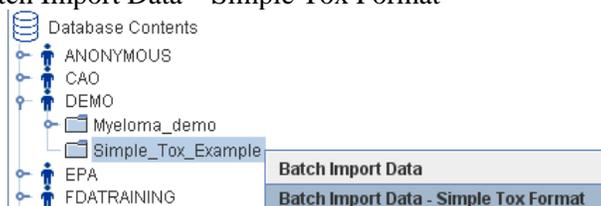


Figure 2-10: batch import in simple tox format

In Figure 2-11 click “Browse” button to locate the hybridization file and data files directory. Then click “OK” button.

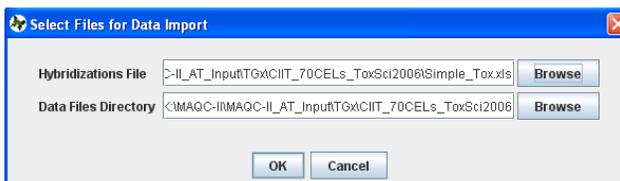


Figure 2-11: select files for data import

Figure 2-12 is an example of hybridization file for SimpleTox format. User can use this as a template to make his own hybridization file. Just make sure that in your own hybridization file, the column titles are exactly same as this template. This example file (sample data) can be downloaded from ArrayTrack help pull-down menu.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Array_ID	SubjectID	Institution	StudyTitle	StudyType	CompoundClass	ClassReference	Compound	CAS	Control	Treatmen	Dose	DoseUnit	HybName	SampleName
792	1-22	Hamner	Hamner Mice Lu	Repeat Dose			Corn Oil		Y	C	0	mg/kg	RT_0_13_22	RT_0_13_22_Lung_CornOil
793	1-24	Hamner	Hamner Mice Lu	Repeat Dose			Corn Oil		Y	C	0	mg/kg	RT_0_13_24	RT_0_13_24_Lung_CornOil
794	1-25	Hamner	Hamner Mice Lu	Repeat Dose			Corn Oil		Y	C	0	mg/kg	RT_0_13_25	RT_0_13_25_Lung_CornOil
795	1-27	Hamner	Hamner Mice Lu	Repeat Dose			Rodent Chow		Y	C	0	ppm	RT_0_13_27	RT_0_13_27_Lung_RodentCho
796	1-28	Hamner	Hamner Mice Lu	Repeat Dose			Rodent Chow		Y	C	0	ppm	RT_0_13_28	RT_0_13_28_Lung_RodentCho
797	1-29	Hamner	Hamner Mice Lu	Repeat Dose			Rodent Chow		Y	C	0	ppm	RT_0_13_29	RT_0_13_29_Lung_RodentCho
798	1-12	Hamner	Hamner Mice Lu	Repeat Dose	Non-lungtumorigen	NTP_No168	N-(1-naphthyl)et	1465-25-4	N	NLT	2000	ppm	RT_2000_13	RT_2000_13_12_Lung_ethylen
799	1-14	Hamner	Hamner Mice Lu	Repeat Dose	Non-lungtumorigen	NTP_No168	N-(1-naphthyl)et	1465-25-4	N	NLT	2000	ppm	RT_2000_13	RT_2000_13_14_Lung_ethylen
800	1-15	Hamner	Hamner Mice Lu	Repeat Dose	Non-lungtumorigen	NTP_No168	N-(1-naphthyl)et	1465-25-4	N	NLT	2000	ppm	RT_2000_13	RT_2000_13_15_Lung_ethylen
801	1-2	Hamner	Hamner Mice Lu	Repeat Dose	lung tumorigen	NTP_No143	1,5-Naphthalene	2243-62-1	N	LT	2000	ppm	RT_2000_13	RT_2000_13_2_Lung_Naphthal

Figure 2-12: template of hybridization file for simple tox format batch import

The following table explains the meanings of each column in the above hybridization file. The red-colored parts are required fields, while black-colored parts are optional fields. If you don't have Array\_ID, you can use subject\_ID as Array\_ID. Subject ID must be unique. If it is not, the user needs to add suffix number to make it unique. Additional fields that may be needed to describe the study can be added to the SimpleTox table.

SimpleTox Column Head	Description	Example
<b>Institution</b>	Laboratory or institution name	NCTR/FDA, EPA
<b>DataFile</b>	Microarray data file	GSM142129.CEL
<b>HybName</b>	User specified identifier for a hybridization name	APAP_D100_T6_Jun04; APAP_D0_T6_Jun04
<b>SampleName</b>	Sample name	APAP_Dose100_Time6; APAP_Dose0_Time6
<b>Array_ID</b>	User-specified identifier for a hybridization	1,2,3 ... or p1002356
<b>Label</b>	RNA label reagent	Biotin; Cy3
<b>ArrayType</b>	Array type	Affymetrix Mouse 430_2
<b>Subject_ID</b>	Subject identifier	
<b>StudyTitle</b>	Study title	6 days repeating toxicity study
<b>Tech_Rep</b>	Technical replicates; microarray specific	A, B or C; 1,2, or 3
<b>Bio_Rep</b>	Biological replicates	A, B or C; 1,2, or 3
<b>HybDate</b>	Hybridization date	2/25/2007
<b>StudyType</b>	Study type	Single Dose Toxicity or Repeat Dose Toxicity
<b>Compound</b>	Compound name	Carbon Tetrachloride; Acetaminophen
.....	.....	.....

In Figure 2-13, select the right array type, then click OK button. During the import users will see the Figure 2-14 A. When import is finished, a summary report will show up, see Figure 2-14B and C.

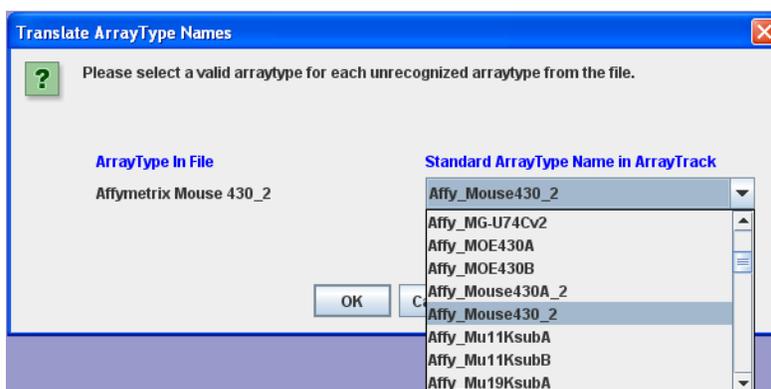


Figure 2-13: select array type for batch import

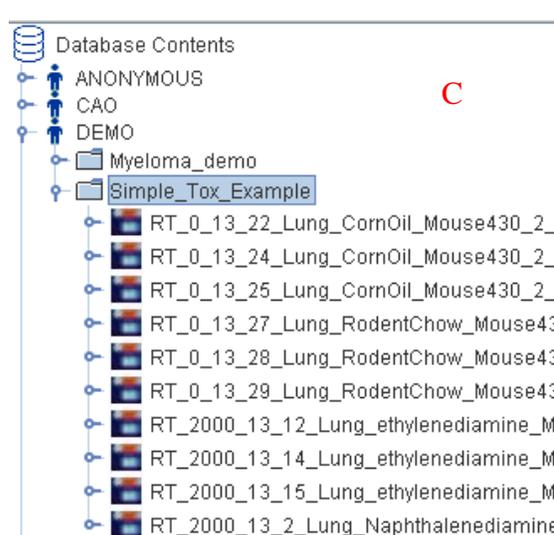
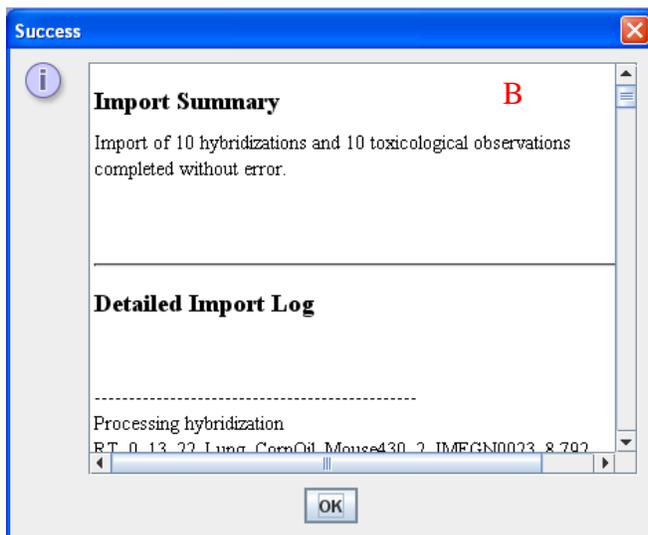
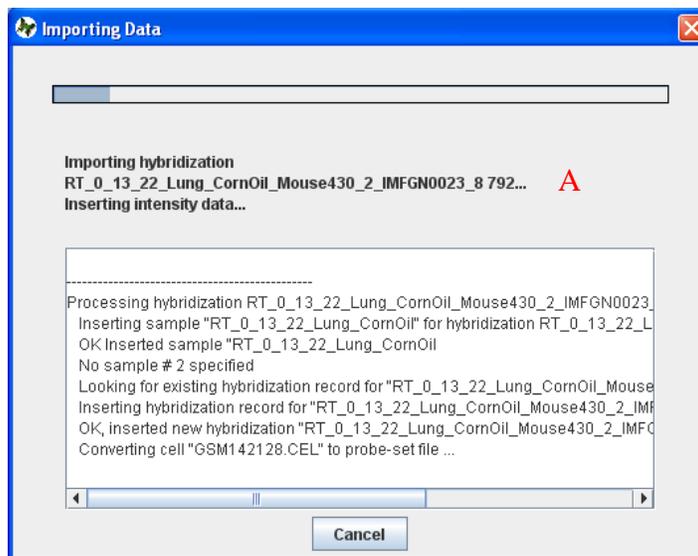


Figure 2-14: import status and import summary

### 2.2.3 View SimpleTox data

To view SimpleTox data, click “Studies...” button (Figure 2-15) to bring out the window below (Figure 2-16).

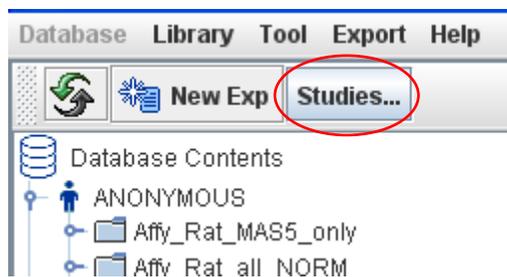


Figure 2-15: view SimpleTox data

In Figure 2-16, user can select the study name (e.g. Simple\_ToFormat) and then click “View Observation” button at the left top of the window. Users can also search data by study or individual observations.

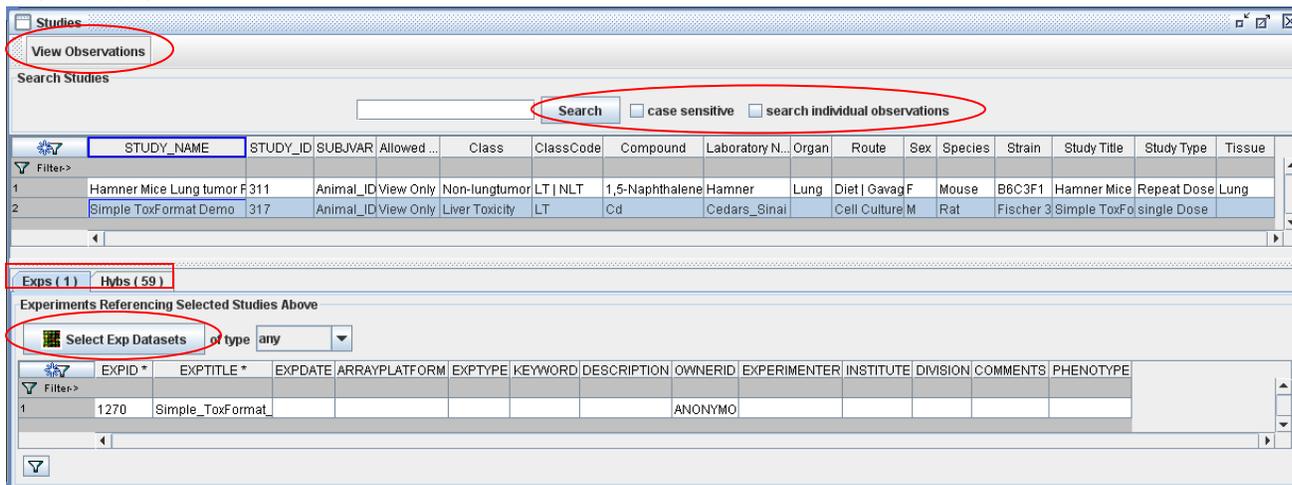


Figure 2-16: view observations of SimpleTox data

There are two tabs in the middle section of the window: Exps and Hybs. Clicking these two tabs will bring out the information for experiment associated with the study and hybridizations for the experiment. See Figure 2-17.

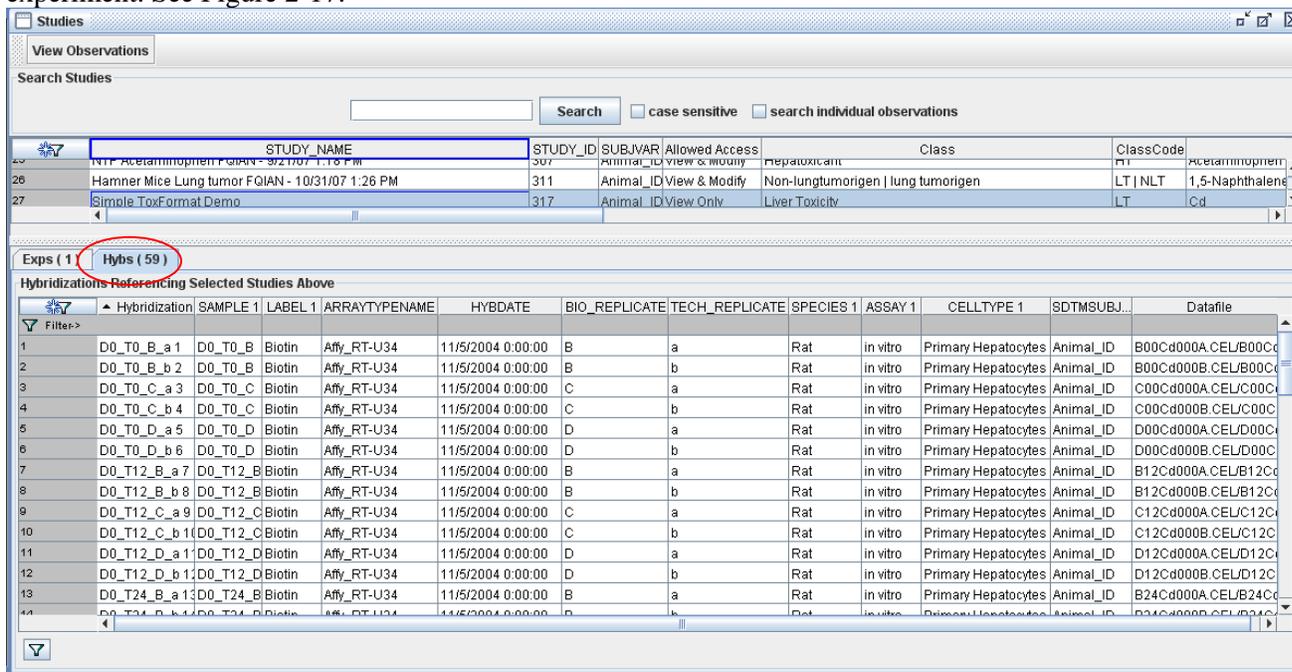


FIGURE 2-17: displaying hybridization information for the experiment

In Figure 2-18, users can highlight the experiment and click “Select Exp Dataset” button to select all the data (or just raw/normalized data) in the data panel for data analysis purpose.

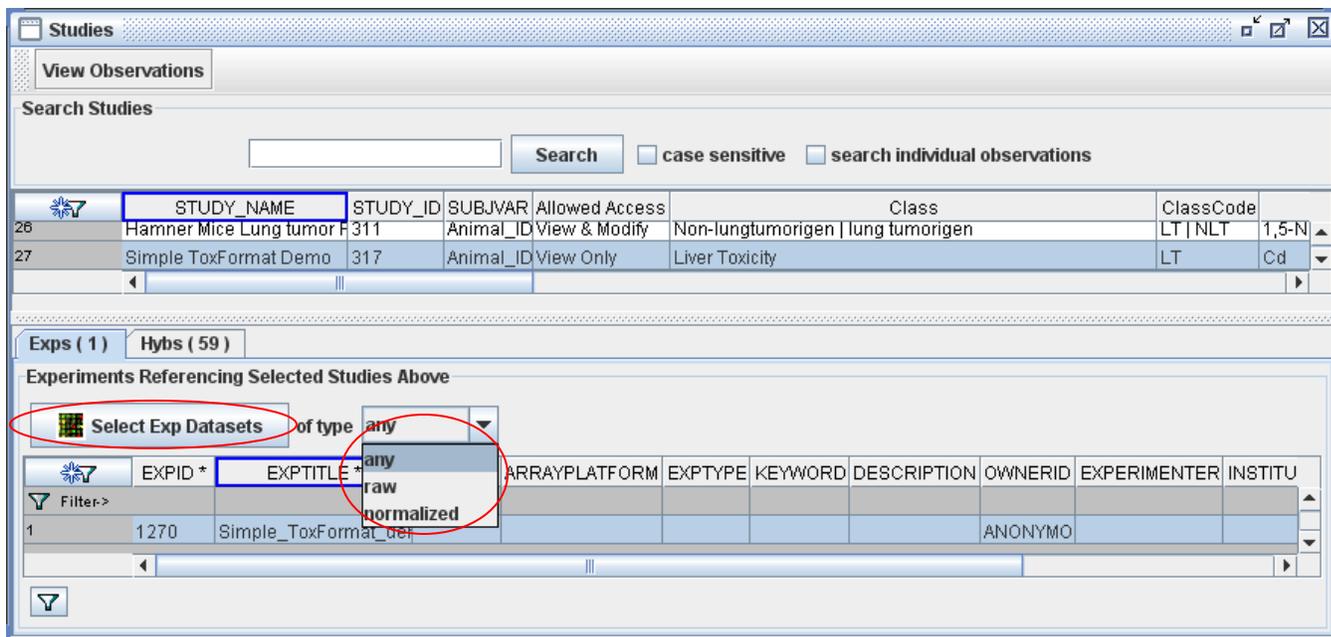


Figure 2-18: select experiment dataset for data analysis purpose

### 2.2.4 Delete SimpleTox data

To delete SimpleTox data you have imported, you need to delete the experiment first and then delete the study. Double-click the experiment name in the data panel to bring out the Input Form, then click “Delete Exp” button.

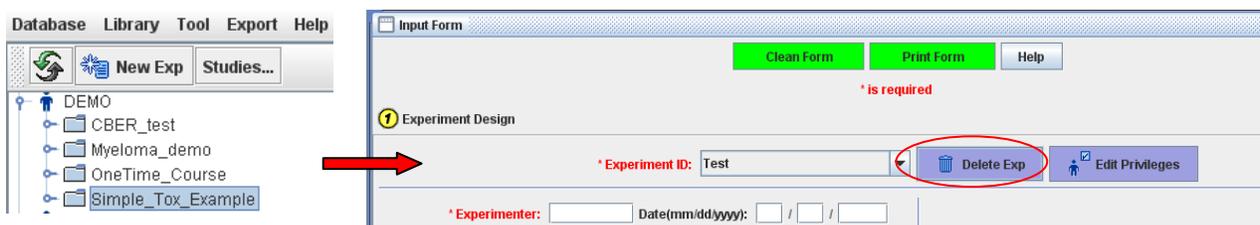


Figure 2-19: delete SimpleTox data

Users will be asked the following question, click “Yes” button to permanently delete the experiment.



Figure 2-20: permanently delete the experiment

After deleting experiment, you need to delete the study. Select the study, right-click, choose “Delete Studies”. See Figure 2-21.

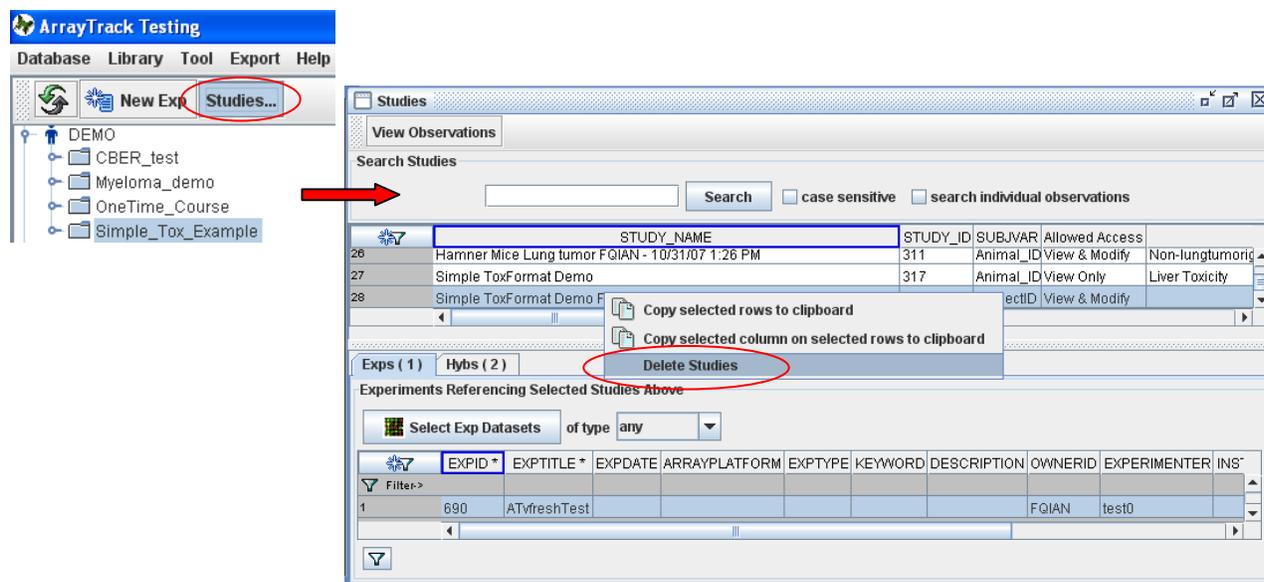


Figure 2-21: delete the study

**Data Export:** for information about exporting data please refer Chapter 10 Data Export.

### 2.2.5 Update Hybridization information

After importing data, users may want to add more information to the hybridization (e.g. dose, date, etc) without overwriting data intensity. ArrayTrack provides this function. The user can right-click the experiment name, then select “Batch Import Data”. See Figure 2-22.

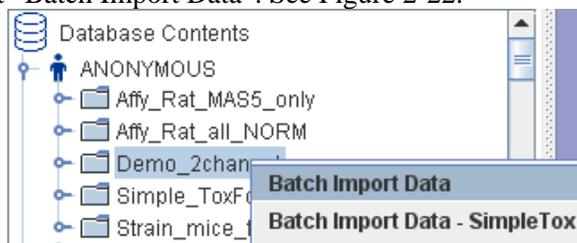


Figure 2-22: update hybridization information



Figure 2-23: update existing hybridizations and samples

In Figure 2-23, there is an option for inserts or updates. Make sure the second option (updates) is selected. Click “Next” button. Then in Figure 2-24, users only need to map the hybridization name and the additional column which is “Bio\_replicate” in this example, leaving other mapping untouched. So the previous imported data will not be changed, only new information (Bio\_replicate) is added. Click “Next” button. The rest steps are the same as regular batch import.

**Batch Import**

**Map Hybridization Info File To Database**

Array Raw Data Files Root Directory :

Hybridization Information File :

HybName	Sample1	Channel1	Sample2	Channel2	Compound	Bio_replicate	Dose	Time	FileName
control_0_1	S0507	Cy5	reference	Cy3	control	1	0	6mont...	2004-12-1...
control_0_2	S0508	Cy5	reference	Cy3	control	2	0	6mont...	2004-12-1...
control_0_3	S0517	Cy5	reference	Cy3	control	3	0	6mont...	2004-12-1...
control_0_4	S0518	Cy5	reference	Cy3	control	4	0	6mont...	2004-12-1...
control_0_5	S0527	Cy5	reference	Cy3	control	5	0	6mont...	2004-12-1...

For each column description below, select a table column above and click the description button to make the assignment. To add more database fields, click right button

[HYBRIDIZATION] hybname

[SAMPLE] samplename for channel 1

[HYBRIDIZATION] labelname1

[SAMPLE] samplename for channel 2

[HYBRIDIZATION] labelname2

[ARRAYRAWDATA] description

[ARRAYRAWDATA] rawdatafilename (txt)

[VIVODOSE] compoundname for channel 1

[VIVODOSE] compoundname for channel 2

[VIVODOSE] doseamount for channel 1

[VIVODOSE] doseamount for channel 2

[VIVODOSE] schedule for channel 1

[VIVODOSE] schedule for channel 2

[HYBRIDIZATION] bio\_replicate

Figure 2-24: Update hybridization information

## 2.3 Create Array Type

### 2.3.1 Overview

One essential item that needs to be input in batch import is ArrayType. ArrayType defines basic information about the arrays with which an experiment is being conducted. In ArrayTrack 3.4 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 4-36 in Chapter 4).

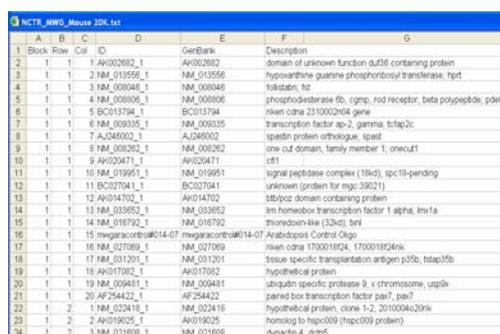
### 2.3.2 Activate Array Type Information

Array type specifies all the information of each spot like gene name (or other unique ID), location (row, column) and other annotation contents. ArrayTrack Chip Library stores over 190 array types (standard and customized array type). Before importing any array data into ArrayTrack, the user needs to check if the related array type is available in the Chip Library, if not then the user needs to create a new array type first.

**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 4 4.9 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 2-25 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.



A	B	C	D	E	F	G
1	1	1	1	AK002692_1	AK002692	domain of unknown function out56 containing protein
2	1	1	2	NM_019556_1	NM_019556	hydroxymethylguanine phosphoribosyl transferase, hprt
3	1	1	3	NM_008046_1	NM_008046	histidine-kt
4	1	1	4	NM_008606_1	NM_008606	phosphodiesterase 8b, comp. rod receptor, beta polypeptide, pde8b
5	1	1	5	BC013794_1	BC013794	riian cdra 2310002h04 gene
6	1	1	6	NM_009335_1	NM_009335	transcription factor ap-2, gamma, tcfap2c
7	1	1	7	A246002_1	A246002	spec1n protein, ortholog, spec1
8	1	1	8	NM_008362_1	NM_008362	onc cut domain, family member 1, onc cut1
9	1	1	9	AK000471_1	AK000471	c81
10	1	1	10	NM_019951_1	NM_019951	signal peptidase complex (19kd), spc18-pending
11	1	1	11	BC027041_1	BC027041	unknown protein for mgs-39211
12	1	1	12	AK014702_1	AK014702	zfp103 domain containing protein
13	1	1	13	NM_033652_1	NM_033652	lim homeobox transcription factor 1 alpha, lim1a
14	1	1	14	NM_016792_1	NM_016792	thrombosin-like (32kd), bcl1
15	1	1	15	imgjarcorn0014-07	imgjarcorn0014-07	Arabidopsis Control Oligo
16	1	1	16	NM_027099_1	NM_027099	riian cdra 1700181d4a, 1700181d4a
17	1	1	17	NM_031201_1	NM_031201	issue specific transplantation antigen p35b, tsap35b
18	1	1	18	AK017082_1	AK017082	hypothetical protein
19	1	1	19	NM_009481_1	NM_009481	ubiquitin specific protease 9, x chromosome, uspb9
20	1	1	20	AF254422_1	AF254422	paired box transcription factor pax7, pax7
21	1	2	1	NM_022418_1	NM_022418	hypothetical protein, clone 1-2, 2010004c20nk
22	1	2	2	AK019025_1	AK019025	homolog to hspc009 (hspc009 protein)
23	1	2	3	NM_021808_1	NM_021808	dynactin 4, dnr5

Figure 2-25: ArrayType Information File.

Generally any array type can be created in ArrayTrack. Following are just two typical examples.

#### 1. Create two channel array type:

To create a new array type, click “Array Type Modification” from Database pull-down menu then choose “Create a New Array Type”. This function is only available for locally-installed ArrayTrack. The on-line version will not allow the user to create/delete new array type.

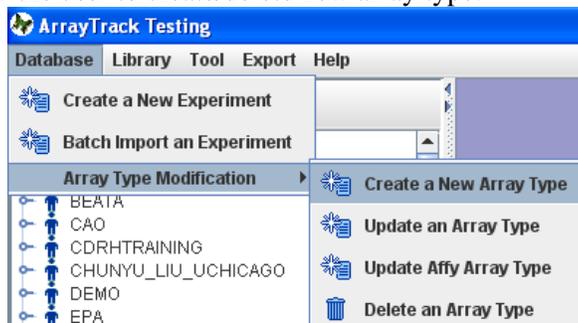


Figure 2-26: Create a new array type

In Figure 2-27 user needs to specify the file name for the ArrayType Information File, the array type name and manufacture name (Agilent, Affymetrix, etc). If the manufacture name is not in the list, user

can create a new one. User also needs to specify the channel type, species and the type of array elements (oligo or cDNA).

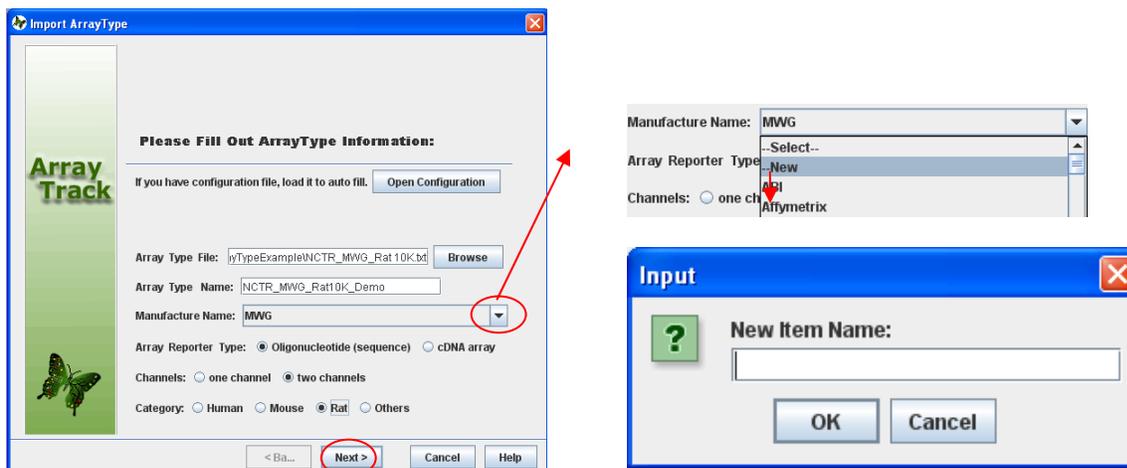


Figure 2-27: Loading ArrayType Information File

In Figure 2-27, users can ignore the button “Open Configuration” if you don’t have one. You will be asked to save the configuration file later. Click “Next” button.

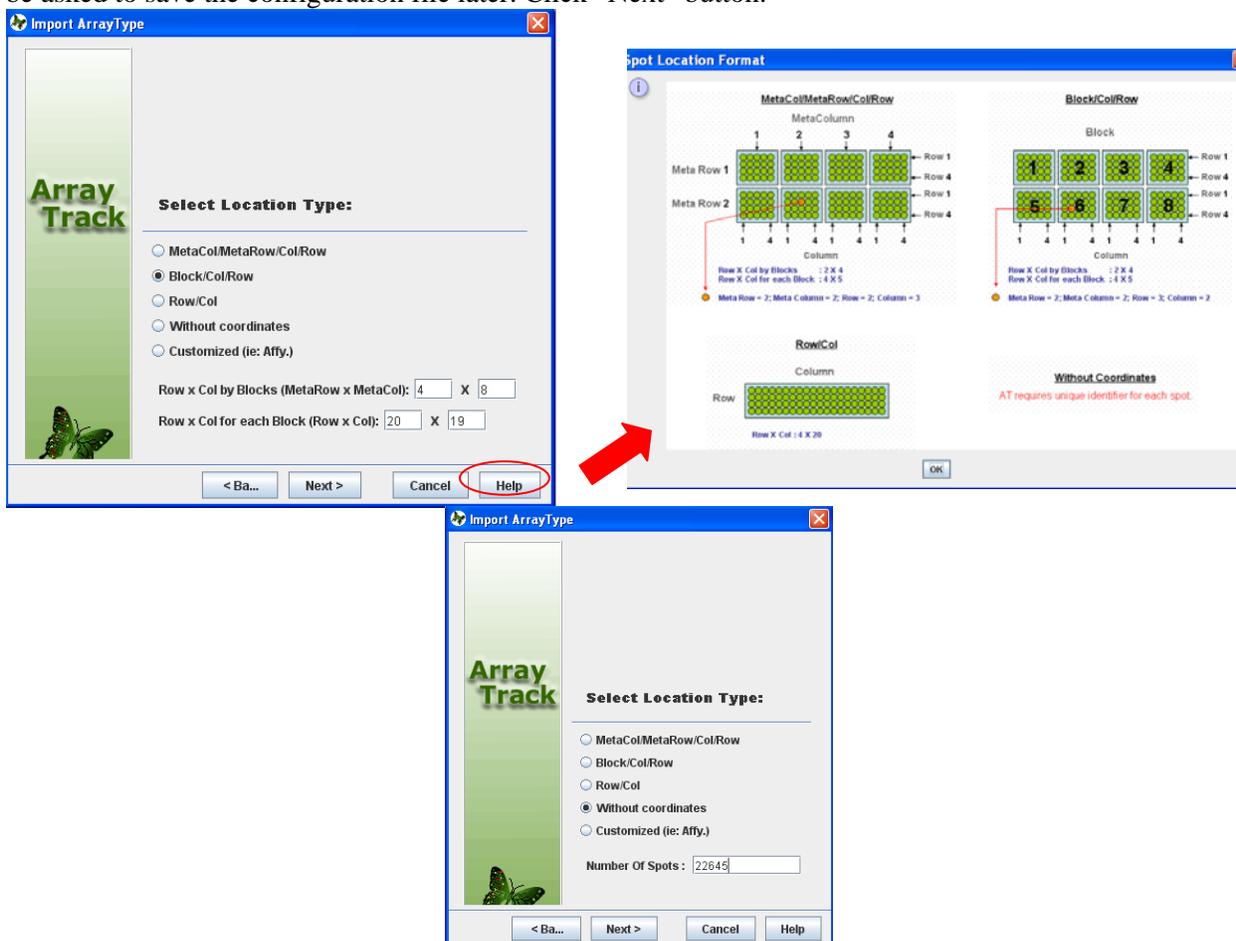


Figure 2-28: select location type for creating array type

In Figure 2-28, users need to specify the coordinate location type and the number of columns and rows. This information can be found in the array type file.

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 total genes.

Caution: For some array type (e.g. Agilent), even the total number of genes is the same but if the coordinates is different (chip's orientation is different), we consider it as a different array type.

Figure 2-29 shows the mapping between the columns of array type file and database fields. If the array type file has several rows of headings, user can click the button "Manually select first data row..." at the top and then click the first row of data. Data columns already mapped will be highlighted in yellow and their corresponding MicroarrayDB data fields will be marked with a ✓. Clicking on ✕ will undo a previous column assignment. After mapping all the necessary columns, click "Import" button. Then user will be asked to save the configuration file for future auto-loading and reference.

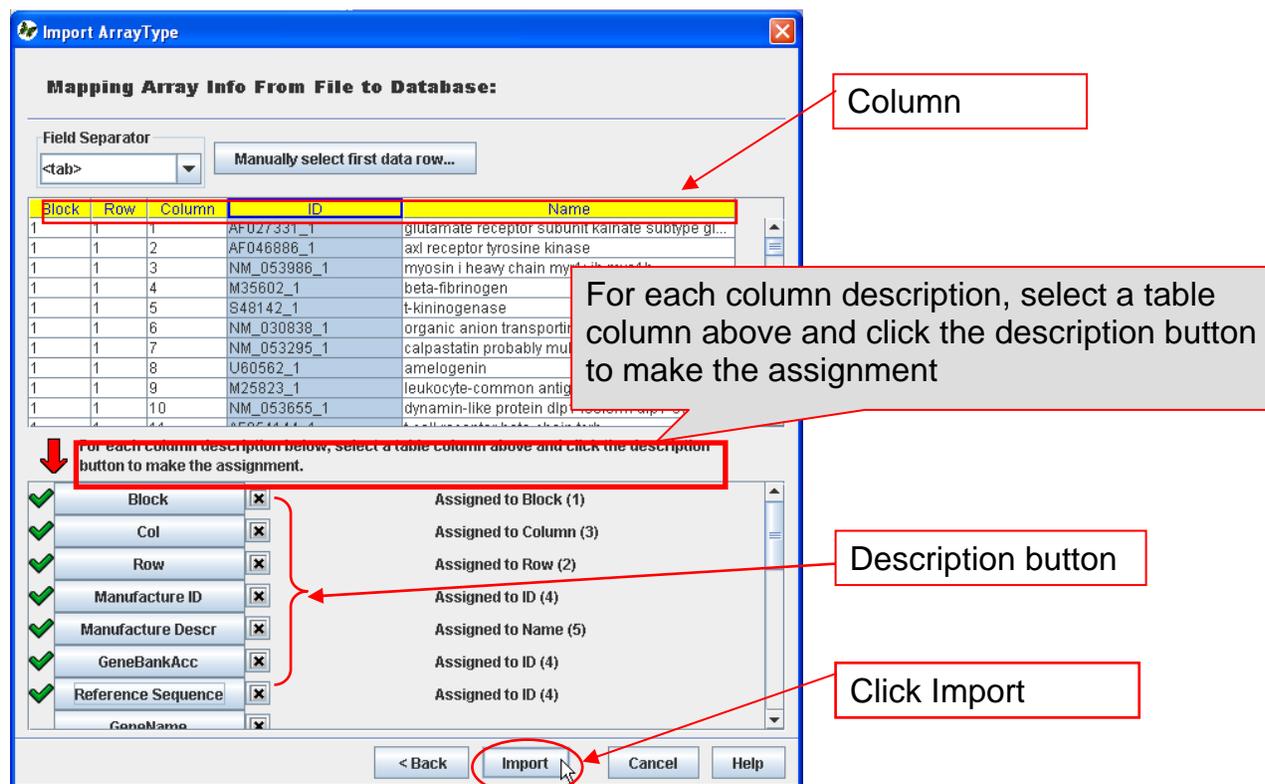


Figure 2-29: mapping array info from file to database fields

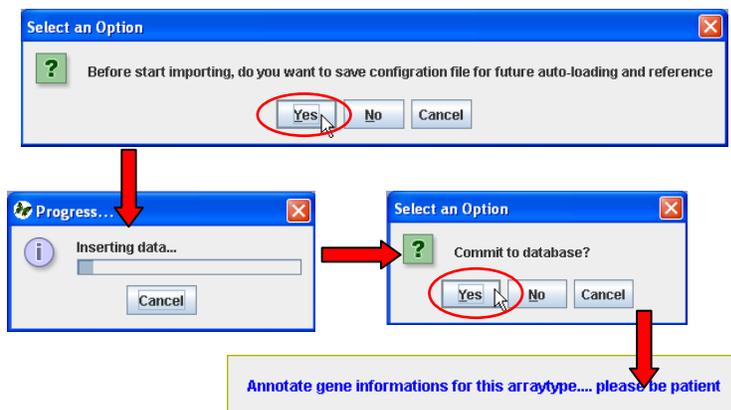


Figure 2-30: save configuration file and finish creating array type

**2. Create a new array type for Affymetrix data**

ArrayTrack Chip library stores most of the standard Affymetrix array types. If the array type for your data does not exist in the Chip Library, then you need to create new array type for Affymetrix data. In Figure 2-28, user must select “customized (ie: Affy.)”. Then a pop up window will let the user to choose the converter that will convert the .CEL file to .TXT file, see Figure 2-31. For Affymetrix data, “AffyLocConverter” should be selected.

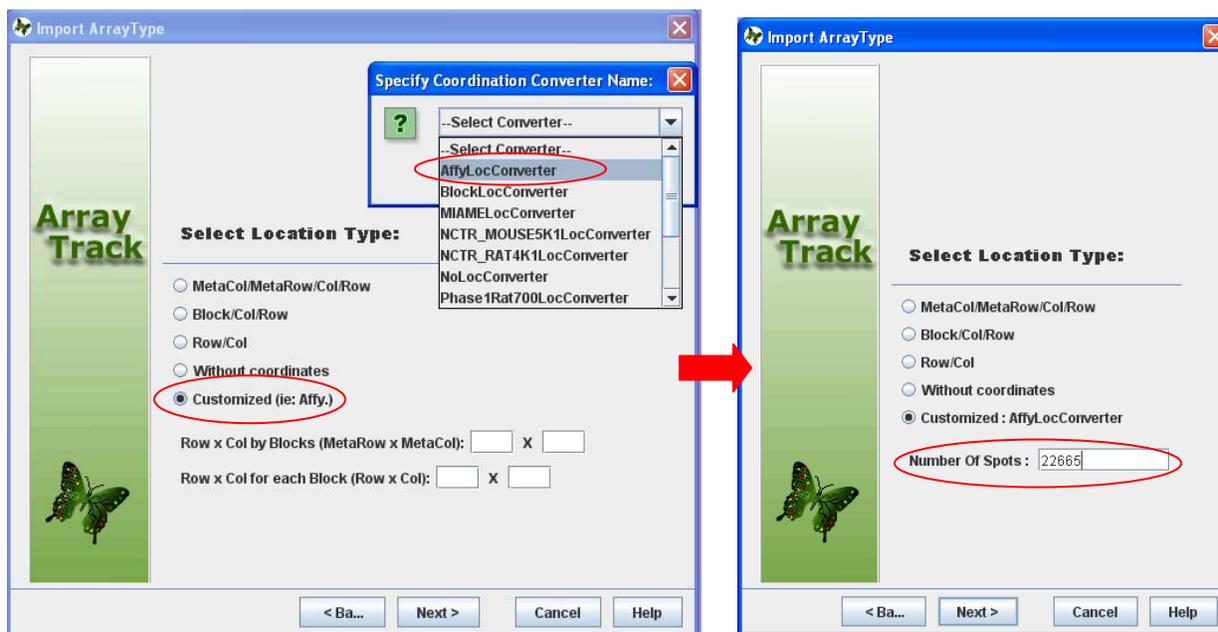


Figure 2-31: Choose the Converter for creating Affymetrix array type and specify the total number of spots.

Same as creating array type for two channel data, user needs to map the Affy array type file columns to database fields. We suggest using the following table as reference when mapping.

Affymetrix annotation CSV file	ArrayTrack chip fields
Probe Set ID	GEN_ID_MFR
Strand	Strand
Sart	BP_Start
End	BP_End
Target Description	GEN_DESCR_MFR
Representative Public ID	GENEBANKACC
Archival UniGene Cluster	
UniGene ID	UNIGENEID
Genome Version	
Alignments	
Gene Title	DESCRIPTION
Gene Symbol	GENENAME
Chromosomal Location	CHROMLOCATION
Unigene Cluster Type	
Ensembl	Ensembl
Entrez Gene	LOCUSID
SwissProt	SWISSPROT_ACC_NUMBER
EC	
OMIM	OMIM
RefSeq Protein ID	PROTEIN_REFSEQ
RefSeq Transcript ID	REFSEQ
Gene Ontology Biological Process	BIOLOGICAL_PROCESS
Gene Ontology Cellular Component	CELULAR_COMPONENT
Gene Ontology Molecular Function	MOLECULAR_FUNCTION
Pathway	PATHWAYS

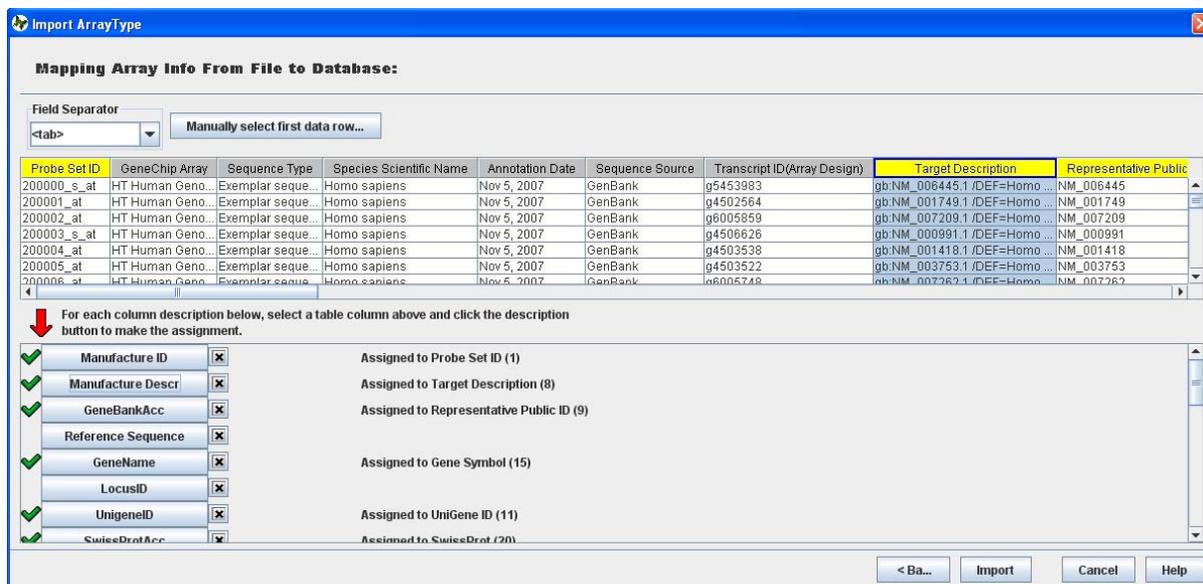


Figure 2-32: Mapping the array info from file to database.

When finished mapping, click “Import” button in Figure 2-32 to start creating array type. If everything goes well, user will see the following pop-up message.

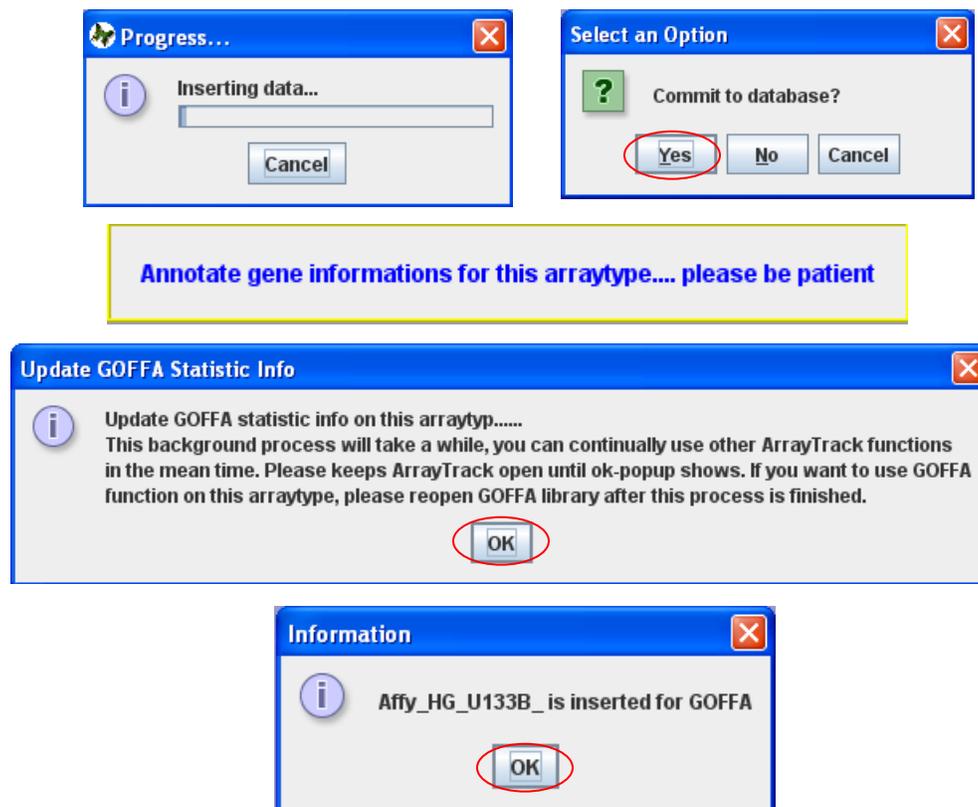


Figure 2-33: Array type is created

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

### 3. Delete array type

This function is not available for on-line version (<http://edkb.fda.gov/webstart/arraytrack/>) of ArrayTrack. Users with locally-installed ArrayTrack or within FDA firewall can delete array type.

To delete an array type, first user has to delete all the data associated with this array type then click Database pull-down menu, select “Array Type Modification” -> “Delete an Array Type”.

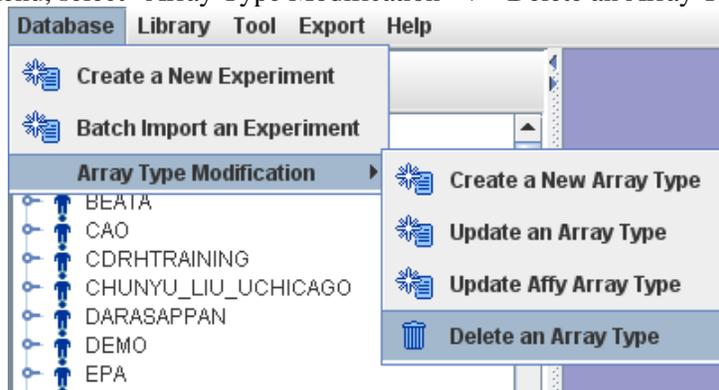


Figure 2-34: delete array type

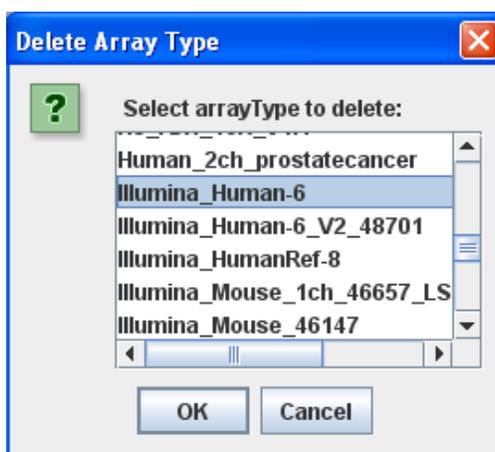


Figure 2-35: select the array type to delete

User needs to select the array type to be deleted, and click “OK” button.

#### 4. Update array type

This function is not available for on-line version (<http://edkb.fda.gov/webstart/arraytrack/>) of ArrayTrack. Users with locally-installed ArrayTrack or within FDA firewall can update array type.

To update an array type, user can click “Database” pull-down menu, select “Array Type Modification”-> “Update an Array Type”.

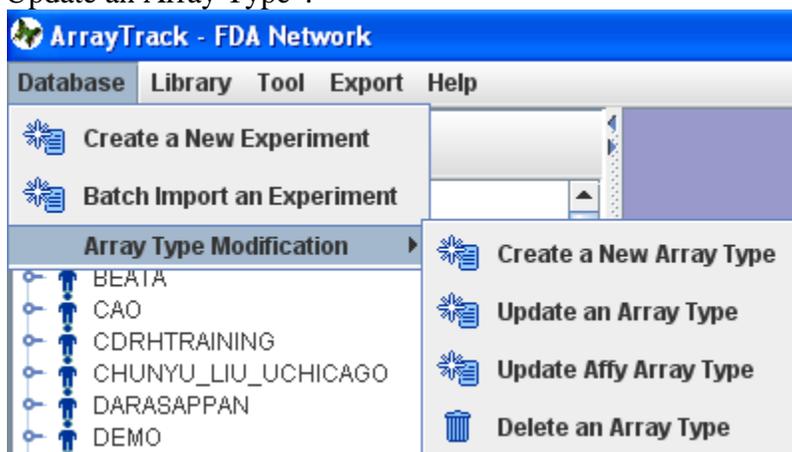


Figure 2-36: Update an array type

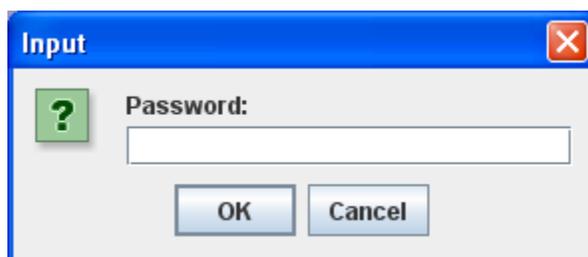


Figure 2-37: type in password to update array type

Users will be asked for password to update array type. The password is the user's username to login his computer. Click OK button.

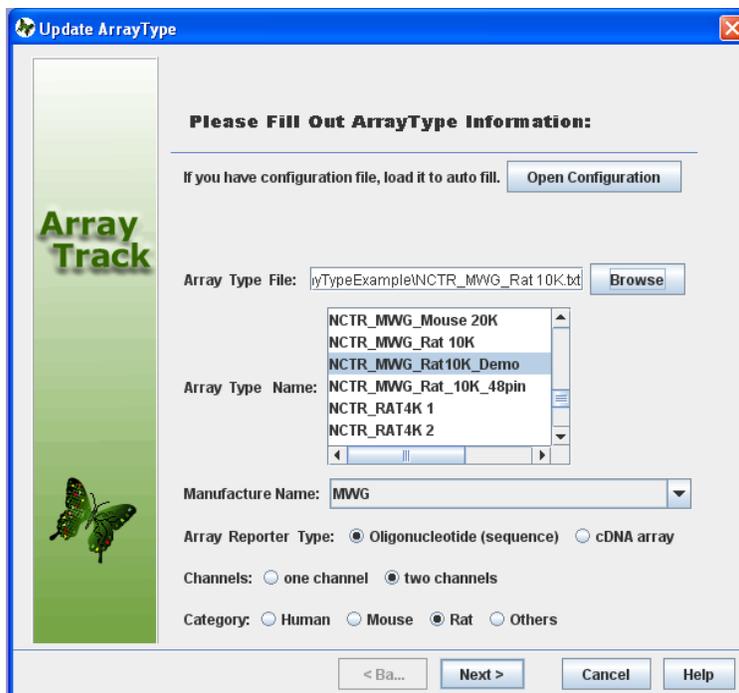


Figure 2-38: fill out array type information to update array type

In Figure 2-38, user first clicks “Browse” button to load the array type file and then selects the array type name from the pull-down list. Click “Next” button.

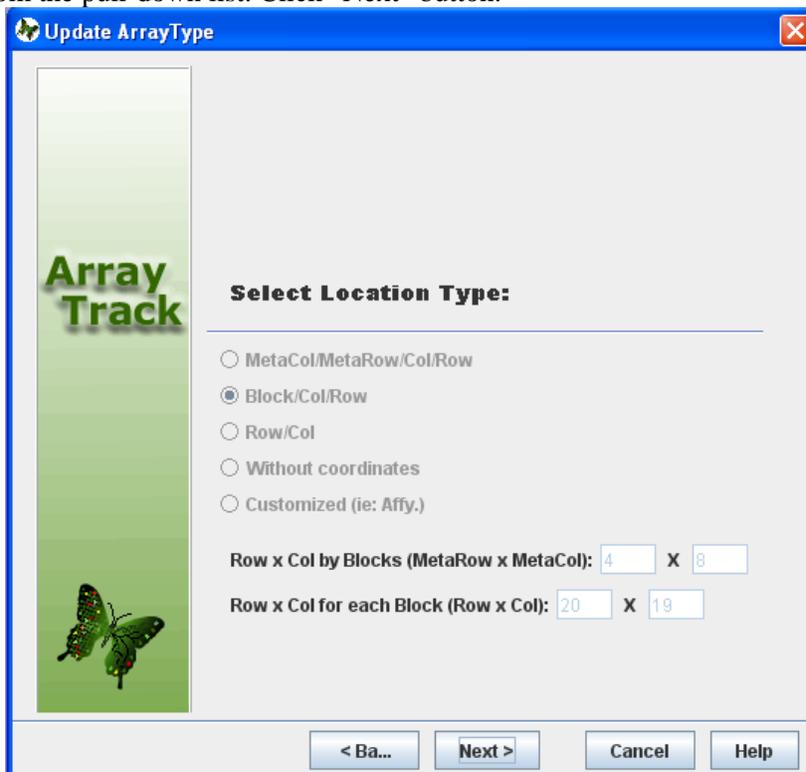


Figure 2-39: select location type

Figure 2-39 shows the spot layout for this array type. All the numbers are grayed out which means user can not change the gene spot layout. Users can only update annotation information, e.g. add more columns to the array type file and map them to other fields.

Figure 2-29 shows the original mapping for this array type. To update this array type, users only need to map the additional info, keeping all the other original mapping untouched. For example, column "Name" was mapped to "Manufacture Descr" before and now users can perform an update and map column "Name" to other field like "Description". The result is that the column "Name" will be mapped to two fields: "Manufacture Descr" and "Description". It is very important to assign "Block", "Col" and "Row" when doing mapping, because they define the spot location. See Figure 2-40. Click "Import" button.

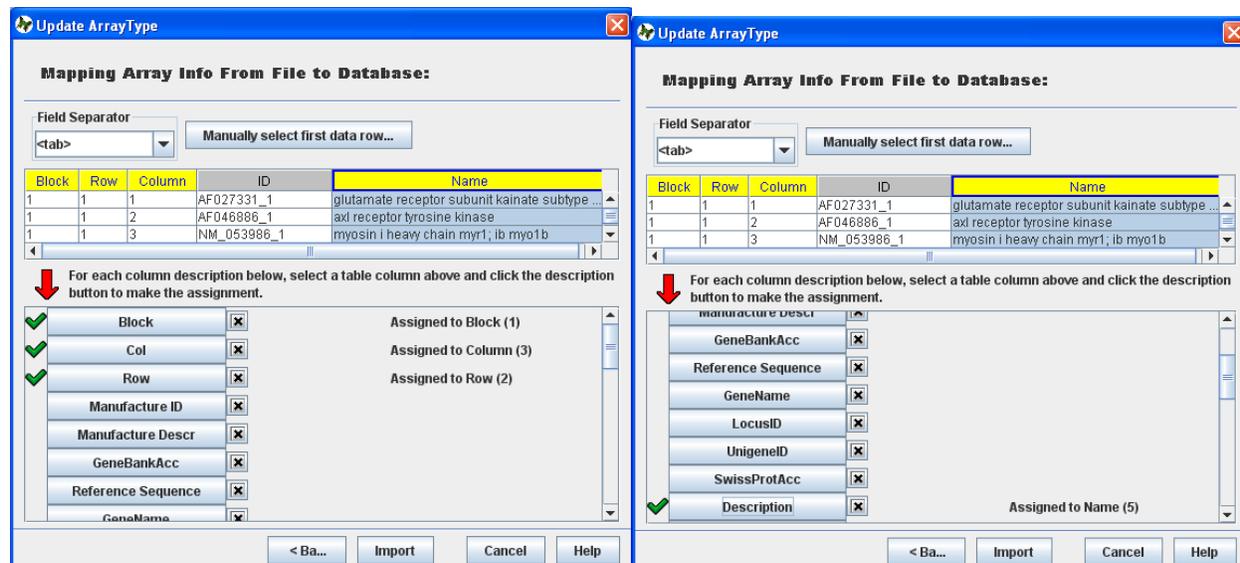


Figure 2-40: update array type

## 2.4 Data Sharing and Security Protection

The owner of the experiment can assign Read/Write/Normalize/Create Genelists/Manage Permission privileges to individuals by clicking on  Edit Privileges of Input Form to share the experiment data with others (Figure 2-41). The following is the explanation of these privileges:

1. Read - Users assigned this privilege can view, copy/paste or export the dataset. They can also perform analyses on the dataset.
2. Create Genelists - users assigned this privilege, in addition to viewing, copying/pasting, exporting and performing analyses, can also create and save gene lists within the experiment.
3. Normalize - users assigned this privilege have the ability to normalize the data in addition to viewing, copying/pasting, exporting and performing analyses on it.
4. Write - users assigned this privilege can do all the above (1~3) and in addition, can make changes to the experiment and can delete the experiment if need be. Also the user can import additional data under the experiment.
5. Manage Permissions - users holding this privilege for an experiment can grant and revoke data permissions from other users for that particular experiment. However, this user does not implicitly have the permissions mentioned above.

The owner of an experiment, when importing array data, will be given privileges to write and manage permissions.

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)

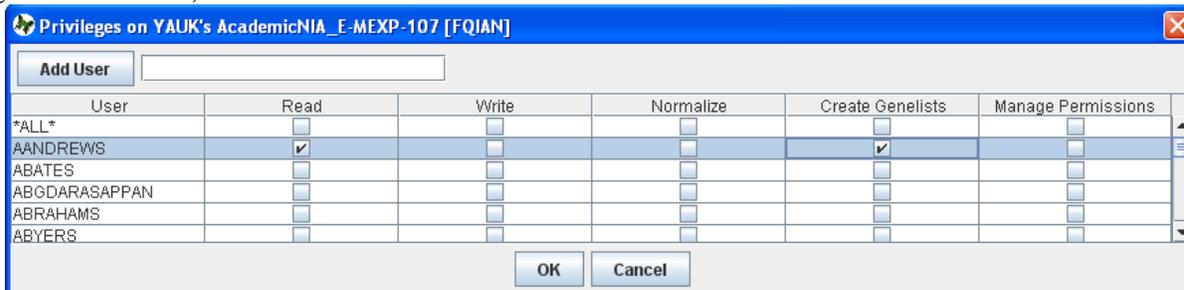


Figure 2-41: Assigning Read/Write privileges to individuals.

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 2-41.

## 2.5 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 2-42. 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:

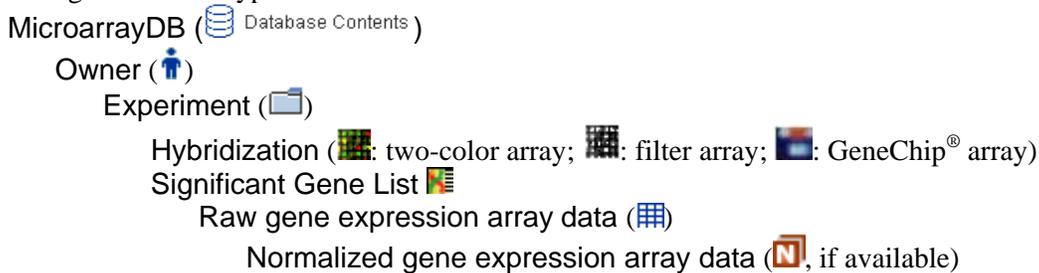


Figure 2-42: Hierarchical Structure of the MicroarrayDB Contents.

**Exploring Database Contents:** The tree-like Database Contents structure allows the user to select and/or open a particular user (User icon), experiment (Folder icon), or hybridization (Grid icons, or GeneChip icon) by toggling on the lock (Lock icon) unlock (Unlock icon) signs. The content structure can also be manipulated by right-click on a particular item (Figure 2-43). 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 2-43. 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 hyname 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.

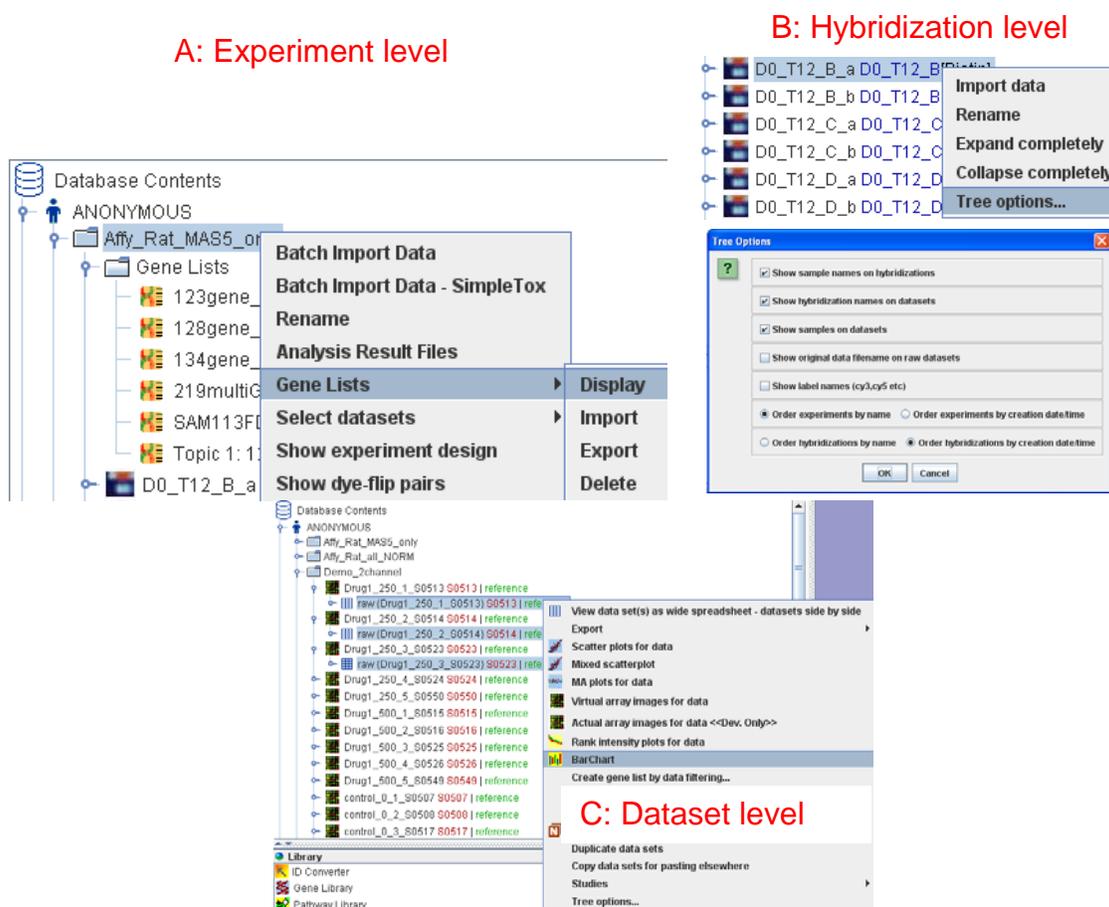


Figure 2-43: Database Contents tree can be expanded or collapsed completely by right-clicking on an item. Depending on the nature of the object that is right-clicked on, a set of applicable functions become accessible. (A) Experiment; (B) Hybridization; and (C) Array Dataset.

**Operating on Experiment:** Double-clicking on an Experiment will bring up the Input Form. Upon right-clicking on an Experiment, in addition to Expand completely and Collapse completely the tree, the user can choose several other options (Figure 2-43A).

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 pairing information is listed in a spreadsheet view (Figure 2-44 left). Hybridizations without matched flip-dye pairing will be displayed in a separate spreadsheet (Figure 2-44 right). This function is based on the sample and hybridization information entered in the Input Form (Figure 2-3). 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).

	HYBNAME	HYBNAME	label	label	sample	sample
1	K01-C	K02-C	Cy3	Cy5	Control Liver 6	Universal Mouse RNA
2	K06-V	K07-V	Cy3	Cy5	Valproic Liver 4	Universal Mouse RNA
3	K08-C	K09-C	Cy3	Cy5	Control Liver 12	Universal Mouse RNA
4	K10-V	K11-V	Cy3	Cy5	Valproic Liver 8	Universal Mouse RNA
5	K12-C	K13-C	Cy3	Cy5	Control Liver 14	Universal Mouse RNA
6	K14-V	K15-V	Cy3	Cy5	Valproic Liver 15	Universal Mouse RNA
7	K16-C	K17-C	Cy3	Cy5	Control Liver 7	Universal Mouse RNA
8	K18-V	K19-V	Cy3	Cy5	Valproic Liver 1	Universal Mouse RNA
9	K22-V	K23-V	Cy3	Cy5	Valproic Liver V1	Universal Mouse RNA
10	K24-C	K25-C	Cy3	Cy5	Control Liver 1	Universal Mouse RNA

	HYBNAME	label 1	sample 1	label 2	sample 2

Figure 2-44: 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-clicking on a Hybridization will bring up the Input Form 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 2-43B). 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 2-43C). Instructions for use of these functions are discussed in Chapters 4 through 8.

## Chapter 3 Gene List

### 3.1 Overview

Gene list is an important concept in ArrayTrack. It can be operated independently. User can get gene list from quality control filtering, data analysis and function analysis. Further analysis and VennDiagram can be performed on gene list as well. ArrayTrack allows the user to Create, Import, Export, Delete and Display a list of genes associated with this Experiment. Double-clicking any *Gene List* will bring up the gene list displayed in a spreadsheet, see Figure 3-1.

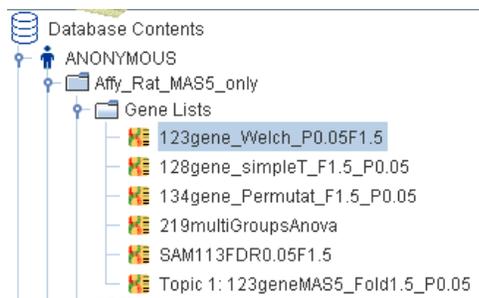


Figure 3-1: Double-click the Gene lists for display

Filter>	GENELIST_NAME*	EXPID*	GENEBANKACC	GENENAME	LOCUSID	SWISSPROT_ACC_NUMBER	FOLD	PVALUE	GEN_ID
1	123gene_Welch_P0.05F1.5	650	L15079	Abcb4	24891	Q08201 // Q64714	0.5458	0	L15079mF
2	123gene_Welch_P0.05F1.5	650	L26267	NfkB1	81736	Q63369	0.4487	0	L26267_at
3	123gene_Welch_P0.05F1.5	650	D87336	Blmh	287552	A1A5L1	0.5661	0.0001	D87336_g
4	123gene_Welch_P0.05F1.5	650	X07365	Gpx1	24404	P04041 // Q6PDW8	0.5872	0.0001	X07365_s
5	123gene_Welch_P0.05F1.5	650	E00778	Cyp11a1	24296	P00185 // Q80UE2	0.2501	0.0002	E00778cds
6	123gene_Welch_P0.05F1.5	650	M55534	Cttnb1	24296	P06762	6.2489	0.0002	J02722cds
7	123gene_Welch_P0.05F1.5	650	M55534mf	Cttnb1	24296	P23928 // Q63136 // Q63137 // Q63138	3.1305	0.0002	M55534mf
8	123gene_Welch_P0.05F1.5	650	AA848563	Cttnb1	24296	P06762	14.9075	0.0003	AA848563
9	123gene_Welch_P0.05F1.5	650	Z75029_s	Cttnb1	24296	P06762	7.5353	0.0004	Z75029_s
10	123gene_Welch_P0.05F1.5	650	rc_AA8186	Cttnb1	24296	P06762	12.8329	0.0004	rc_AA8186
11	123gene_Welch_P0.05F1.5	650	U01344_g	Cttnb1	24296	P06762	0.5856	0.0005	U01344_g
12	123gene_Welch_P0.05F1.5	650	AA108277	Cttnb1	24296	P06762	286	0.0006	AA108277
13	123gene_Welch_P0.05F1.5	650	M99169	Rps6ka1	817	Q08201 // Q64714	41	0.0006	M99169_a
14	123gene_Welch_P0.05F1.5	650	AI22965	Ctdsp1	363	Q08201 // Q64714	5	0.0007	rc_AI22965

Figure 3-2: Display of Significant Gene List information associated with an experiment

Users can open, or export gene lists. But to create, delete, rename, copy or move a gene list the user needs to be granted the privilege of creating gene list or he must be the owner of the experiment.

### 3.2 Create Gene List

The user can create a gene list from the result of T-test (see page 93) or by right-clicking the selected dataset and selecting “Create Gene List by data filtering...”.

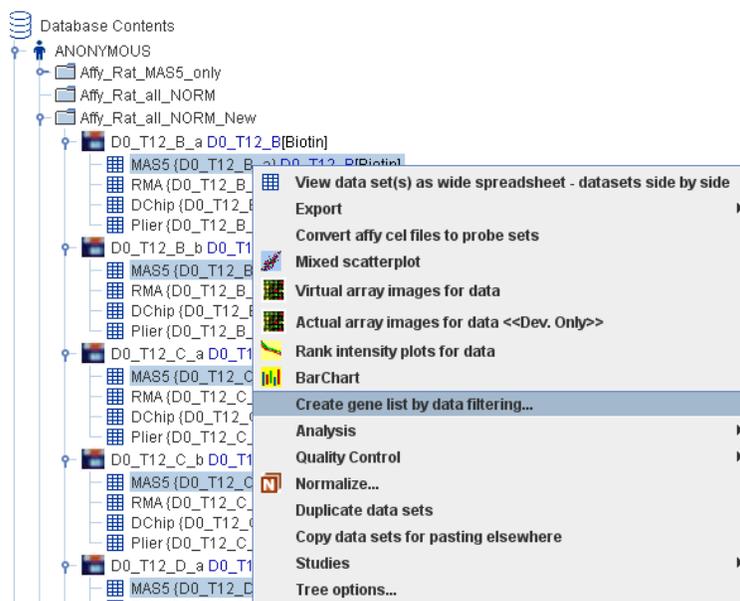


Figure 3-3: Create significant gene list from selected dataset

User can set the criteria (like flag, intensity value) to create a gene list, e.g find out all the genes which intensity is greater than 3.0 in at least 6 out of 12 hybridizations.

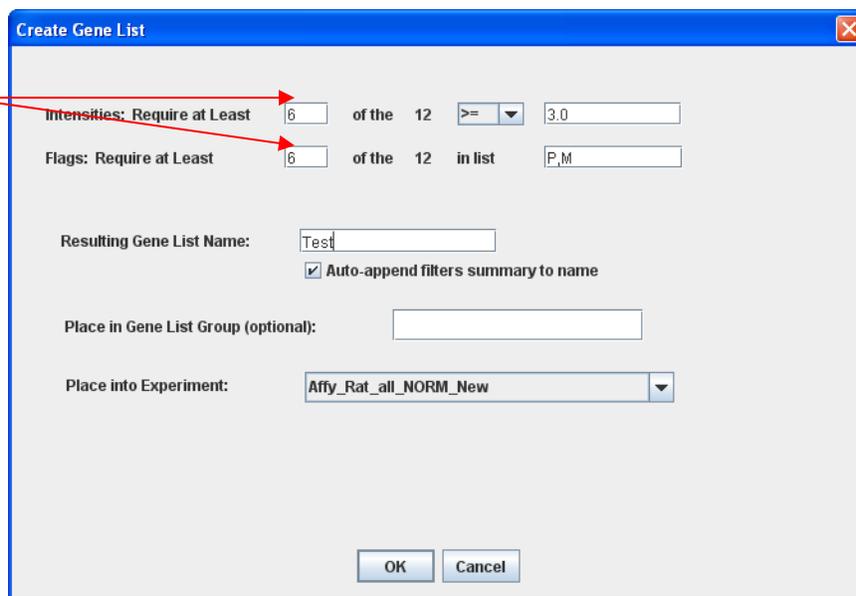


Figure 3-4: Set the criteria for the significant gene list

In

Figure 3-4, the user can filter out low intensity value by setting the criteria, and has the option to place the gene list under any existing experiments. Click OK button then the significant gene list will be created and can be viewed from the Significant Gene List folder. See Figure 3-1.

### 3.3 Display Gene List

Double-clicking the name of a Gene List  will bring out the spreadsheet displaying the gene list. Another way to bring out gene list is right-clicking the experiment-> choosing Gene List-> choosing Display, then a pop-up window will show up and list all the gene lists under the experiment. The user can

choose the gene list to display, see Figure 3-5. If the user selects all the gene lists to display, then the multiple lists will be shown in one spreadsheet.

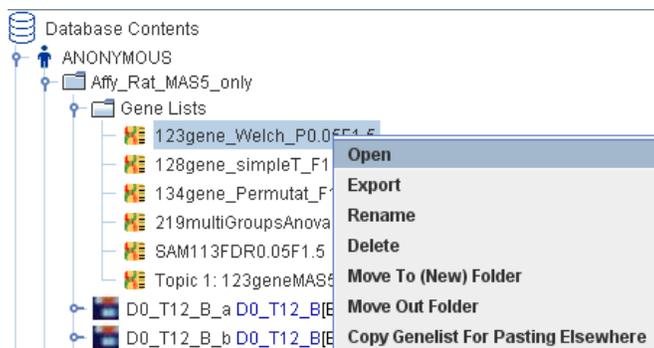
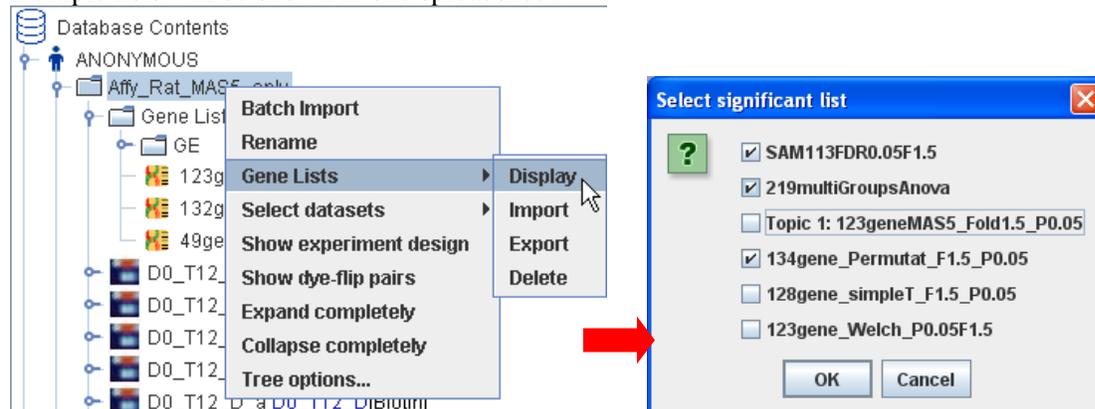


Figure 3-5: select Significant Gene List to display

Filter->	GENELIST_NAME *	EXPID *	GENE	KEGG	LOCUSID	SWISSPROT_ACC_NUMBER
1	123gene_Welch_P0.05F1.5	650	L1507		24891	Q08201 /// Q64714
2	123gene_Welch_P0.05F1.5	650	L2626		81736	Q63369
3	123gene_Welch_P0.05F1.5	650	D87336		287552	A1A5L1
4	123gene_Welch_P0.05F1.5	650	X07365	Gpx1	24404	P04041 /// Q6PDW8
5	123gene_Welch_P0.05F1.5	650	E00778	Cyp1a1	24296	P00185 /// Q80UE2
6	123gene_Welch_P0.05F1.5	650	J02722	Hmox1	24451	P06762
7	123gene_Welch_P0.05F1.5	650	M55534	Cryab	25420	P23928 /// Q63136 /// Q63137 /// Q63138
8	123gene_Welch_P0.05F1.5	650	AA848563	Hspa1a /// Hspa1b	24472	Q07439 /// Q6LA95
9	123gene_Welch_P0.05F1.5	650	Z75029	Hspa1b	294254	Q6LA95
10	123gene_Welch_P0.05F1.5	650	AA818604	Hspa1a /// Hspa1b	24472	Q07439 /// Q6LA95
11	123gene_Welch_P0.05F1.5	650	U01344	Nat1	116631	P50297 /// P50298 /// Q45G59 /// Q45G71
12	123gene_Welch_P0.05F1.5	650	AA108277	Hsph1	288444	Q66HA8
13	123gene_Welch_P0.05F1.5	650	M99169	Rps6ka1	81771	Q63531
14	123gene_Welch_P0.05F1.5	650	A1229655	Ctdsp1	363249	Q3B8P1

Figure 3-6: The gene list is opened in a spreadsheet

In Figure 3-6 there are some buttons at the top for further information associated with selected genes, e.g. KEGG, PathArt, GeneGo MetaCore, etc. If user selects “KEGG”, the following pop-up window will show up. Click “Yes” button to see the pathway for the selected genes.

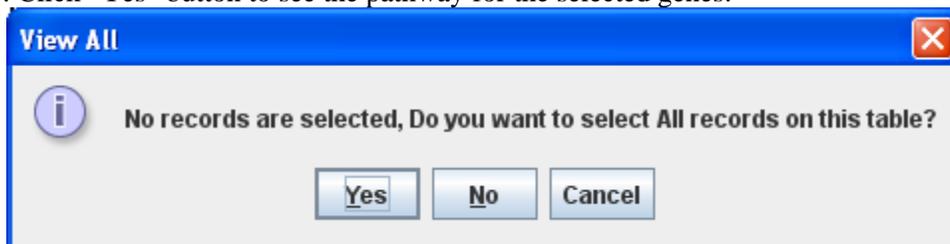


Figure 3-7: from gene list to KEGG

Gene name(LocusID)	Map	Category	Fisher P Value
<ul style="list-style-type: none"> <li>↓ Dusp6(116663)</li> <li>↓ Dusp7(300980)</li> <li>↑ Hspa1a(24472)</li> <li>↓ Map3k1(116667)</li> <li>↓ Mapk14(81649)</li> <li>↓ Mapk9(50658)</li> <li>↓ Mapkapk3(315994)</li> <li>↓ Nfkb1(81736)</li> <li>↓ Nras(24605)</li> <li>↓ Rps6ka1(81771)</li> <li>↓ Taok1(286993)</li> <li>↓ Tgfb3(25717)</li> <li>↓ Tgfbr2(81810)</li> <li>↓ Tnfrsf1a(25625)</li> <li>↓ Trp53(24842)</li> </ul>	MAPK signaling pathway(rno0401...	Regulatory pathway	0.00002427
<ul style="list-style-type: none"> <li>↓ Bcl2l1(24888)</li> <li>↓ Ccnd1(58919)</li> <li>↓ Mapk9(50658)</li> <li>↓ Nfkb1(81736)</li> <li>↓ Tgfb3(25717)</li> <li>↓ Tgfbr2(81810)</li> <li>↓ Trp53(24842)</li> <li>↓ Vegfa(83785)</li> </ul>	Pancreatic cancer(rno05212)	Regulatory pathway	0.00004327
<ul style="list-style-type: none"> <li>↓ Bcl2l1(24888)</li> <li>↓ Ccnd1(58919)</li> <li>↓ Nfkb1(81736)</li> <li>↓ Nfkb1a(25493)</li> <li>↓ Nras(24605)</li> <li>↓ Tgfb3(25717)</li> <li>↓ Tgfbr2(81810)</li> </ul>	Chronic myeloid leukemia(rno05...	Regulatory pathway	0.00007117

Input genes = 121, 65 genes found, 56 not found, Total 97 pathway maps.

Figure 3-8: KEGG pathway for the selected gene list

The user can open multiple 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, (Figure 3-9) 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 3-10.

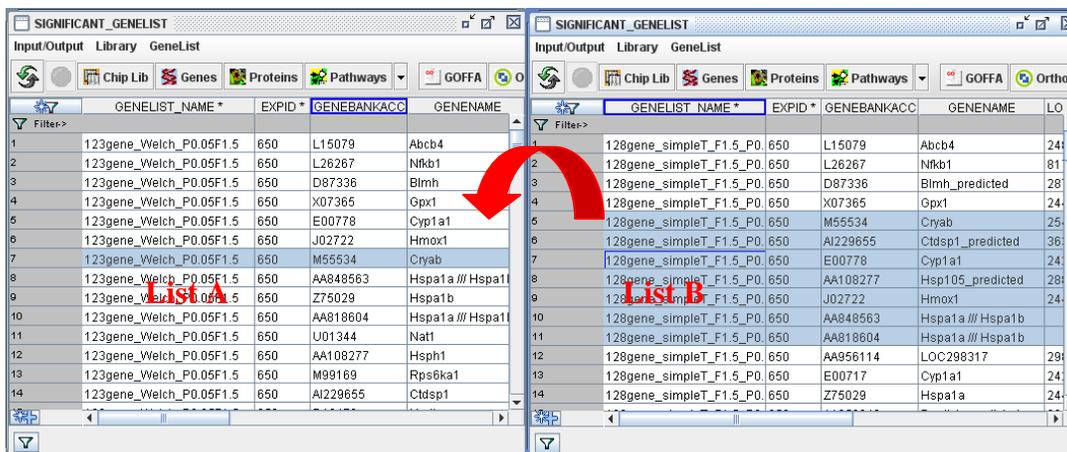


Figure 3-9: combine two lists to get common gene list

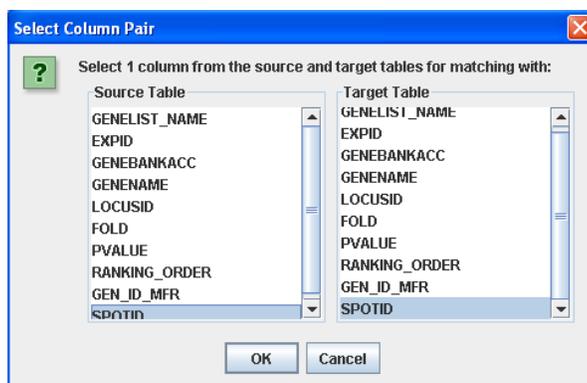


Figure 3-10: choose the column SPOTID for mapping two lists

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** under GeneList pull-down menu and then typing in the explanation. Bar Chart (Chapter 6) can be accessed by right-clicking on selected gene records.

If the gene list has only gene bank accession number, but the user needs to get other ID (e.g. LocusID, etc) to link to KEGG, he can click GeneList pull-down menu and choose 'Sig. gene complete annotation' to re-display the gene list with other kind of IDs. Just be aware that this will take a while if the gene list is very big (>10000).

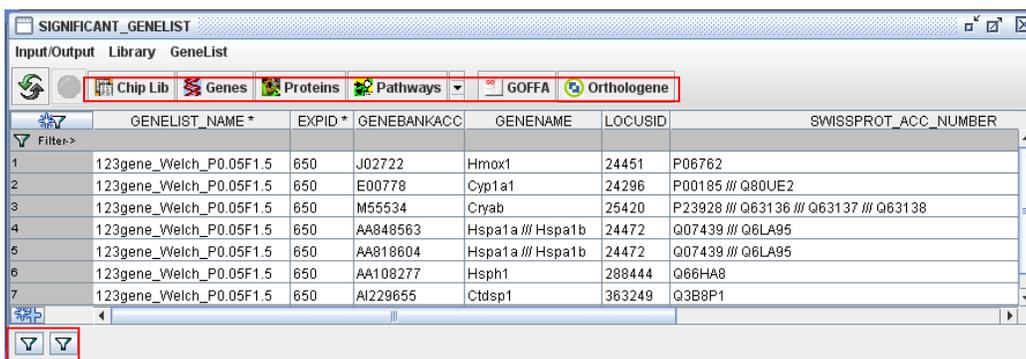


Figure 3-11: The combined gene list result

Figure 3-11 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  $0 < \text{fold} < 1$ , it will be  $-1/\text{Fold}$ ) in Ingenuity analysis results.



Figure 3-12: Fold converting options

See VennDiagram in page 135 for an easier way to obtain common gene list.

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.

### 3.4 Import Gene List

In Figure 3-5 the user can right-click an experiment name and choose Gene List ->Import to import a gene list under the experiment, if he is the owner of the experiment or assigned privilege of creating gene list. The text file for Import can be loaded (Figure 3-13) and the data columns in the text file can be mapped to ArrayTrack database fields. The process of importing Gene List is quite similar to that of importing new array type (Figure 2-29).

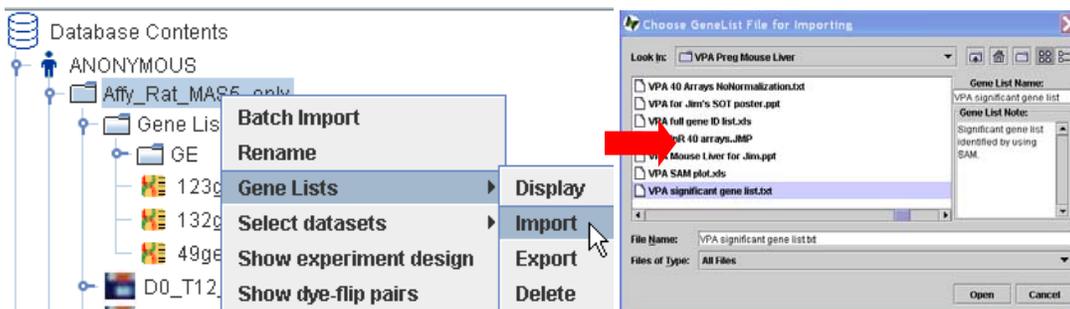


Figure 3-13: Loading a text file with information about Gene List associated with an Experiment.

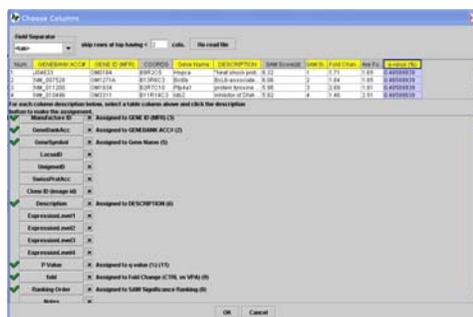


Figure 3-14: Mapping Significant Gene List information to ArrayTrack database fields.

### 3.5 Export Gene List

By right-clicking the gene list->export.

### 3.6 Delete Gene List

By right-clicking the gene list->delete.

The significant gene list can also be moved to another folder, moved out folder, or copied to paste elsewhere. See Figure 3-15.

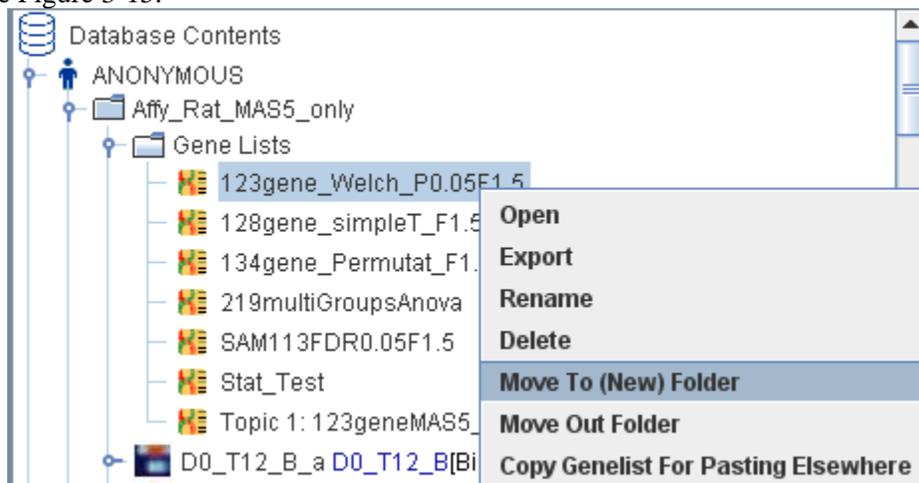


Figure 3-15: Move significant gene list to other location

To perform the above job on gene list, the user has to be the owner of the experiment or assigned the privilege of creating gene lists.

## Chapter 4 Working with Libraries

### 4.1 Overview

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 Library panel (Figure 4-1) 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.

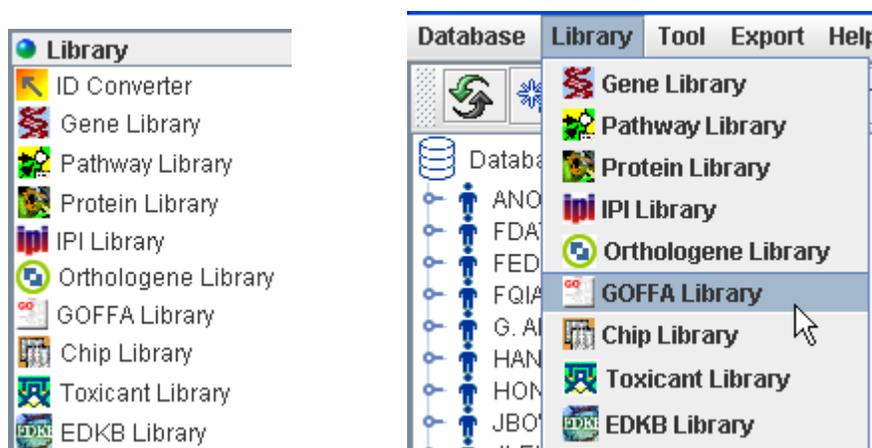


Figure 4-1: Libraries can be accessed from the Library panel or the Library pull-down menu.

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 4-2). 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 4-2: Gene Search panel displays the number of records in Gene Library.

Figure 4-2: Gene Search panel displays the number of records in Gene Library.

## 4.2 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

The main part of the Gene Library is an Excel-like spreadsheet table. As shown in Figure 4-2, 109,483 records are present in the table. 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 4-2).

**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 (🔍) (or pressing the Enter key) will display all the records meeting the filter criteria. In the example shown in Figure 4-3, 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 4-3. Newly inserted filtering rows can be deleted by selecting it, right-clicking on them and then choosing “remove this filter row”.

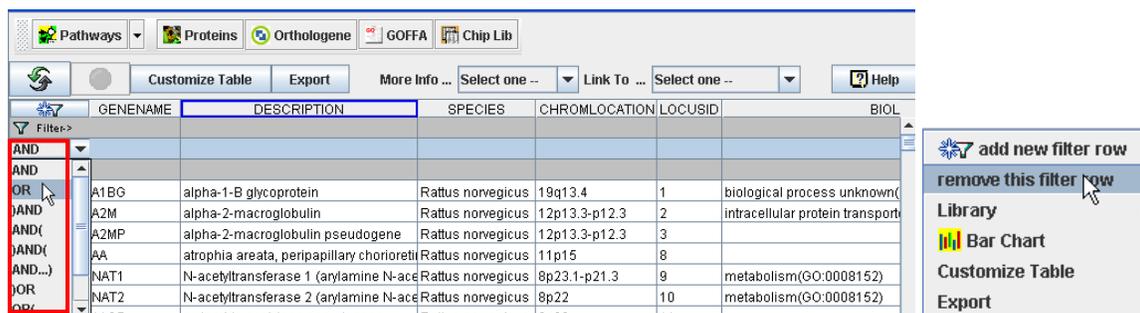


Figure 4-3: Use of a combination of filters to find out interesting genes.

**Sort Table by Column:** The table can be sorted (▲ or ▼) by toggling on the column header and then pressing . 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 4-4A), whereas LOCUSID is searched by numerical order (Figure 4-4B). An additional sorting column can be added to a previously sorted table.

**A**

	GENENAME
3861	39L15
3862	3B1
3863	3P18S
3864	3P18T
3865	3S
3866	3T
3867	3a4
3868	3b11
3869	3f1
3870	3f3
3871	40.MMHAP76FLF4
3872	401e9Sp6
3873	40J4

**B**

	GENENAME	DESCRIPTION	SPECIES	CHROMLOCATION	LOCUSID	
1	A1BG	alpha-1-B glycoprotein	Rattus norvegicus	19q13.4	1	biological proce
2	A2M	alpha-2-macroglobulin	Rattus norvegicus	12p13.3-p12.3	2	intracellular prot
3	A2MP	alpha-2-macroglobulin pseudog	Rattus norvegicus	12p13.3-p12.3	3	
4	AA	atrophia areata, peripapillary ch	Rattus norvegicus	11p15	8	
5	NAT1	N-acetyltransferase 1 (arylamine	Rattus norvegicus	8p23.1-p21.3	9	metabolism(GO
6	NAT2	N-acetyltransferase 2 (arylamine	Rattus norvegicus	8p22	10	metabolism(GO
7	AACP	arylamine acetylase pseudogen	Rattus norvegicus	8p22	11	
8	SERPINA3	serpin peptidase inhibitor, clade	Rattus norvegicus	14q32.1	12	acute-phase res
9	AADAC	arylacetamide deacetylase (este	Rattus norvegicus	3q21.3-q25.2	13	metabolism(GO
10	AAMP	angio-associated, migratory cell	Rattus norvegicus	2q35	14	cell motility(GO
11	AANAT	arylalkylamine N-acetyltransfera	Rattus norvegicus	17q25	15	circadian rhythr

Figure 4-4: 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

There are three simple steps to find functional information for a list of genes using the functions in Box 2 as shown in Figure 4-2:

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 1). 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 4-5 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.

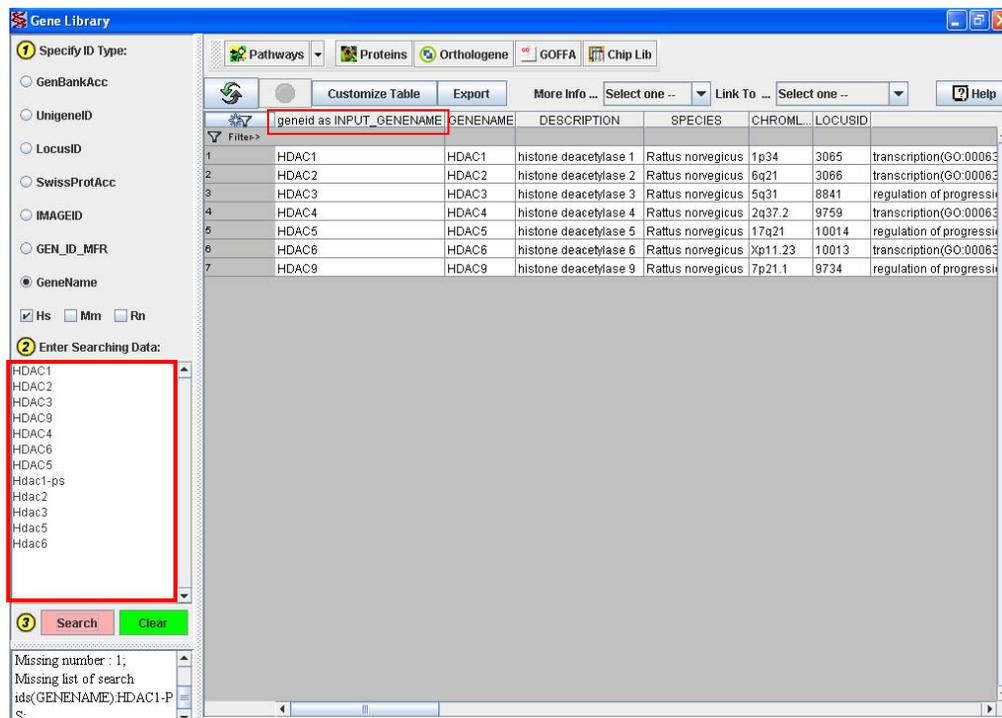


Figure 4-5: Gene Search results by searching against GeneSymbol and Homo sapiens.

**More Information about a Gene 4:** 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 4-6), 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.

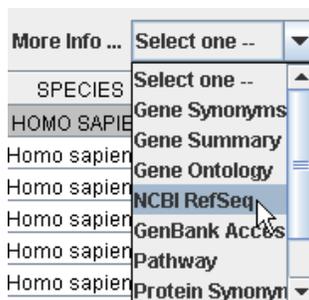


Figure 4-6: More information can be displayed for a selected gene record.

**Link to Other Public Databases 5:** “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 4-7).

	GENENAME	DESCRIPTION		CHROMLOCATIC
1	A1BG	alpha-1-B glycoprotein	Rat	
2	A2M	alpha-2-macroglobulin	Rat	p12.3
3	A2MP	alpha-2-macroglobulin pseudogene	Rat	p12.3
4	AA	atrophia areata, peripapillary chorioretinal degeneration	Rat	
5	NAT1	N-acetyltransferase 1 (arylamine N-acetyltransferase)	Rat	21.3
6	NAT2	N-acetyltransferase 2 (arylamine N-acetyltransferase)	Rattus norvegicus	8p22
7	AACP	arylamide acetylase pseudogene	Rattus norvegicus	8p22
8	SERPINA3	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), m	Rattus norvegicus	14q32.1
9	AADAC	arylacetamide deacetylase (esterase)	Rattus norvegicus	3q21.3-q25.2
10	AAMP	angio-associated, migratory cell protein	Rattus norvegicus	2q35
11	AANAT	arylalkylamine N-acetyltransferase	Rattus norvegicus	17q25
12	AARS	alanyl-tRNA synthetase	Rattus norvegicus	16q22
13	AAVS1	adeno-associated virus integration site 1	Rattus norvegicus	19q13 19q13-qter

Figure 4-7: Official web sites are linked for a selected gene record.

**Sort Genes Based on Their Common Pathway and Proteins 6:** The user can highlight any number of genes in the table and click pull-down triangle and select or , etc to reorder the highlighted genes based on their shared common pathways. Details are discussed under Pathway Library.

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

**Customize Table 7:** The table that displays search results can be customized by clicking on Customize Table button in Figure 4-2. The user can select particular fields to be displayed in the table (see Figure 4-8).

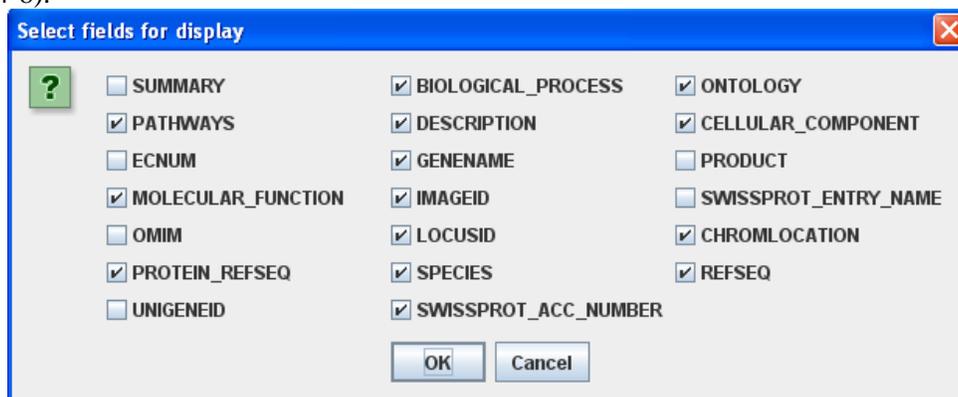


Figure 4-8: Table items that can be chosen for display in the Gene Search results table.

**Export Table Contents to a Local File 8:** 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 4-9 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.

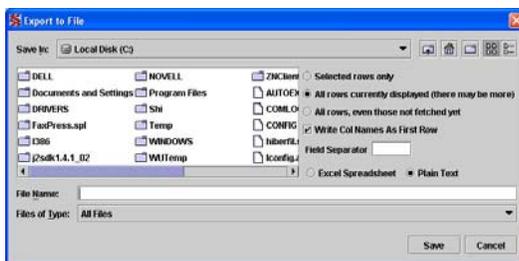


Figure 4-9: Export of search results into a local text or Excel file.

**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 “Copy selected columns on selected rows to clipboard”. See Figure 4-10.

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

**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 4-10).

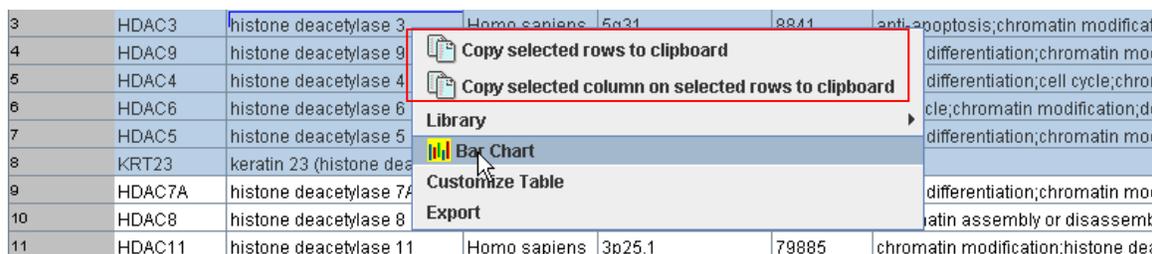


Figure 4-10: Launch Bar Chart for selected gene records.

User can also click Help  to get information about the libraries (Figure 4-11).

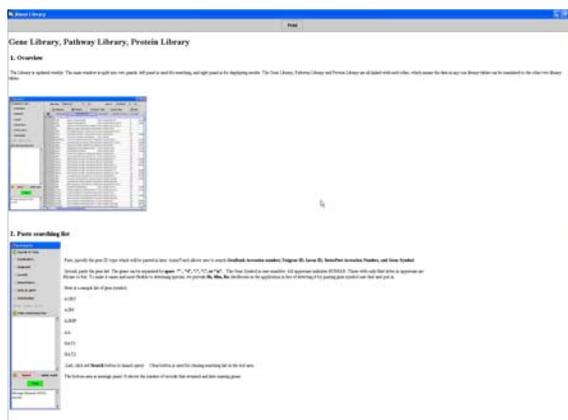


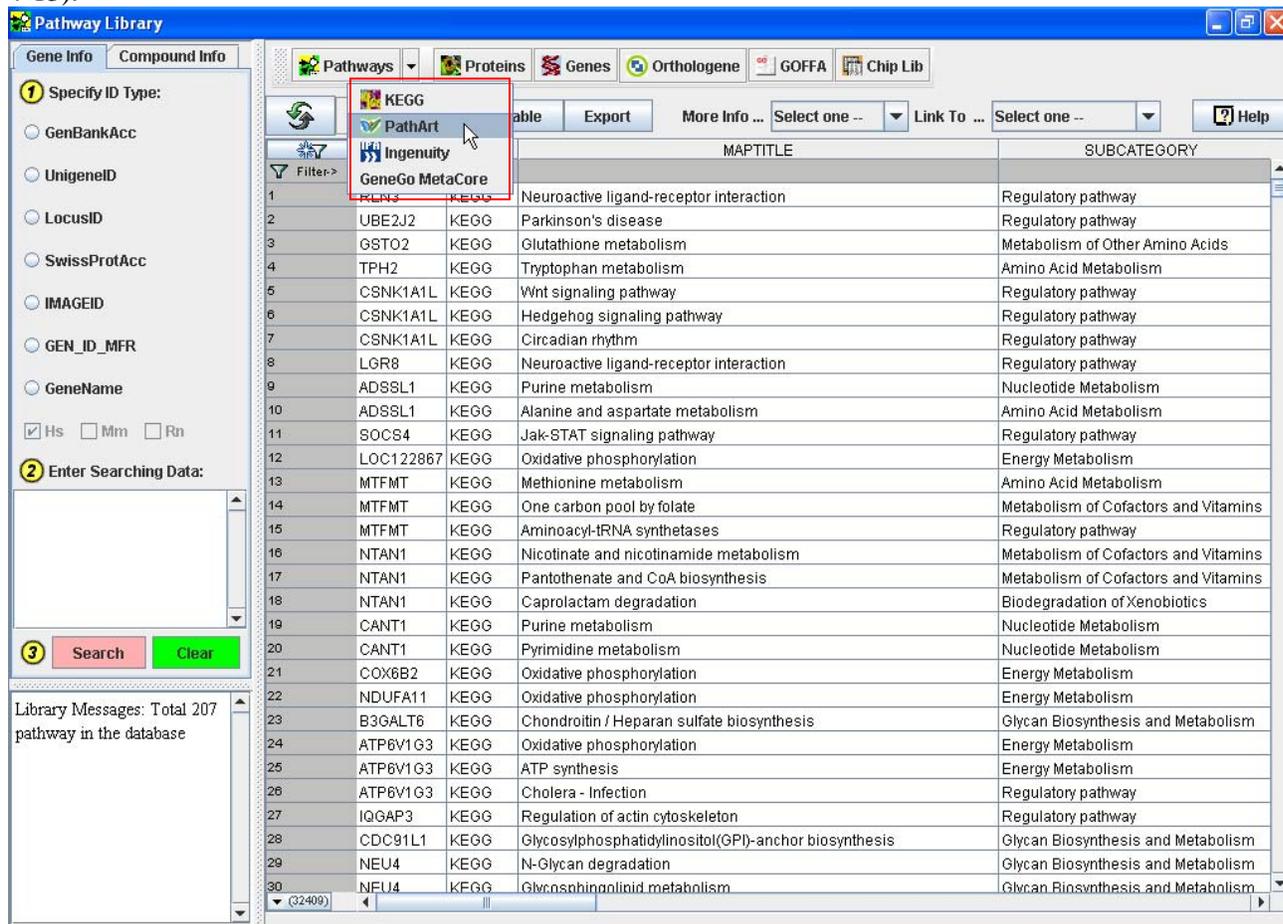
Figure 4-11: Information on the libraries under Help menu

### 4.3 Pathway Library

The Pathway Search panel can be activated by double-clicking on  Pathway Library in the Library panel or Library pull-down menu (Figure 4-1). But users usually access Pathway Library after creating significant gene list for further biological interpretation, see Chapter 3 about Gene list.

Figure 4-12 shows that pathway information is available in ArrayTrack. This library contains pathway information from KEGG, ParthArt, Ingenuity and GeneGo MetaCore. ArrayTrack only provides the portal for Ingenuity and GeneGo, users need to contact these two companies to get his account to access these two tools.

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 4-2). 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 4-13).



The screenshot shows the Pathway Library window with the following components:

- Left Sidebar:**
  - Specify ID Type:** Radio buttons for GenBankAcc, UnigeneID, LocusID, SwissProtAcc, IMAGEID, GEN\_ID\_MFR, GeneName. Checkboxes for Hs, Mm, Rn.
  - Enter Searching Data:** A text input field with Search and Clear buttons.
  - Library Messages:** Total 207 pathway in the database.
- Top Navigation:** Pathways (selected), Proteins, Genes, Orthologene, GOFFA, Chip Lib.
- Table Headers:** MAPTITLE, SUBCATEGORY.
- Table Content (Sample Rows):**

ID	Source	MAPTITLE	SUBCATEGORY
1	KEGG	Neuroactive ligand-receptor interaction	Regulatory pathway
2	KEGG	Parkinson's disease	Regulatory pathway
3	KEGG	Glutathione metabolism	Metabolism of Other Amino Acids
4	KEGG	Tryptophan metabolism	Amino Acid Metabolism
5	KEGG	Wnt signaling pathway	Regulatory pathway
6	KEGG	Hedgehog signaling pathway	Regulatory pathway
7	KEGG	Circadian rhythm	Regulatory pathway
8	KEGG	Neuroactive ligand-receptor interaction	Regulatory pathway
9	KEGG	Purine metabolism	Nucleotide Metabolism
10	KEGG	Alanine and aspartate metabolism	Amino Acid Metabolism
11	KEGG	Jak-STAT signaling pathway	Regulatory pathway
12	KEGG	Oxidative phosphorylation	Energy Metabolism
13	KEGG	Methionine metabolism	Amino Acid Metabolism
14	KEGG	One carbon pool by folate	Metabolism of Cofactors and Vitamins
15	KEGG	Aminoacyl-tRNA synthetases	Regulatory pathway
16	KEGG	Nicotinate and nicotinamide metabolism	Metabolism of Cofactors and Vitamins
17	KEGG	Pantothenate and CoA biosynthesis	Metabolism of Cofactors and Vitamins
18	KEGG	Caprolactam degradation	Biodegradation of Xenobiotics
19	KEGG	Purine metabolism	Nucleotide Metabolism
20	KEGG	Pyrimidine metabolism	Nucleotide Metabolism
21	KEGG	Oxidative phosphorylation	Energy Metabolism
22	KEGG	Oxidative phosphorylation	Energy Metabolism
23	KEGG	Chondroitin / Heparan sulfate biosynthesis	Glycan Biosynthesis and Metabolism
24	KEGG	Oxidative phosphorylation	Energy Metabolism
25	KEGG	ATP synthesis	Energy Metabolism
26	KEGG	Cholera - Infection	Regulatory pathway
27	KEGG	Regulation of actin cytoskeleton	Regulatory pathway
28	KEGG	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	Glycan Biosynthesis and Metabolism
29	KEGG	N-Glycan degradation	Glycan Biosynthesis and Metabolism
30	KEGG	Glycosaminoglycan metabolism	Glycan Biosynthesis and Metabolism

Figure 4-12: Pathway Library window

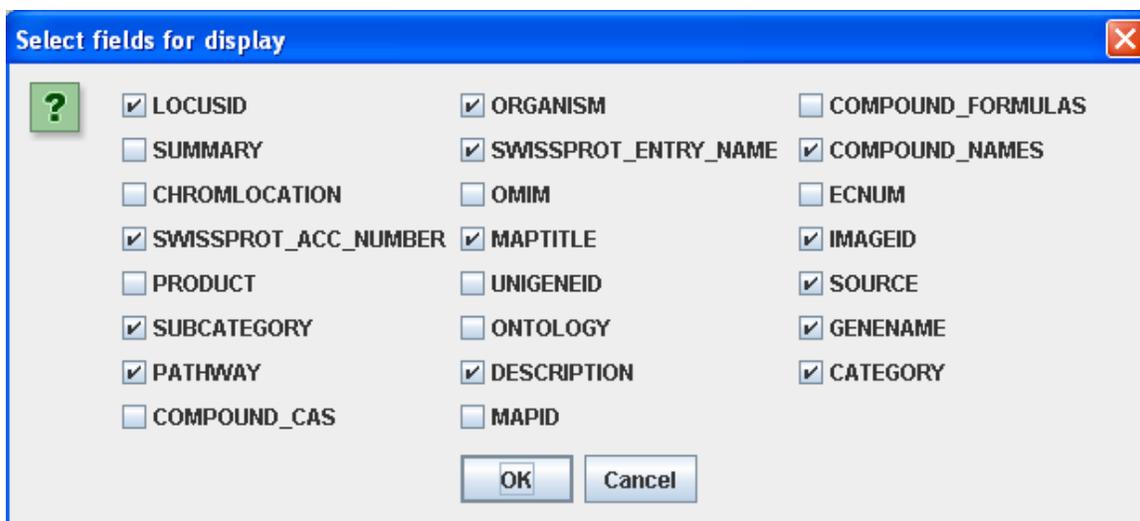


Figure 4-13: Select fields for display

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 4-14).

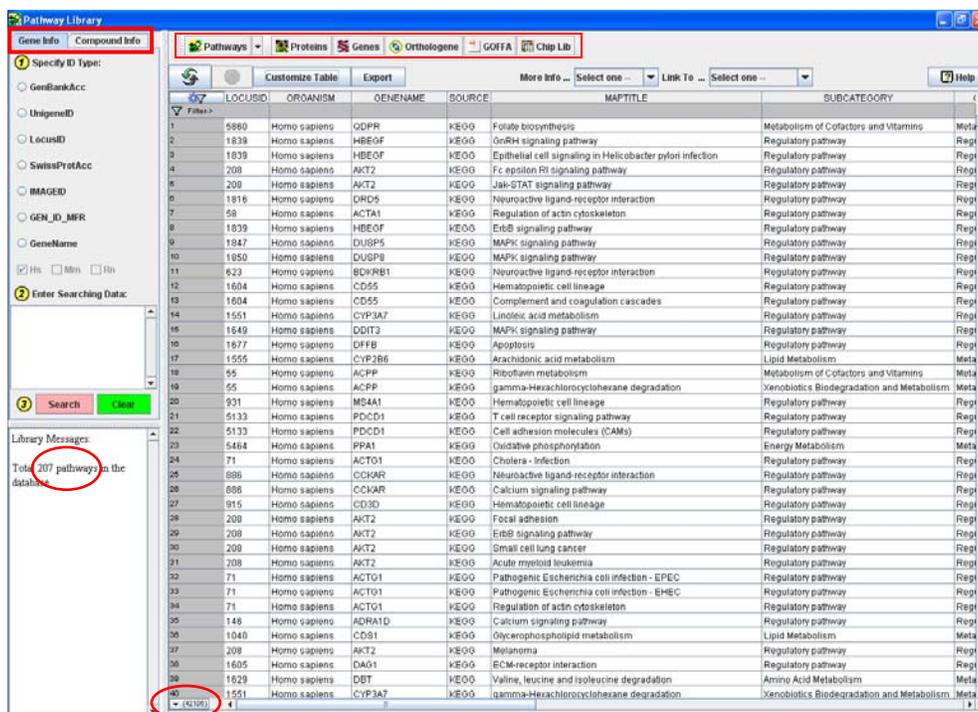


Figure 4-14: Pathway Search panel for gene information.

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 4-15.

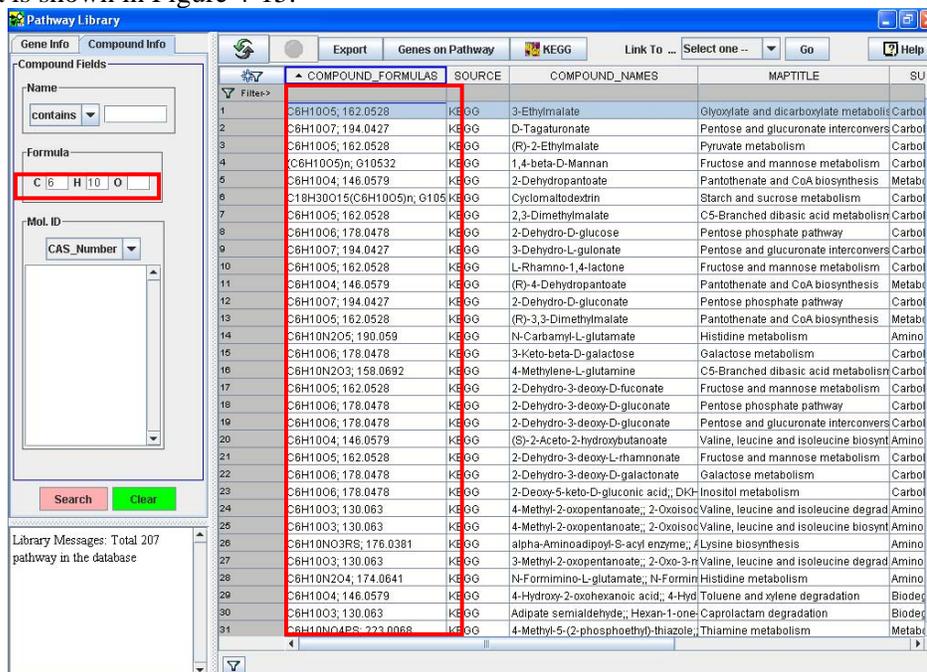


Figure 4-15: Pathway searching panel for compound information

As for the Gene Library, the user can rearrange the table columns by dragging and moving the column header. 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 Pathways pull-down triangle and choose Kegg, or PathArt button; since Kegg and ParthArt are different libraries, probably different search results can be expected.

In Figure 4-16, type in “citrate cycle” in the first row under “MAPTITLE”, then click fresh button to get a list of filtered genes.

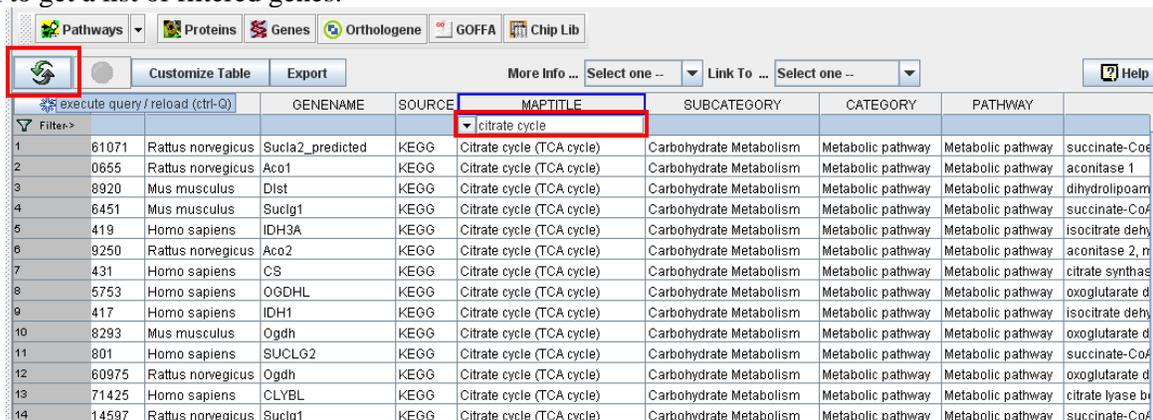


Figure 4-16: type “citrate cycle” in the first row to filter genes

In Figure 4-17, highlight some of the filtered genes then click “Pathway” button to select KEGG.

	GENENAME	SOURCE	MAPTITLE	SUBCATEGORY
1	61071 Rattus norvegicus	Sucla2_predicted	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
2	0655 Rattus norvegicus	Aco1	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
3	8920 Mus musculus	Dist	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
4	6451 Mus musculus	Suc1g1	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
5	419 Homo sapiens	IDH3A	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
6	9250 Rattus norvegicus	Aco2	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
7	431 Homo sapiens	CS	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
8	5753 Homo sapiens	OGDHL	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
9	417 Homo sapiens	IDH1	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
10	8293 Mus musculus	Ogdh	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
11	801 Homo sapiens	SUCLG2	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
12	60975 Rattus norvegicus	Ogdh	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
13	71425 Homo sapiens	CLYBL	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
14	14597 Rattus norvegicus	Suc1g1	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
15	4159 Rattus norvegicus	Acly	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
16	420 Homo sapiens	IDH3B	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
17	5929 Mus musculus	ldh3g	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
18	6945 Mus musculus	Sdha	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
19	4551 Mus musculus	Pck2	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
20	69951 Mus musculus	ldh2	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
21	98596 Rattus norvegicus	Sdhb_predicted	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
22	391 Homo sapiens	SDHC	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
23	83398 Homo sapiens	LOC283398	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
24	8563 Mus musculus	Pcx	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
25	8 Homo sapiens	ACO1	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
26	967 Homo sapiens	OGDH	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism
27	61042 Rattus norvegicus	Pck2_predicted	KEGG Citrate cycle (TCA cycle)	Carbohydrate Metabolism

Figure 4-17: Select genes for Pathway (KEGG) Search.

All pathway information related to these genes is displayed in three separated tables respectively for human, rat and mouse, as shown in Figure 4-18. At the bottom of the pathway table, a summary on pathway information is provided: *Total submitted genes, number of genes found, number of genes not found, Total number of pathway maps*. The Pathway table is organized by gene name, Locus ID, pathway map, the category of the pathway and Fisher P value.

Gene name(LocusID)	Map	Category	Fisher P Value
CS(1431)			
IDH3A(3419)	Citrate cycle (TCA cycle)(hsa00020)	Carbohydrate Metabolism/Metabolism	2.9E-7
OGDHL(55753)			
CS(1431)	Glyoxylate and dicarboxylate...	Carbohydrate Metabolism/Metabolism	0.00993605
OGDHL(55753)	Lysine degradation(hsa00030)	Amino Acid Metabolism/Metabolism	0.02821369
OGDHL(55753)	Tryptophan metabolism(hsa00003)	Amino Acid Metabolism/Metabolism	0.04211932

Input genes = 8, 3 genes found, 5 not found, Total 4 pathway maps.

Gene name(LocusID)	Map	Category	Fisher P Value
Dist(78920)	Citrate cycle (TCA cycle)(mmu00020)	Carbohydrate Metabolism/Metabolism	0.00004187
Suc1g1(56451)			
Suc1g1(56451)	Propanoate metabolism(mmu00030)	Carbohydrate Metabolism/Metabolism	0.01460012
Dist(78920)	Lysine degradation(mmu00030)	Amino Acid Metabolism/Metabolism	0.01605423

Input genes = 8, 2 genes found, 6 not found, Total 3 pathway maps.

Gene name(LocusID)	Map	Category	Fisher P Value
Aco1(50655)			
Aco2(79250)	Reductive carboxylate cycle (CO2 fi...)	Energy Metabolism/Metabolic p...	1E-8
Sucla2_predicted(361071)			
Aco1(50655)			
Aco2(79250)	Citrate cycle (TCA cycle)(rno00020)	Carbohydrate Metabolism/Metabolism	2.0E-7
Sucla2_predicted(361071)			
Aco1(50655)			
Aco2(79250)	Glyoxylate and dicarboxylate metab...	Carbohydrate Metabolism/Metabolism	0.00001416

Input genes = 8, 3 genes found, 5 not found, Total 4 pathway maps.

Figure 4-18: Search results from KEGG



In Figure 4-20, the user can highlight and double-click one record, (got error message, will add screen shot later)

Click the pathway name under the “Result” tab in the left panel, the pathway will show up in the right panel. See Figure 4-21.

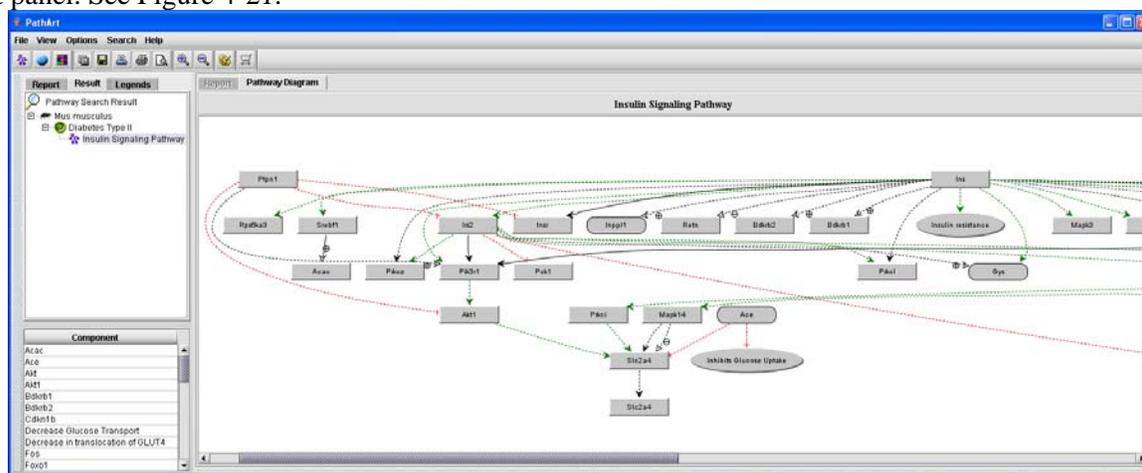


Figure 4-21: Pathway from PathArt

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 4-18.

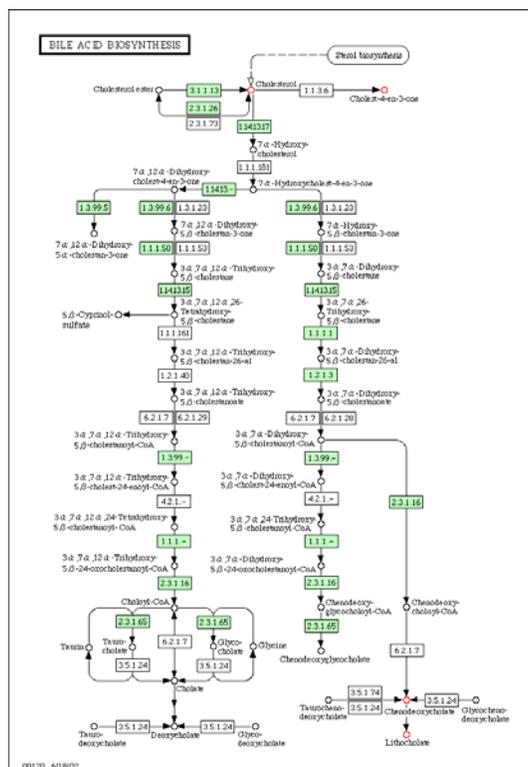


Figure 4-22: KEGG pathway map (hsa00120) on Bile acid biosynthesis. Query compounds are highlighted in red cycles.

### 4.5 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 4-23 shows that protein information is available for 41,795 records in ArrayTrack. Also same as mentioned in Pathway Library, users usually access Protein Library after creating significant gene list for further biological interpretations.

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.

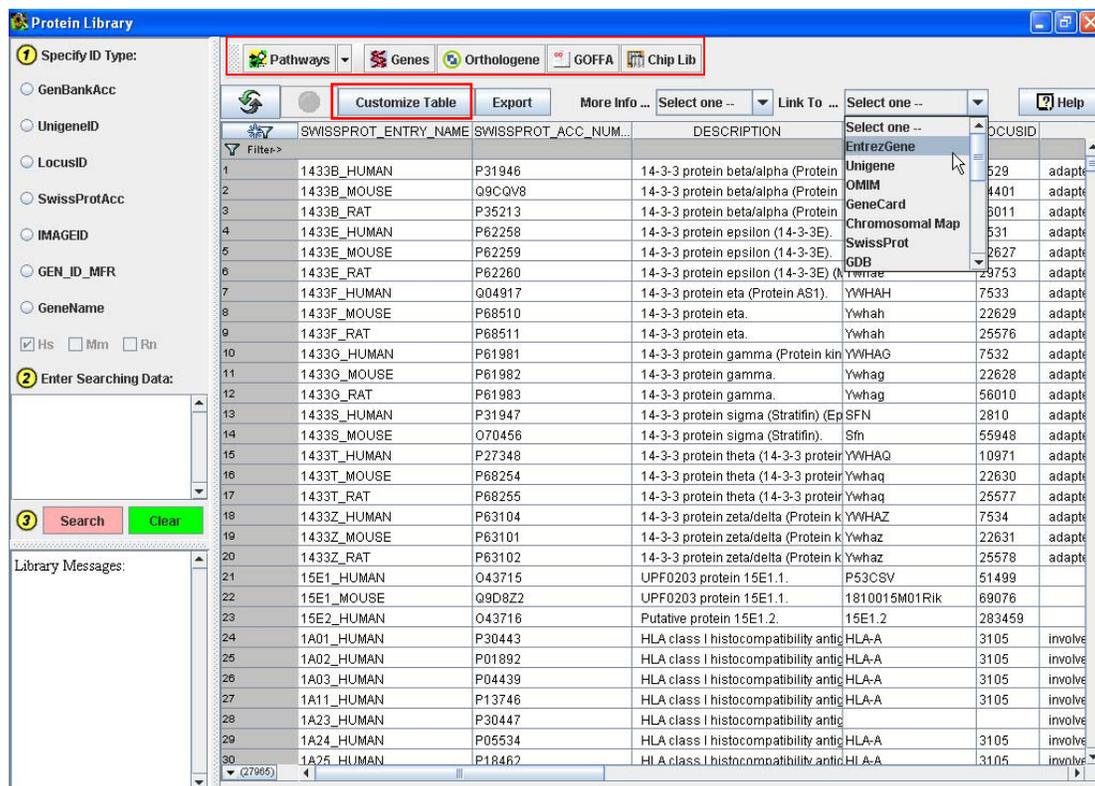


Figure 4-23: Protein Search panel.

The contents of the table can be customized by selecting data items shown in the Figure 4-24.

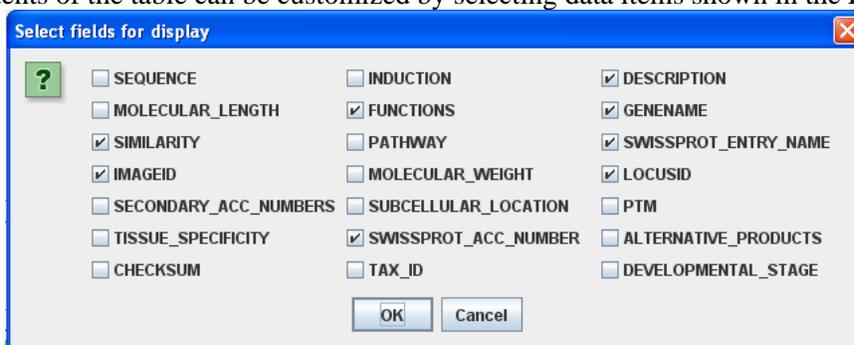


Figure 4-24: Items that can be chosen for display in the Protein Search result table.

**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 4-23).

#### 4.6 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 4-25) is similar to that of Gene Search and Protein Search.

The screenshot shows the IPI Library search panel. On the left, there is a 'Specify ID Type:' section with radio buttons for IPI, GenBankAcc, UnigeneID, LocusID, SwissProtAcc, IMAGEID, GEN\_ID\_MFR, and GeneName. Below this is an 'Enter Searching Data:' field with 'Search' and 'Clear' buttons. The main area is a table with columns: IPINAME, DESCRIPTION, SWISS, and SUBCELLULAR. A dropdown menu is open over the table, showing options: Select one --, Gene Synonyms, Gene Summary, Gene Ontology, NCBI RefSeq, GenBank Access, Pathway, and Protein Synonym. The table contains 30 rows of protein data.

IPINAME	DESCRIPTION	SWISS	SUBCELLULAR
1	IPI00000011	HOMOLOG.	
2	IPI00000005	GTPASE NRAS.	P011
3	IPI00000006	GTPASE HRAS.	P011
4	IPI00000012	ZINC TRANSPORTER 8.	
5	IPI00000013	CATHEPSIN L2 PRECURSOR.	O609
6	IPI00000015	SPLICING FACTOR, ARGININE/SERINE-F	Q08170
7	IPI00000017	ORF1 5' TO PD-ECGF/TP PROTEIN.	
8	IPI00000020	ORF3 5' OF PD-ECGF/TP PROTEIN.	
9	IPI00000021	ORF2 5' TO PD-ECGF/TP PROTEIN.	
10	IPI00000023	GAMMA-AMINOBUTYRIC ACID A RECEPT	
11	IPI00000024	SPLICE ISOFORM 1 OF PROTOCADHERIN	Q08174
12	IPI00000026	1G.	Q96NK5
13	IPI00000027	PITUITARY ADENYLATE CYCLASE ACTIV	P18509
14	IPI00000030	KDA REGULATORY SUBUNIT, DELTA ISO	Q14738
15	IPI00000033	PROTEIN AF-9.	P42568
16	IPI00000035	VOMERONASAL TYPE-1 RECEPTOR 3.	Q9BXE9
17	IPI00000041	RHO-RELATED GTP-BINDING PROTEIN	P62745
18	IPI00000043	VERY HYPOTHETICAL BLYM-1 PROTO-O	P01124
19	IPI00000044	PLATELET-DERIVED GROWTH FACTOR	P01127
20	IPI00000045	PRECURSOR.	P18510
21	IPI00000046	BWRT PROTEIN.	
22	IPI00000047	HYPOTHETICAL PROTEIN (FRAGMENT).	
23	IPI00000048	INTERLEUKIN-20 PRECURSOR.	Q9NYY1
24	IPI00000049	TNFSF10 PROTEIN.	P50591
25	IPI00000051	PREFOLDIN SUBUNIT 1.	O60925
26	IPI00000057	CONSERVED OLIGOMERIC GOLGI COM	Q14746
27	IPI00000058	INWARD RECTIFIER POTASSIUM CHANN	O60928
28	IPI00000059	LFA-3 (FRAGMENT).	
29	IPI00000060	SUBUNIT.	Q9Y3K8
30	IPI00000070	LOW-DENSITY LIPOPROTEIN RECEPTO	P01130

Figure 4-25: IPI Library search panel

#### 4.7 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.

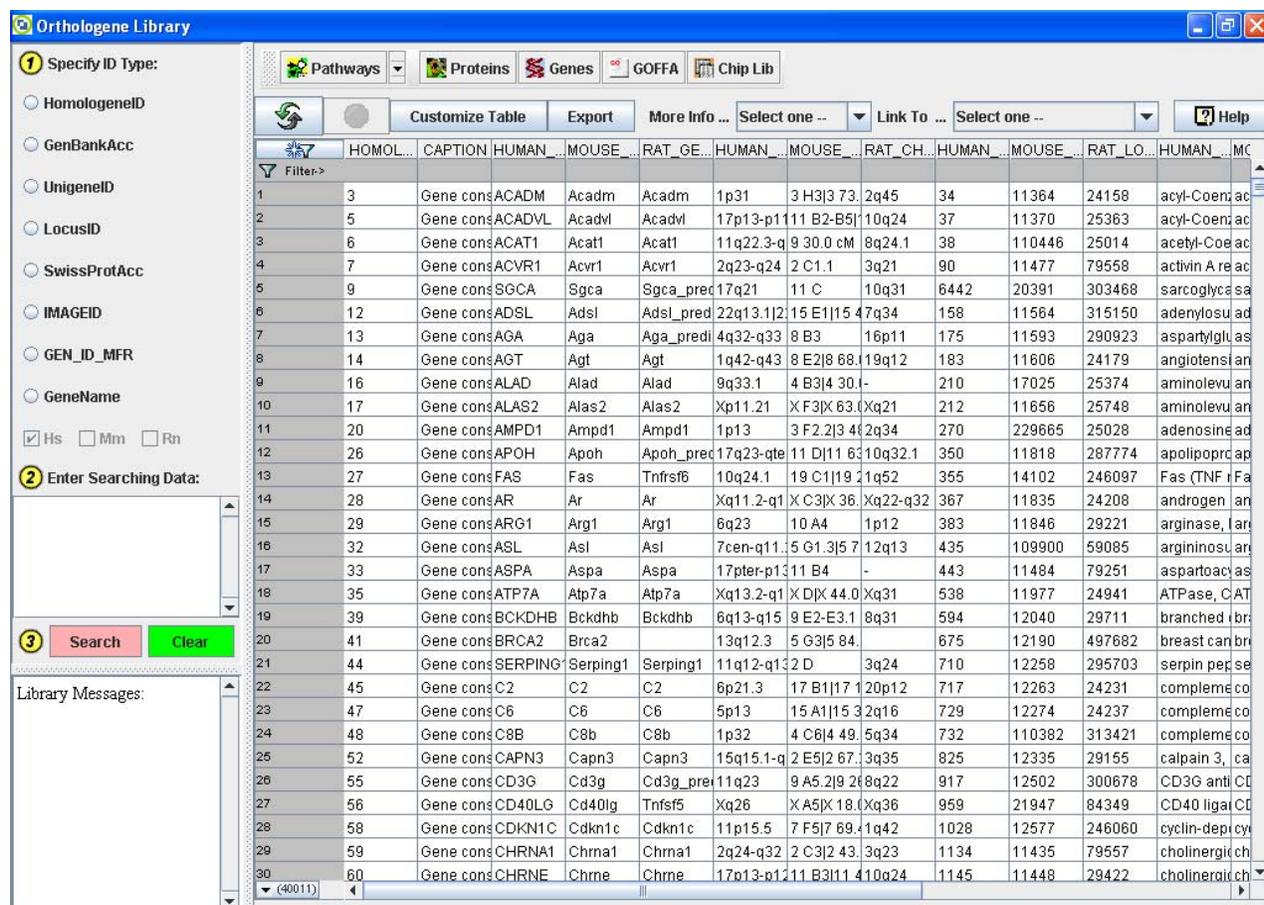


Figure 4-26: Orthologene search panel.

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 are different, and consist of website databases related to cross-species, chromosome location across species, gene homology and species-specific genome information (see Figure 4-27).

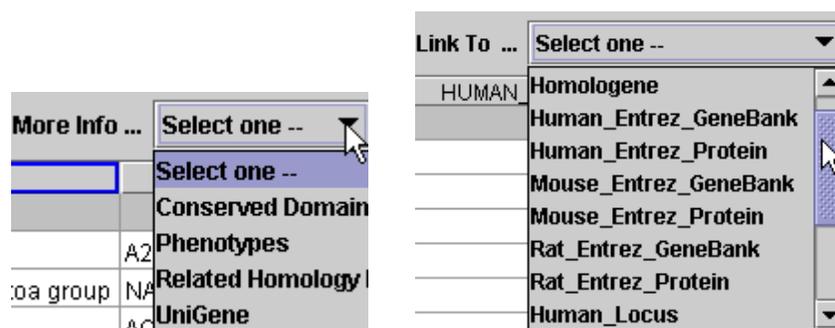


Figure 4-27: More links to other resources

### 4.8 GOFFA Library

The GOFFA (Gene Ontology For Functional Analysis) 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 or common way to access GOFFA is through Gene Library or through gene list.

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 4-28.

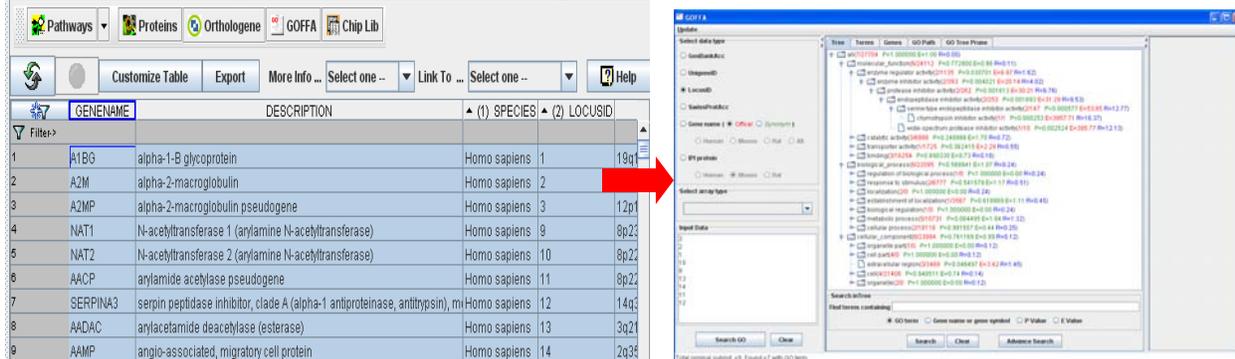


Figure 4-28: Accessing GOFFA via Gene Library

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 five tabs categorized as: 1) Tree. 2) Term Clustering. 3) All Genes. 4) GO Path Plot. 5) Go Tree Prune.

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 4-29.

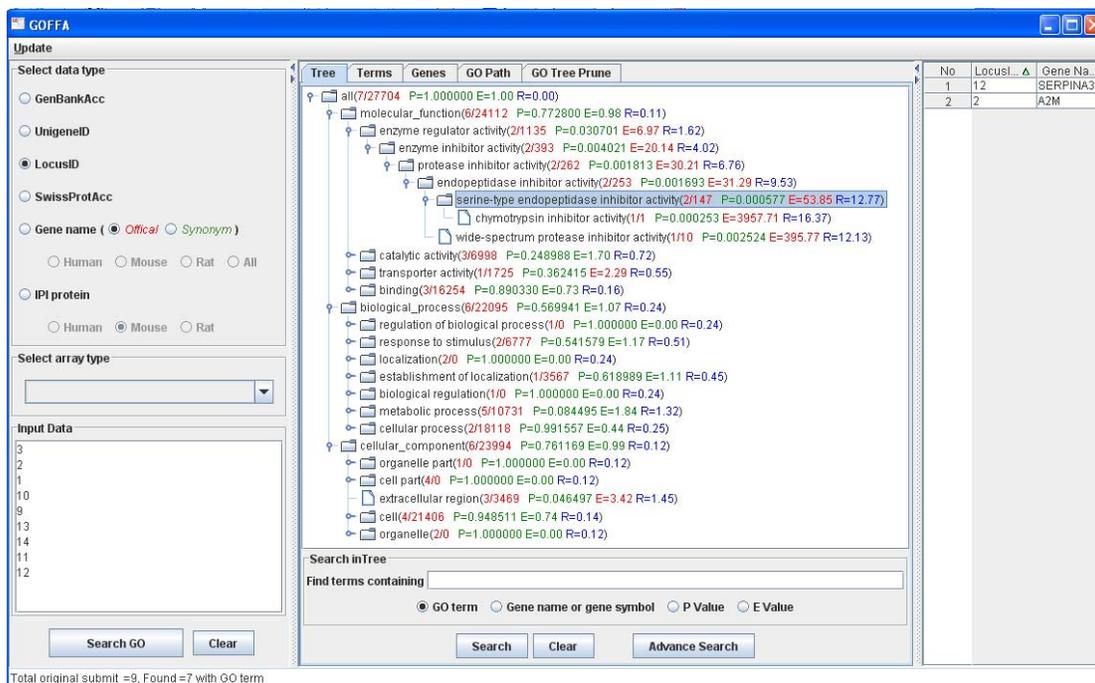


Figure 4-29: GOFFA Library search panel

$$\text{Enrichment factor} = (n_i/N) / (g_i/G) \quad (1.1)$$

Where  $n_i$  is the number of hit genes in term  $i$ .  $N$  is the number of input genes.  $g_i$  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 Tree tab panel, the user can enter a term in the “Find terms containing” text box, choose search by “GO term” or “Gene name or gene symbol”, then click search button, the term contained at all levels of the tree will be highlighted in blue color. See Figure 4-29. Also the user can search by filtering out some terms by setting the P-value or E value criteria. More than that, the user can do advance search by clicking button “Advance Search”. See Figure 4-30.

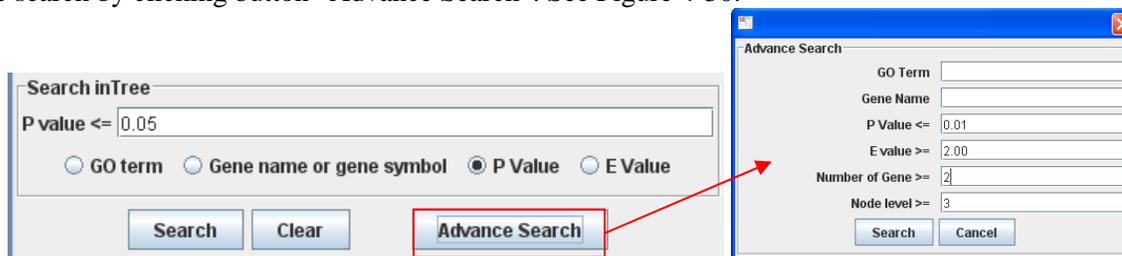


Figure 4-30: Search tree according to P-value or E value

Under Term Clustering tab: there are three sub-categorized tabs (molecular function, biological process and cellular component) each having a spreadsheet with 7 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, Gene Hits (the number of the gene products associated with the term) and E(enrichment) value. 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	GO Tree Prune			
Molecular function		Biological process		Cellular component			
No	Term	GO ID	Average Level	Average P value Δ	Gene Hits	E value	
1	N-acetyltransferase activity	GO:0008080	8.00	0.000063	2.00	160.99	▲
2	N-acyltransferase activity	GO:0016410	7.00	0.000083	2.00	140.55	
3	acetyltransferase activity	GO:0016407	7.00	0.000111	2.00	121.30	
4	aralkylamine N-acetyltransferase activity	GO:0004059	9.00	0.000226	1.00	4427.33	
5	chymotrypsin inhibitor activity	GO:0030569	7.00	0.000226	1.00	4427.33	
6	acyltransferase activity	GO:0008415	6.00	0.000615	2.00	51.48	
7	transferase activity, transferring groups other th...	GO:0016747	5.00	0.000629	2.00	50.89	
8	transferase activity, transferring acyl groups	GO:0016746	4.00	0.000703	2.00	48.12	
9	alanine-tRNA ligase activity	GO:0004813	7.00	0.000903	1.00	1106.83	
10	arylamine N-acetyltransferase activity	GO:0004060	9.00	0.001129	1.00	885.47	
11	tRNA binding	GO:0000049	5.00	0.004059	1.00	245.96	
12	deacetylase activity	GO:0019213	4.00	0.005859	1.00	170.28	
13	heparin binding	GO:0008201	6.00	0.019937	1.00	49.75	
14	tRNA ligase activity	GO:0004812	6.00	0.021047	1.00	47.10	
15	ligase activity, forming aminoacyl-tRNA and rel...	GO:0016876	5.00	0.021047	1.00	47.10	
16	RNA ligase activity	GO:0008452	5.00	0.021047	1.00	47.10	
17	ligase activity, forming carbon-oxygen bonds	GO:0016875	4.00	0.021047	1.00	47.10	
18	ligase activity, forming phosphoric ester bonds	GO:0016886	4.00	0.023485	1.00	42.17	
19	glycosaminoglycan binding	GO:0005539	5.00	0.025919	1.00	38.17	
20	polysaccharide binding	GO:0030247	4.00	0.029008	1.00	34.06	

Figure 4-31: Under Term Clustering tab in GOFFA library

Under All Genes tab: there are 7 columns same as under term clustering tab, plus a Gene column for each spreadsheet categorized under molecular function, biological process and cellular component. The 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. From here users can also create significant gene list by highlighting some gene records and right-clicking -> choosing "Create significant gene list". Double-click on any row will bring up tree view with highlighted gene number in blue on any associated term.

Tree	Terms	Genes	GO Path	GO Tree Prune			
Molecular function		Biological process		Cellular component			
No	Gene	Term	GO ID	Level (Average)	P value(Average) Δ	Gene Hits	E value
1	Ephx2	phosphoric ester hy...	GO:0042578	5.00	0.009057	5.00	3.44
2	Ctdsp1	phosphoric ester hy...	GO:0042578	5.00	0.009057	5.00	3.44
3	Hmox1	phosphoric ester hy...	GO:0042578	5.00	0.009057	5.00	3.44
4	Dusp7	phosphoric ester hy...	GO:0042578	5.00	0.009057	5.00	3.44
5	Dusp6	phosphoric ester hy...	GO:0042578	5.00	0.009057	5.00	3.44
6	Abcb4	drug transporter acti...	GO:0015238	3.00	0.014467	2.00	8.26
7	Abcb1b	drug transporter acti...	GO:0015238	3.00	0.014467	2.00	8.26
8	Abcb4	xenobiotic transport...	GO:0042910	3.00	0.014467	2.00	8.26
9	Abcb1b	xenobiotic transport...	GO:0042910	3.00	0.014467	2.00	8.26
10	Abcb4	multidrug transporte...	GO:0015239	4.00	0.014467	2.00	8.26
11	Abcb1b	multidrug transporte...	GO:0015239	4.00	0.014467	2.00	8.26
12	Abcb4	xenob...	GO:0042910	3.00	0.014467	2.00	8.26
13	Abcb1b	xenob...	GO:0042910	3.00	0.014467	2.00	8.26
14	Nfkb1	heat s...	GO:0031072	4.00	0.015504	4.00	3.67
15	Nfkb1a	heat s...	GO:0031072	4.00	0.015504	4.00	3.67
16	Dnaja1	heat shock protein b...	GO:0031072	4.00	0.015504	4.00	3.67
17	Dnaja4	heat shock protein b...	GO:0031072	4.00	0.015504	4.00	3.67
18	Ephx2	phosphoric monoes...	GO:0016791	6.00	0.033600	4.00	3.00
19	Ctdsp1	phosphoric monoes...	GO:0016791	6.00	0.033600	4.00	3.00
20	Dusp7	phosphoric monoes...	GO:0016791	6.00	0.033600	4.00	3.00
21	Dusp6	phosphoric monoes...	GO:0016791	6.00	0.033600	4.00	3.00
22	Ephx2	hydrolase activity, ac...	GO:0016788	4.00	0.049021	7.00	1.99
23	Pold1	hydrolase activity, ac...	GO:0016788	4.00	0.049021	7.00	1.99
24	Ctdsp1	hydrolase activity, ac...	GO:0016788	4.00	0.049021	7.00	1.99
25	Acot7	hydrolase activity, ac...	GO:0016788	4.00	0.049021	7.00	1.99
26	Hmox1	hydrolase activity, ac...	GO:0016788	4.00	0.049021	7.00	1.99

Figure 4-32: 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.

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

Where  $P_i$  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 4-33.

In Figure 4-33A and B, 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 4-33C.

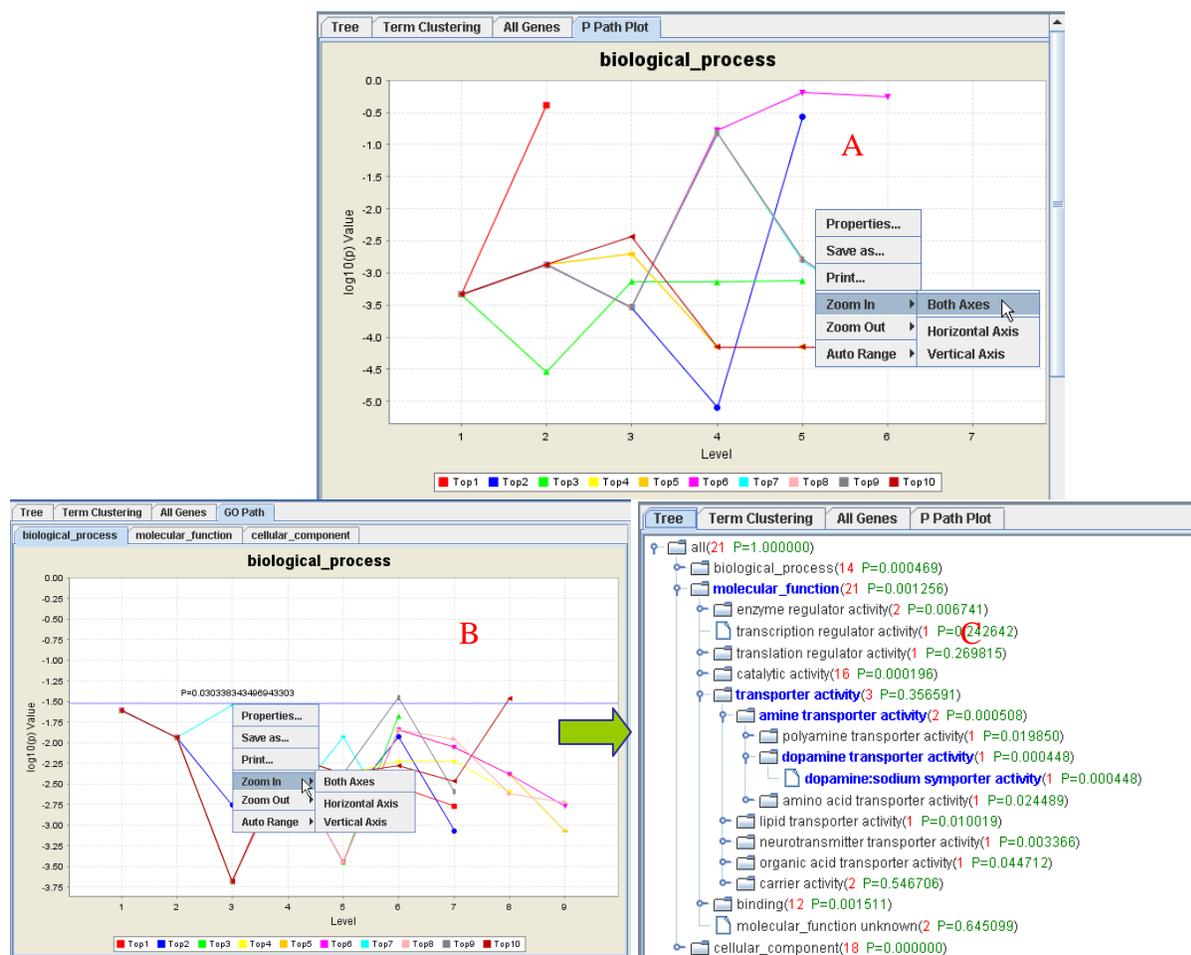


Figure 4-33: Top ten paths with lowest average P value

Under GO Tree Prune tab:

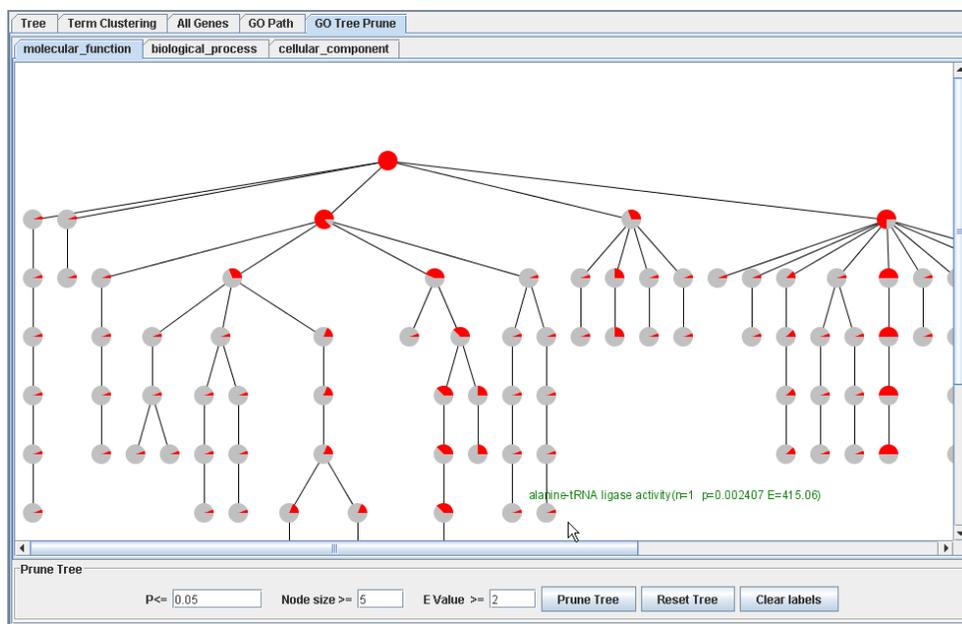


Figure 4-34: Go tree plot

This is a graphic version of the Tree tab which is a tree in text version. Each pie represents a term. Move mouse over a pie will show the term for this pie. Double-clicking a pie will bring the term to display, and double-clicking again will hide the term. The user can also click “Clear labels” button to hide all the displaying term. The pie can be dragged to any designated place.

The user can prune the tree by setting some criteria, for example, p-value, E-value at the bottom of the tree plot. The pruned tree can be further truncated to smaller tree. Clicking “Reset Tree” button will bring the truncated tree back to the full tree.

## 4.9 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 4-35) 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.

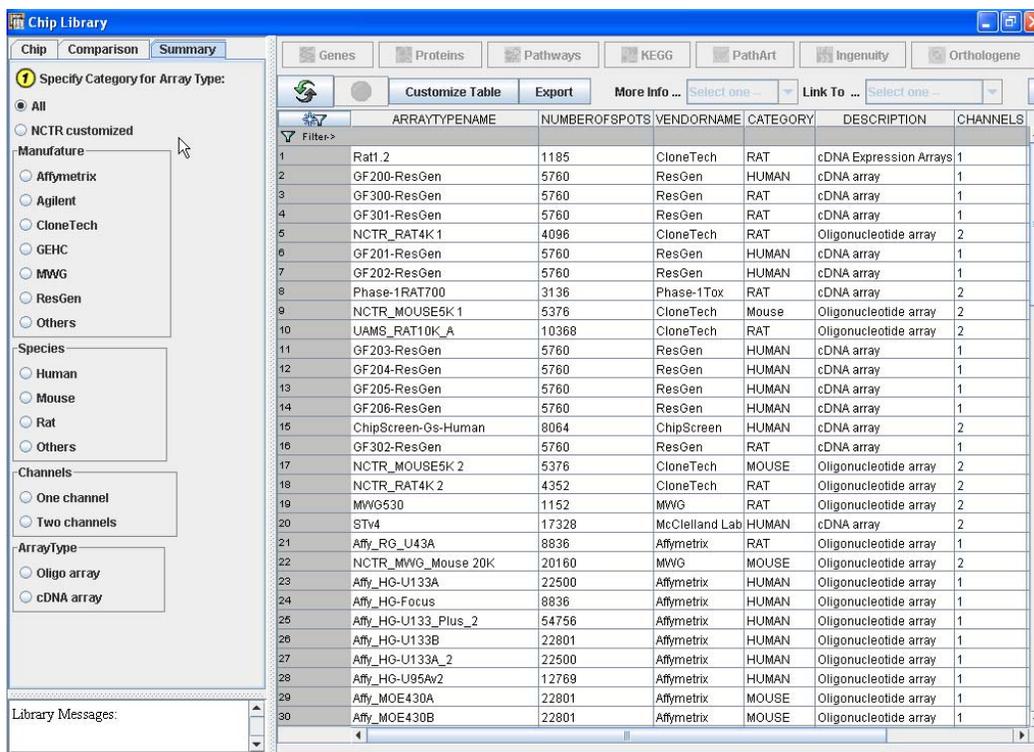


Figure 4-35: Default display of Chip Search panel.

However, the user can also view the arraytype information for a particular array type (e.g. different manufacture, species, number of channels, etc.) by specifying it under the **1 Specify Category for Array Type:**. Figure 4-36 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.

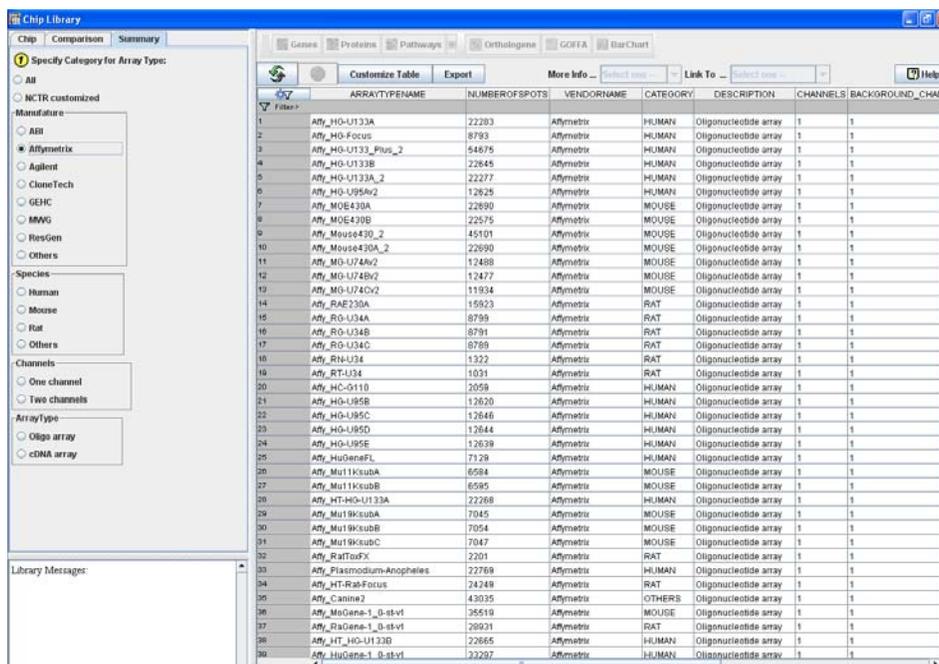


Figure 4-36: Display all the Affymetrix arraytype.

**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 “VennDiagram” button, then decide the common ID for the two array types. The Venndiagram will show in a separate window, see Figure 4-37. By selecting any part of the Venndiagram and right-clicking, the user can save the image, change the color, and see the original data for overlapped genes.

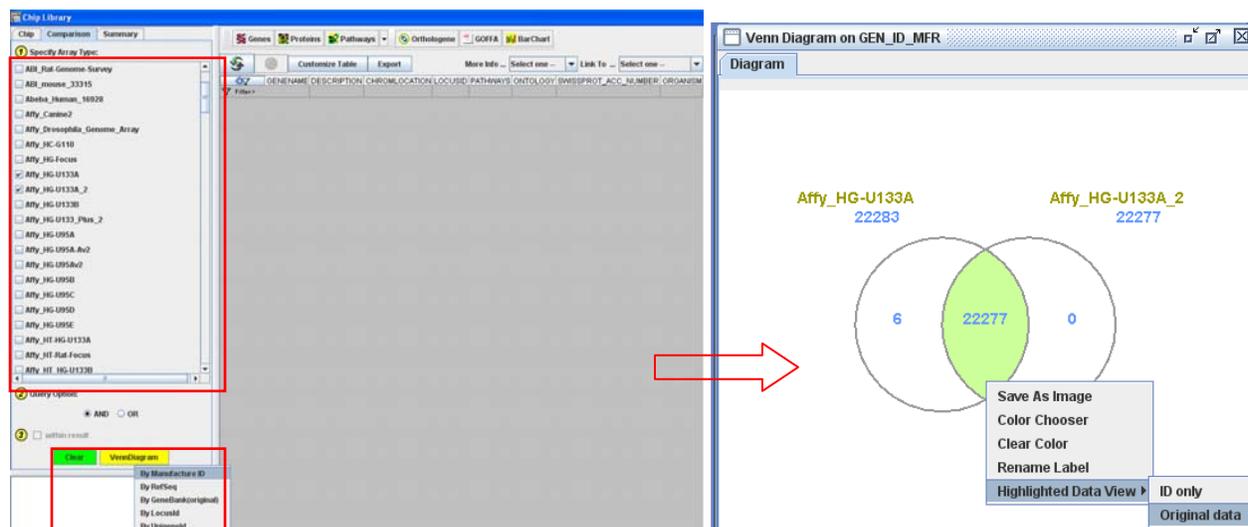


Figure 4-37: Overlapped gene list between Affy\_HG\_133A array and Affy\_HG\_133A\_2 array.

If the user want to see some specific genes for an array type, she/he can double-click the array type in the Chip Library, then highlight those genes and right click (Figure 4-38) and choose

**Mark selected spots in open viewers**. The highlighted genes will be marked in the pre-opened array viewer, see Figure 4-39.

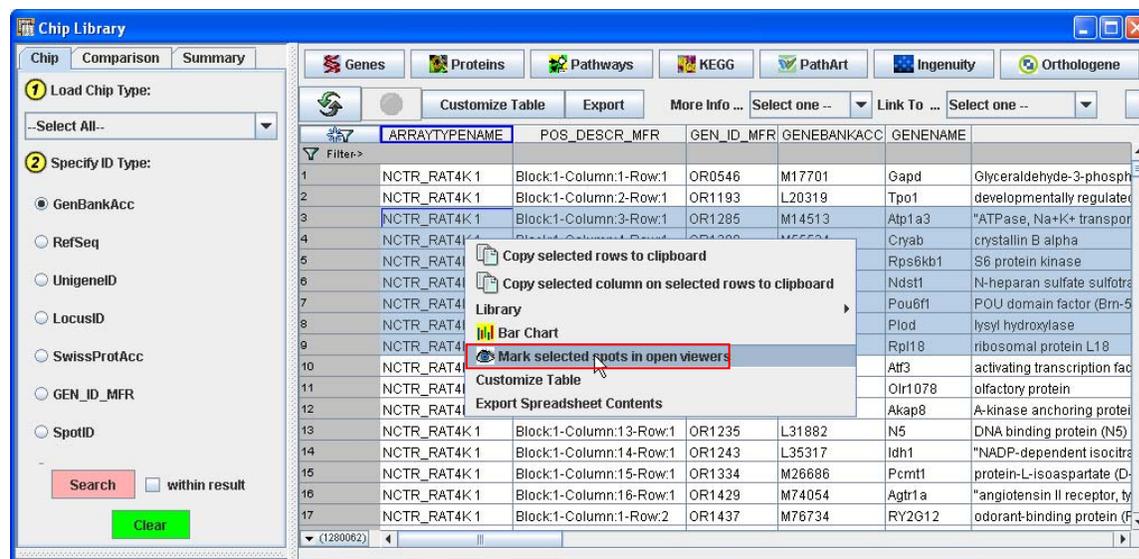


Figure 4-38: Choose specific gene records to be marked in the viewer

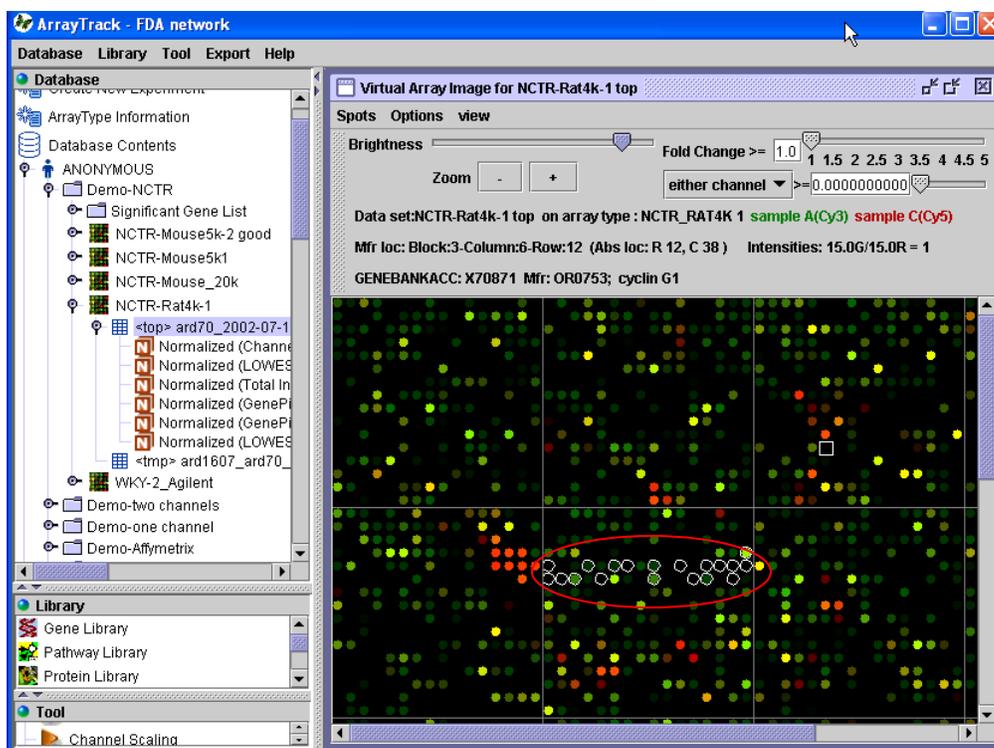


Figure 4-39: the marked genes chosen in the Chip Library

**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.

#### 4.10 Toxicant Library

Toxicant Library contains toxicological data and chemical structures (Figure 4-40). 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 4-40). 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 4-43, using pull down menus, icons and options available by left or right-clicking the chemical structure.

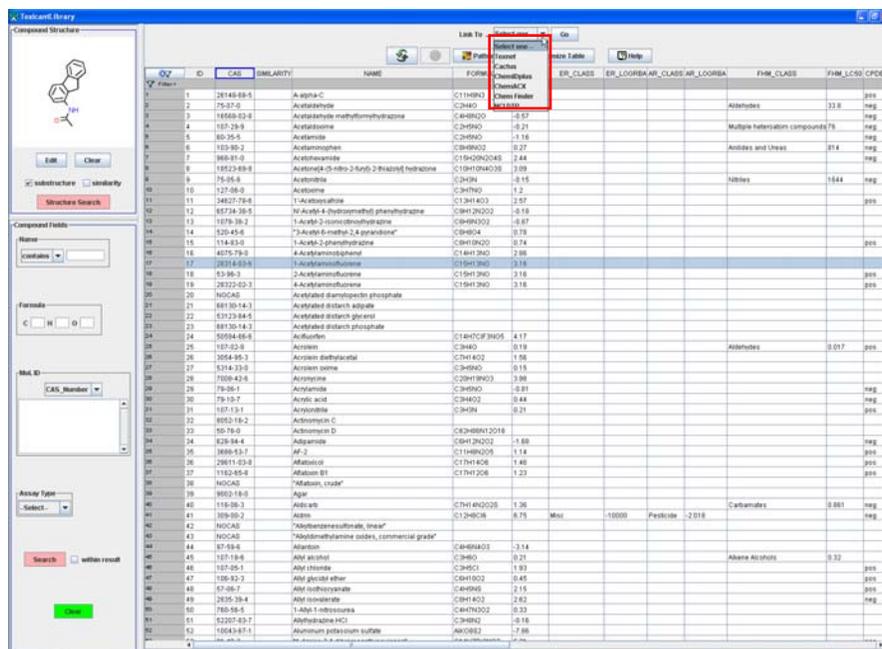


Figure 4-40: Toxicant Library panel.

The user can also highlight a group of chemicals and click  **Pathways ...** button and will see how many compounds and pathway information have been found, see Figure 4-41.

CAS	NAME	Pathway	Category
75-07-0	Acetaldehyde; Ethanal	Glycolysis / Gluconeogenesis(mo00010)	Carbohydrate Metabolism/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Benzoate degradation via hydroxylation(mo00362)	Biodegradation of Xenobiotics/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Aminophosphonate metabolism(mo00440)	Metabolism of Other Amino Acids/Metabolic pathway
107-29-9	Acetaldehyde oxime, Acetaldoxime, Aldoxime	Cyanoamino acid metabolism(mo00460)	Metabolism of Other Amino Acids/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Glycerolipid metabolism(mo00561)	Lipid Metabolism/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Pyruvate metabolism(mo00620)	Carbohydrate Metabolism/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Tetrachloroethene degradation(mo00625)	Biodegradation of Xenobiotics/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Fluorene degradation(mo00628)	Biodegradation of Xenobiotics/Metabolic pathway
75-07-0	Acetaldehyde; Ethanal	Ethylbenzene degradation(mo00642)	Biodegradation of Xenobiotics/Metabolic pathway

Input compounds = 30, 2 compounds found, 28 not found, Total 9 pathway maps.

Figure 4-41: Compound pathway

In Figure 4-41, click any record will bring up the pathway map. See Figure 4-42.

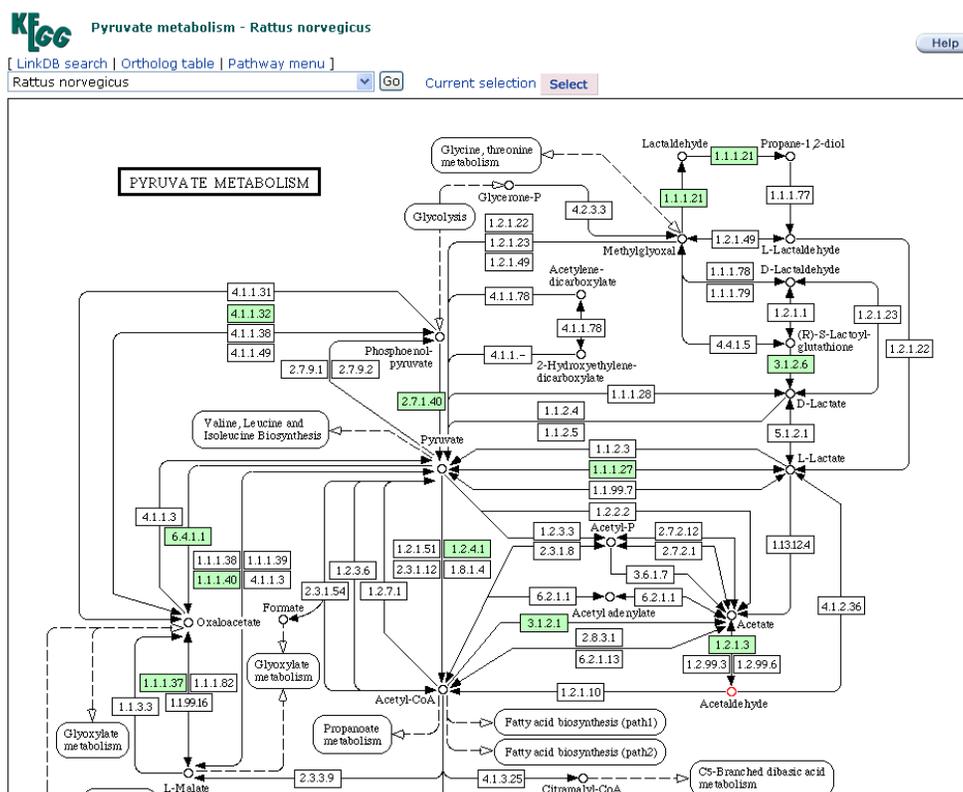


Figure 4-42: Pathway map

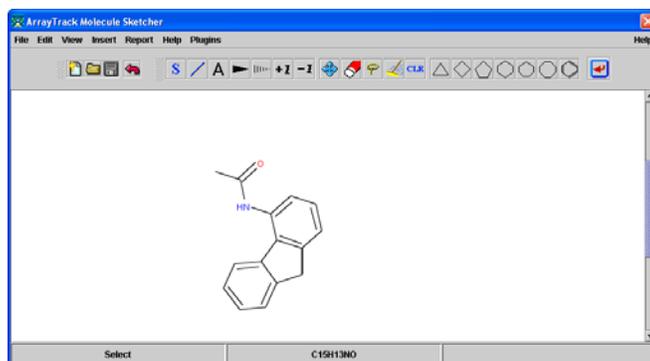


Figure 4-43: Molecular structure editing

### 4.11 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, uterotrophic weight gain, E-screen and combined receptor-reporter gene data (Figure 4-44). 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 4-44). 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 4-45).

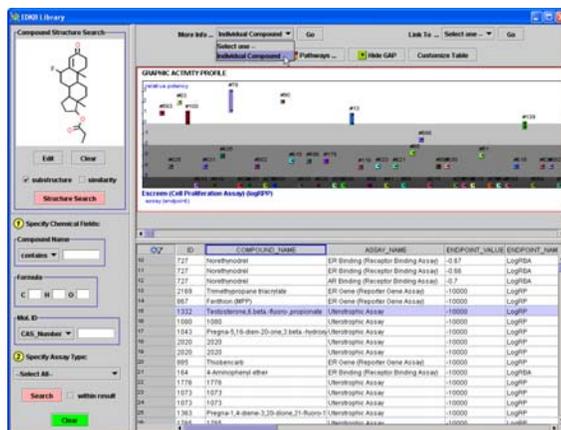


Figure 4-44: EDKB Library panel.

The user can also highlight a group of chemicals and click  Pathways ... button and will get the chemicals pathway information.

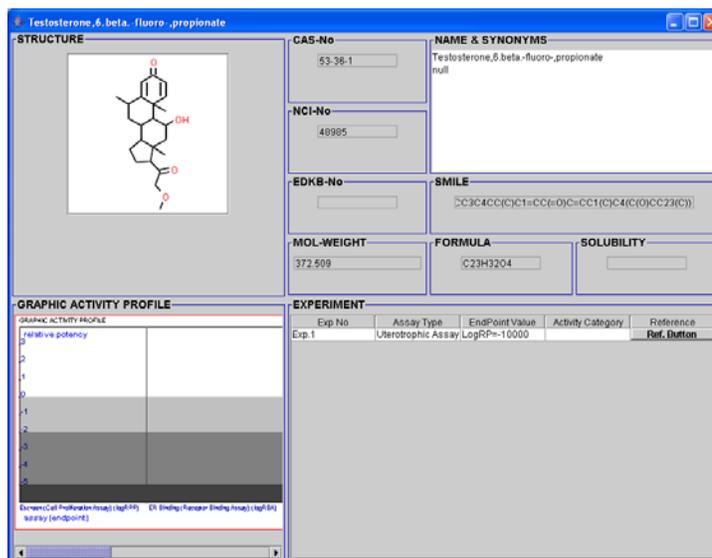


Figure 4-45: Individual compound information

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

### 4.12 ID Converter

ID Converter is a very useful tool for converting one kind of ID to another kind for searching. The user can activate ID Converter by clicking the icon  in the Library panel, see Figure 4-46.



Figure 4-46: 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 and click Convert button. The converted results will be shown in the right panel. See Figure 4-47. 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.

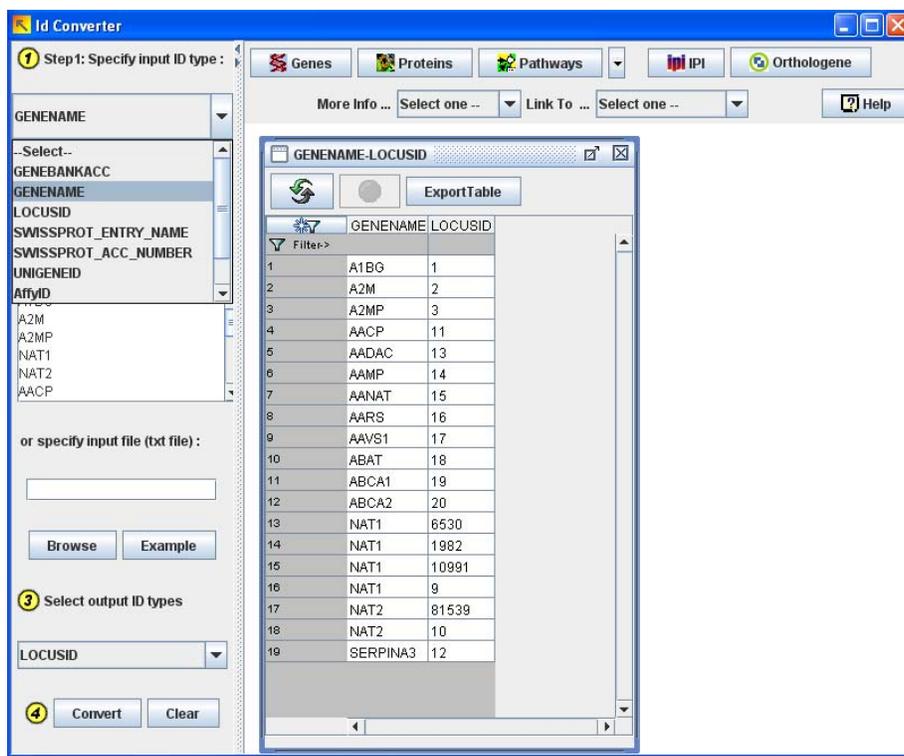


Figure 4-47: ID Converter tool

The ID types that ArrayTrack accepts 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 that ArrayTrack can convert 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, Affymetrix ID and Agilent ID.

## Chapter 5 Working with Tools: Quality Control

### 5.1 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 (Figure 5-1A) or Tool pull-down menu (Figure 5-1B).

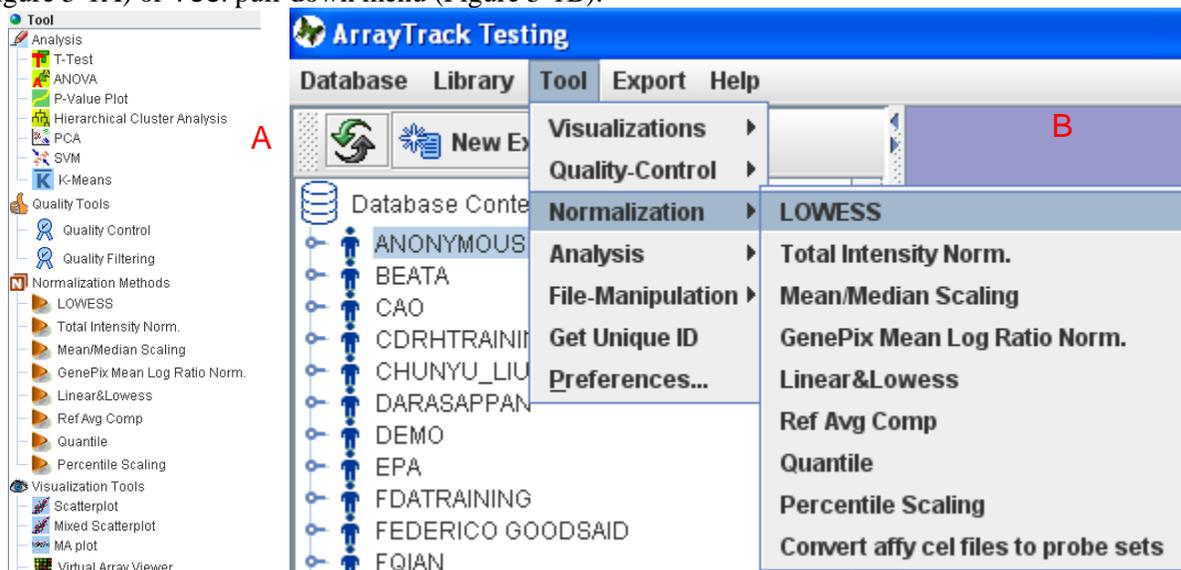


Figure 5-1: Access Tools from Tool panel or from Tool pull-down menu

#### From Tool panel

The TOOL functionalities are classified into four categories: Quality Tools, Normalization Methods, Analysis and Visualization Tools. As has been pointed out earlier, these tools can also be accessed by right-clicking on selected array data sets (see Figure 5-2).

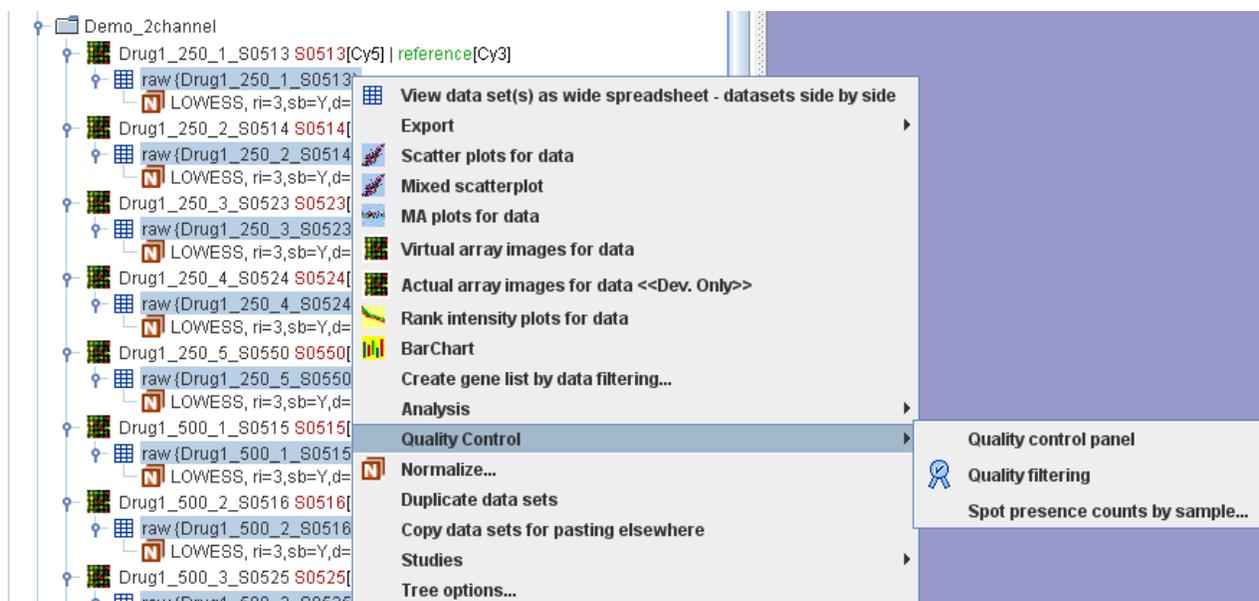


Figure 5-2: Quality Tools can also be accessed from right-clicking on selected array(s).

Quality Tools is discussed in this Chapter; Normalization and Visualization are discussed in Chapter 6 and Chapter 7, respectively.

### **From Tool pull-down menu**

From Tool pull-down menu, the user can also access these four tools plus some other miscellaneous tools. These tools include Split File, Combine Files and set preferences for HCA and PCA. These miscellaneous tools will be addressed in Chapter 9.

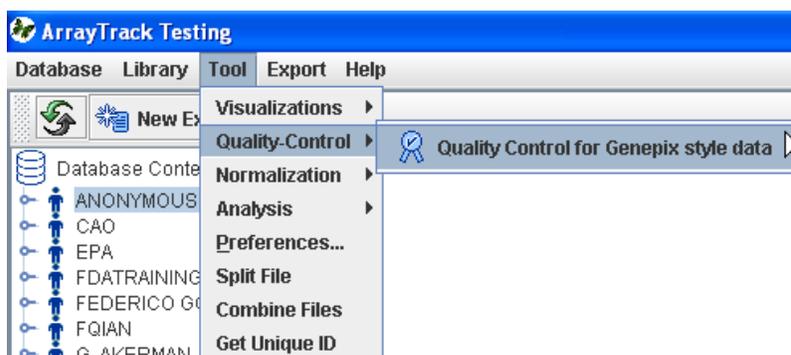


Figure 5-3: Accessing Quality Control from Tool pull-down menu

## **5.2 Overview of Quality Control**

Quality Control provides various visual plots and numerical parameters for measuring the quality of 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.

## **5.3 Launch of Quality Control**

**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. A Quality Control panel will be launched for each of the selected arrays, see Figure 5-2.

**From the TOOL Panel:** If Quality Control is launched from the TOOL panel (Figure 5-1A), the user is asked to select array(s)/hybridization(s) for Quality Control view from a list of hybridizations (Figure 5-4), 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 clicked.

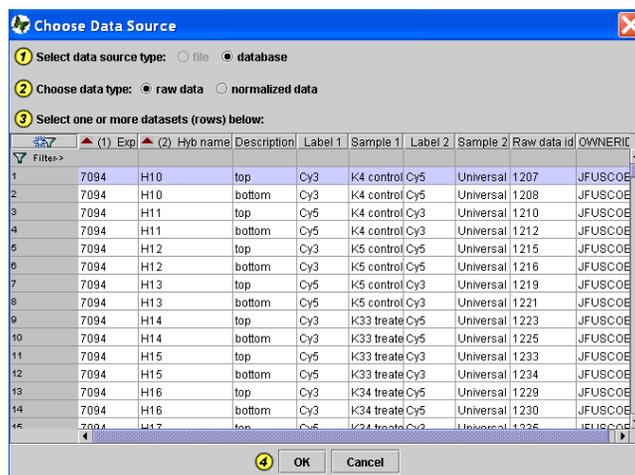


Figure 5-4: Select arrays for Quality Control view

**From the TOOL Pull-down Menu:** When Quality Control is launched from the TOOL pull-down menu, a Quality Control view will be opened for each of the currently selected arrays under the MicroarrayDB Contents tree structure (Figure 2-43C). 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 5-5).



Figure 5-5: The Quality Control panel can not be launched from the pull-down menu until one or more arrays are selected from the MicroarrayDB Contents tree.

#### 5.4 Contents of Quality Control View

Figure 5-6 and Figure 5-7 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 has the options of background subtraction, showing flagged spots, and switching between Scatterplot and MA Plot. Details about Scatterplot can be found in Chapter 8.

**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 8.

**Quality Control Parameters:** Various quality control parameters are shown in the middle of the panel (Figure 5-6 and Figure 5-7) 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 5-6 has a much higher signal-to-noise ratio than the one shown in Figure 5-7.

Quality control parameters

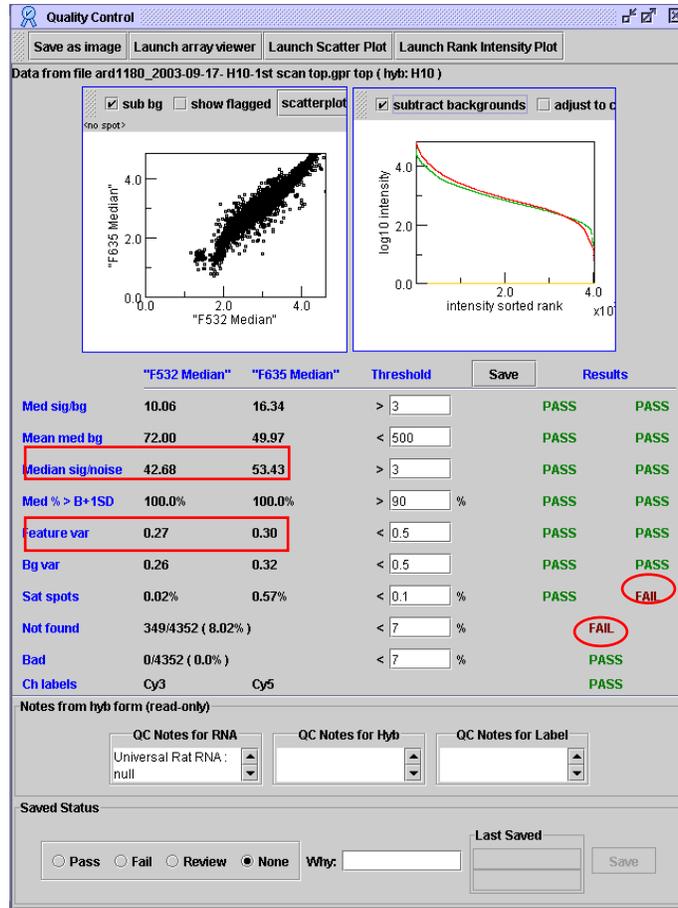


Figure 5-6: Quality Control view and parameters – a successful hybridization with much higher signal-to-noise ratio.

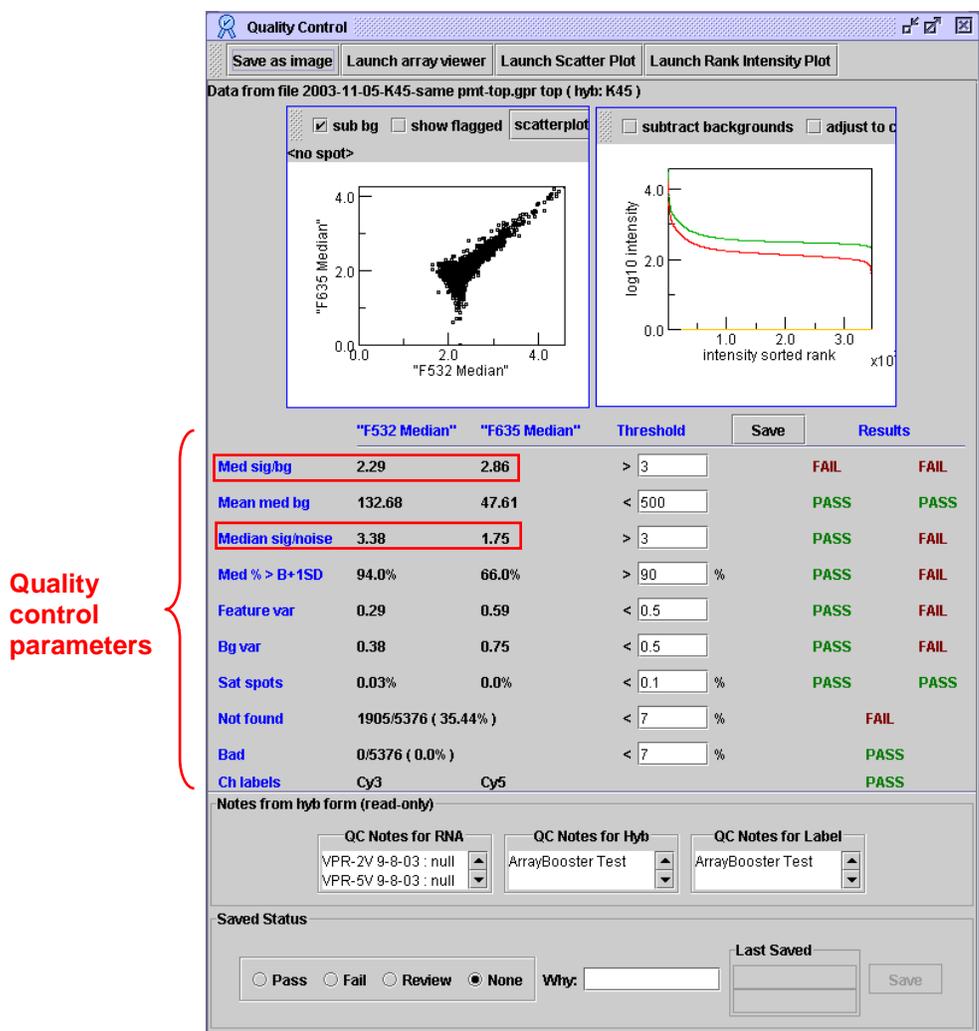


Figure 5-7: Quality Control plots and parameters – a failed hybridization with very low signal-to-noise ratio.

**Function Buttons for Quality Control:** On the top of the Quality Control view, there are four function buttons (Figure 5-8). 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 8 on Visualization.



Figure 5-8: Functional options available within the Quality Control view.

### 5.5 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.

### 5.5.1 Launch of Quality Filtering

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

### 5.5.2 Contents of Quality Filtering

Figure 5-9 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 5-9).

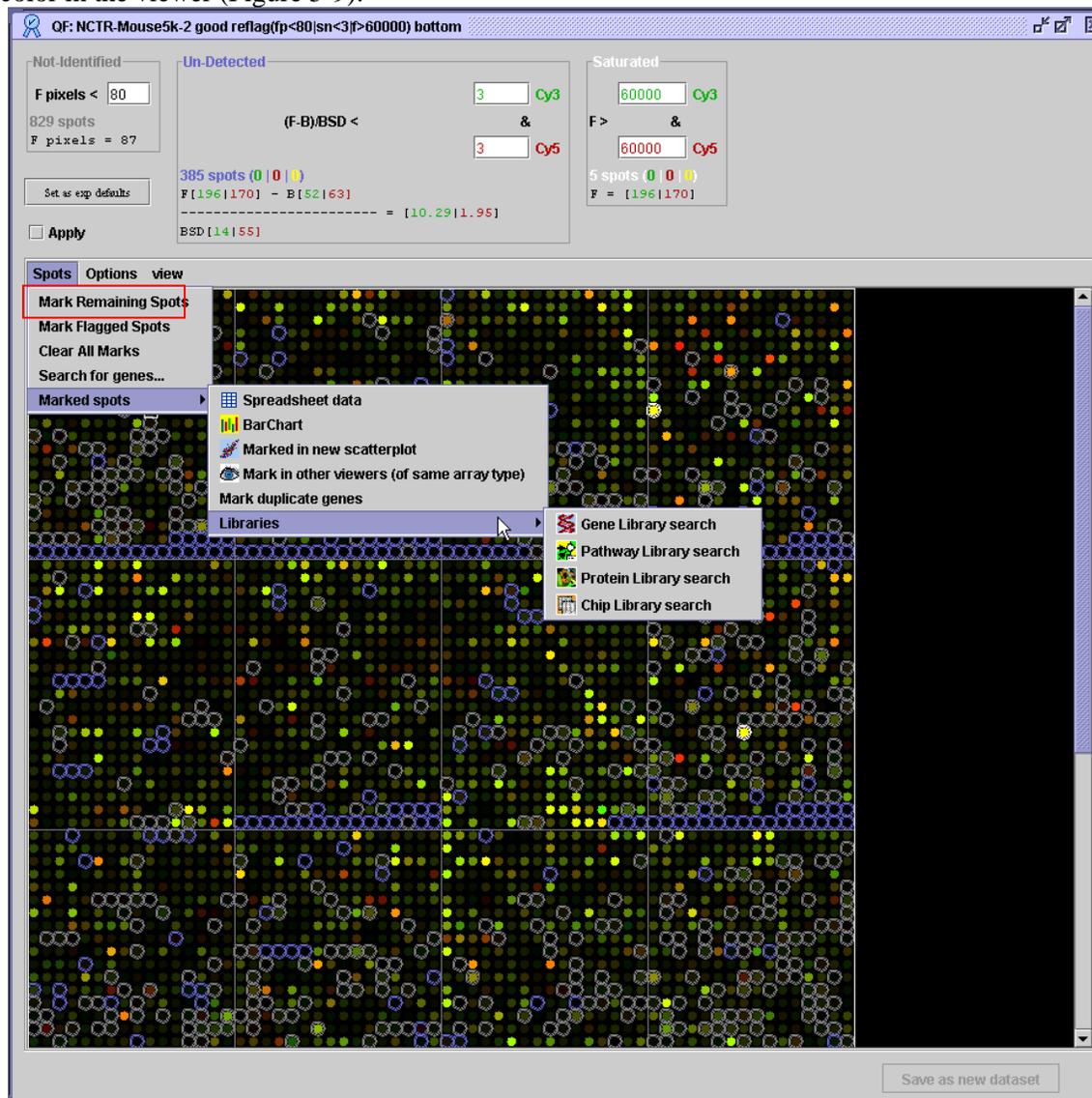


Figure 5-9: Functional options available within the **Quality Filtering** view

In Figure 5-9, 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 5-9.

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 5-10.

The screenshot displays the ArrayTrack 3.4 software interface. The title bar reads "QF: NCTR-Mouse5k-2 good reflag(fp<80|sn<3|f>60000) reflag(fp<80|sn<3|f>60000) bottom". The interface is divided into several sections:

- Not-Identified:** Shows "F pixels < 80" and "829 spots". A button "Set as exp defaults" and an "Apply" checkbox are present.
- Un-Detected:** Shows "(F-B)/BSD < 3" and "3" with "Cy3" and "Cy5" labels. Below this, it displays "385 spots (0 | 0 | 0)" and "F [10257|4450] - B [57|70]" with a calculation result "[318.75|60.00]".
- Saturated:** Shows "F > 60000" and "5 spots (0 | 0 | 0)" with "F = [10257|4450]".
- Spots Options view:** A red box highlights this menu. It includes "Brightness", "Zoom" (-, +), "Fold Change >= 1.0" (with a slider from 1 to 5), and "either channel" (with a dropdown menu).
- Tool bar:** A red box highlights this section, containing text: "Data set:NCTR-Mouse5k-2 good reflag(fp<80|sn<3|f>60000) reflag(fp<80|sn<3|f>60000) bottom on array type : NCTR\_MOUSE5K 2", "Mfr loc: Block:3-Row:8-Col:1 (Abs loc: R 8, C 33) Intensities: 10200.0G/4380.0R = 2.329", and "GENEBANKACC: NM\_009082 Mfr: OM1782A; ribosomal protein L29".
- Spot Map:** A large grid of spots, each represented by a colored dot (green, yellow, orange, red, blue, white) on a dark background. A horizontal line is drawn across the middle of the grid.
- Save as new dataset:** A button at the bottom right of the interface, highlighted with a red box.

Figure 5-10: 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 6 Working with Tools: Normalization

### 6.1 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 6-1).

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

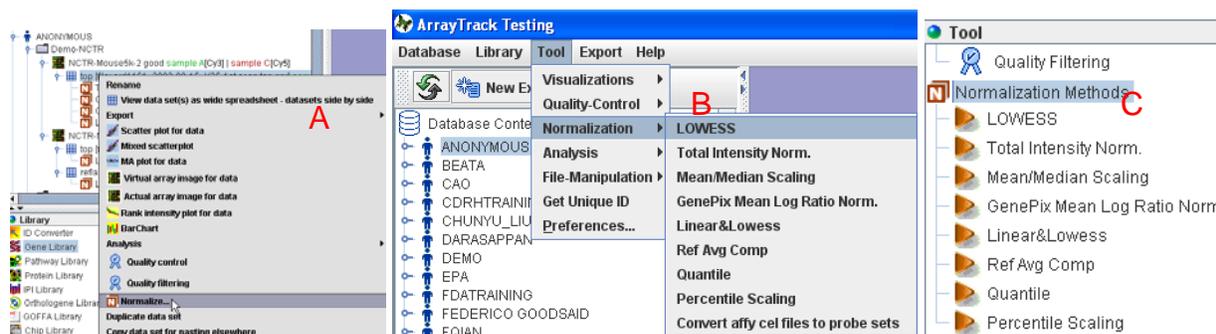


Figure 6-1: Several normalization methods have been implemented in ArrayTrack and can be accessed from (A) Right-click on selected array(s); (B) Tool pull-down menu; and (C) TOOL panel.

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 6-2).

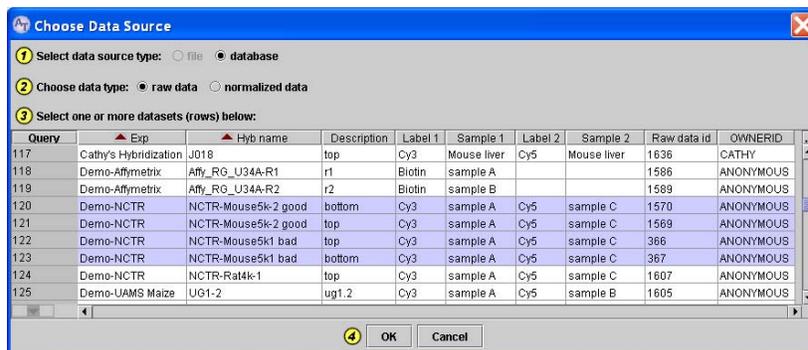


Figure 6-2: Normalization method applies to a set of selected arrays.

In our experience, however, the user is more likely to use the Database Contents tree (see Figure 2-43. 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 6-3). By right-clicking on any of the selected

raw arrays, a set of functions including **Normalize...** can be applied (Figure 6-4) and the user can choose one of the four normalization methods (Figure 6-1C).

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.

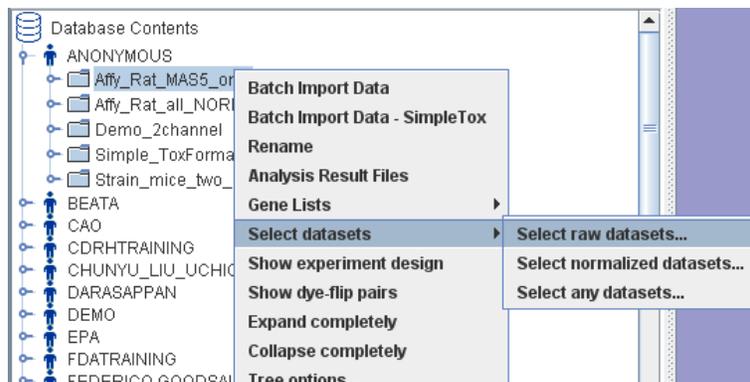


Figure 6-3: All raw data (arrays) under an experiment can be conveniently selected by right-click on the name of the experiment.

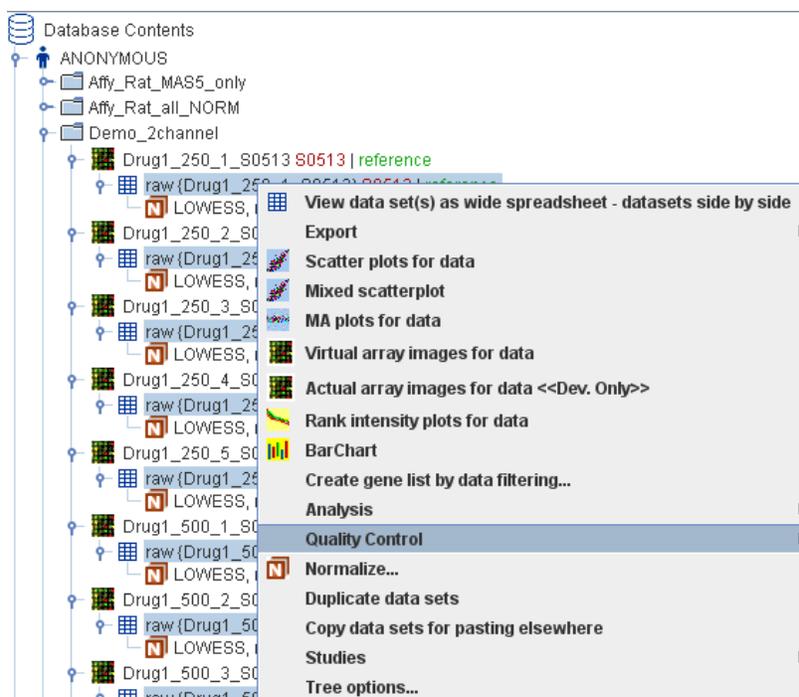


Figure 6-4: Normalization methods can be applied to a set of selected arrays under the Database Contents tree.

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**.

## 6.2 Lowess

Lowess (Loess) refers to Locally Weighted regression and soothing 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 6-5).

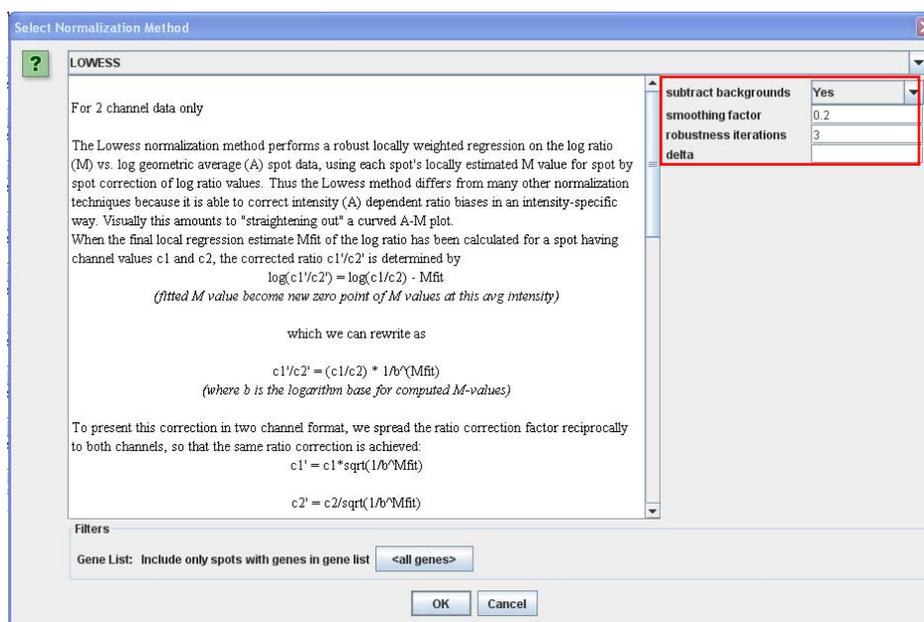


Figure 6-5: Parameter settings for Lowess normalization.

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

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

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

which can be rewritten as

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

(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:

$$c1' = c1 \sqrt{\frac{1}{b^{M_{fit}}}}$$
$$c2' = c2 \sqrt{\frac{1}{b^{M_{fit}}}}$$

There are three parameters that need to be set for a Lowess (Figure 6-5).

**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  $0 < f \leq 1$  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  $\text{delta} > 0$  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  $0.01 \times \text{the range of A values}$ , 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 ( $0.01 \times \text{the range of A values}$ ) as shown in Figure 6-5.

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

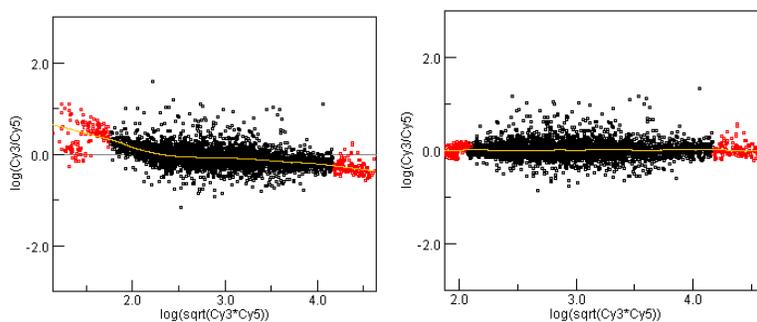


Figure 6-6: Systematic, intensity-based ratio bias (Left) is corrected by Lowess (Right). The yellow line shows the Lowess fitted values, Mfit.

### 6.3 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.,  $r = (\text{sum ch1 vals})/(\text{sum ch2 vals})$ ;
- 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.

### 6.4 Mean/Median Scaling

Mean/Median 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 6-7). 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.

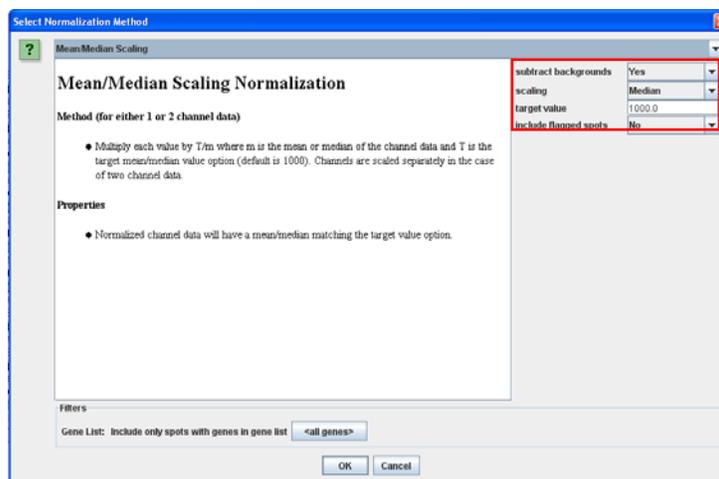


Figure 6-7: Channel Scaling normalization.

## 6.5 GenePix Mean Log Ratio Normalization

GenePix Mean Log Ratio Norm. only applies to two-channel platforms (Figure 6-8). 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.

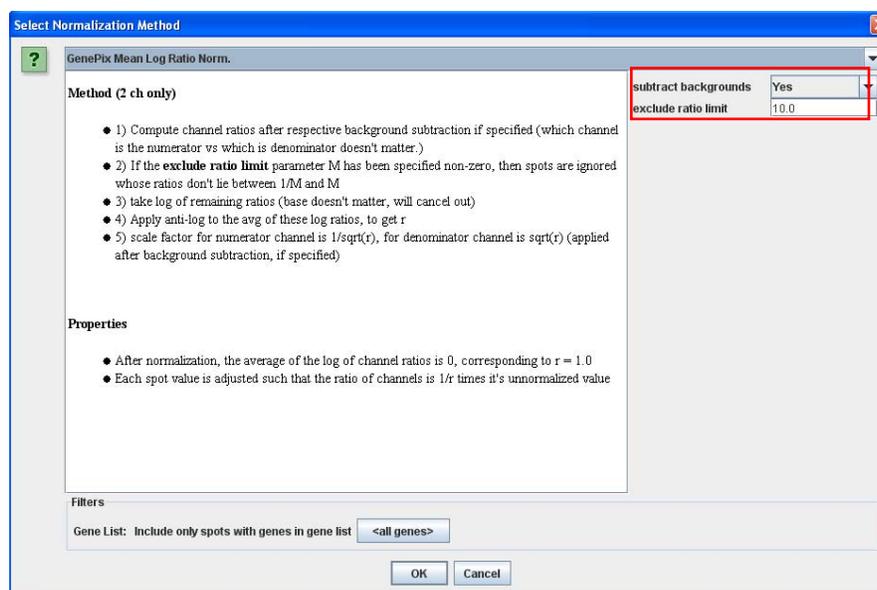


Figure 6-8: GenePix Mean Log Ratio Normalization

## 6.6 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 6-9. 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.

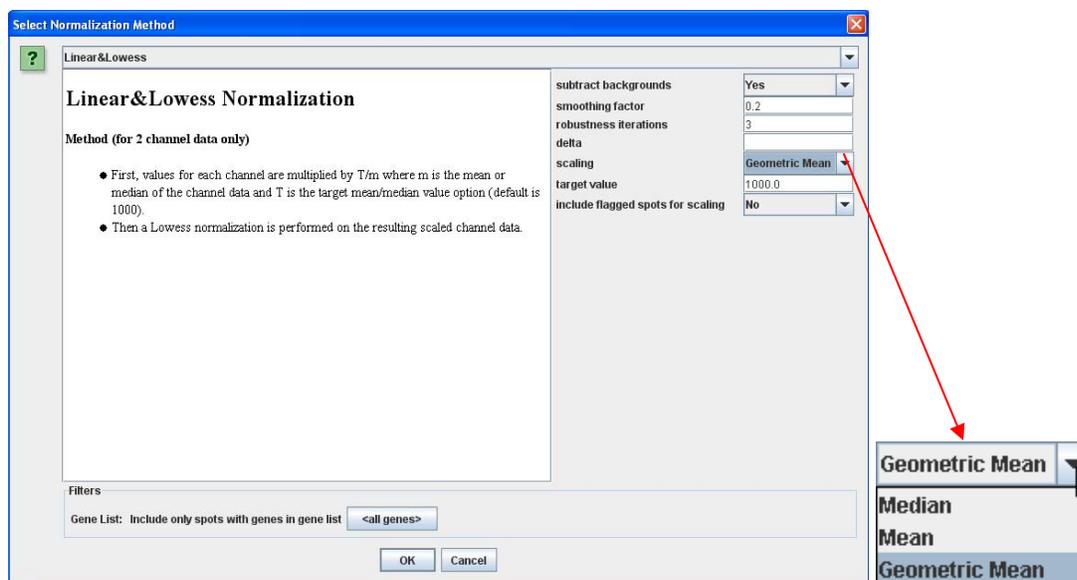


Figure 6-9: Linear and Lowess Normalization

### 6.7 Quantile Normalization

This normalization is for one channel data only.

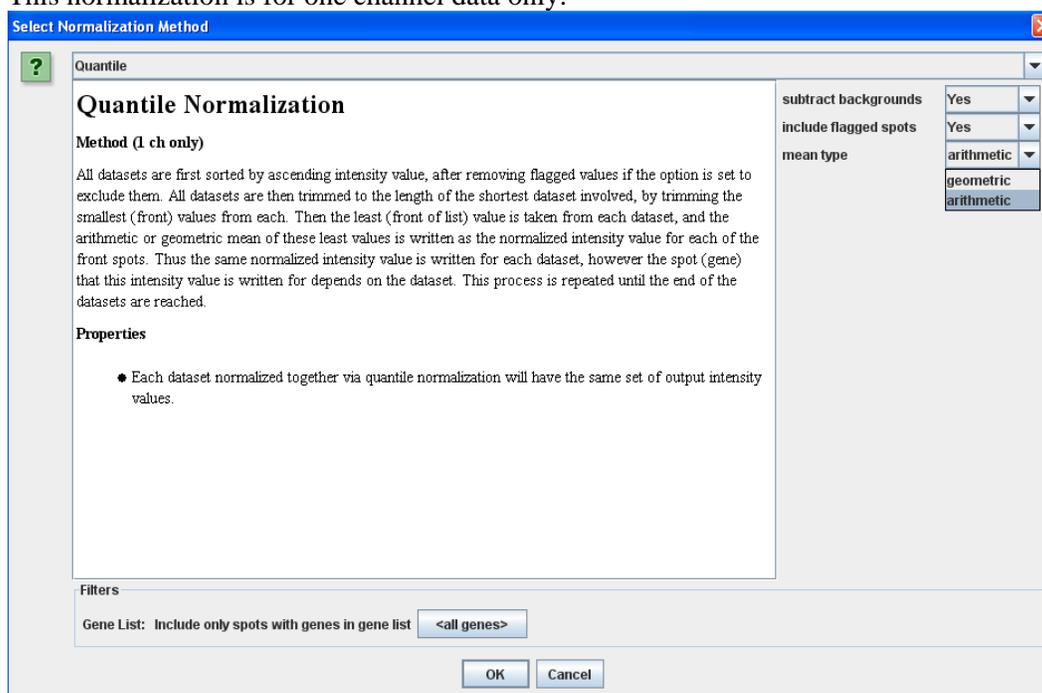
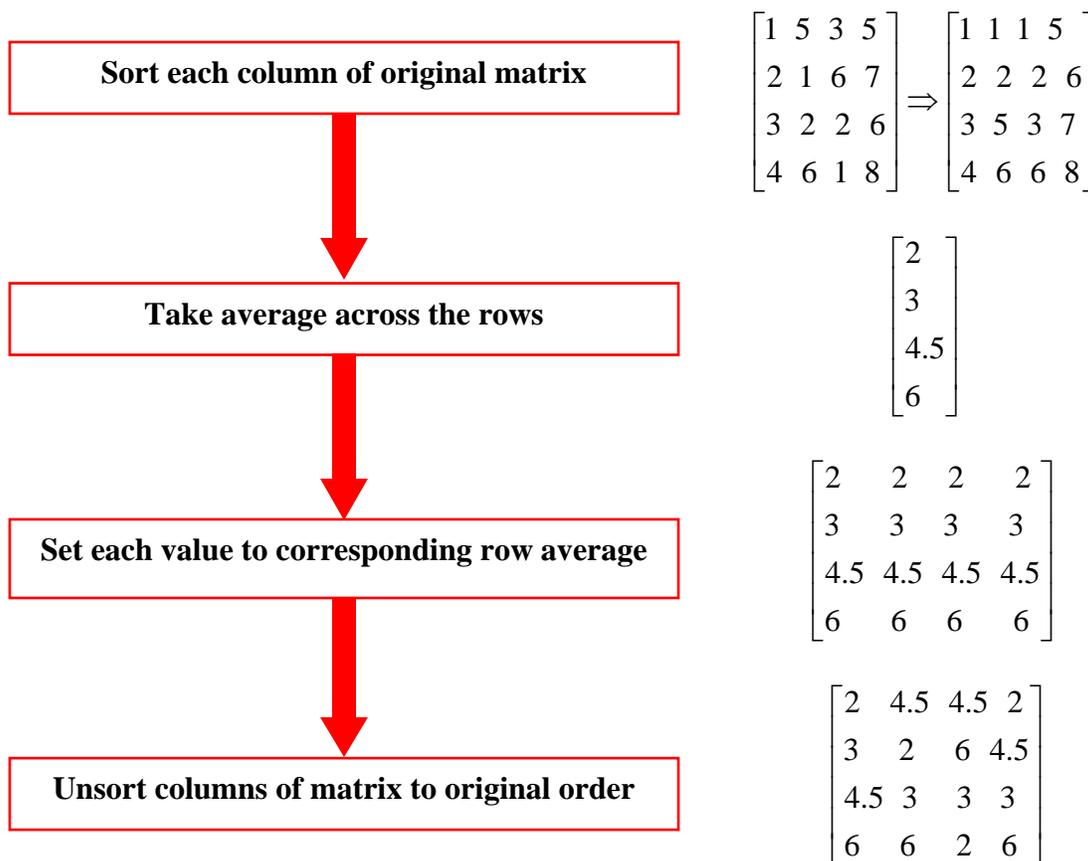


Figure 6-10: Quantile normalization

All datasets are first sorted by ascending intensity value, after removing flagged values if the option is set to exclude them. All datasets are then trimmed to the length of the shortest dataset involved, by trimming the smallest (front) values from each. Then the least (front of list) value is taken from each dataset, and the arithmetic or geometric mean of these least values is written as the normalized intensity value for each of the front spots. Thus the same normalized intensity value is written for each dataset;

however the spot (gene) that this intensity value is written for depends on the dataset. This process is repeated until the end of the datasets is reached. An example of the Quantile normalization is as following:



### 6.8 Reference Average Comparison Normalization

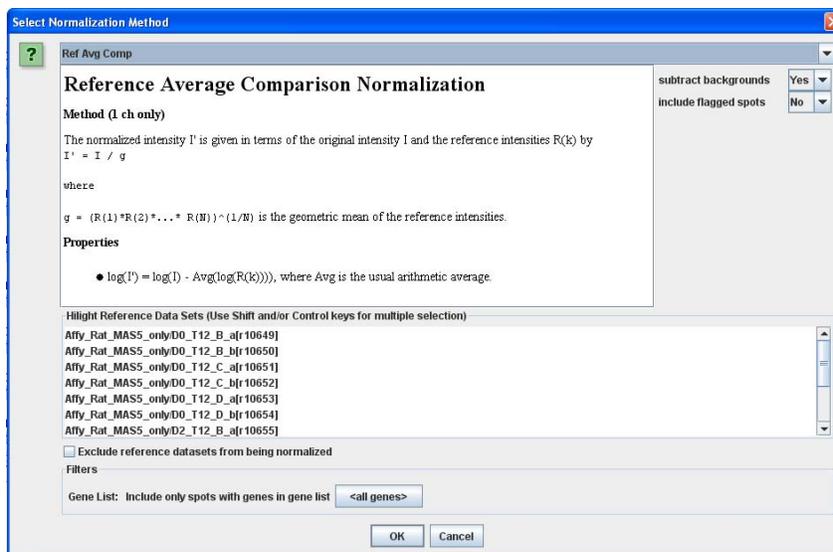


Figure 6-11: Reference average comparison normalization

Reference average comparison normalization only applies to one channel. In Figure 6-11, users need to choose the hybridizations for normalization. The “all genes” button allows the user to include only spots within the gene list for the 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 near mean zero and variance one) before being systematically compared and analyzed by other data analysis software, depending on the particular analysis methods.

## Chapter 7 Working with Tools: Analysis Tools

### 7.1 Overview

The analysis tools can be accessed within the “Tool” window (Figure 7-1A), or from tool menu (Figure 7-1B), or from database with data selected and right-clicking (Figure 7-1C). Analysis tools are provided to perform eight mathematical/statistical operations: 1) T-test, 2)ANOVA, 3) P-value, 4) Clustering, 5) Principal Components Analysis (PCA), 6)Correlation matrix, 7)SAM, 8)K-Means.

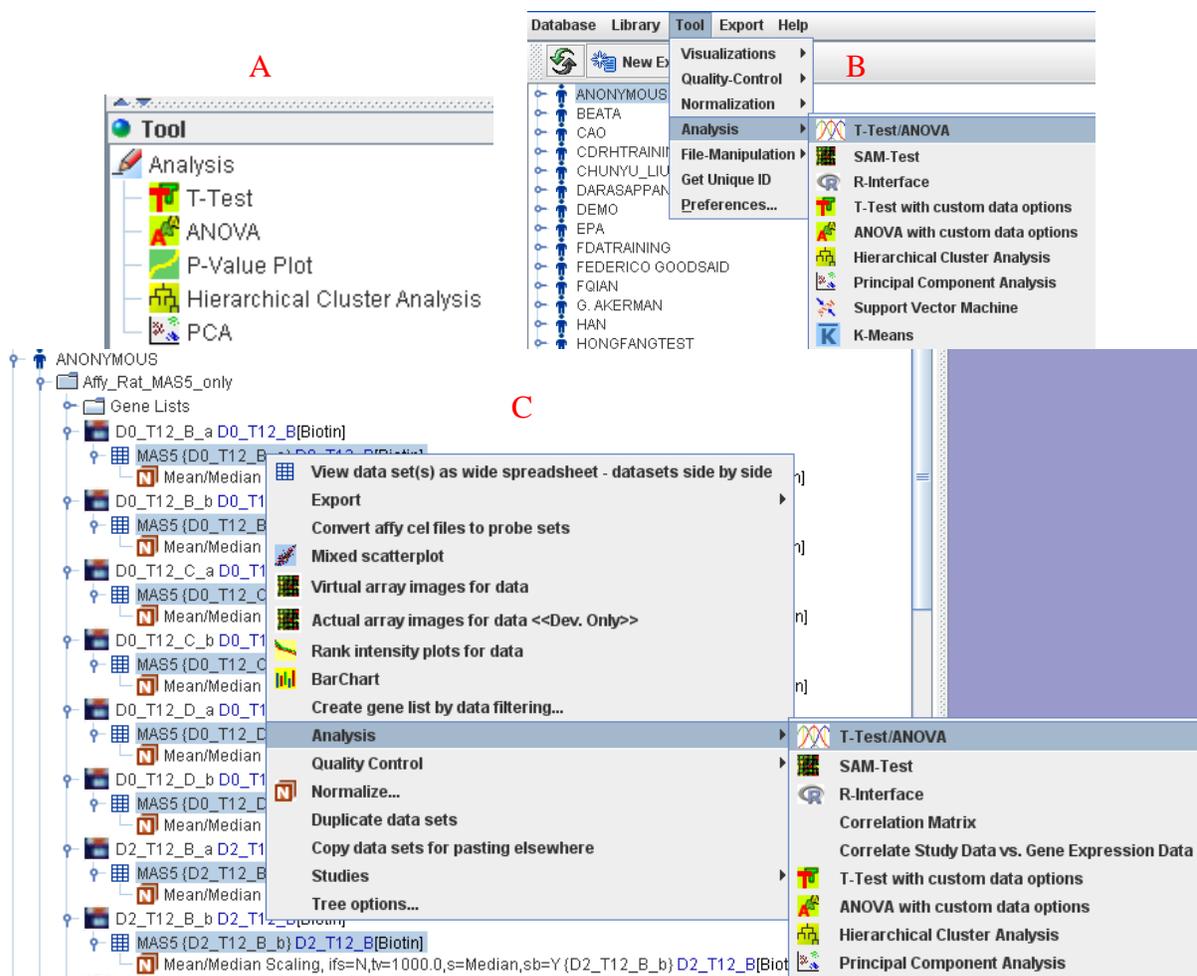


Figure 7-1: Access analysis tools

Before proceeding to explain the use of each analysis tool, we want to mention that the various analysis tools are interlinked. For example, after applying the T-test to two sets of data, users can perform additional analysis using other tools (more about this will be explained further below).

### 7.2 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 do T-tests only 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 7-2A) 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. So this data file can come from other sources. It does not have to be stored in ArrayTrack database. 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 7-2B.

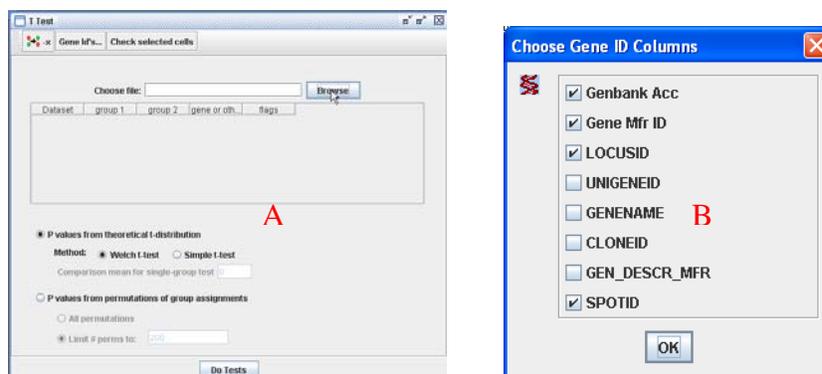


Figure 7-2: Choosing the data after activating T-test through the Tools panel

2) First choose the dataset from database panel as shown in (Figure 7-1C), and then right click and choose Analysis -> “T-test ...”. Accessing in this way, the user does not need to combine the data files. But the dataset has to be imported in ArrayTrack first.

3) Choose datasets from database panel and then click Tool pull-down menu ->choose “Analysis” -> “T-test ...”.

As shown in Figure 7-1B &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.

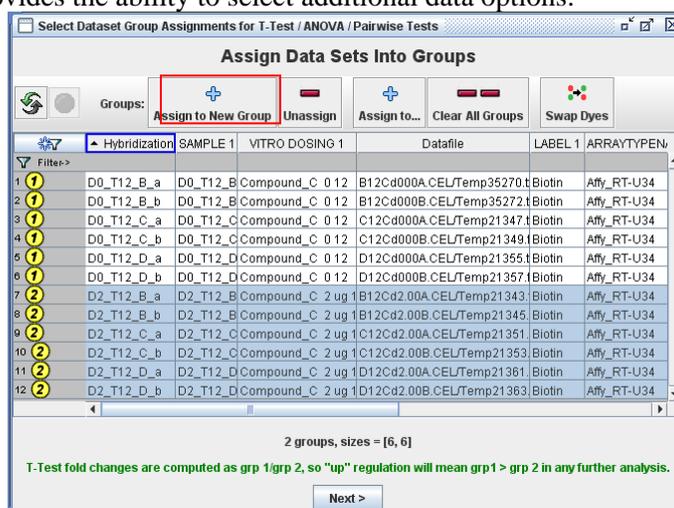


Figure 7-3: Assign dataset to groups

Doing T-test/ANOVA

If user selects “T-test/ANOVA”, then he will be asked to assign datasets to group. In Figure 7-3, 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  and typing the group number. The user can unassign the group by clicking the “Unassign” button. The “Swap Dyes”  button is only for two-channel data. The T-test computes the fold change as group1/ group2 (or Cy5/Cy3), if “Swap Dyes” button is clicked, then the fold change will swapped as group2/group1 (or Cy3/Cy5). This is useful when the user try to compare two groups that have opposite sample label. For example, for group1 the channel 1 is labeled as Cy3 and channel 2 is labeled as Cy5 but for group 2 the channel 1 is labeled as Cy5 and channel 2 is labeled as Cy3. Using “Swap Dyes” button can swap either group making the two groups have consistent dye labeling and comparable. (I'm not sure about this part.)

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. In Figure 7-3, users can click “Next” button after assigning the dataset to groups. Then following window will pop up to let the users to select the options for T-test.

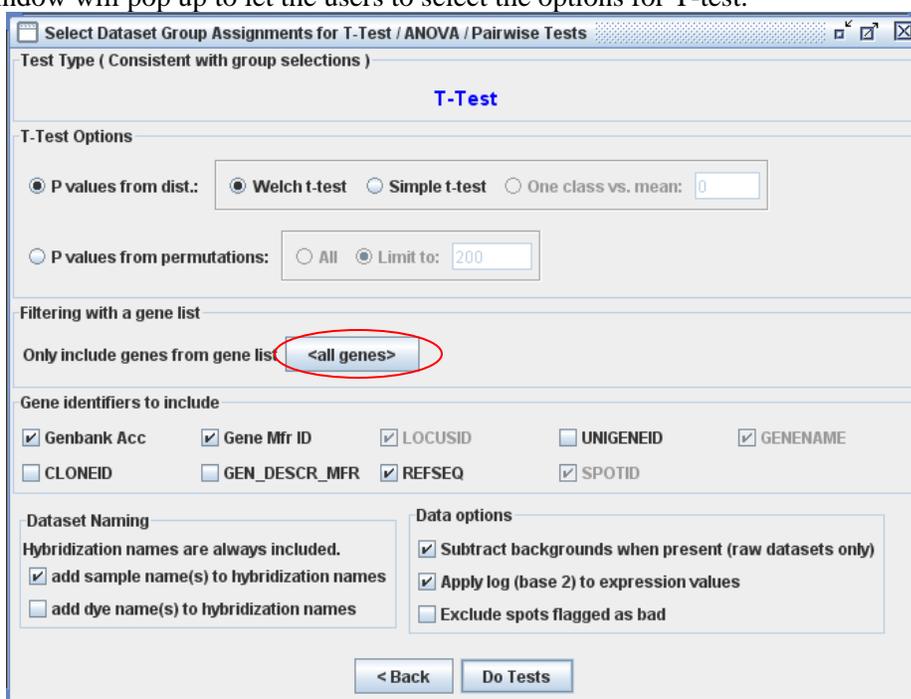


Figure 7-4: Set the criteria of the T-test

The user can choose to run T-test on all genes or on selected gene list by clicking the button . A window will pop up letting the user choose the gene list for the T-test.

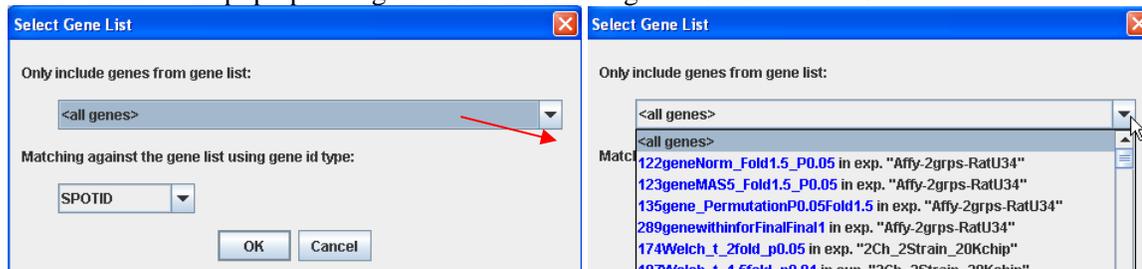


Figure 7-5: Select gene list for T-test analysis

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:

$$P_{\text{permutation}} = \frac{\text{Numberof}(P_k < P_0)}{\text{NumberofPermutation}}$$

$$\text{where } {}_n P_k \equiv \frac{n!}{(n-k)!}$$

The user can choose to calculate  $P_{\text{permutation}}$  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  $P_{\text{permutation}}$  using  $\text{NumberofPermutation} = 924$  or limit the number of permutation to any number that is less than 924. See Figure 7-4. Click “Do tests” button, the results of T-test/ANOVA will be shown, see Figure 7-8.

### Doing T-test with custom data options

If user selects “T-test with custom data options”, then following window (Export Options) will pop up.

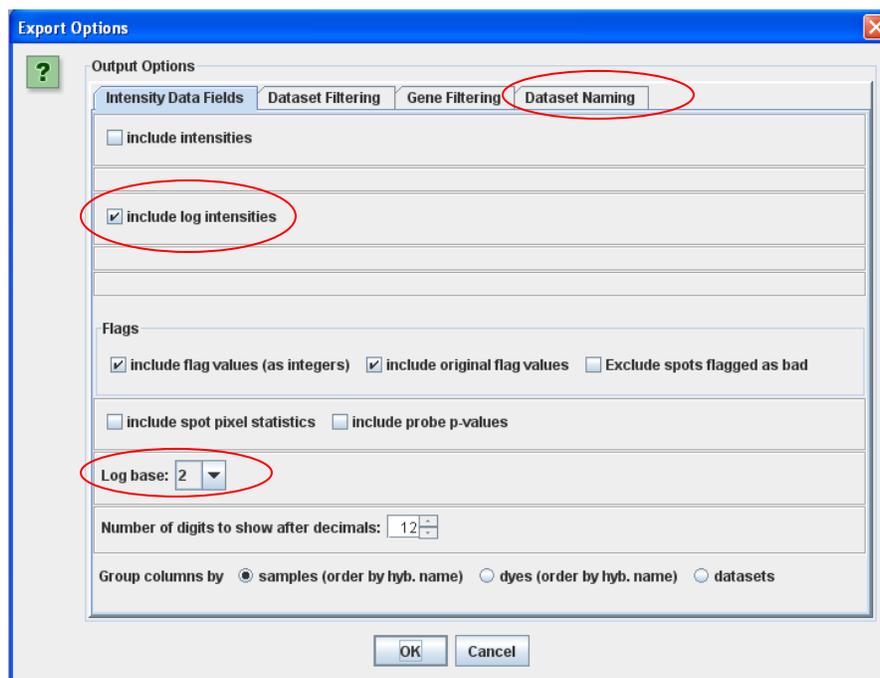


Figure 7-6: options for T-test

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 7-6, only log intensity data will be exported in log base 2, and flag values will be included as integers and the original flag value will be included. 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, click OK button. Then “T Test” window will display, see Figure 7-7A. In this window the user clicks the boxes to assign each data element to either the first or the second group. Clicking the “Gene Id’s” button opens another window (see Figure 7-7B) 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, with both “include log intensities” and “include intensities” selected, both types of the data will be included in the data to be exported. The data options that are 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.

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 7-8. 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 7-8 and further explained below.

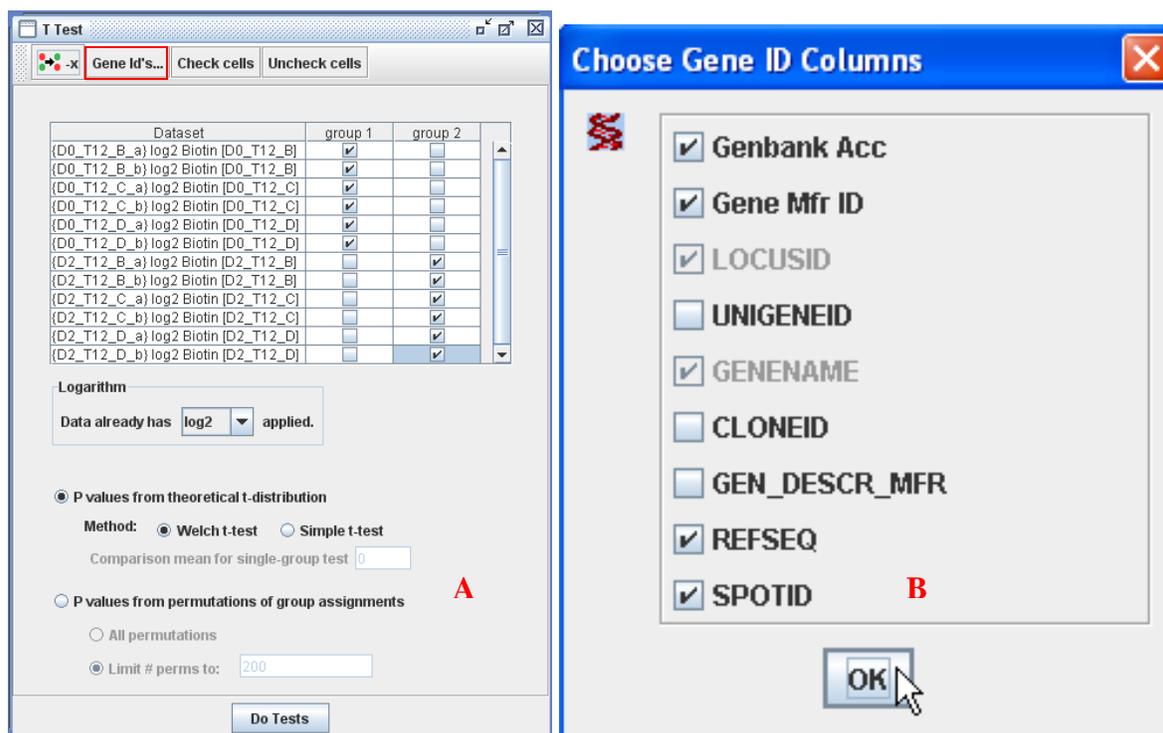


Figure 7-7: assign data in two groups (A) and choose Gene ID types (B)

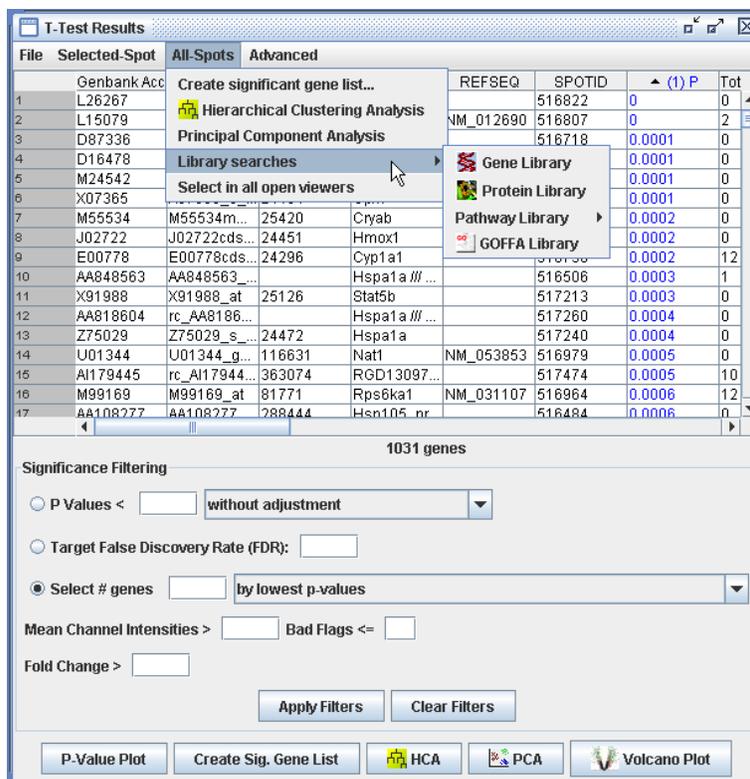


Figure 7-8: T-test results window

In Figure 7-8, the users can get the value of the T-test result such as mean of group 1, mean group 2 and mean difference (group 2 - group 1), etc. Furthermore, ArrayTrack provides several types of filters and interactive graphic tools to aid the user in evaluating the data generally, and in choosing a significant set of genes particularly. As shown in Figure 7-8, the user can perform significant gene 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 several options (by lowest P-values, by lowest P-values with equal # of up and down regulated, by largest symmetric fold change and by largest symmetric fold change with equal # of up and down regulated). Text boxes are also provided for removing spots below a specified mean channel intensity or/and below a specified minimum fold-change. Multiple filtering criteria can be applied in parallel.

From the T-test result, users can use Bar Chart to see some genes expression across multiple arrays within the same experiment or across different experiments by highlighting some genes (less than 5) and choosing "Selected Spot" pull-down menu -> Create Bar Chart. See Chapter 8 for detail about Bar Chart.

Clicking the **Volcano Plot** button causes a volcano plot of the selected data to be produced and displayed in a new window. Notice that in the example of volcano plot in Figure 7-9, 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 7-9 will comprise the significant gene list, that is, the spots denoted by red in Figure 7-9.

The volcano plot produced by ArrayTrack has a number of features and interactive capabilities providing particular utility for 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 7-7. 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 7-9); 2) the P-value cutoff alone (Areas A, B and C of Figure 7-9); and, 3) the fold-change cutoff alone (areas A, C, D and F in Figure 7-9); the number of non-significant genes is also displayed (area E).

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 corresponds 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 7-9.

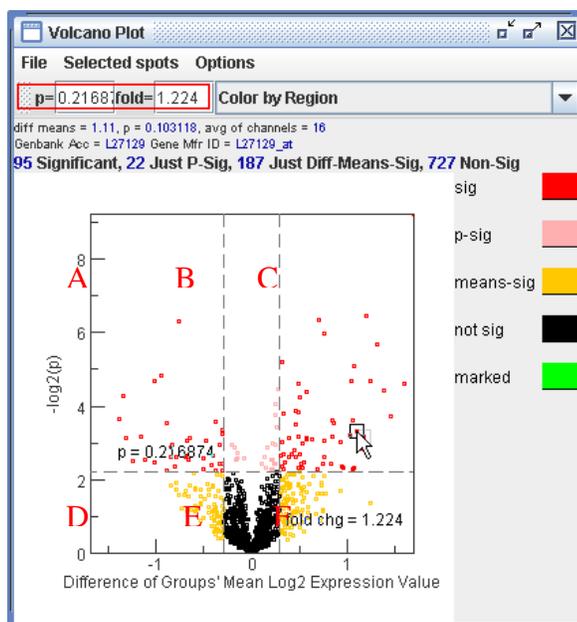


Figure 7-9: Volcano Plot

- 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 7-10: Selected spots pull-down menu

### 7.3 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 7-8).

Also P-value Plot tool can be activated by clicking  P-value Plot under the Tool/Analysis section. See Figure 7-1A. A pop up window with title "Choose Data Source" allows the user to choose the data file (Figure 7-11). 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.

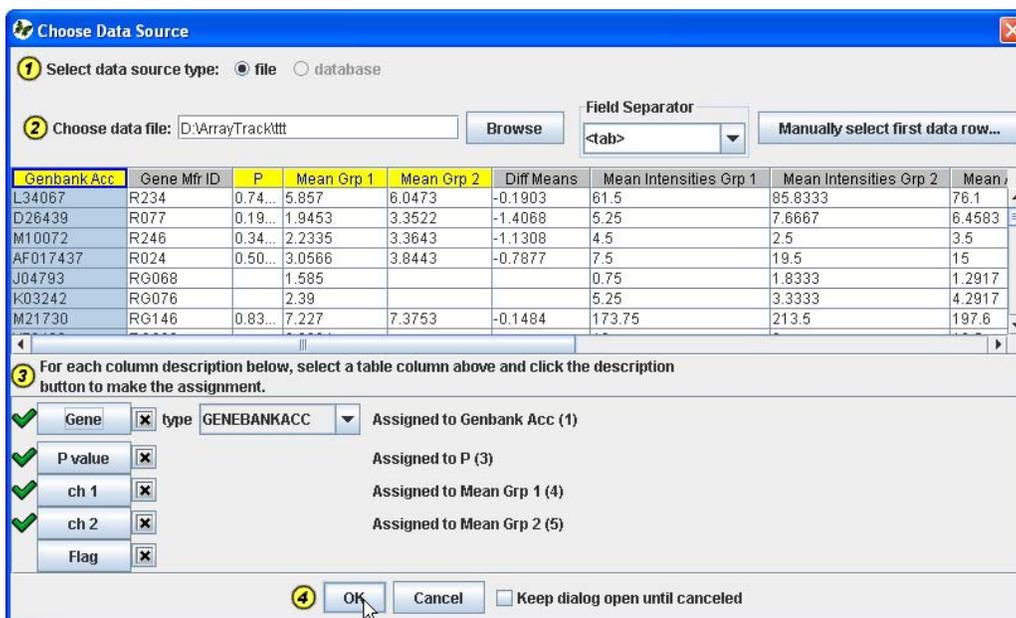


Figure 7-11: Choose data file to get p-value plot

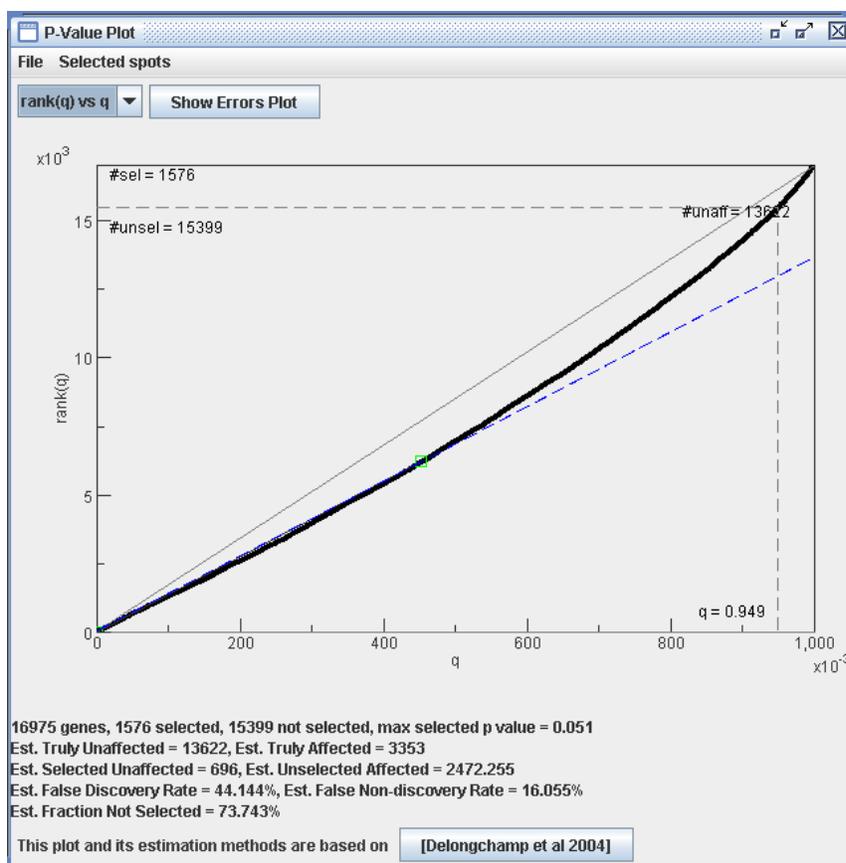


Figure 7-12: P-value plot

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 7-13.

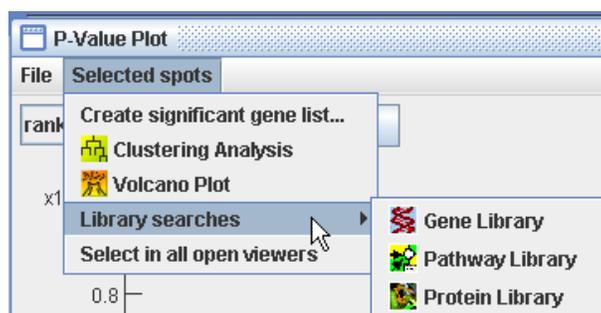


Figure 7-13: Access other functions from p-value plot

## 7.4 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 7-1.

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 7-14A). The text file will be shown after the user click “Open” button (see Figure 7-14B). Then the user can do data analysis by clicking “Analysis” in the menu bar or clicking HCA icon . The result is shown in Figure 7-16.

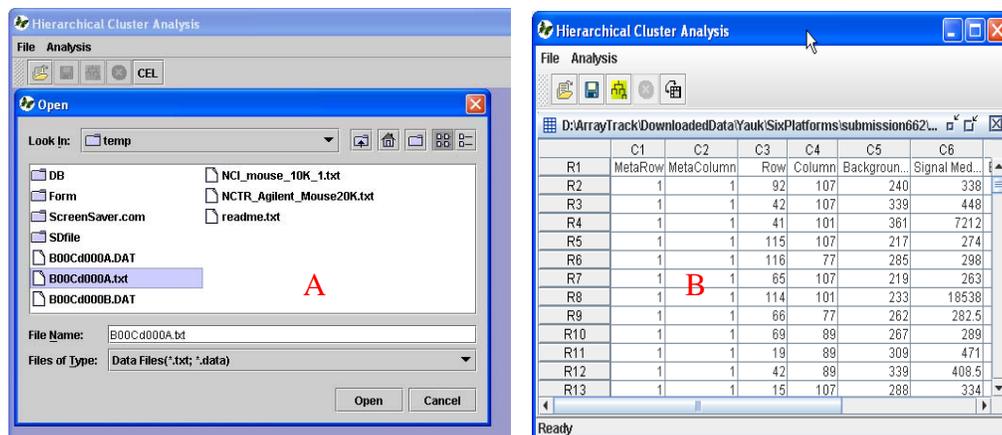


Figure 7-14: Select data file to do data analysis

2) If HCA is activated in the second way, the “Export options” window pops up (Figure 7-15). Clicking OK will bring out the HCA plot, see Figure 7-16.

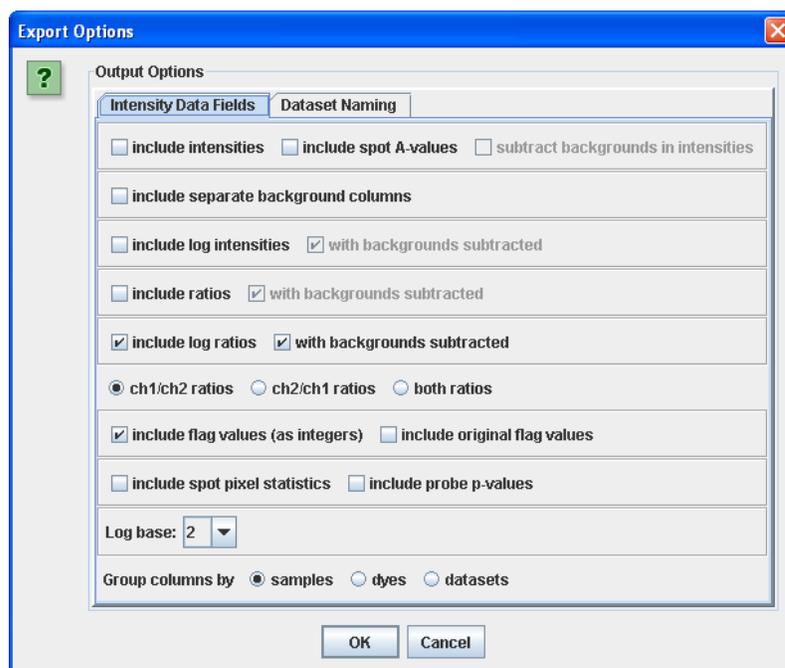


Figure 7-15: Export Option window

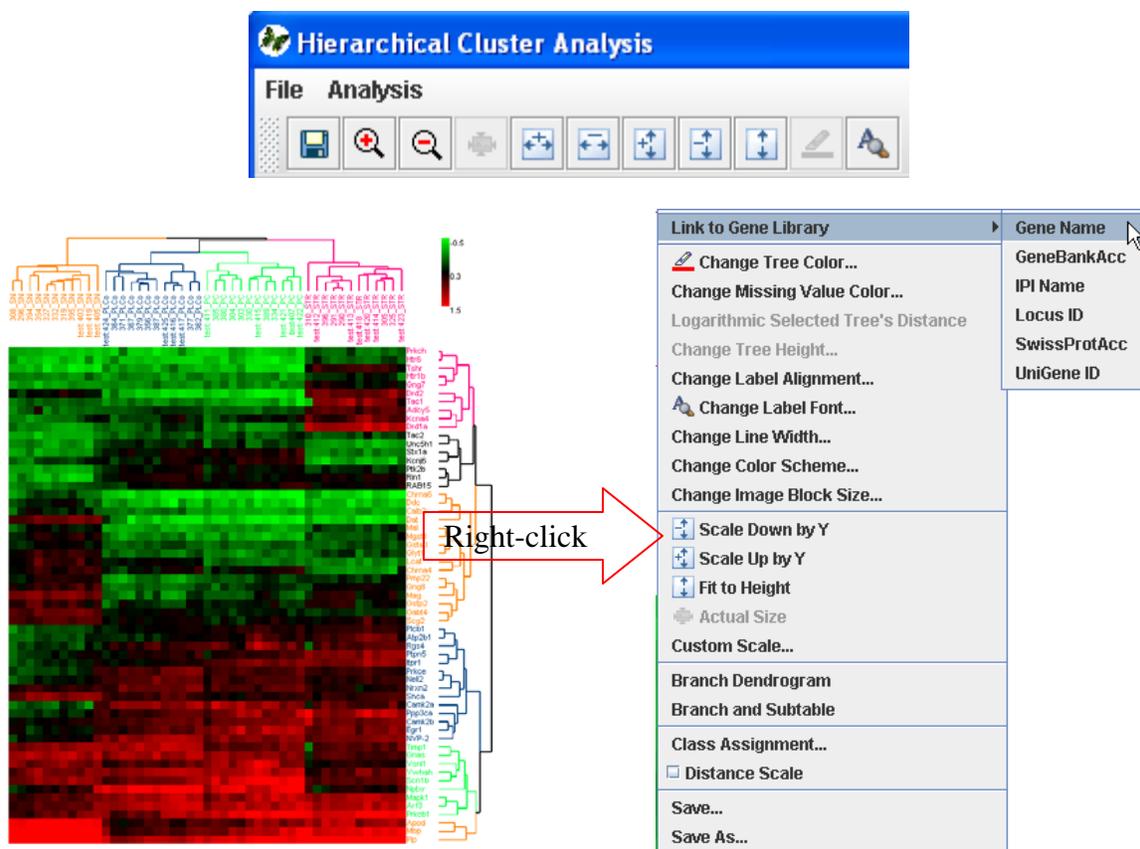


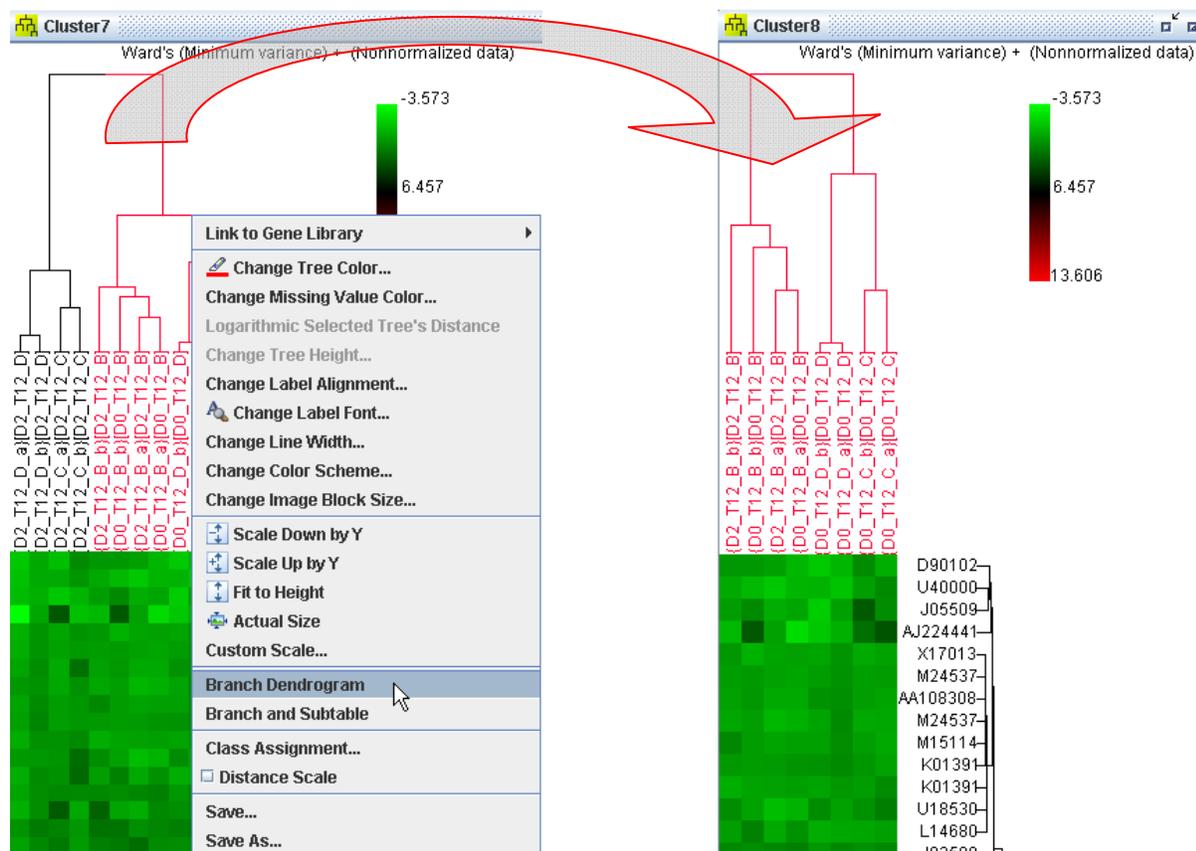
Figure 7-16: HCA analysis result

From the HCA results the user can do the following maneuvers:

--Zoom in and zoom out: 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 select a branch of the tree and open the branch in a new window by choosing “Branch Dendrogram”.

--The user can save the HCA image to the local drive, also the cluster results can be saved as .cls file format. See Figure 7-16.

## 7.5 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 dimensionality 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 7-17.

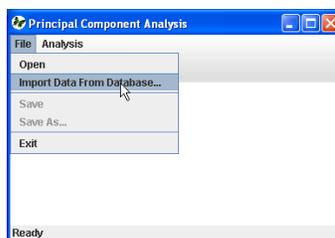


Figure 7-17: Open a file or import data from database for PCA

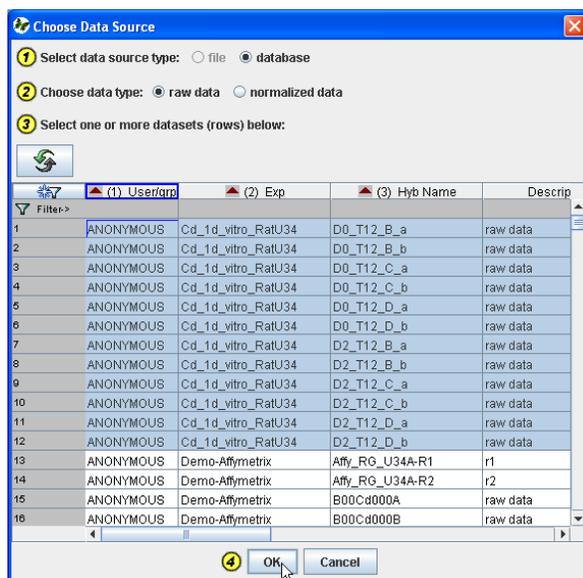


Figure 7-18: Choose data from the database to do PCA

In Figure 7-18, multiple hybridization data can be selected. Clicking OK will bring up the “Export option” window (Figure 7-15). Click OK button, then the PCA plot will show up (see Figure 7-19A).

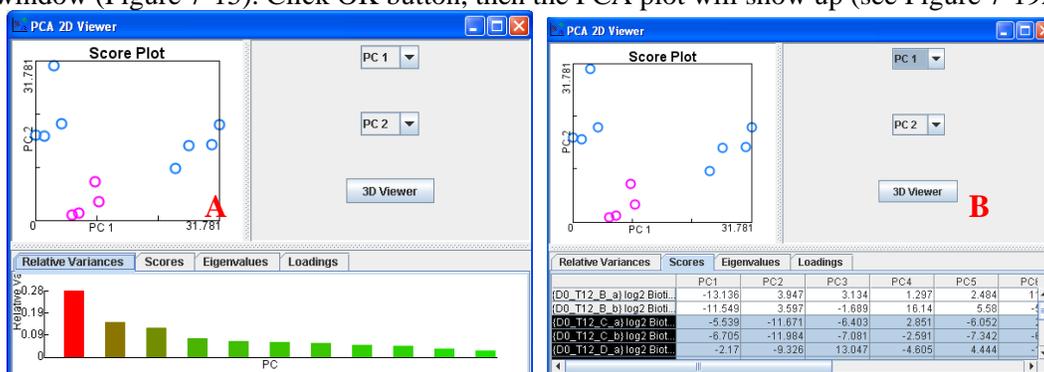


Figure 7-19: PCA plot

3) If the user activates PCA in the second way, the PCA options window will pop up (Figure 7-20).

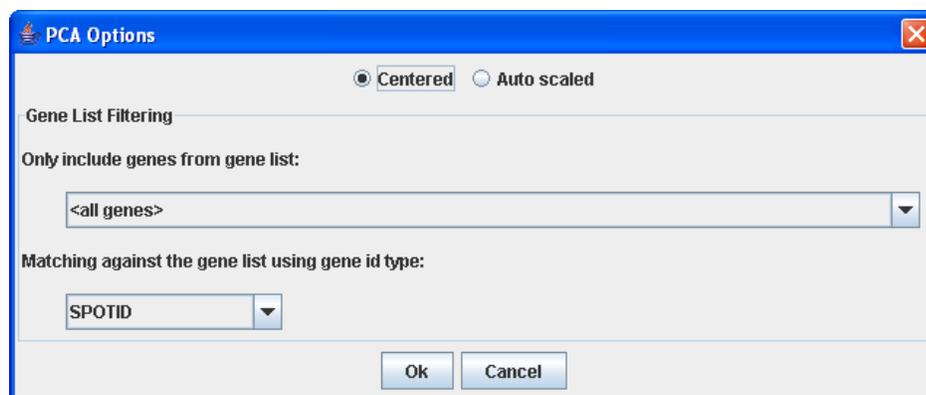


Figure 7-20: PCA options

There are two algorithm method options: Centered and Auto scaled. The user can choose only include genes from a specific gene list by clicking the pull-down list. The user needs to specify the gene ID type for matching. Click OK then PCA results will show up.

In Figure 7-19A, 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 7-19B). 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 7-21. The spot color can be changed by selecting the spot and then right-clicking-> choose new color or highlight the record and right-clicking. The spot shape can be changed to cube or sphere.

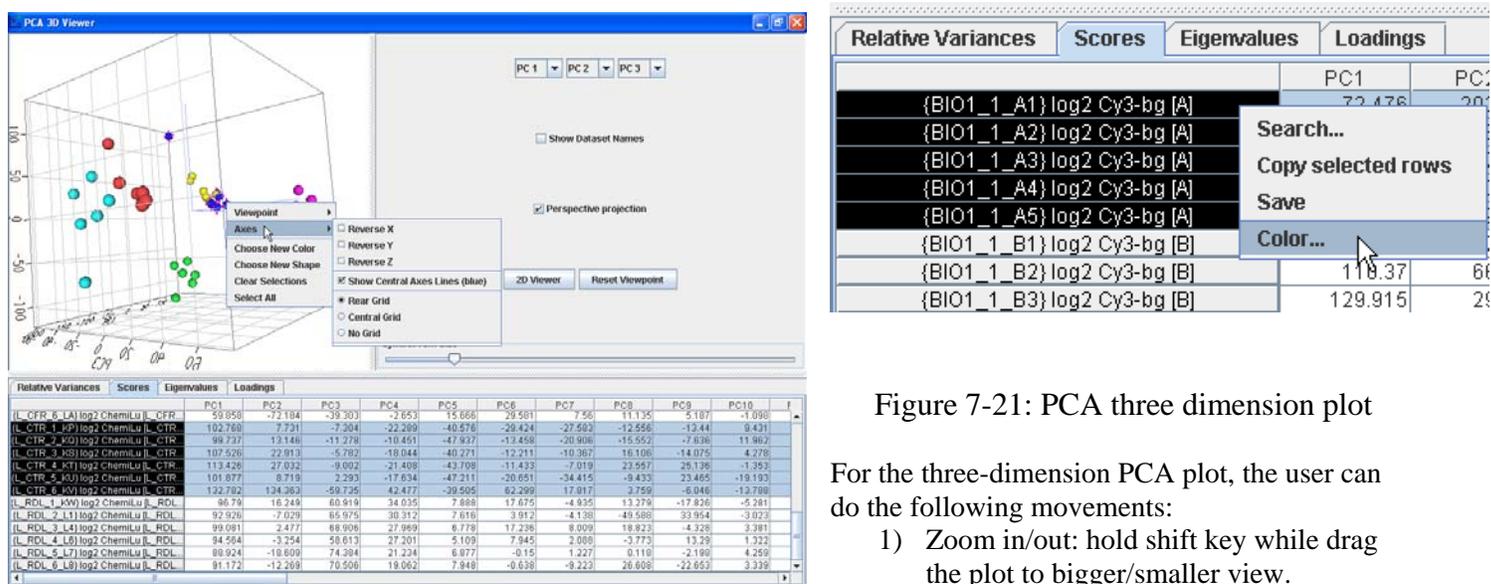


Figure 7-21: PCA three dimension plot

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.

## 7.6. Correlation Matrix

The correlation matrix shows the correlation between column i and column j of the original matrix. It is used in ArrayTrack to visually show the correlation between two groups of data.

Right-click the selected data, choose “Analysis ->correlation matrix. Assign the data into two groups.

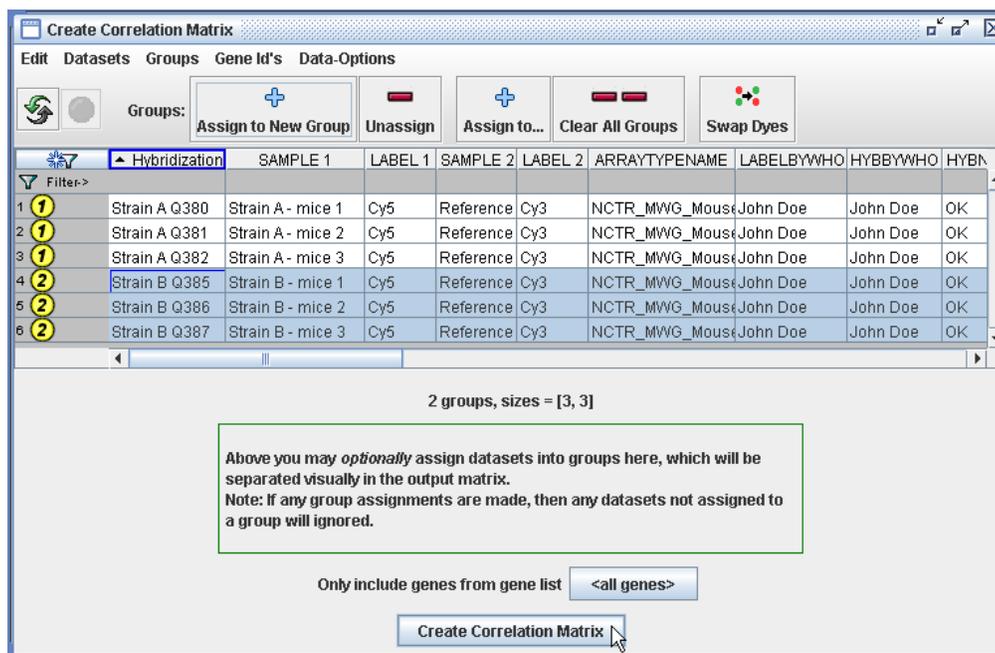
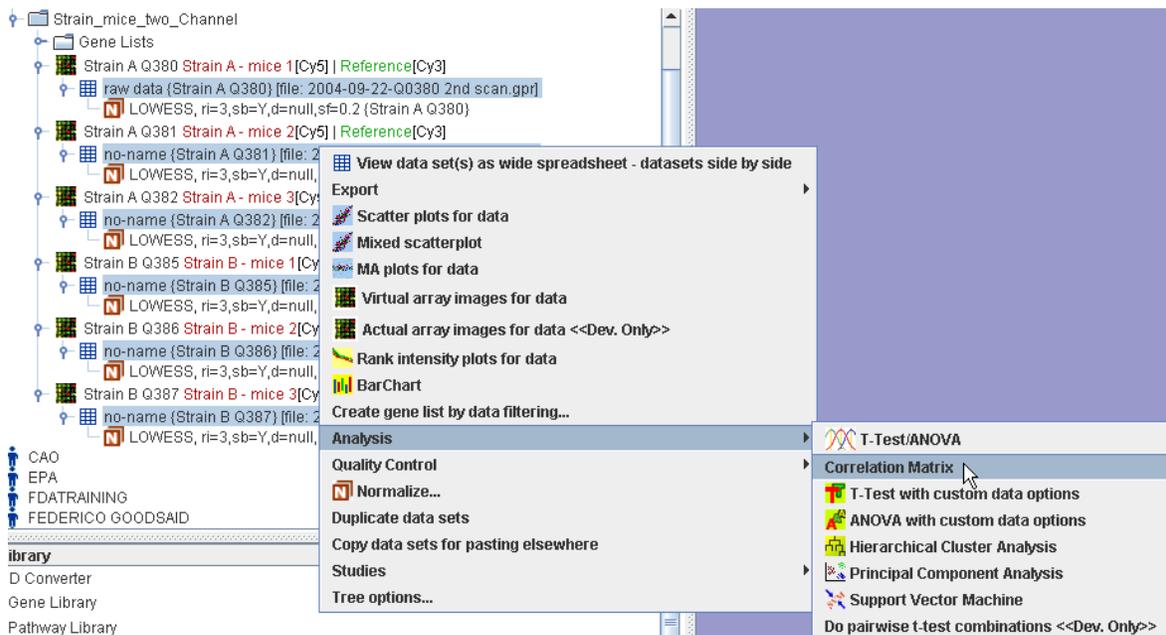


Figure 7-22: correlation matrix

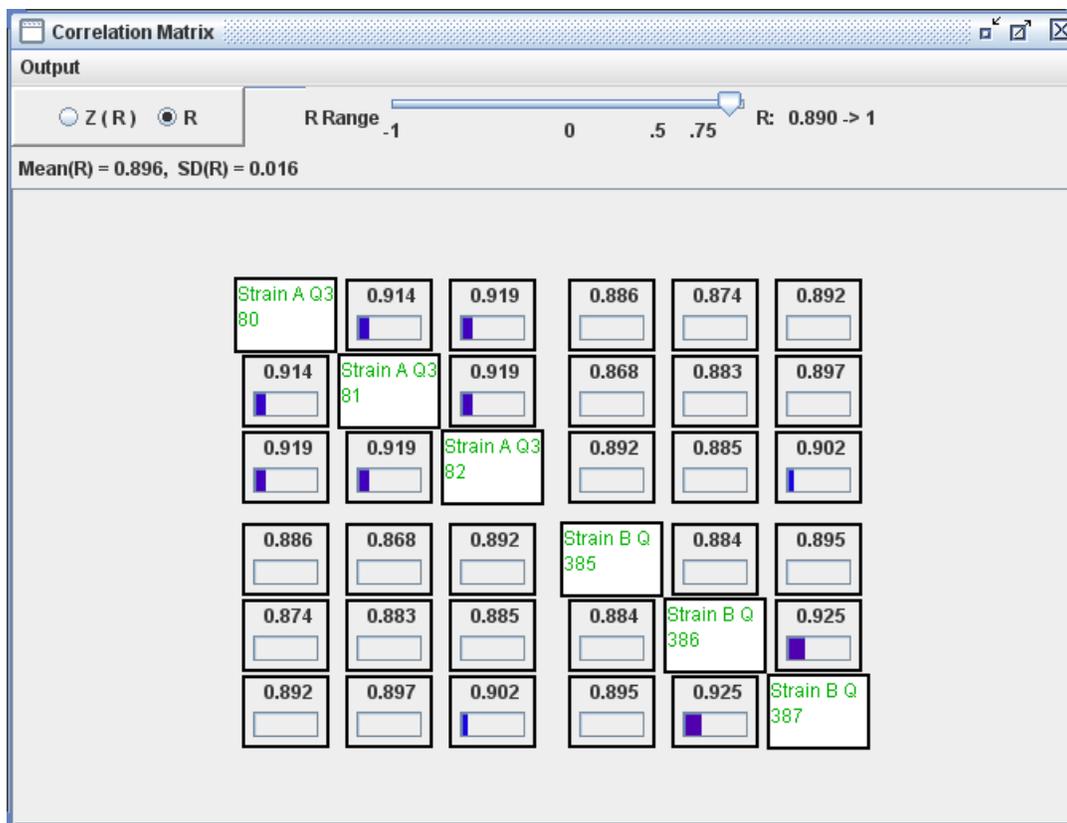


Figure 7-23: correlation matrix result

In Figure 7-23, the diagonal elements of the correlation matrix will be 1 since they are the correlation of a column with itself. The correlation matrix is also symmetric since the correlation of column  $i$  with column  $j$  is the same as the correlation of column  $j$  with column  $i$ . In the matrix, R value is displayed on the top. At the top of the Figure 7-23, there is a sliding bar for correlation factor R. Users can change the R value, and those parts with R value greater than the correlation R will be marked in color.

## 7.7. SAM

SAM (Significance Analysis of Microarrays) is an analysis tool for identifying statistical significant genes in a set of microarray experiments. Before using SAM-test tool in ArrayTrack, users should read SAM manual from website: <http://www-stat.stanford.edu/~tibs/SAM/sam.pdf>. The SAM basic concept and algorithm are not described here. Only instruction for using SAM in ArrayTrack will be addressed as following.

ArrayTrack's SAM is in R version 2.5.1. If you want to compare ArrayTrack SAM results with R SAM results, please make sure you are using the same version of R (2.5.1). The SAM tool in ArrayTrack includes following several analysis types:

- Two class unpaired
- One class
- Multi class
- Survival
- Two class paired
- One class timecourse
- Two class unpaired timecourse
- Two class paired timecourse

### Two class unpaired

First select all the dataset, right-click then choose “Analysis->SAM-Test”, see Figure 7-24.

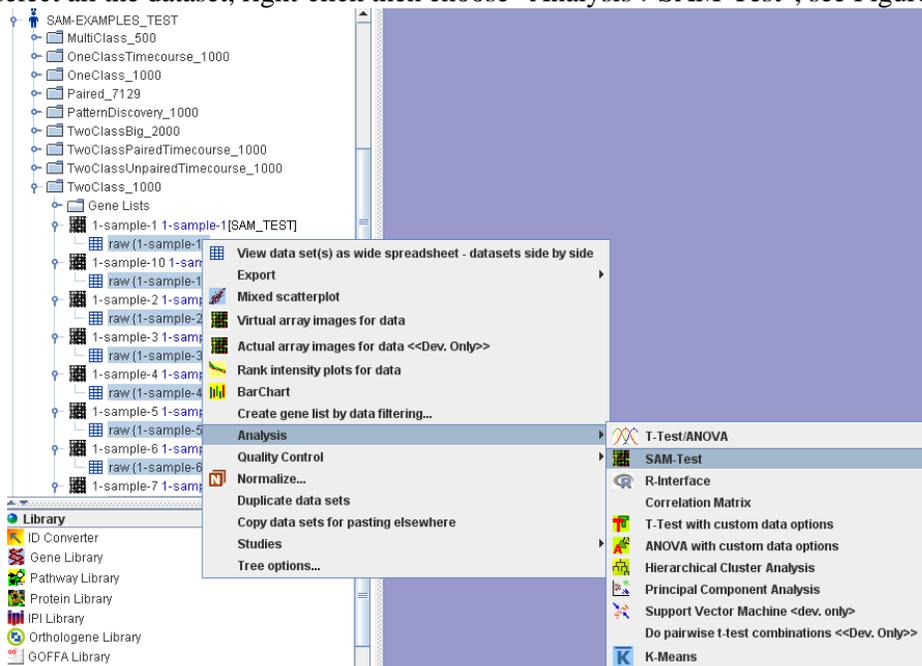


Figure 7-24: SAM test – two class unpaired

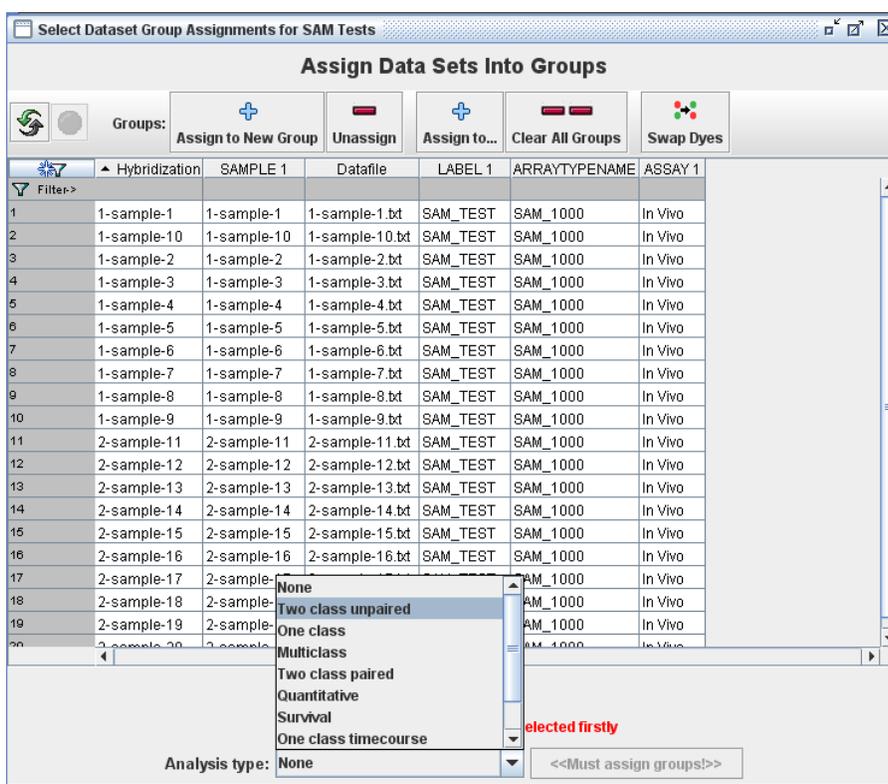


Figure 7-25: SAM test - select analysis type

In Figure 7-25, user needs to select the analysis type which is two class unpaired in this case.

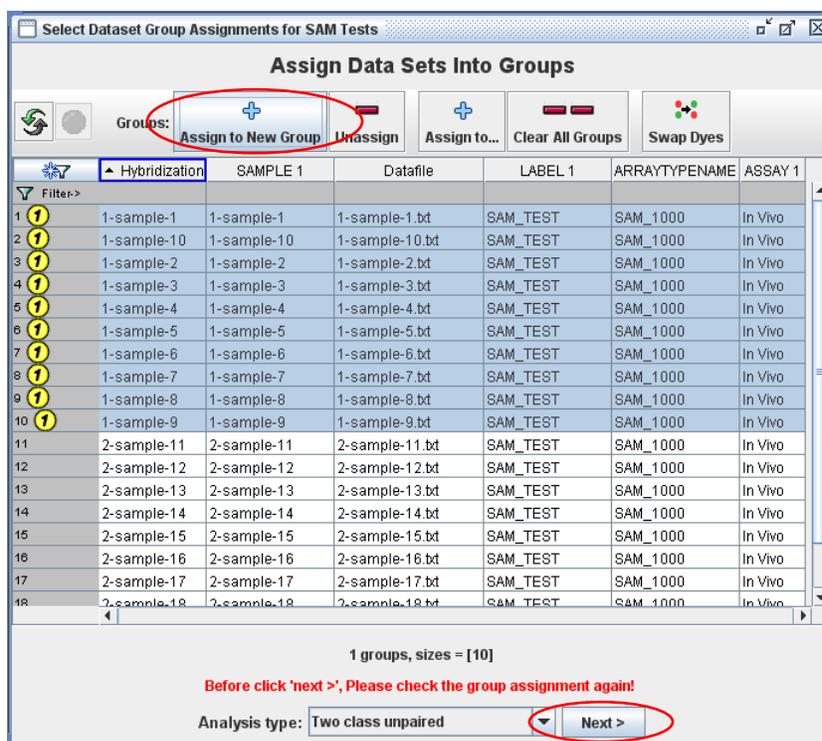


Figure 7-26: assign datasets to groups

In Figure 7-26, user needs to select a group of data, then click “Assign to New Group” button. Repeat the same steps to assign the second group, then click “Next” button.

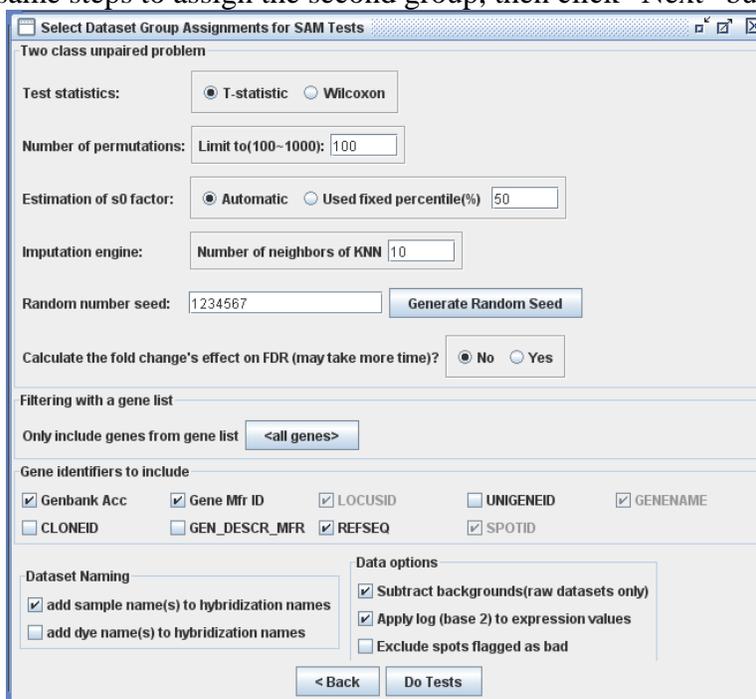


Figure 7-27: option settings for SAM test

There are many options for doing SAM test, Figure 7-27 shows the default settings. At the bottom there are a few options for data, if user wants to use the sample data to compare with the SAM results in Excel, he must uncheck the option “Apply log(base 2) to expression values.”

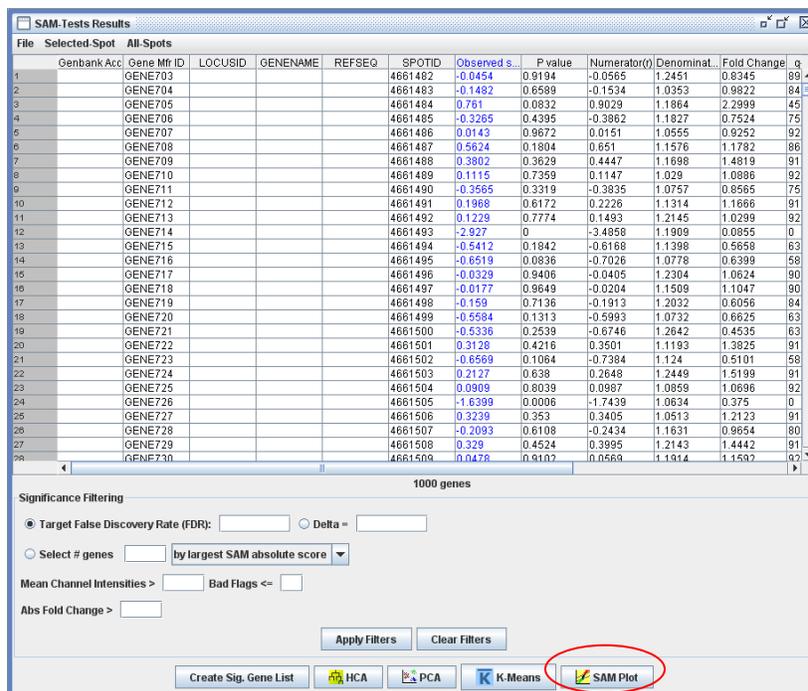


Figure 7-28: SAM test result

Figure 7-28 shows the result of SAM test. This interface is similar with that of T-test result. From the result window user can set criteria to make significance filtering or go further to PCA, HCA, SAM plot by clicking the buttons at the bottom of the window. Figure 7-29 shows the SAM plot.

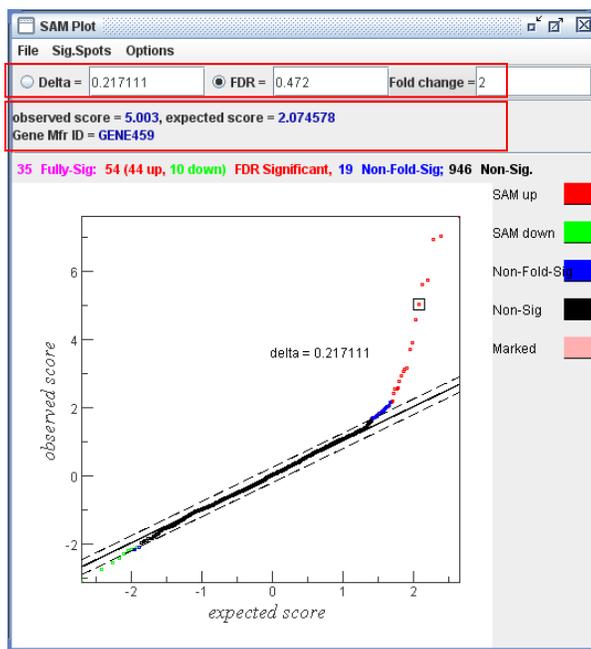


Figure 7-29: SAM plot

In Figure 7-29, users can set values for Delta, FDR and fold change by typing in numbers in the blank boxes at the top of the window. Dragging the dash line in the plot will also change the delta, FDR value. By moving mouse over a spot in the plot, user will be able to see the score value and gene name for that spot that are displayed above the plot.

**One Class**

Select all the dataset, right-click->choose “Analysis” ->”SAM-Test”

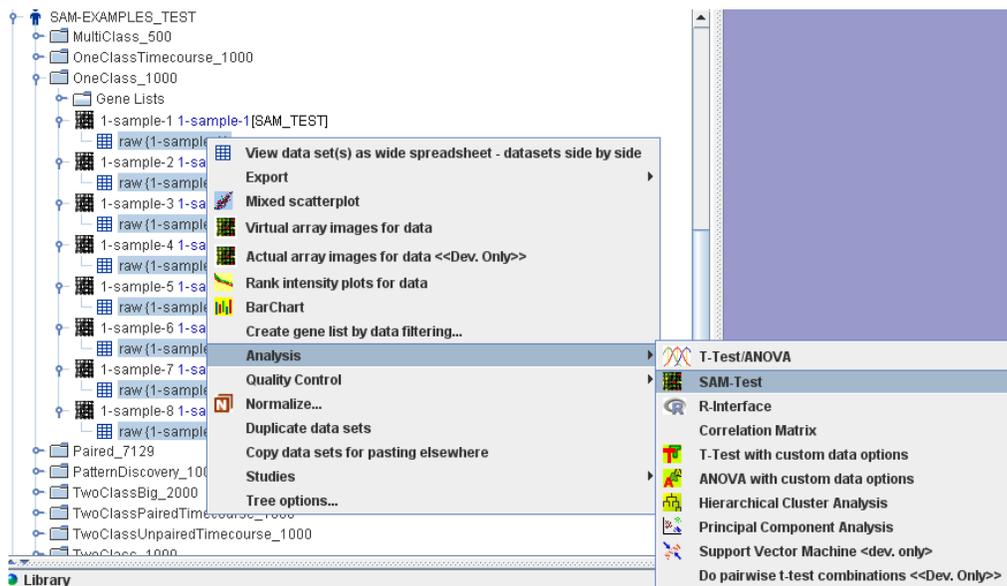


Figure 7-30: SAM test one class

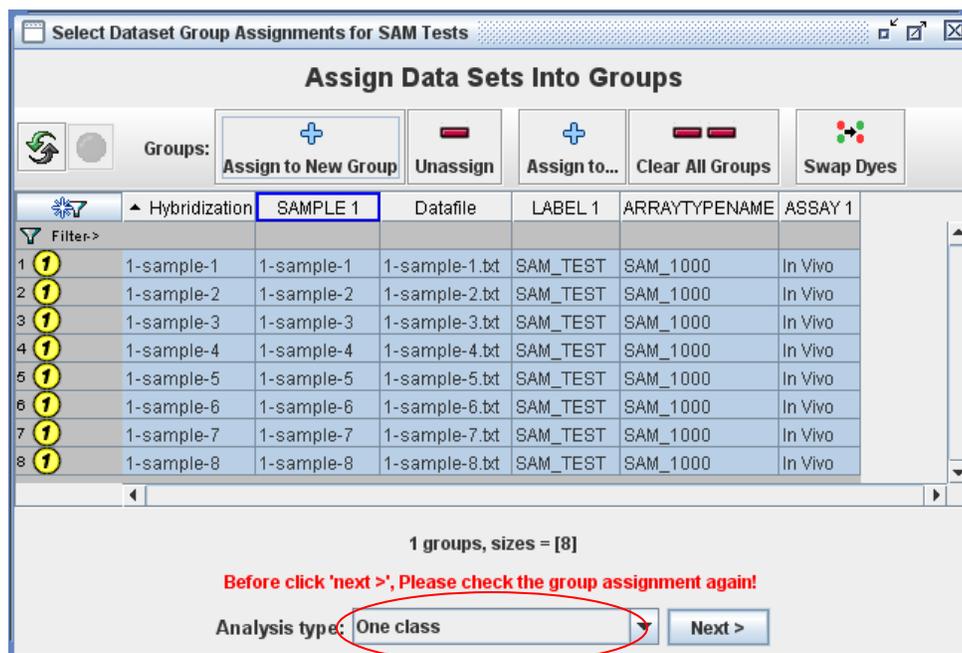


Figure 7-31: Assign all the data as one group

In Figure 7-31, assign all the data into one group, click “Next” button. The rest step is the same as two class unpaired.

**Multi class**

The steps for doing multi class test are similar to one class test, except that the datasets are assigned to multiple groups instead of one group.

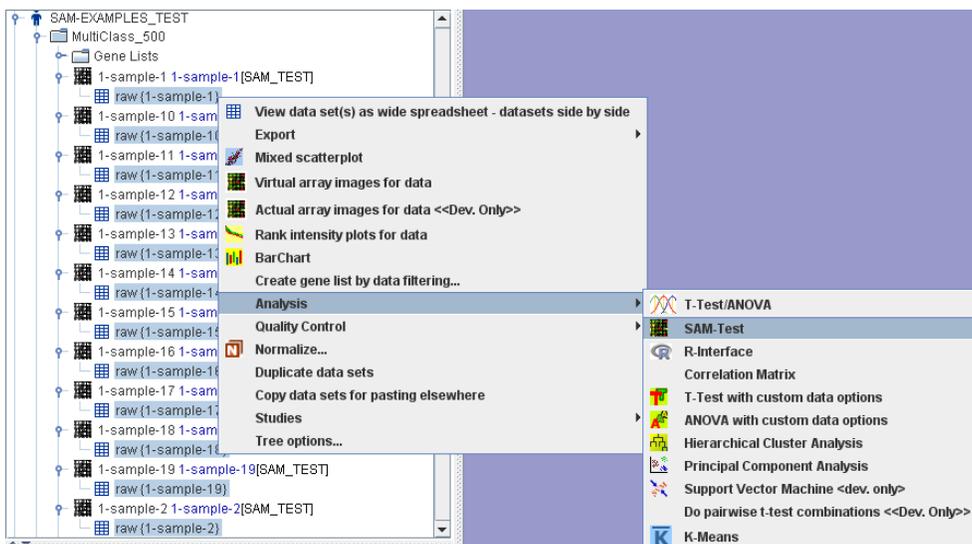


Figure 7-32: select data for SAM multi class test

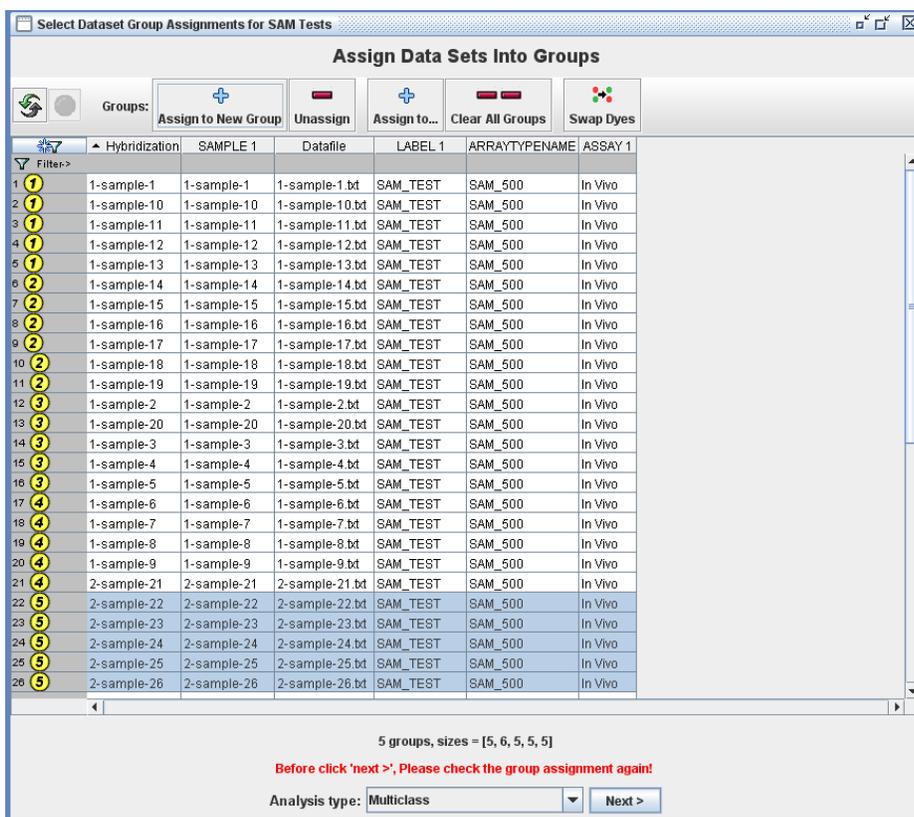


Figure 7-33: assign dataset to multiple groups

Two class paired

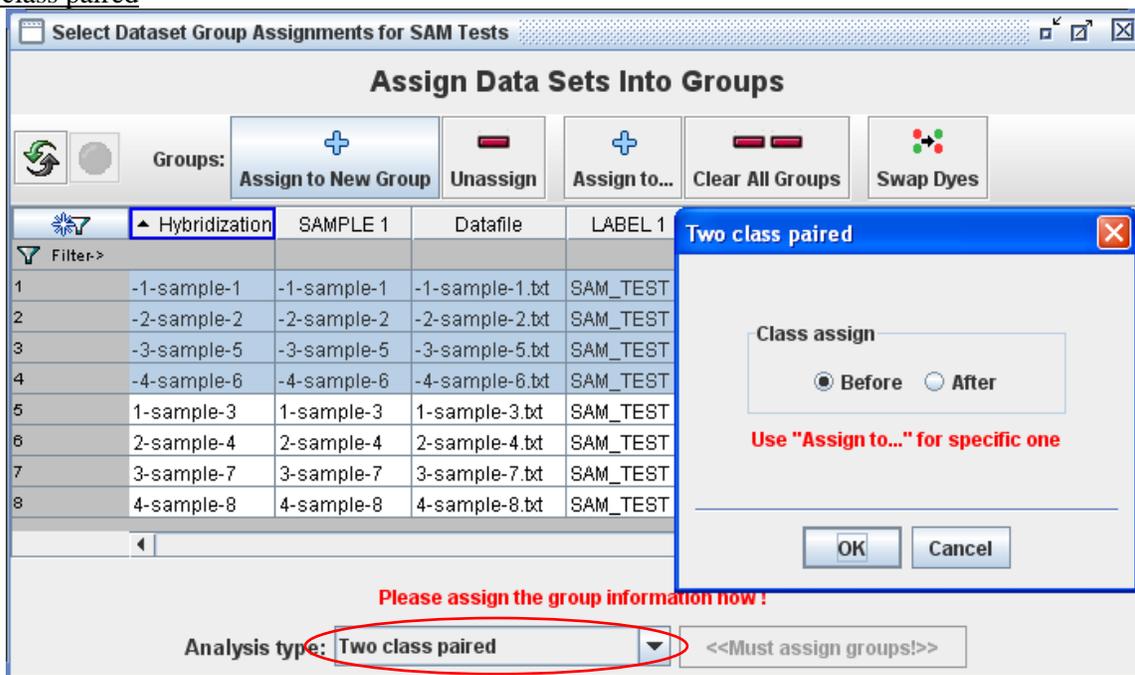


Figure 7-34: assign data for two class paired SAM test

To do two class paired test, select the analysis type first (see Figure 7-34). Select data for the first group, then click button “Assign to New Group”. A pop-up window will let the user choose “Before” or “After”. Choose “Before” for group one, choose “After” for group two. Figure 7-35 shows the assigned two groups. Click “Next” button. The rest steps are the same as other SAM test.

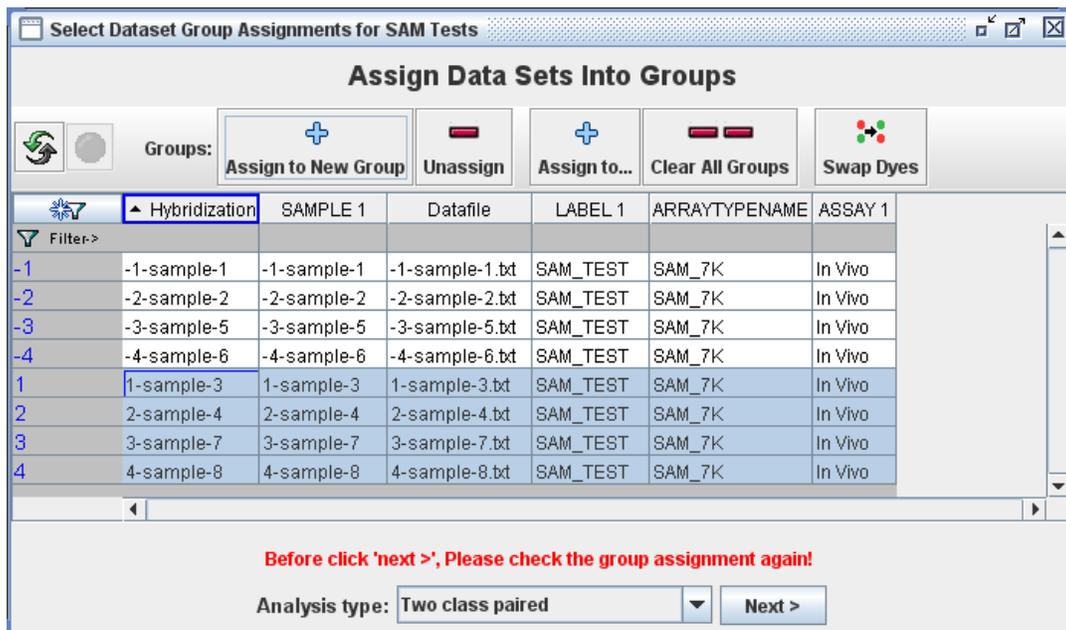


Figure 7-35: assign data to two groups

One class timecourse

Before assign the groups, users need to choose the analysis type (one class timecourse), then highlight the first data group and click “Assign to New Group” button, see Figure 7-36. After the time point window popping up (see inlay), users need to select radio button “Start” and type 1 in “Time” text box. Click OK button.

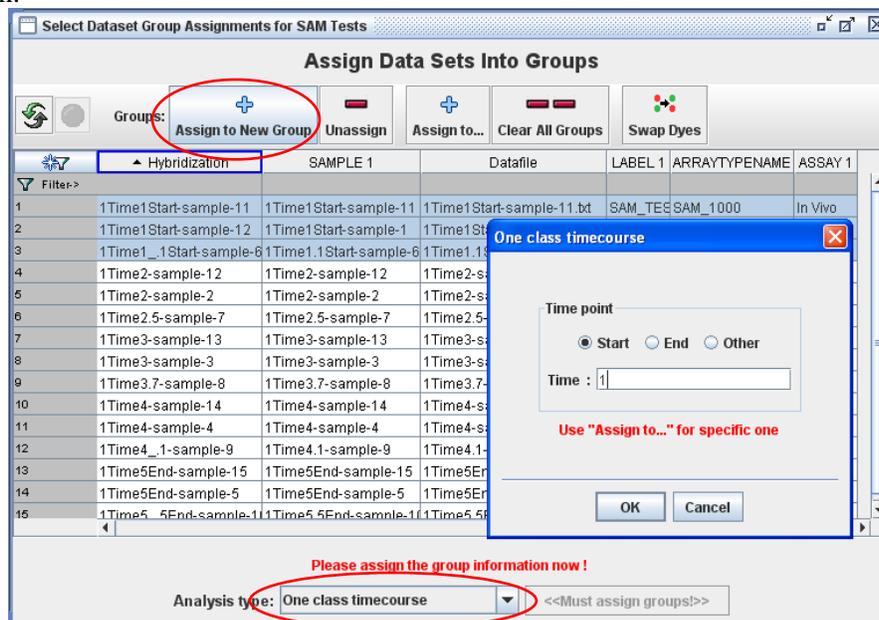


Figure 7-36: one class timecourse – assign data to time point start

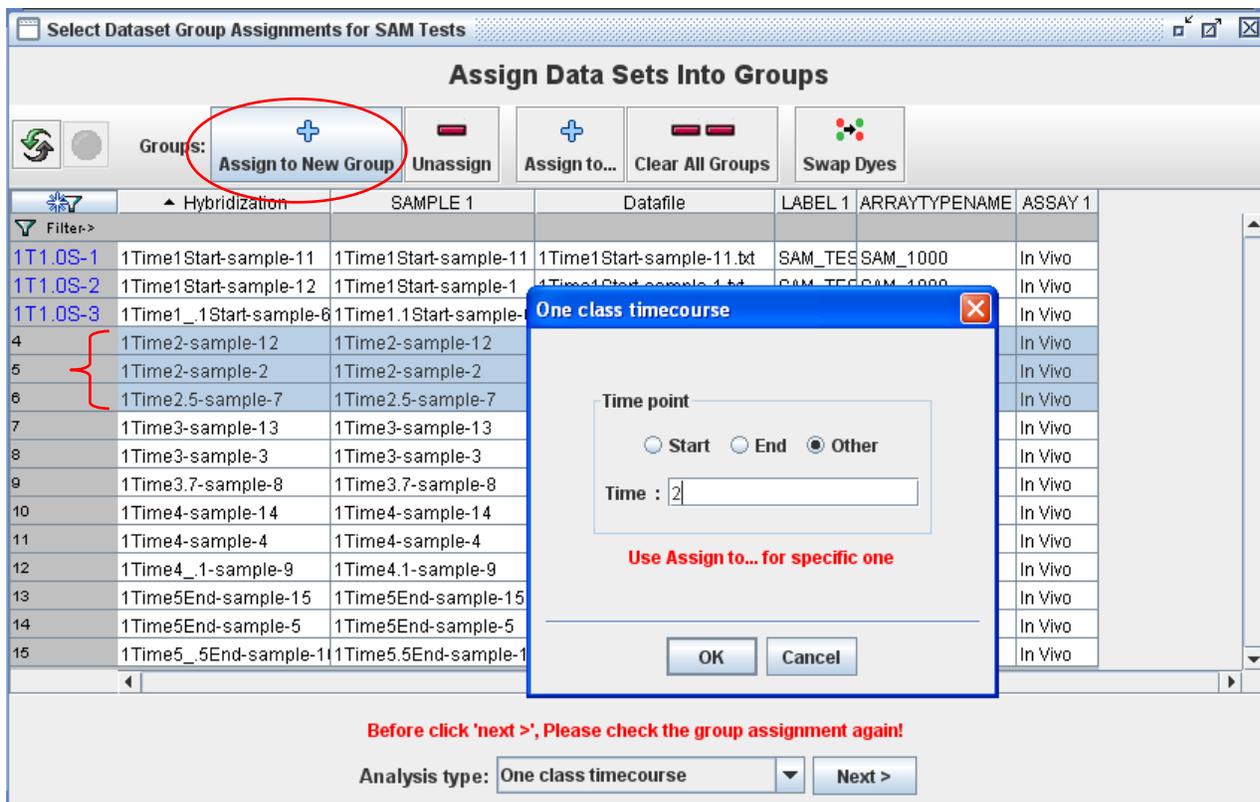


Figure 7-37: one class timecourse – assign data to time point 2

Be aware that in Figure 7-37 the time point option is “other” for the second group.

Repeat the steps to assign the other groups. But for the last group, make sure the option for “End” is checked. Click OK button. See Figure 7-38.

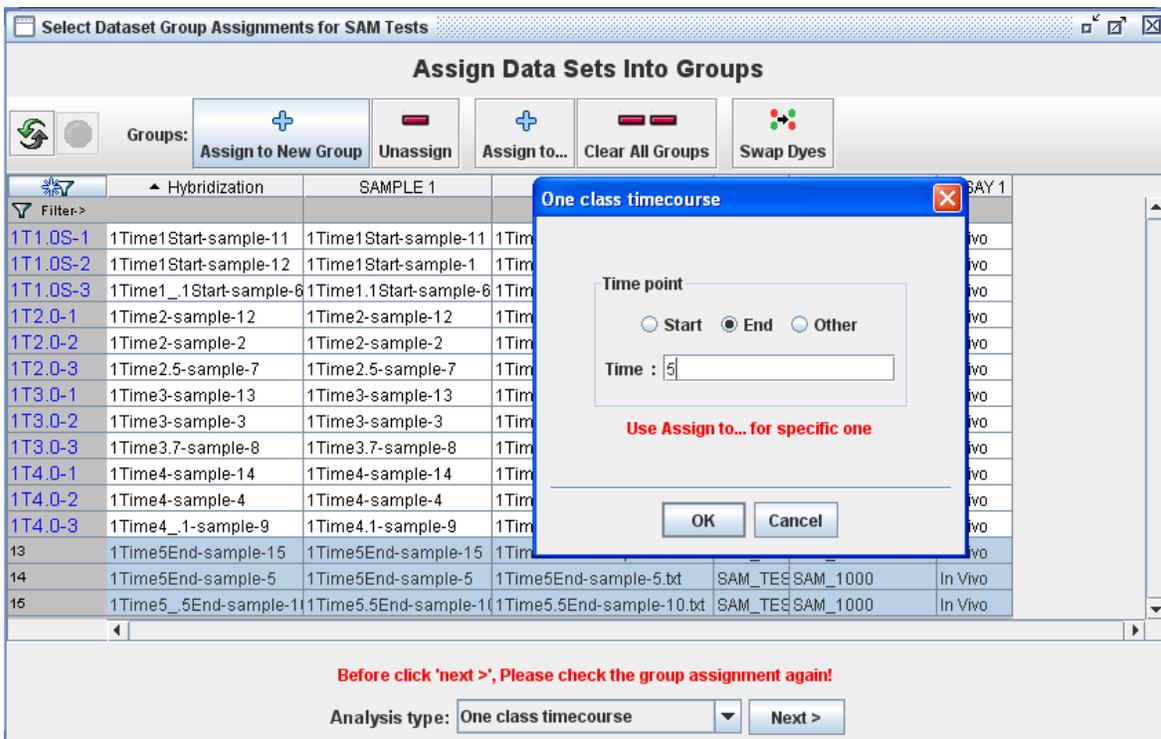


Figure 7-38: one class timecourse – assign data for the end time point

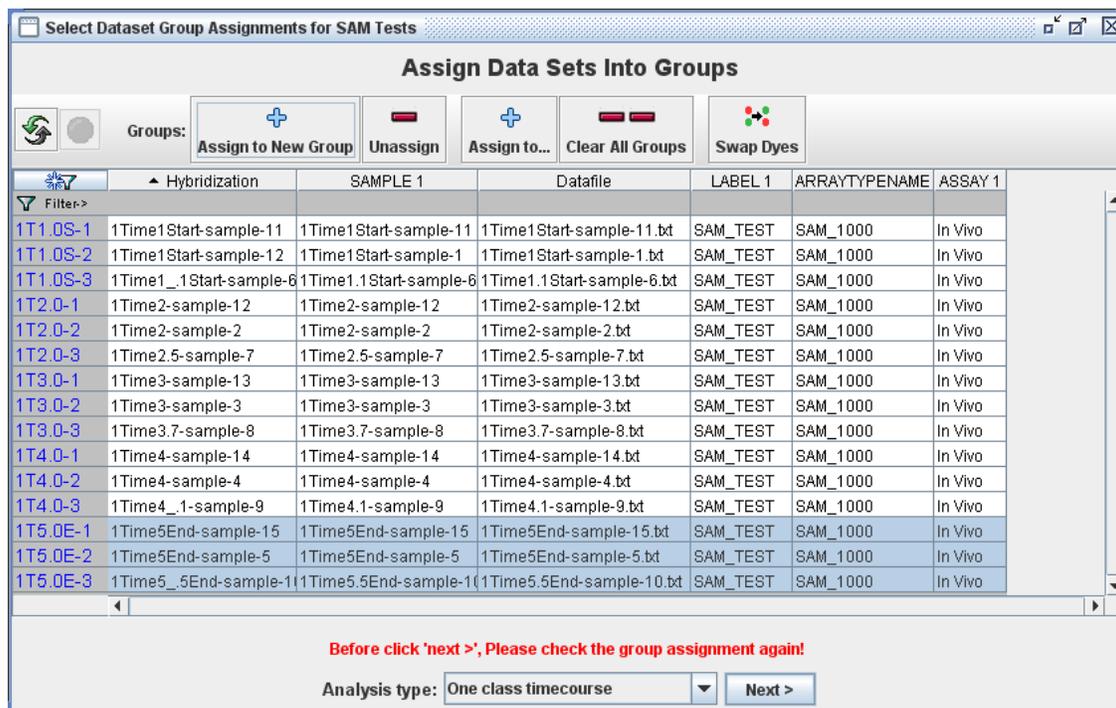


Figure 7-39: one class timecourse – all the data are assigned

Figure 7-39 shows the final look after all the data assigned to groups. Click “Next” button. The rest steps are similar with other SAM test. Figure 7-40 shows the default value for one class timecourse. Click “Do Tests” button will get results.

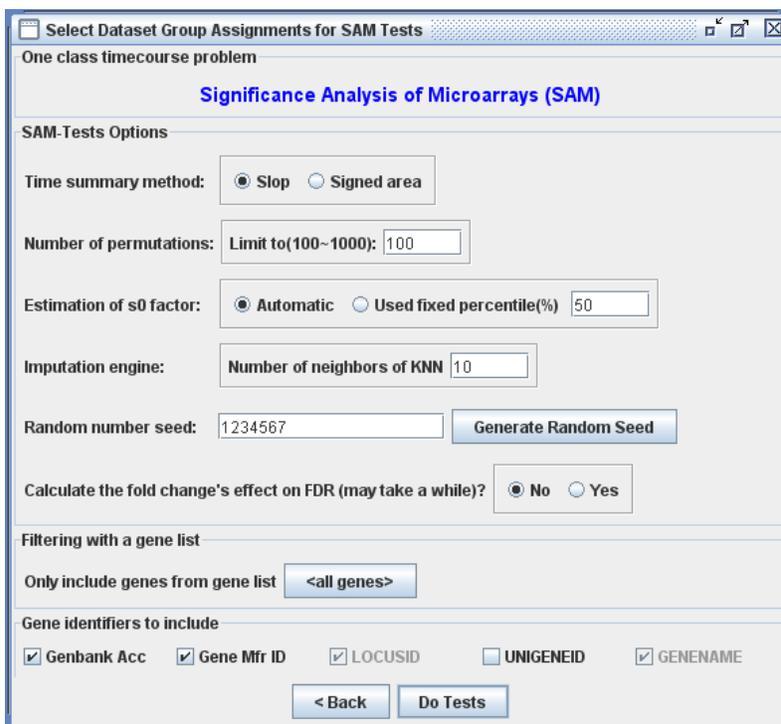


Figure 7-40: default value for one class timecourse

Two class unpaired timecourse

Right-click the selected dataset, choose “Analysis” ->SAM-Test.

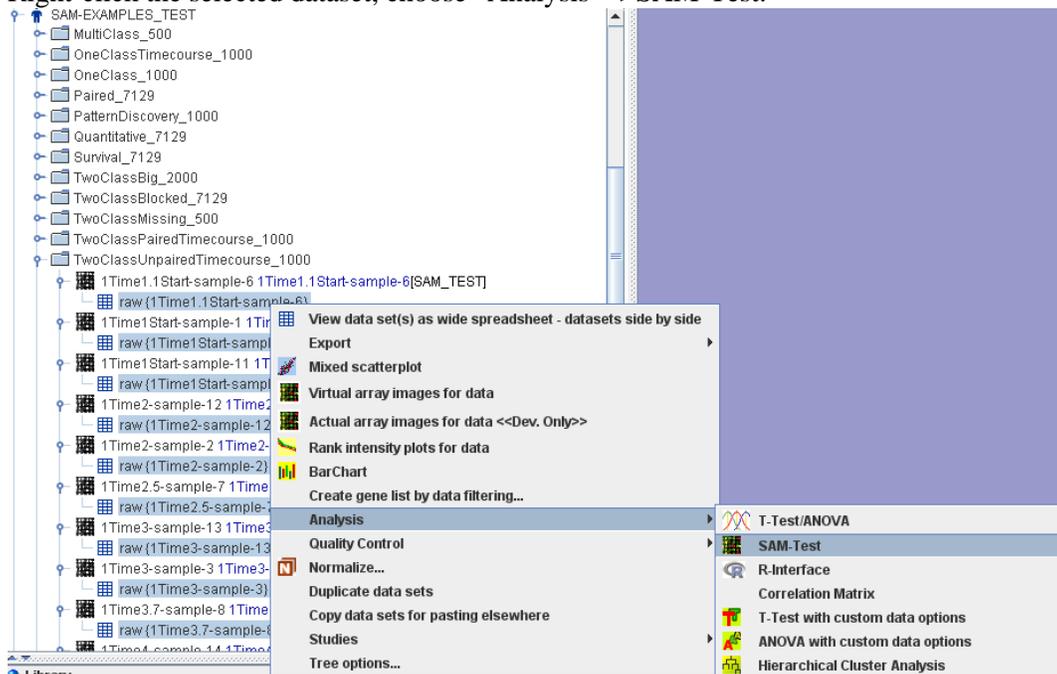


Figure 7-41: two class unpaired timecourse

Make sure that the analysis type is selected before assigning dataset. For two class unpaired timecourse, the steps are similar as one class time course. Users assign start time point, middle point and end time point for class one. And then repeat the steps for class two. Be aware that if class one is assigned control group, then class two should be assigned to treated group. Figure 7-42 shows three hybridizations are assigned to control group time point 1. For time point option, "Start" is selected.

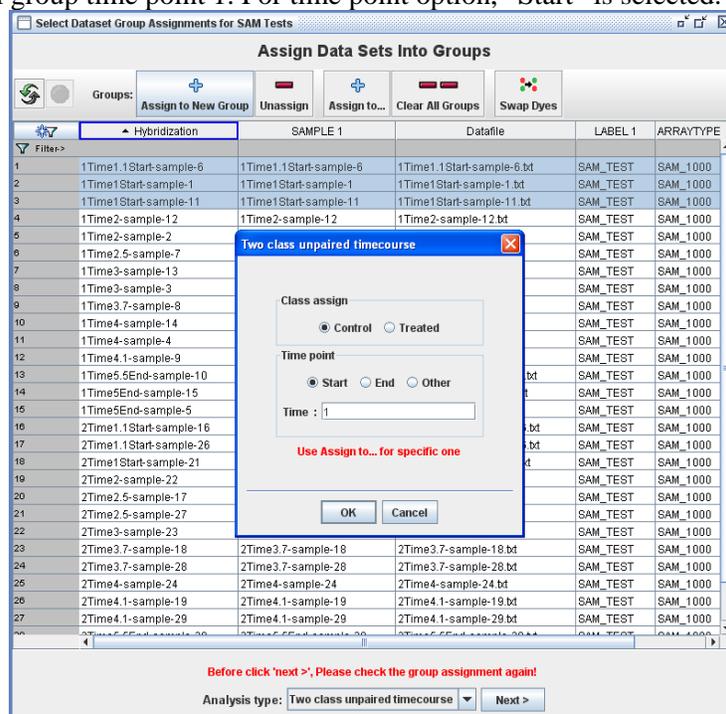


Figure 7-42: two class unpaired timecourse – assign group one (control)

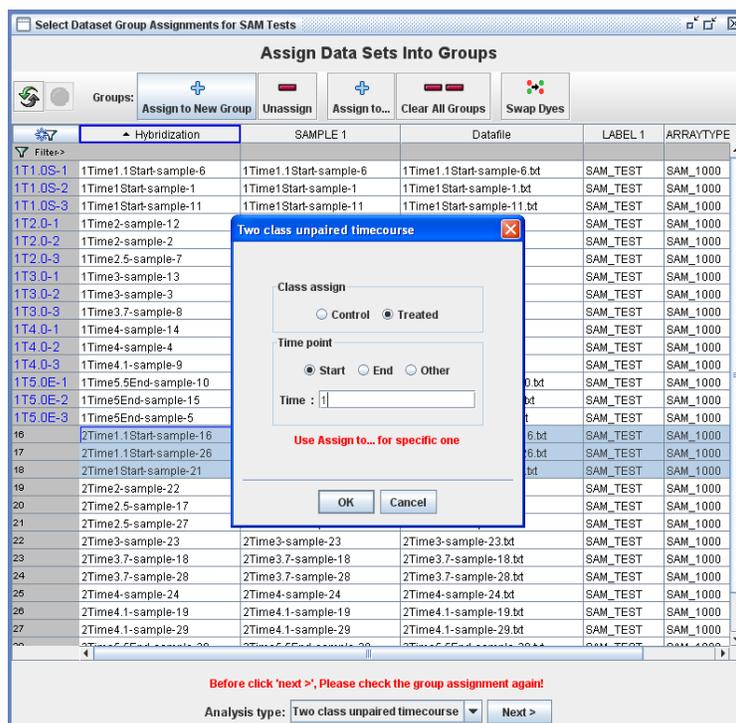


Figure 7-43: two class unpaired timecourse – assign group two (treated)

Figure 7-43 shows that control groups assignment is finished. Then select data for treated group. Repeat the steps to assign group two from start point to end time point. Figure 7-44 shows the final look after assignment is finished. Click “Next” button to get SAM result.

**Assign Data Sets Into Groups**

Groups:

Filter	SAMPLE 1	Datafile	LABEL 1	ARRAYTYPENAME	ASSAY 1	
1T1.OS-1	1Time1.1Start-sample-6	1Time1.1Start-sample-6	1Time1.1Start-sample-6.bt	SAM_TEST	SAM_1000	In Vivo
1T1.OS-2	1Time1Start-sample-1	1Time1Start-sample-1	1Time1Start-sample-1.bt	SAM_TEST	SAM_1000	In Vivo
1T1.OS-3	1Time1Start-sample-11	1Time1Start-sample-11	1Time1Start-sample-11.bt	SAM_TEST	SAM_1000	In Vivo
1T2.O-1	1Time2-sample-12	1Time2-sample-12	1Time2-sample-12.bt	SAM_TEST	SAM_1000	In Vivo
1T2.O-2	1Time2-sample-2	1Time2-sample-2	1Time2-sample-2.bt	SAM_TEST	SAM_1000	In Vivo
1T2.O-3	1Time2.5-sample-7	1Time2.5-sample-7	1Time2.5-sample-7.bt	SAM_TEST	SAM_1000	In Vivo
1T3.O-1	1Time3-sample-13	1Time3-sample-13	1Time3-sample-13.bt	SAM_TEST	SAM_1000	In Vivo
1T3.O-2	1Time3-sample-3	1Time3-sample-3	1Time3-sample-3.bt	SAM_TEST	SAM_1000	In Vivo
1T3.O-3	1Time3.7-sample-8	1Time3.7-sample-8	1Time3.7-sample-8.bt	SAM_TEST	SAM_1000	In Vivo
1T4.O-1	1Time4-sample-14	1Time4-sample-14	1Time4-sample-14.bt	SAM_TEST	SAM_1000	In Vivo
1T4.O-2	1Time4-sample-4	1Time4-sample-4	1Time4-sample-4.bt	SAM_TEST	SAM_1000	In Vivo
1T4.O-3	1Time4.1-sample-9	1Time4.1-sample-9	1Time4.1-sample-9.bt	SAM_TEST	SAM_1000	In Vivo
1T5.OE-1	1Time5.5End-sample-10	1Time5.5End-sample-10	1Time5.5End-sample-10.bt	SAM_TEST	SAM_1000	In Vivo
1T5.OE-2	1Time5End-sample-15	1Time5End-sample-15	1Time5End-sample-15.bt	SAM_TEST	SAM_1000	In Vivo
1T5.OE-3	1Time5End-sample-5	1Time5End-sample-5	1Time5End-sample-5.bt	SAM_TEST	SAM_1000	In Vivo
2T1.OS-1	2Time1.1Start-sample-16	2Time1.1Start-sample-16	2Time1.1Start-sample-16.bt	SAM_TEST	SAM_1000	In Vivo
2T1.OS-2	2Time1.1Start-sample-26	2Time1.1Start-sample-26	2Time1.1Start-sample-26.bt	SAM_TEST	SAM_1000	In Vivo
2T1.OS-3	2Time1Start-sample-21	2Time1Start-sample-21	2Time1Start-sample-21.bt	SAM_TEST	SAM_1000	In Vivo
2T2.O-1	2Time2-sample-22	2Time2-sample-22	2Time2-sample-22.bt	SAM_TEST	SAM_1000	In Vivo
2T2.O-2	2Time2.5-sample-17	2Time2.5-sample-17	2Time2.5-sample-17.bt	SAM_TEST	SAM_1000	In Vivo
2T2.O-3	2Time2.5-sample-27	2Time2.5-sample-27	2Time2.5-sample-27.bt	SAM_TEST	SAM_1000	In Vivo
2T3.O-1	2Time3-sample-23	2Time3-sample-23	2Time3-sample-23.bt	SAM_TEST	SAM_1000	In Vivo
2T3.O-2	2Time3.7-sample-18	2Time3.7-sample-18	2Time3.7-sample-18.bt	SAM_TEST	SAM_1000	In Vivo
2T3.O-3	2Time3.7-sample-28	2Time3.7-sample-28	2Time3.7-sample-28.bt	SAM_TEST	SAM_1000	In Vivo
2T4.O-1	2Time4-sample-24	2Time4-sample-24	2Time4-sample-24.bt	SAM_TEST	SAM_1000	In Vivo
2T4.O-2	2Time4.1-sample-19	2Time4.1-sample-19	2Time4.1-sample-19.bt	SAM_TEST	SAM_1000	In Vivo
2T4.O-3	2Time4.1-sample-29	2Time4.1-sample-29	2Time4.1-sample-29.bt	SAM_TEST	SAM_1000	In Vivo
2T5.OE-1	2Time5.5End-sample-20	2Time5.5End-sample-20	2Time5.5End-sample-20.bt	SAM_TEST	SAM_1000	In Vivo
2T5.OE-2	2Time5.5End-sample-30	2Time5.5End-sample-30	2Time5.5End-sample-30.bt	SAM_TEST	SAM_1000	In Vivo
2T5.OE-3	2Time5End-sample-25	2Time5End-sample-25	2Time5End-sample-25.bt	SAM_TEST	SAM_1000	In Vivo

Before click 'next >', Please check the group assignment again!

Analysis type:

Figure 7-44: finished assignment for two class unpaired timecourse

Survival

Right-click the selected data set, choose “Analysis”->SAM-test.

SAM-EXAMPLES\_TEST

- MultiClass\_500
- OneClassTimecourse\_1000
- OneClass\_1000
- Paired\_7129
- PatternDiscovery\_1000
- Quantitative\_7129
- Survival\_7129
- Gene Lists
- (1,3,1)-sample-6 (1,3,1)-sample-6[SAM\_TEST]
- raw((1,3,1)-sample-6)
- (1151,0,0)-sample-2 (1)
- raw((1151,0,0)-sam
- (1326,3,1)-sample-1 (1)
- raw((1326,3,1)-sam
- (145,0,1)-sample-4 (14)
- raw((145,0,1)-samo
- (506,1,1)-sample-7 (50)
- raw((506,1,1)-samo
- (57,8,1)-sample-5 (57)
- raw((57,8,1)-samo
- (605,4,0)-sample-3 (60)
- raw((605,4,0)-samo
- (623,0,1)-sample-9 (62)
- raw((623,0,1)-samo
- TwoClassBig\_2000
- TwoClassBlocked\_7129
- TwoClassMissing\_500
- TwoClassPairedTimecourse

View data set(s) as wide spreadsheet - datasets side by side

Export

Mixed scatterplot

Virtual array images for data

Actual array images for data <<Dev. Only>>

Rank intensity plots for data

Bar Chart

Create gene list by data filtering...

Analysis

- T-Test/ANOVA
- SAM-Test**
- R-Interface
- Correlation Matrix
- T-Test with custom data options
- ANOVA with custom data options
- Hierarchical Cluster Analysis

Quality Control

Normalize...

Duplicate data sets

Copy data sets for pasting elsewhere

Studies

Tree options...

Figure 7-45: SAM-test Survival

Remember to select analysis type before assigning. This data set has two groups. The first number in parenthesis represent time, and the second number in parenthesis (1 or 0) represents “died” (1) or “censored” (0) group.

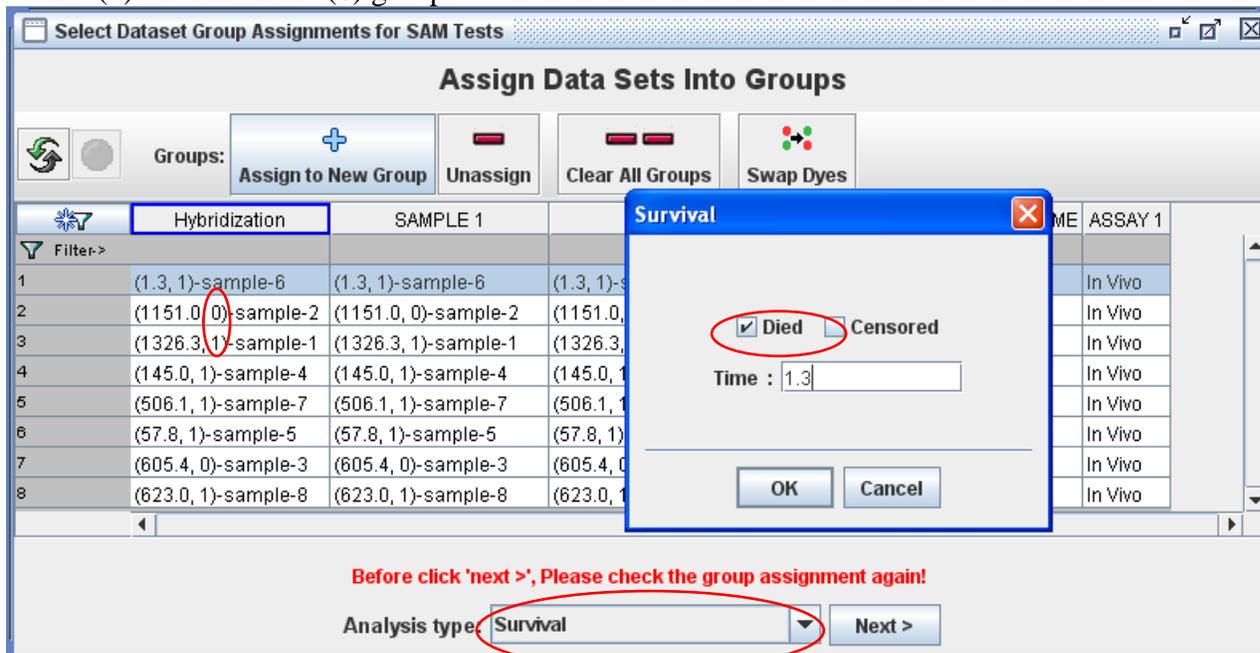


Figure 7-46: SAM-test survival – assign group

In Figure 7-46, the first data “(1.3, 1)-sample-6” of “died” group is selected. After clicking “Assign to New Group” button, the user needs to type in time for the first data (e.g. 1.3). Click OK button. Repeat the previous steps to assign the other data of group 1. See Figure 7-47.

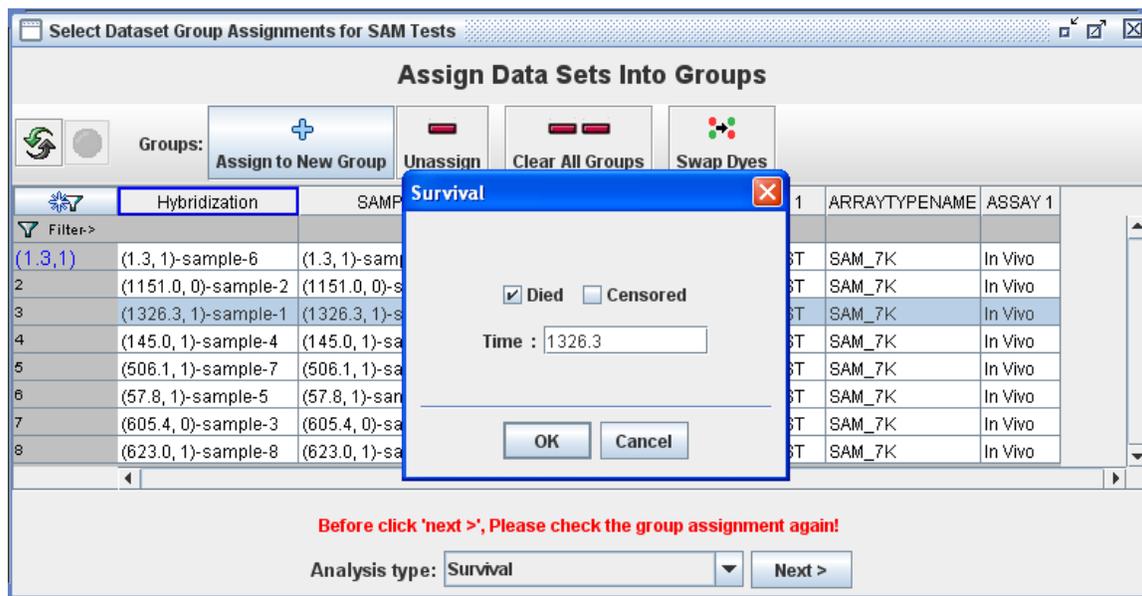


Figure 7-47: SAM-test survival -assign other data

Figure 7-48 shows that all the data of group 1 has been assigned. Then select data in group 0 (censored) to repeat the same steps to assign (see Figure 7-49).

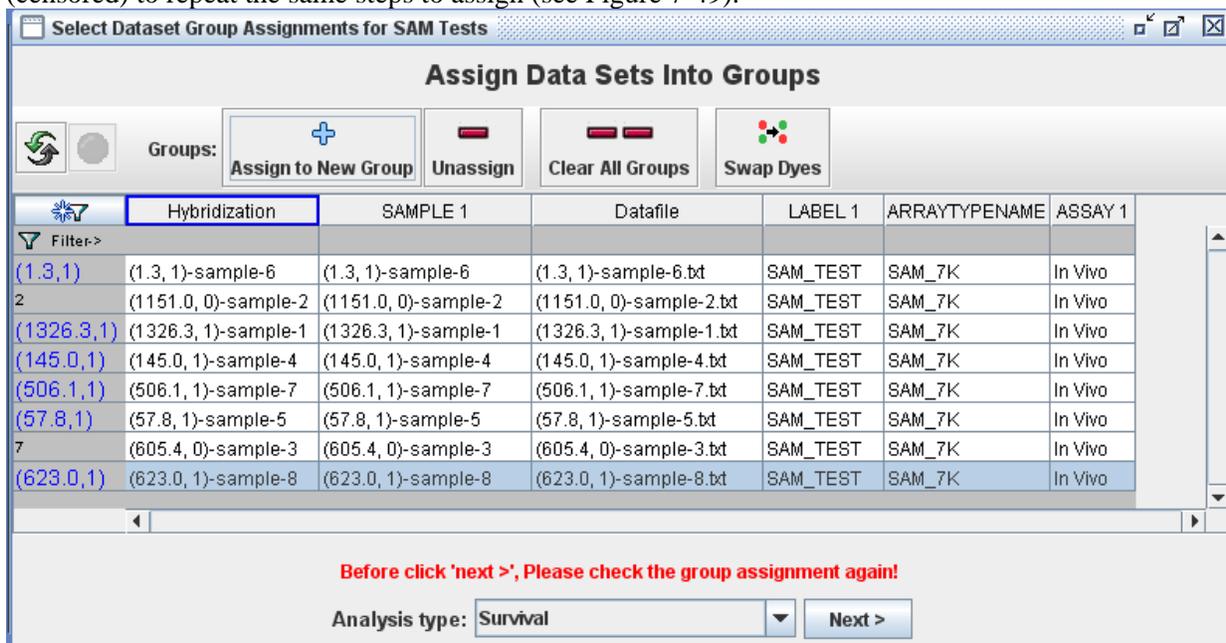


Figure 7-48: SAM-test survival – group 1 has been assigned

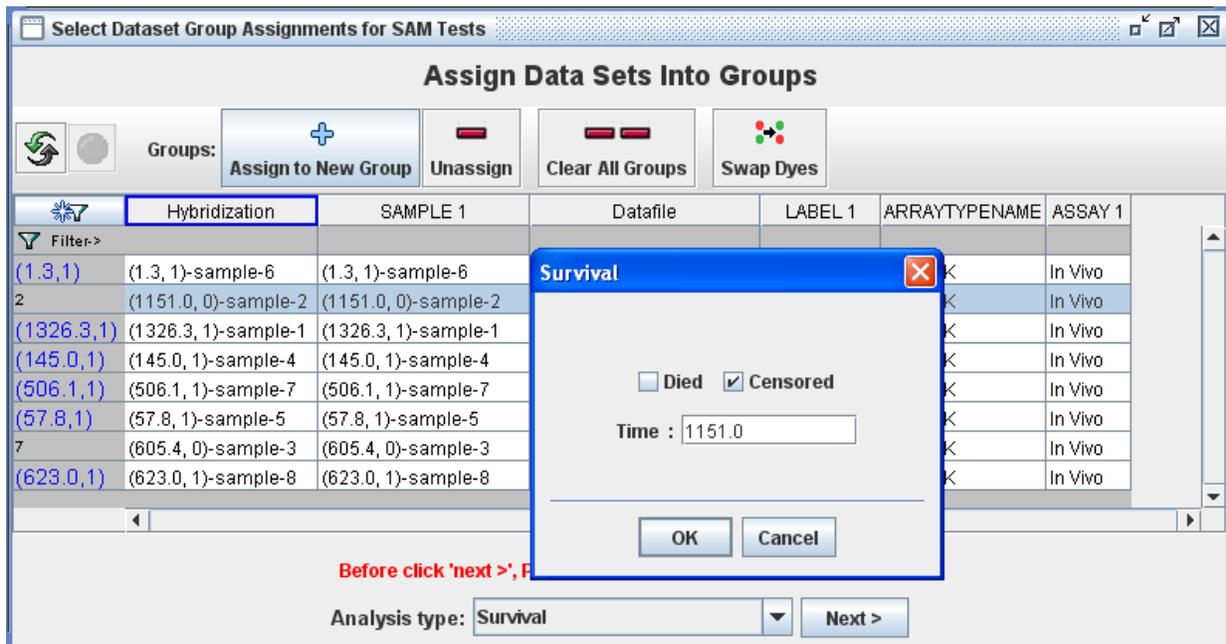


Figure 7-49: SAM-test survival – assign group 0

Figure 7-50 shows that all the data are assigned. Click “Next” button. The rest steps will be same as other SAM test.

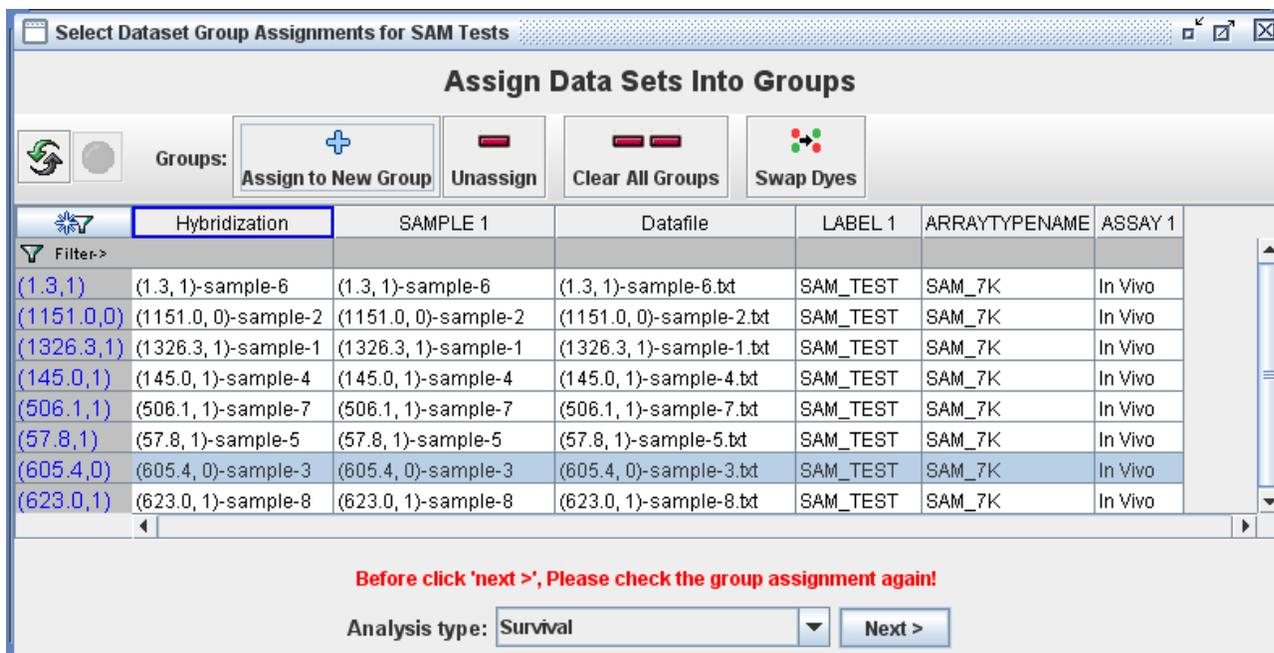


Figure 7-50: finished assignment for survival

### 7.8 K-Means

K-Means is a new feature in ArrayTrack 3.4 version. There are several ways to access K-means: 1) from T-test/Anova result (see Figure 7-51), 2) from SAM-test result (see Figure 7-28), 3) right-clicking the selected data sets, choose “Analysis” -> “K-Means”. See Figure 7-52.

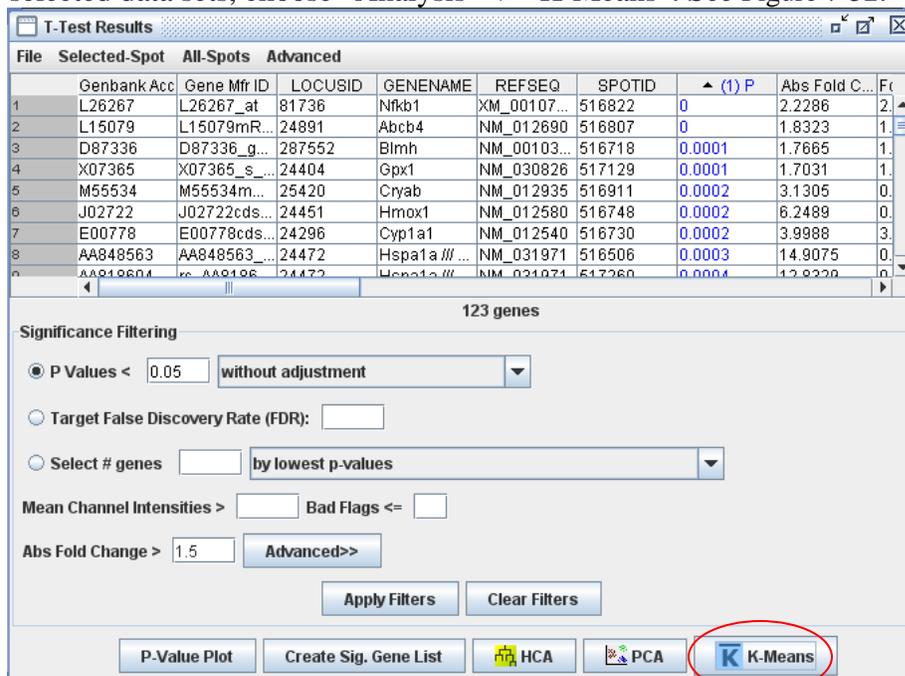


Figure 7-51: access K-Means from T-test result

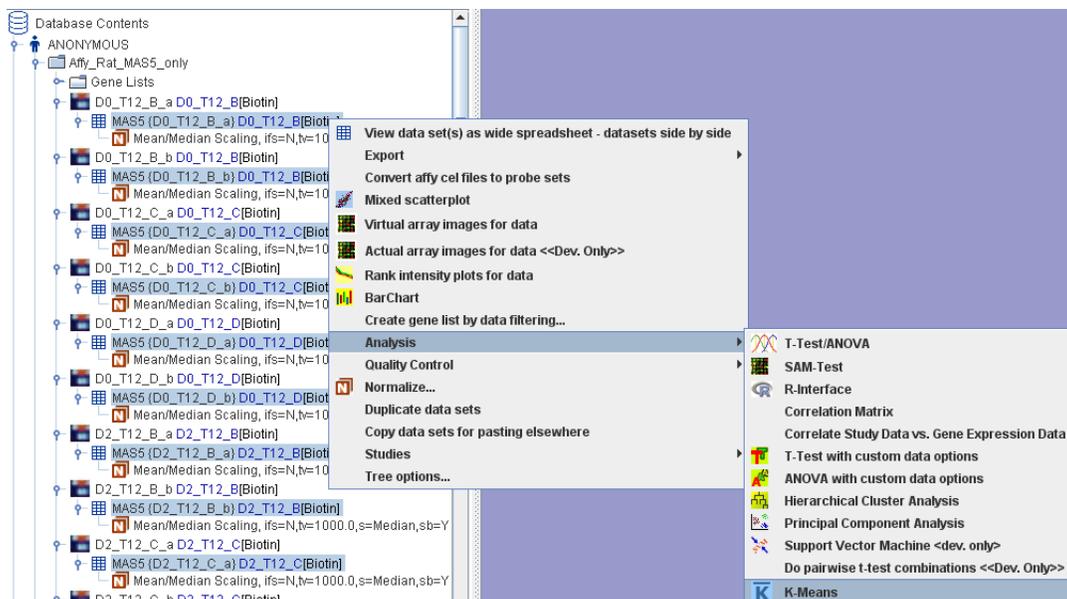


Figure 7-52: K-means

If users access K-Mean by the above three ways, the following window (Figure 7-53) will pop-up, giving the options for K-mean with default values. Users can select gene list so only these genes will be included in K-Means. At the bottom there are options for dataset naming. Sample names or dye names can be added to hybridization names.

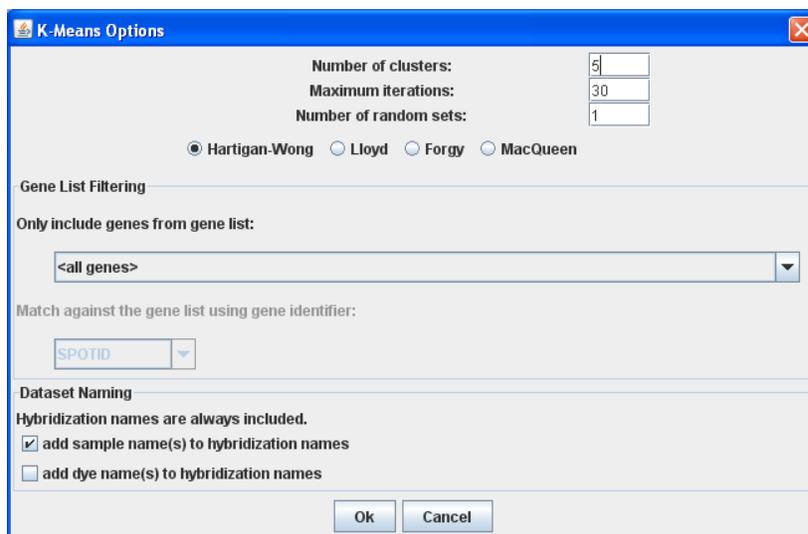


Figure 7-53: K-Means options

Figure 7-54 shows the result of K-mean via T-test window. The left part is the heat map of K-mean, while the right side is the table listing the result values. At the bottom of the left part, there are options for displaying 2D PCA view or 3D PCA view. Figure 7-55 shows the 2D PCA view and 3D PCA view plot of genes by K-mean analysis.

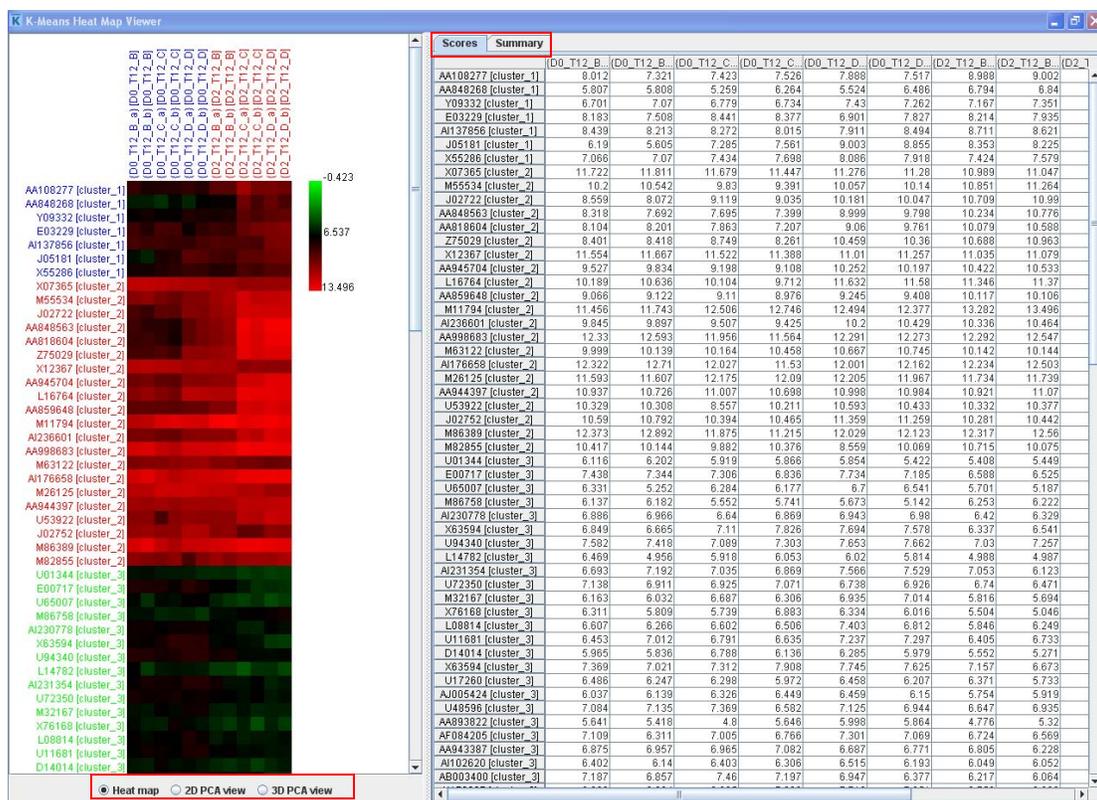


Figure 7-54: Heatmap of K-means

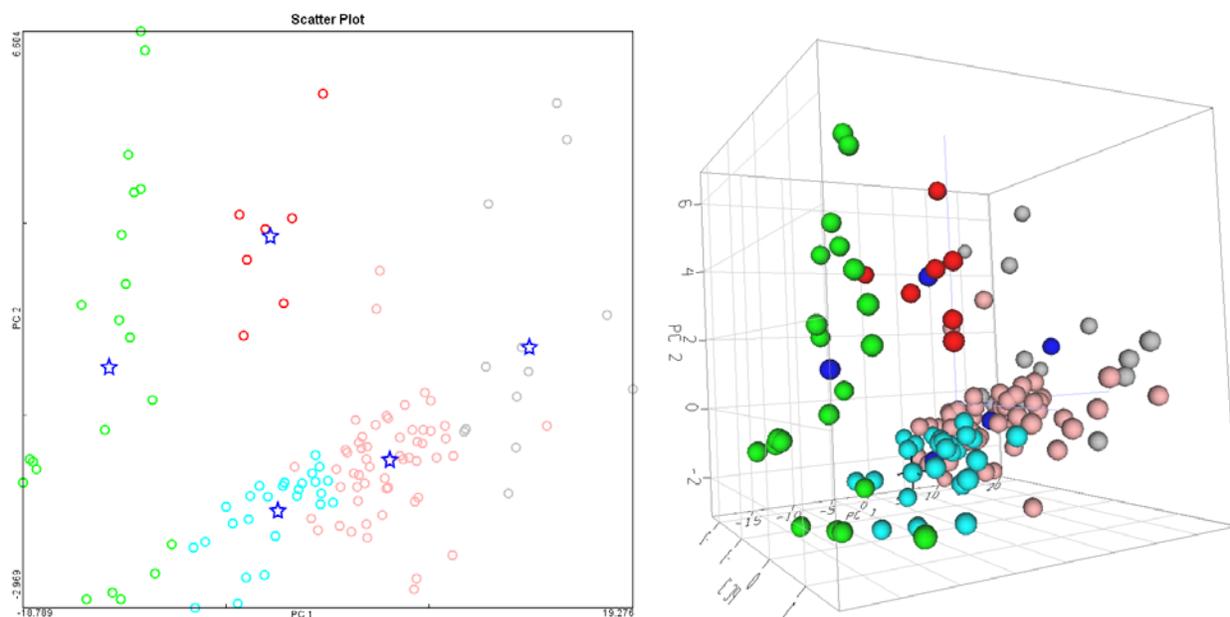


Figure 7-55: K-mean 2D and 3D view

AT the right side of Figure 7-54, there are two tabs: Scores and Summary. Clicking Summary tab will bring out the summary for the K-Means, see Figure 7-56.

Scores	Summary
Algorithm for K-Means: Hartigan-Wong	
Number of Clusters: 5	
Number of maximum iterations: 30	
nNumber of random sets: 1	
The within-cluster sum of squares for each cluster:	
Cluster 1: 48.913	
Cluster 2: 354.598	
Cluster 3: 358.372	
Cluster 4: 226.021	
Cluster 5: 244.098	

Figure 7-56: summary of K-means

## Chapter 8 Working with Tools: Visualization

### 8.1 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 8-1): (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.

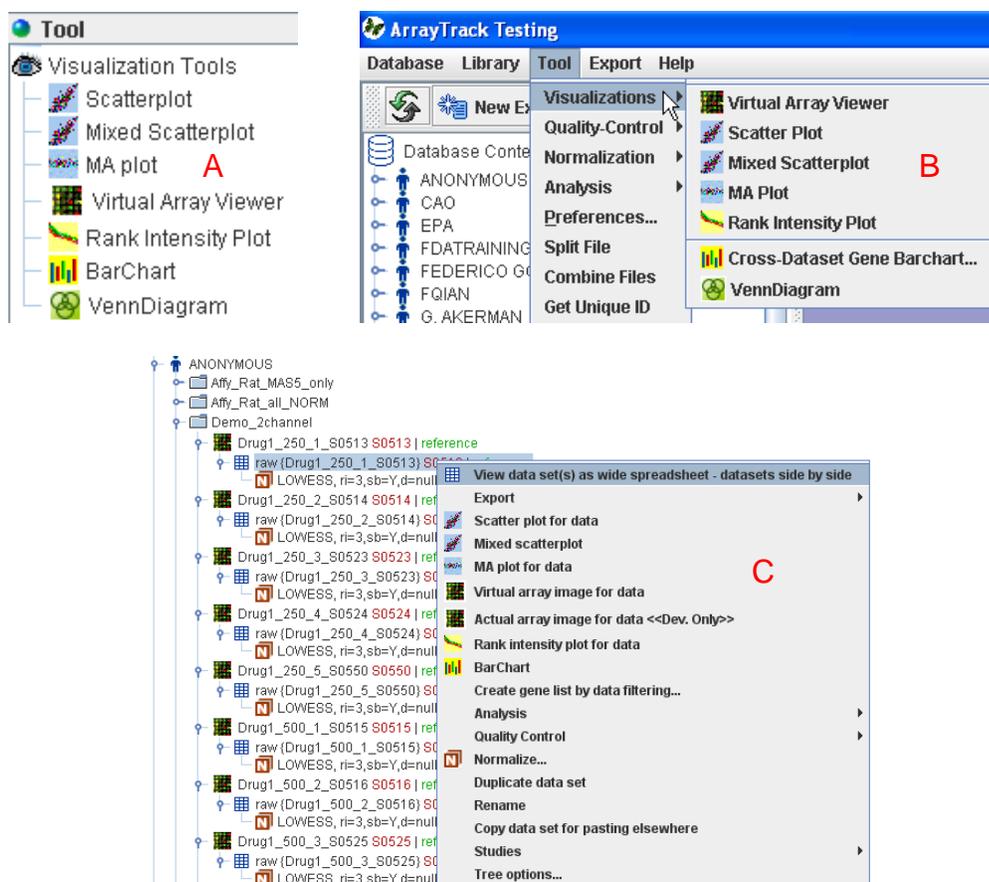


Figure 8-1: Three ways of accessing many Visualization Tools implemented in ArrayTrack: (A) TOOL panel; (B) Tool pull-down menu; and (C) Right-clicking on one of the selected arrays.

## 8.2 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 8-2 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 8-2). 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.

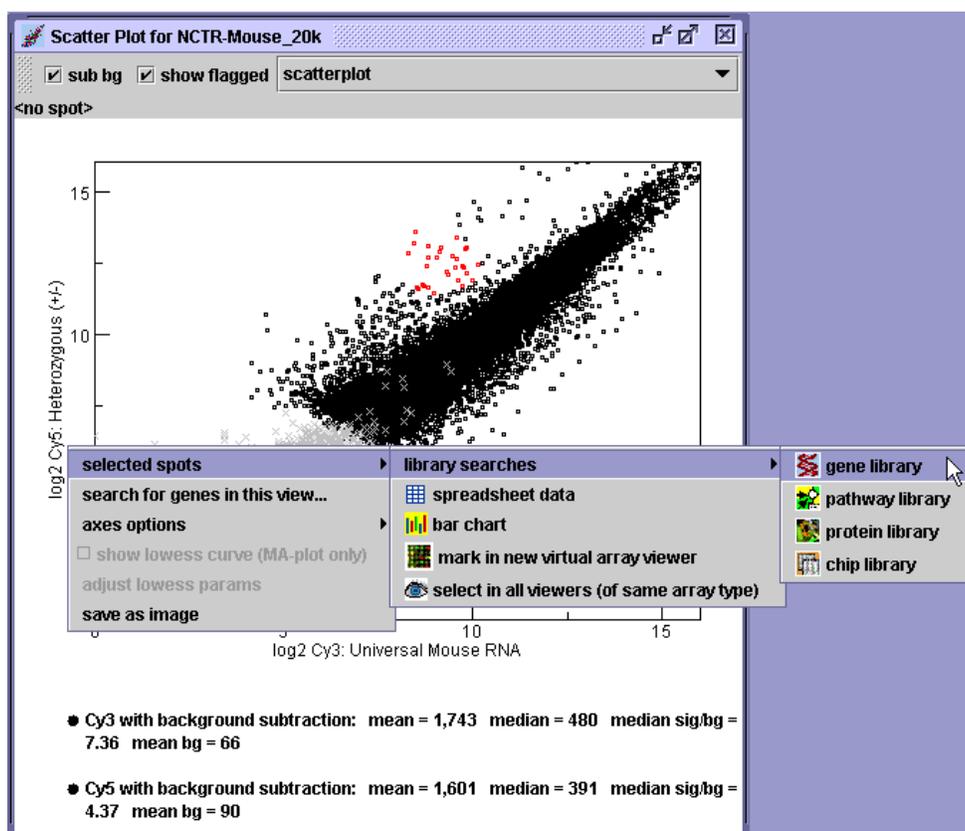


Figure 8-2: Scatter Plot showing the fluorescence intensities for the two hybridization channels.

Right-click on the plot area will pop up a list of actions that can be taken on the scatter plot (Figure 8-2). 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 10: 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 8-3A). 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 8-3C); 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 8-3B).

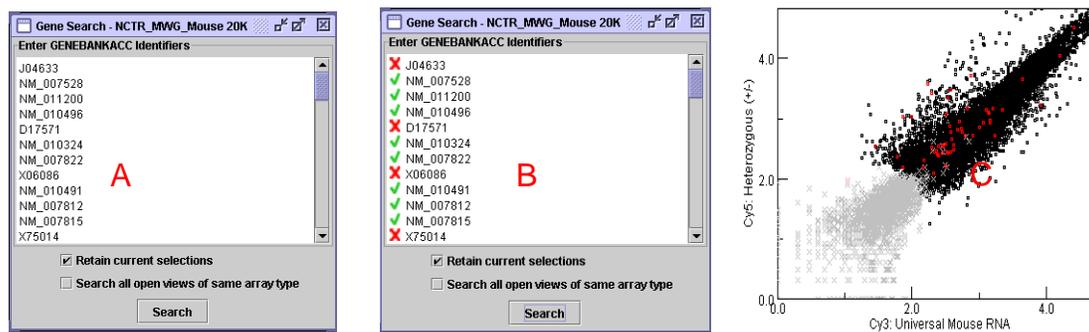


Figure 8-3: Locate a list of interested genes (by GenBank accession numbers) on the scatter plot.

### 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 8-4).
- 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.

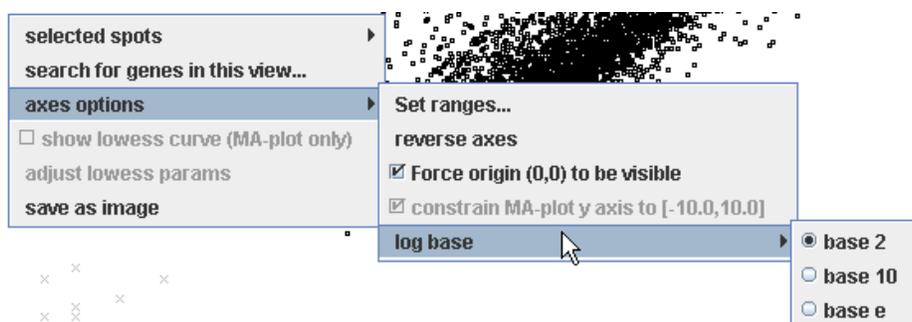


Figure 8-4: Axes Options for scatter plot.

**Show Lowess curve** (MA plot only): The Lowess normalization curve is displayed in yellow (Figure 8-5). For details about Lowess, see Chapter 5 on Normalization Methods.

**Adjust Lowess parameters** (MA plot only): The Lowess Parameters panel (Figure 8-6) 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 8-5B), resulting from 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.

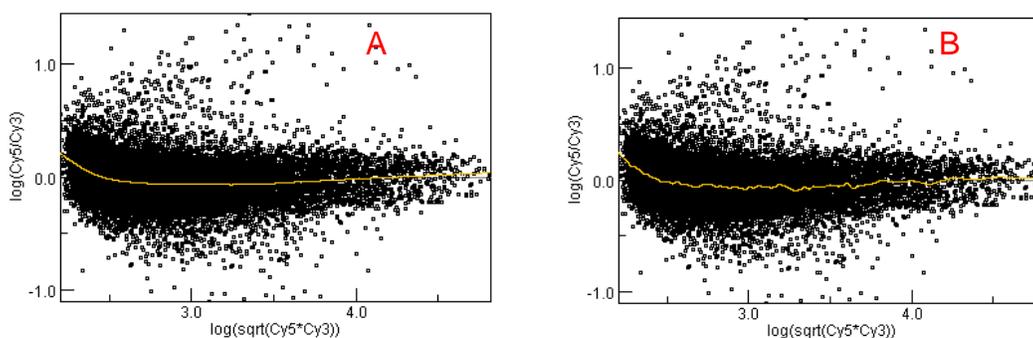


Figure 8-5: Lowess curve is displayed in yellow. A: Default Lowess parameter settings (smoothing factor = 0.2); B: Smoothing factor = 0.01.

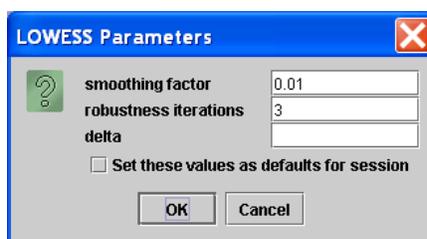


Figure 8-6: Adjusting Lowess parameter settings for the current MA plot.

### 8.3 MA Plot

MA Plot only applies to two-channel data (Figure 8-7). It is a special form of Scatter Plot in which the X axis is the log geometric average (addition) of the intensity values of the two channels and the Y axis is the log fold change (minus of log intensities):

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

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

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 8-7) and applicable only to two-channel microarray data.

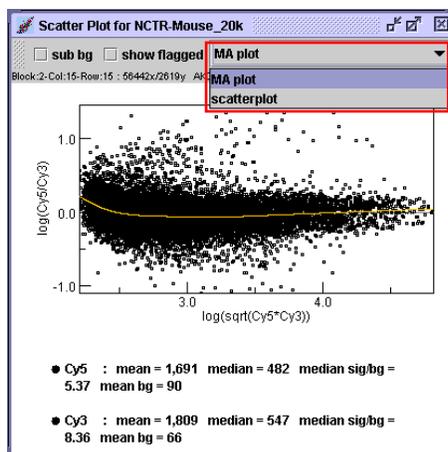


Figure 8-7: MA Plot and Scatter Plot are interchangeable.

### 8.4 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 8-8). 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 8-8 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.

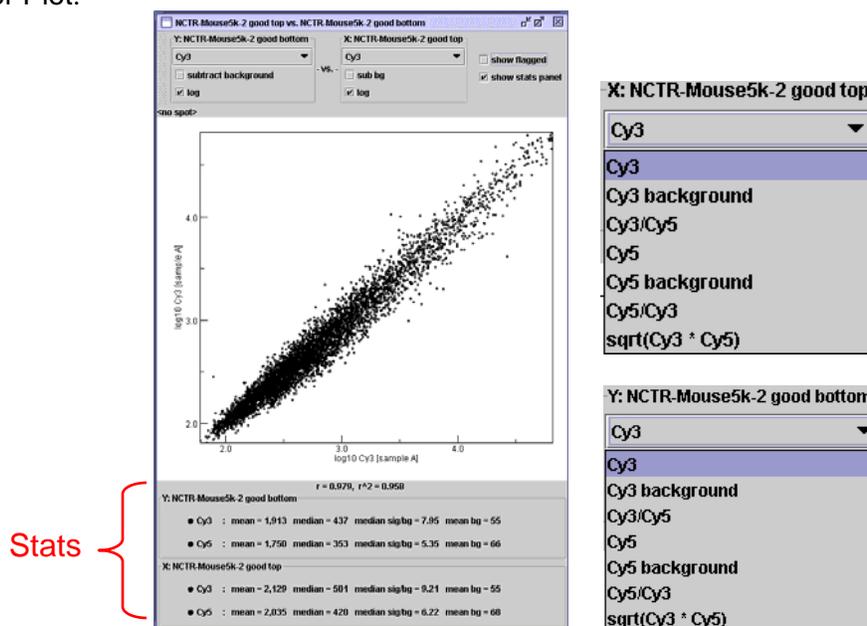


Figure 8-8: Mixed Scatter Plot (Left) and options available for plotting (Right).

### 8.5 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 8-9, the X axis represents the intensity sorted rank (from low to high) and the Y axis is the corresponding log<sub>10</sub> intensity.

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

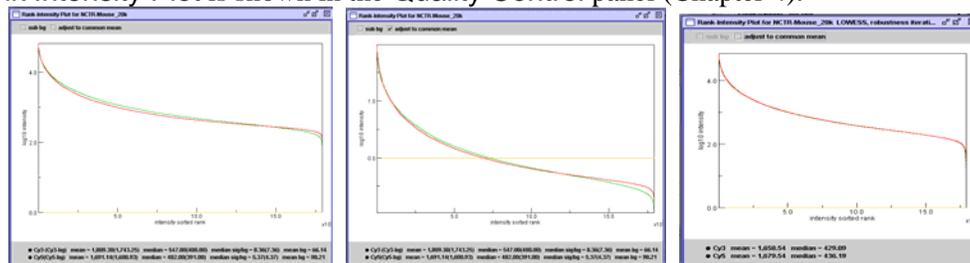


Figure 8-9: Rank Intensity Plot. In the middle plot, mean intensities for Cy3 and Cy5 channels are adjusted to equal mean of zero. In the right plot, Lowess normalized data are plotted.

### 8.6 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 8-10). 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 8-10), a local disk file can be input for plotting (Scatter Plot, Mixed Scatter Plot, MA Plot, and Rank Intensity Plot), as shown in Figure 8-11. The table columns in the local file can be mapped to the required data fields (e.g. the X and Y coordinates) for plotting.

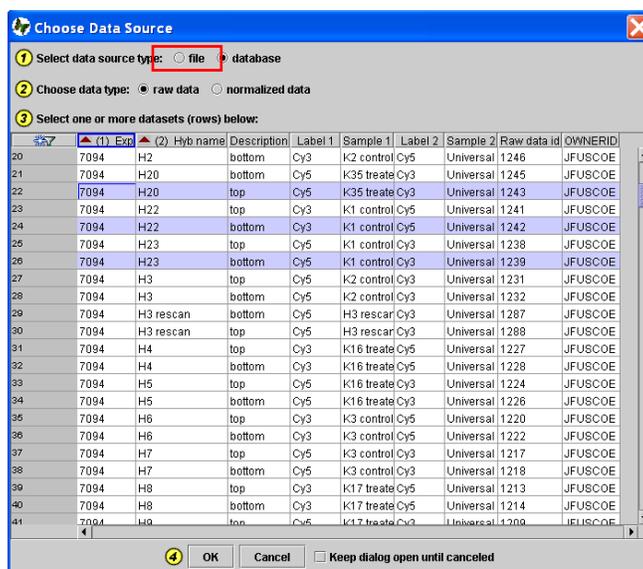


Figure 8-10: List of arrays in MicroarrayDB. Selected arrays will be plotted after click on OK.

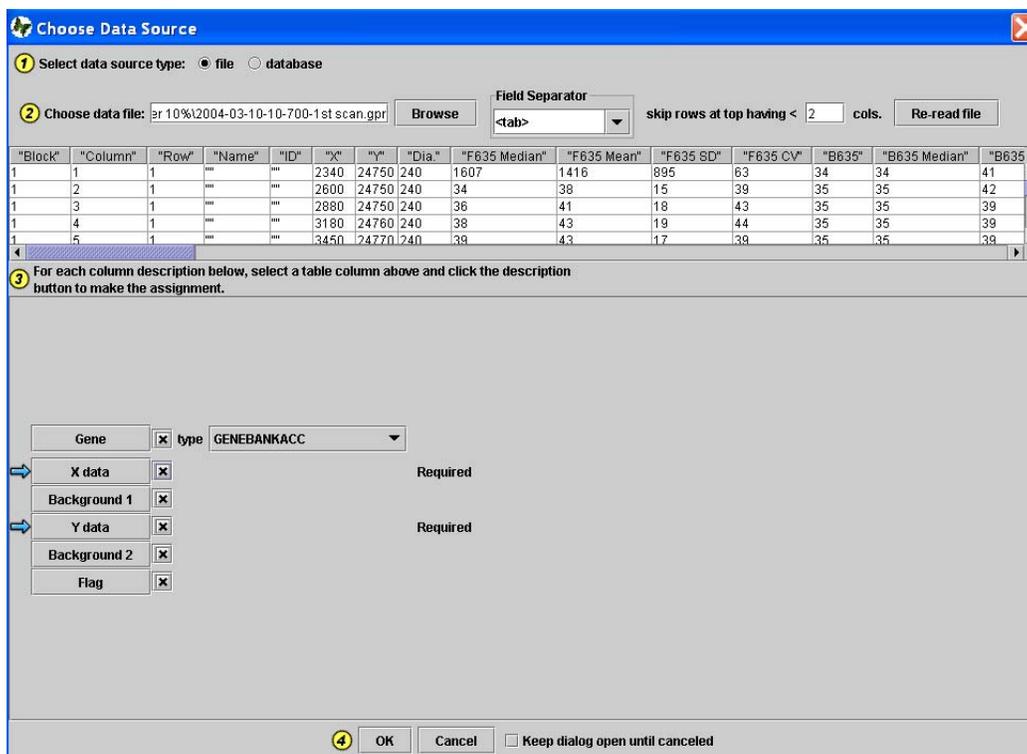


Figure 8-11: Inputting local data file for plotting.

### 8.7 Virtual Array Viewer

**Overview:** The Virtual Array Viewer (Figure 8-12) 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 **Brightness** located at the top-left of Figure 8-12.

**Zoom In and Out of Virtual Array Image:** This is done by clicking on the + or - sign **Zoom** (Figure 8-12).

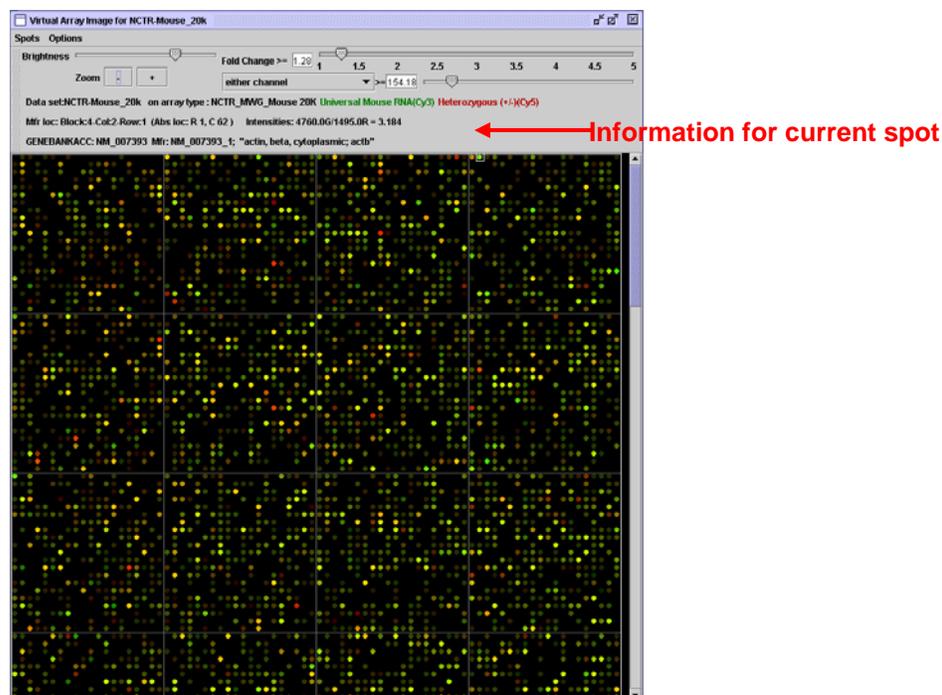


Figure 8-12: Virtual Array Viewer showing a pseudo microarray image reconstructed from fluorescence intensities of a two-channel array. The location of the current gene (*actb*) on the microarray slide is shown on top of the figure.

**Filter Spots:** There are two slider controls for filtering out unwanted spots (Figure 8-13).

**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.

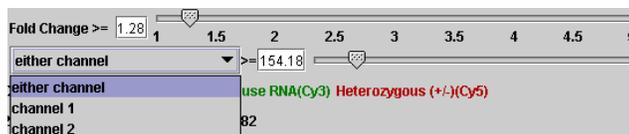


Figure 8-13: Fold Change and Intensity slide controls for filtering out spots for display.

**Pull-Down Functions on Spots:** There are five pull-down functions under Spots (Figure 8-14): **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 8-3 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).

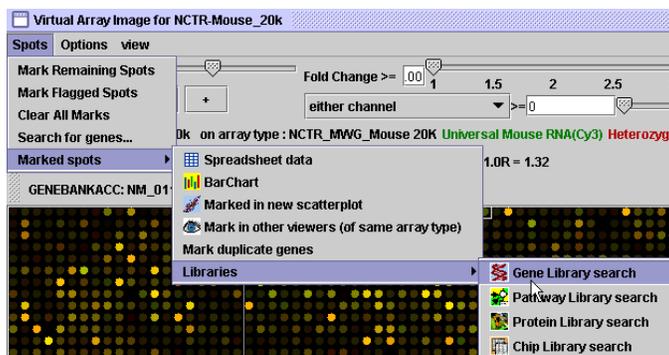


Figure 8-14: Pull-down functions for Spots and Marked Spots.

**Pull-down Options:** Several other options including flag and background handling are available (Figure 8-15A). The color of the virtual array image can also be set in four different ways (Figure 8-15C). 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 8-15B) and the color of marking.

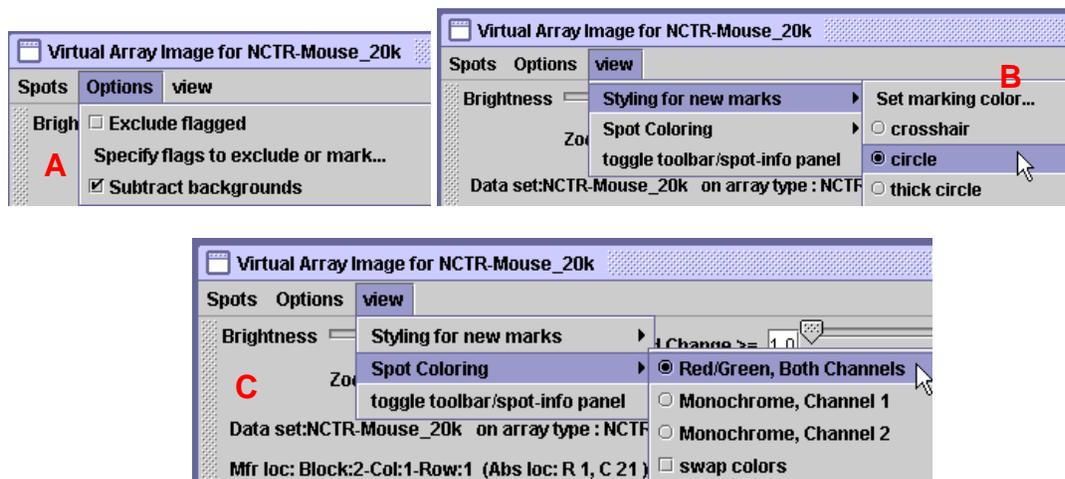


Figure 8-15: Options allow the setting of spot color by channel intensities.

**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 8-12) 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 8-14 and Figure 8-16).

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 8-3 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 8-16) that are the same as those accessible from the **Spots** pull-down menu (Figure 8-14).

**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.

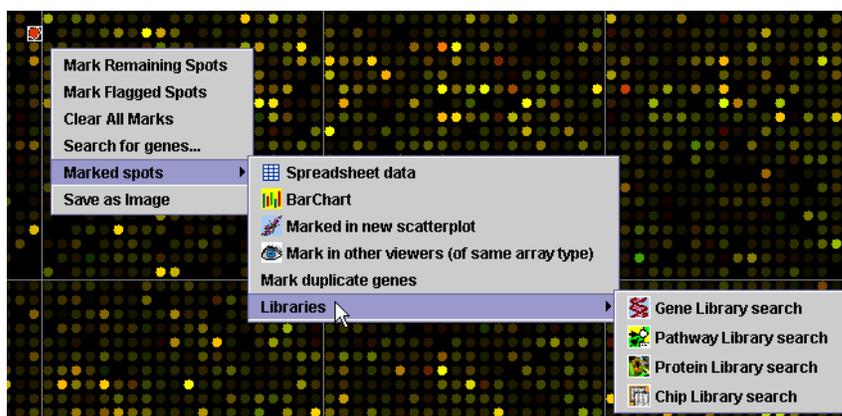


Figure 8-16: Right-click functions and those applicable to Marked Spots.

## 8.8 Actual Array Viewer

**Actual Array Viewer** (Figure 8-17) displays the original image from the scanner and loaded into ArrayTrack from the **Data Import** part (Figure 2-1) of **Input Form**.

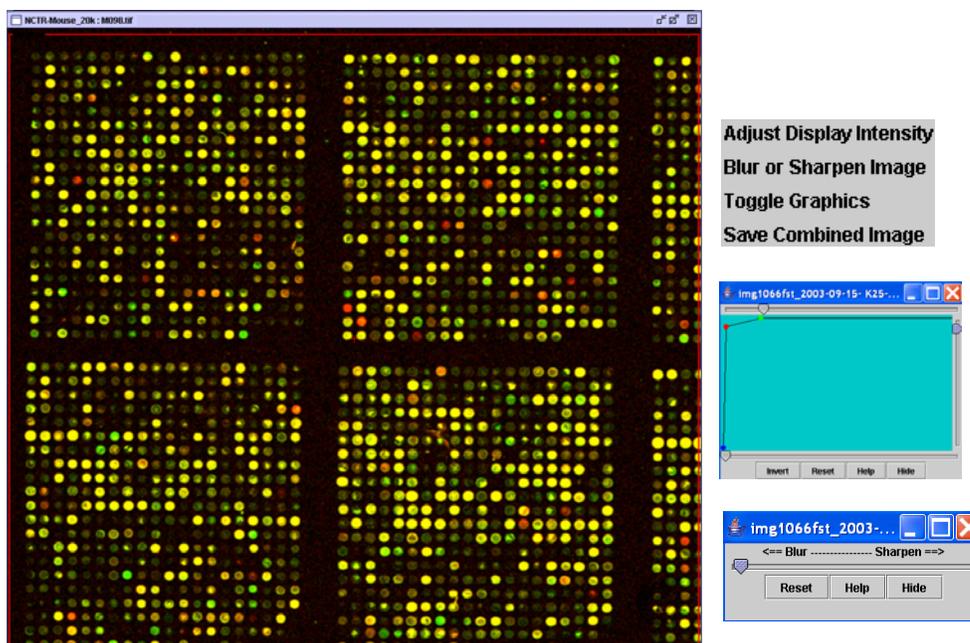


Figure 8-17: Actual Array Viewer displays the original microarray image.

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.

## 8.9 Bar Chart

**Overview:** Bar Chart (Figure 8-18) 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 8-18).

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).

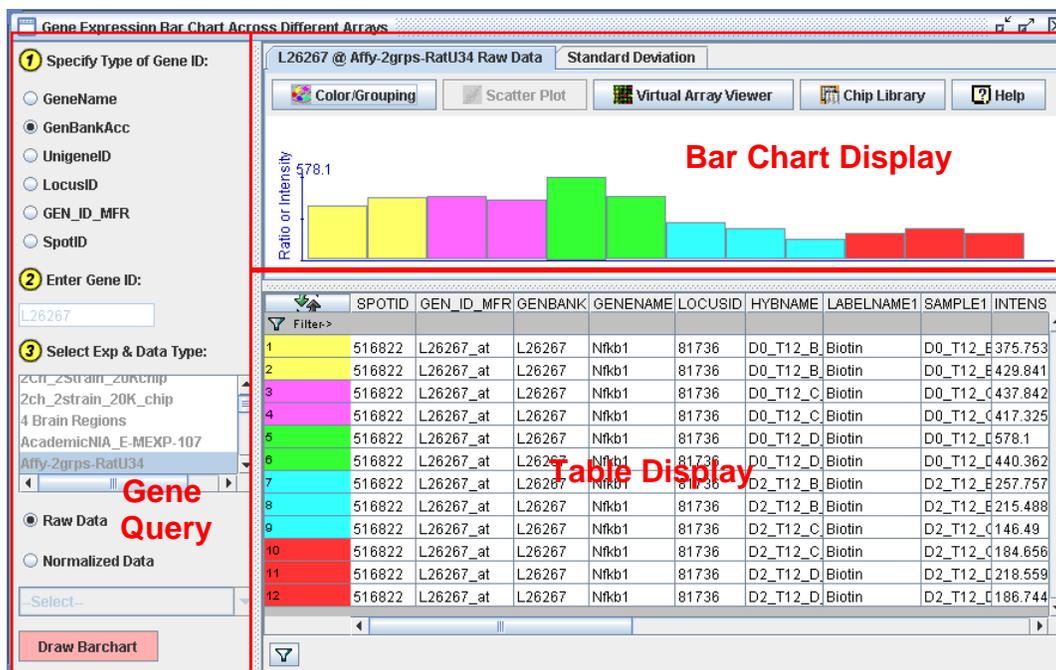


Figure 8-18: Arrangement of Bar Chart panel.

**Toolbars:** Under the first tab, user can see seven tool buttons (Figure 8-19):

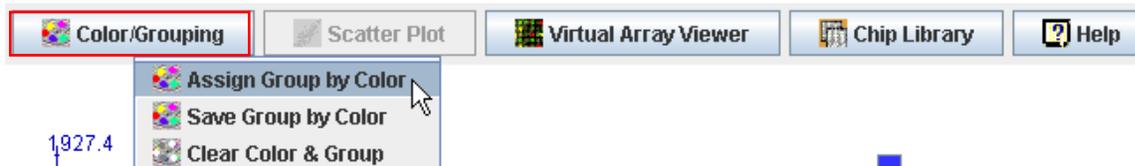


Figure 8-19: Toolbars for Bar Chart

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 8-21) 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 8-22). Also the records will be marked in the corresponding colors (see Figure 8-20). 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.

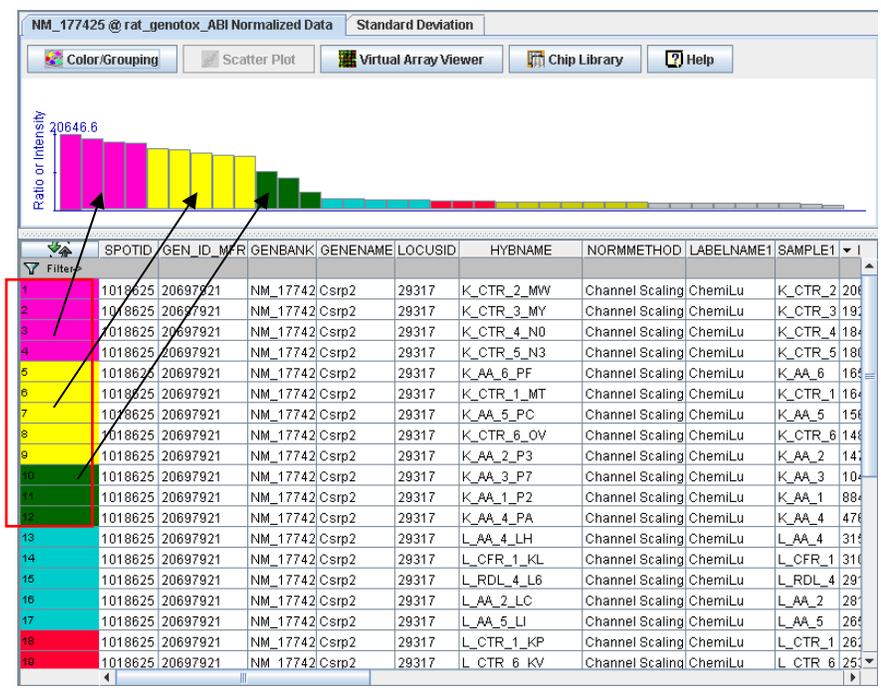


Figure 8-20: After saving the color group each record is marked in the corresponding color

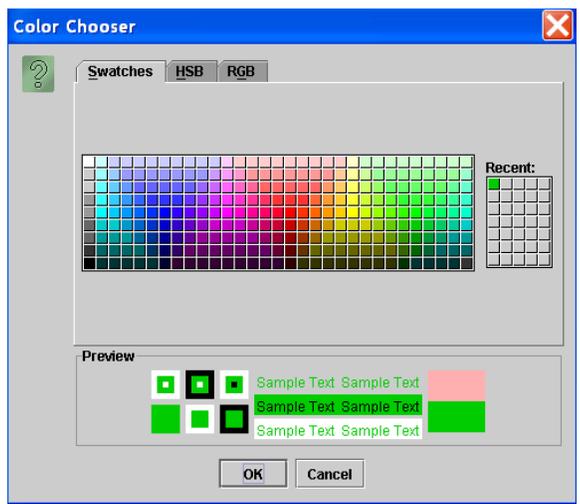


Figure 8-21: Color Chooser for assigning the chosen color as the current color.

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.

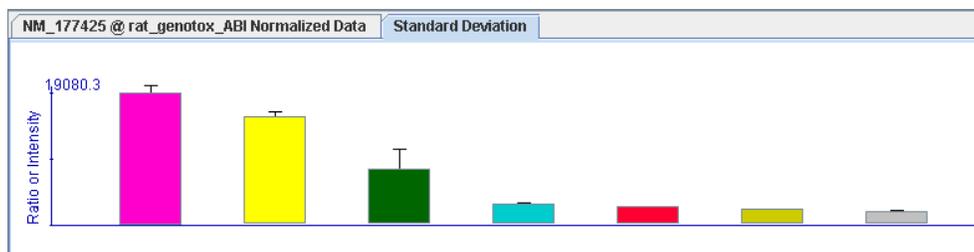


Figure 8-22: Standard Deviation chart

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 8-23 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 8-23, 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.

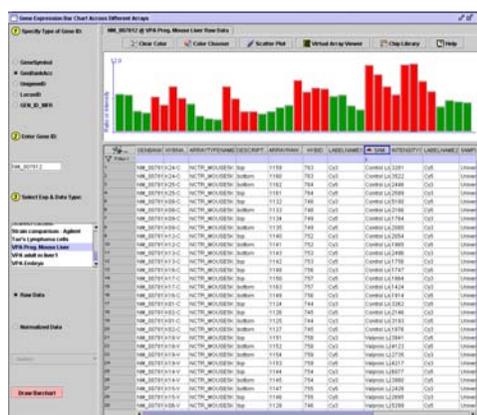


Figure 8-23: Bar Chart displays gene expression data for a gene across multiple arrays. Control samples are colored in green and chemical treated samples are colored in red.

## 8.10 VennDiagram

The Venn diagram is used to display the overlapping part of multiple data set ( $C_1, C_2, C_3, \dots, C_n$ ) which may have some (but not all) elements in common. ArrayTrack can draw the VennDiagram with  $n \leq 3$ . There are 4 ways to activate VennDiagram: a) select 2 or 3 significant gene lists and right-click, choose VennDiagram then select the method for the mapping, b) from the Tool panel, click VennDiagram icon, c) from the Tool pull-down menu, d) from Chip library, under Comparison tab (see Figure 4-37).

### 8.10.1 Draw VennDiagram by common ID

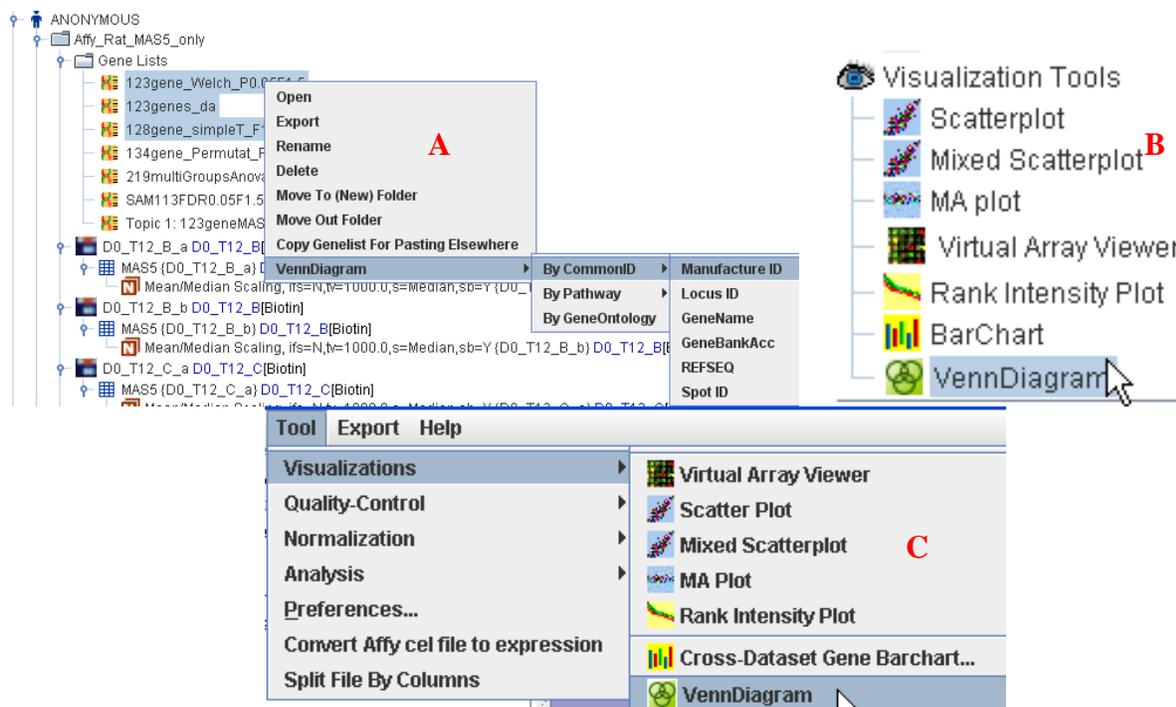


Figure 8-24: Access VennDiagram function

a) If you activate VennDiagram in the first way you will see the Venn diagram (See Figure 8-25, an example of the common genes from the 3 gene lists). The blue number represents the number of the unique IDs in the significant gene lists, and the text above the number in the diagram represents the name of the gene list. 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. Selecting ID only will bring up the table of common ID only. See Figure 8-26.

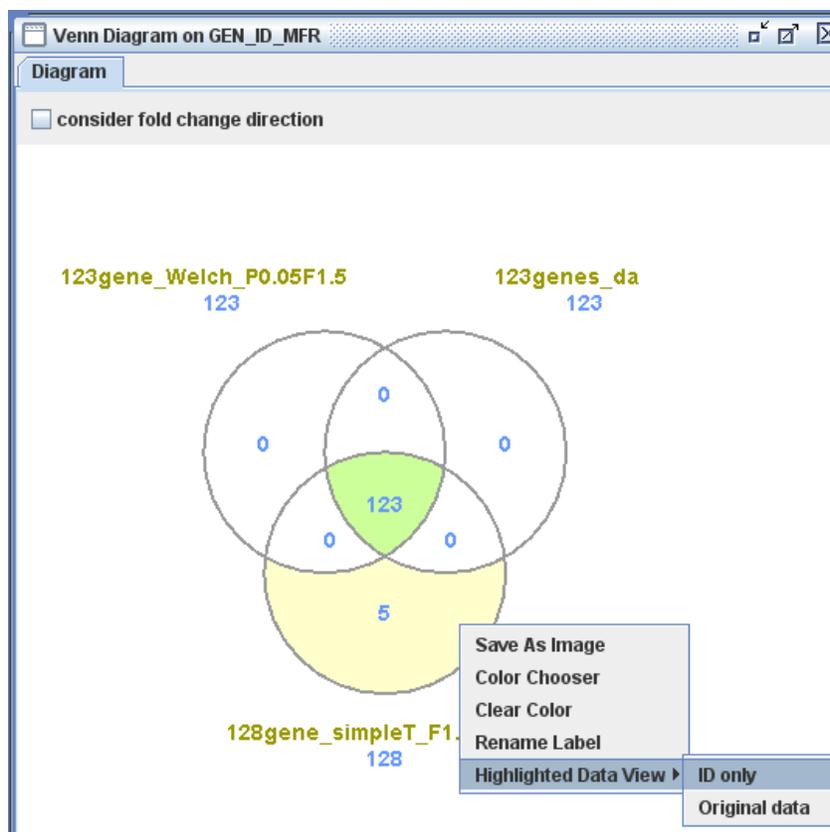


Figure 8-25: The result of the Venn diagram for the common genes

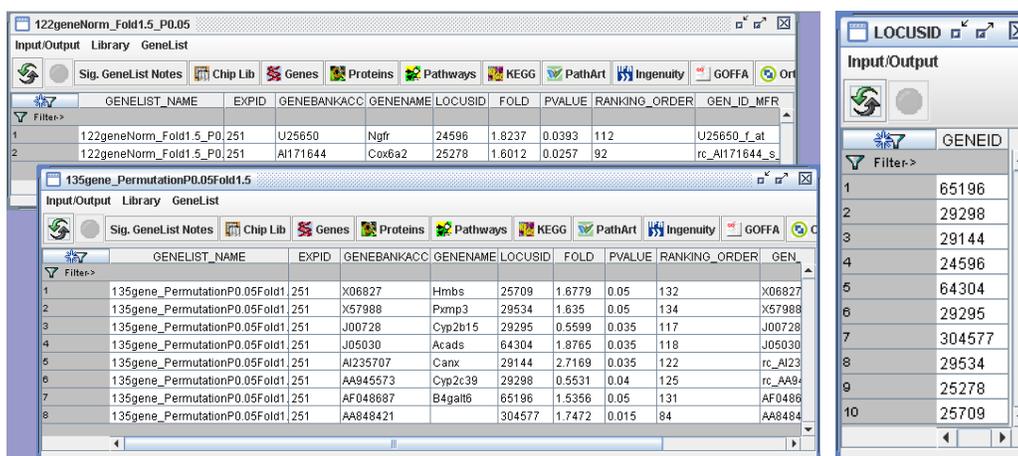


Figure 8-26: The original data view of the common gene list and original data with ID only

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 8-27. Click “Open File” button to load the gene lists and select the column of common ID for the three gene lists (Locus ID is chosen as the common ID), then click “Draw Venn” button. The result of the VennDiagram will be displayed, see Figure 8-28.

In Figure 8-27, the user can also copy IDs from other place and paste here by clicking “Paste ID” button. Clicking “Clear Input” button at the bottom will clear the current contents in all the three tables. The

user has options for matching by common ID or by pathway (KEGG, PATHART) or by geneOntology. These options are also available by activating VennDiagram the first way, see Figure 8-24A.

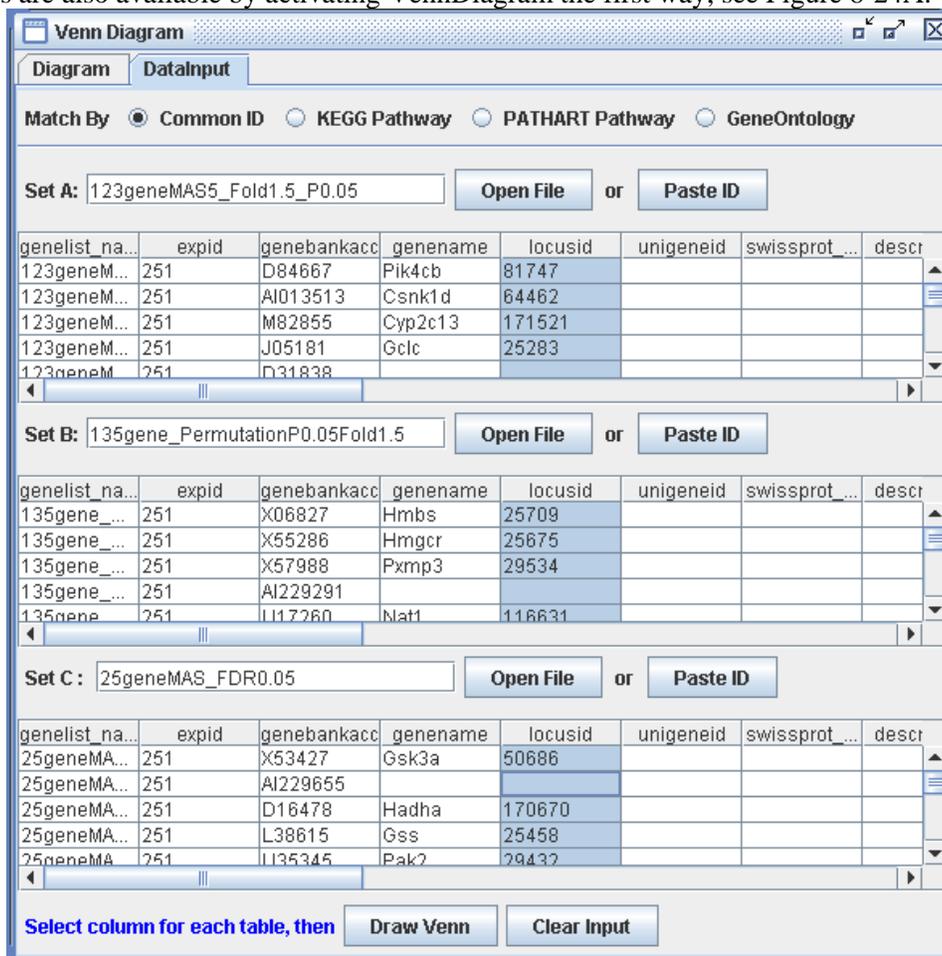


Figure 8-27: open gene lists from local drive

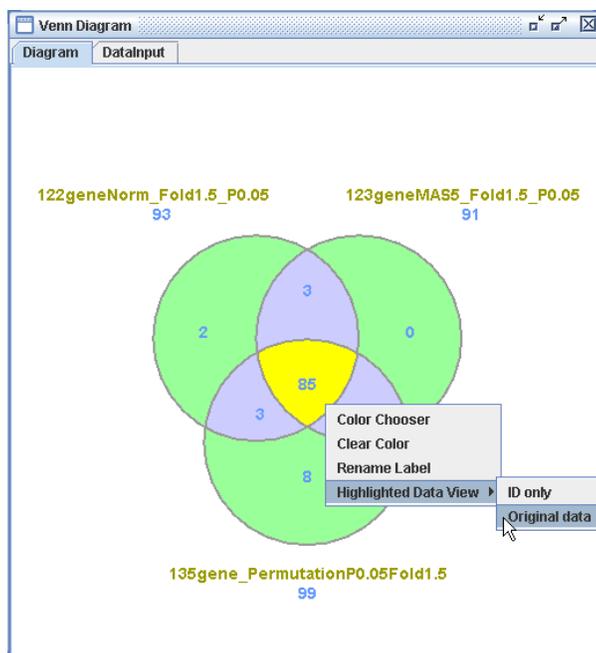


Figure 8-28: VennDiagram for three gene lists

In Figure 8-28, right-clicking any highlighted area in the VennDiagram and choosing HighlightedData View->Original data will bring up the tables (Figure 8-29) 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”.

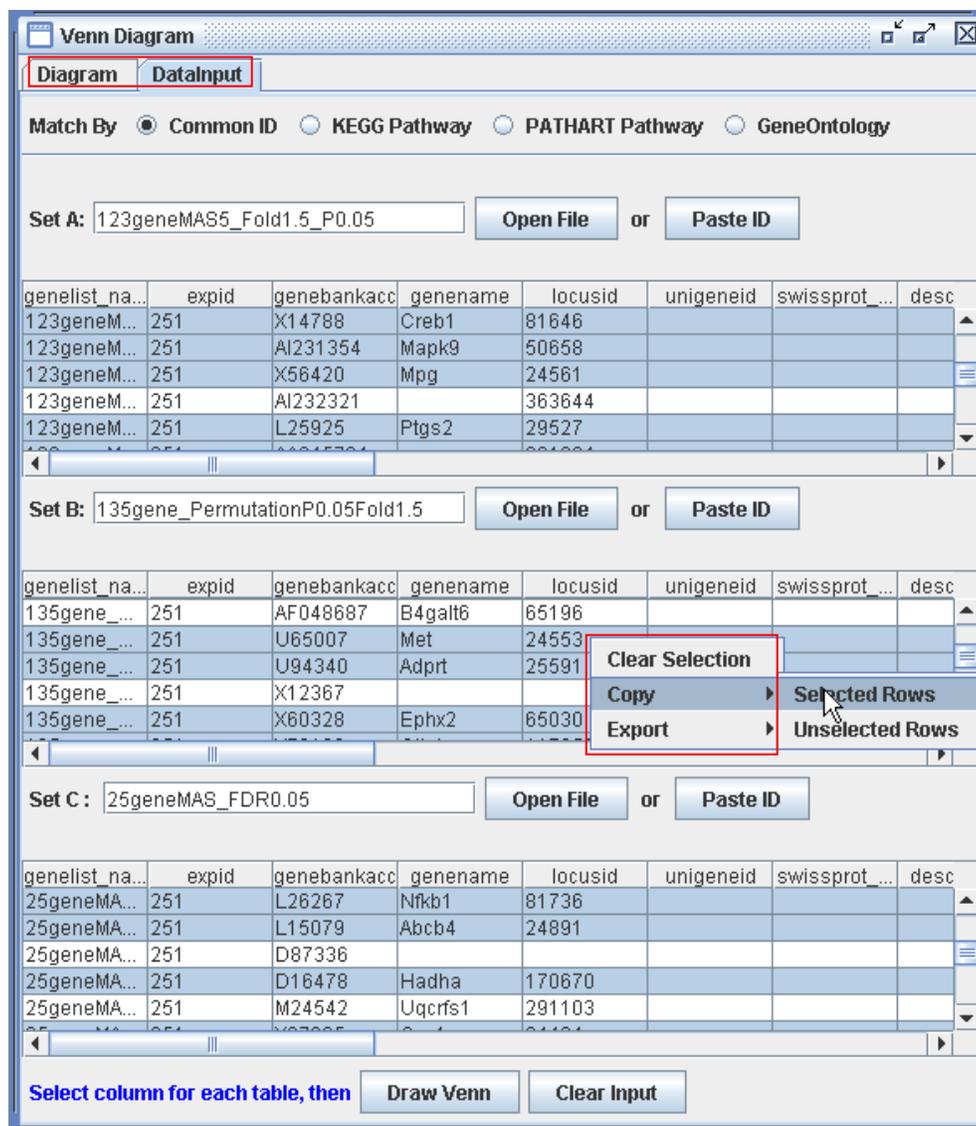


Figure 8-29: Original data view with highlighted rows corresponding to the highlighted area in VennDiagram

c) Activate VennDiagram from the Tool menu.

If user activates VennDiagram from the Tool menu (Figure 8-30), the VennDiagram window will pop up.

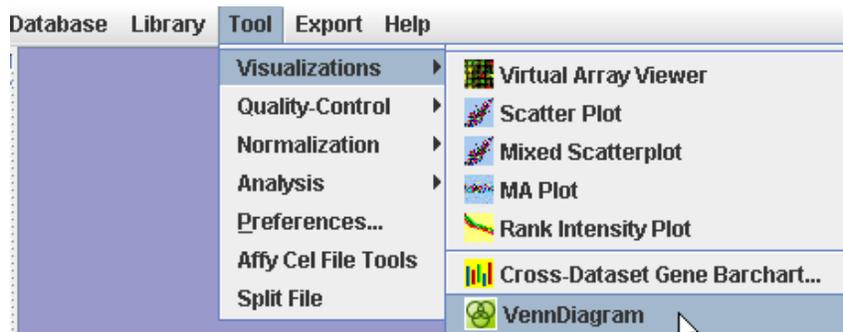


Figure 8-30: Activate Venn Diagram from Tool pull-down menu

In Figure 8-30), the user need to open the gene list or paste ID in by clicking “Open File” button or “Paste ID” button (**Error! Reference source not found.**). The list can be gene list, protein list or compound list. For example, in **Error! Reference source not found.** one gene list with one protein list and one compound list are put in. The user can select matching by common ID, KEGG pathway, PATHART pathway or GeneOntology. First choose ID column from each table (data set A, B, C) for matching, then click “Draw Venn” button.

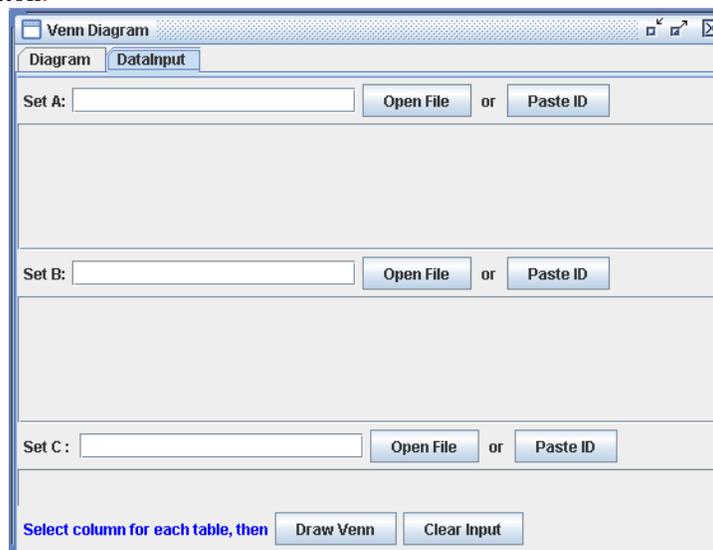


Figure 8-31: The user needs to input the gene list or paste ID

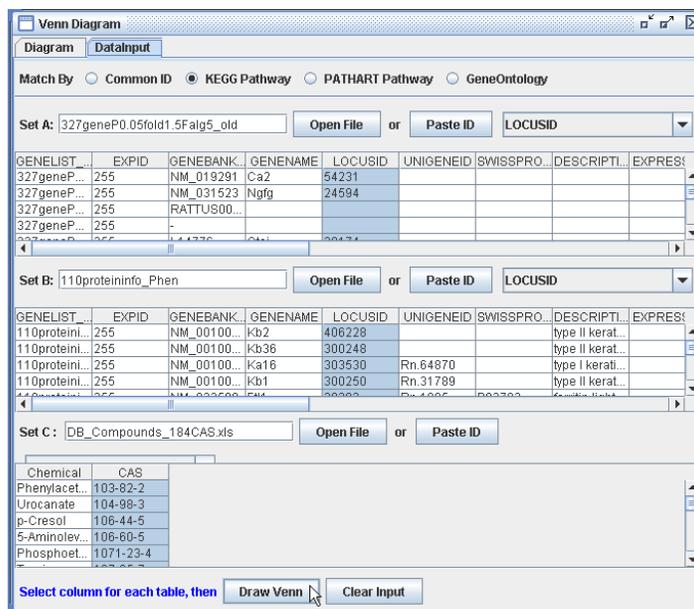


Figure 8-32: The input list can be gene list, protein list or compound list

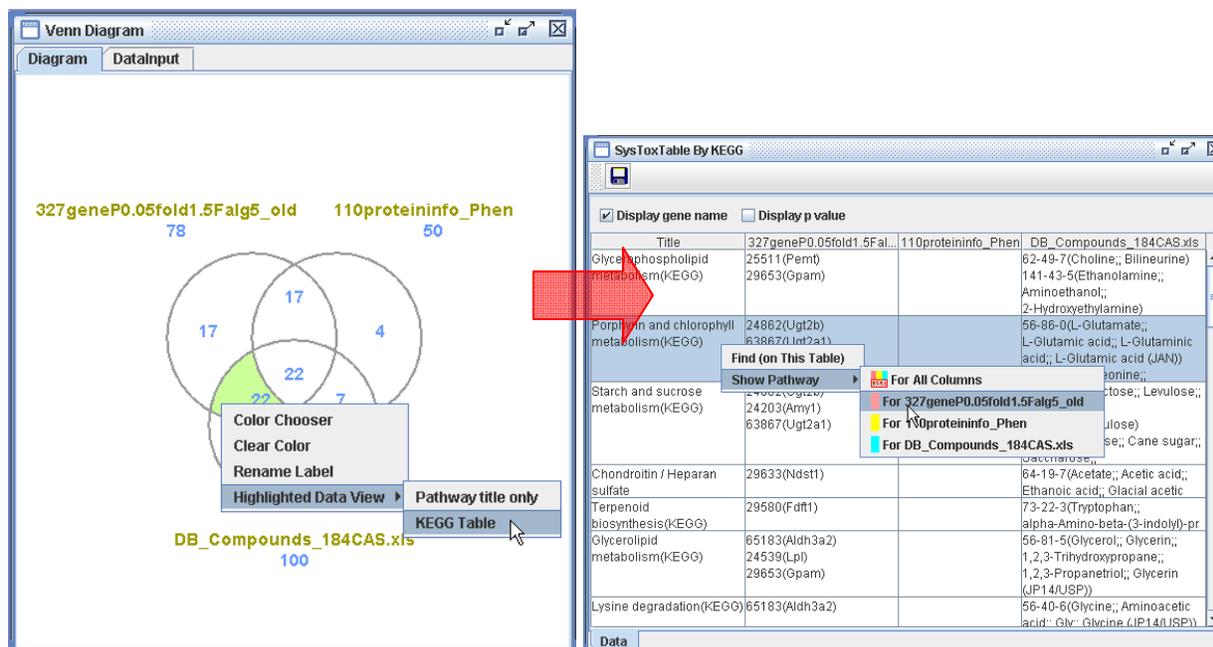


Figure 8-33: Draw VennDiagram using different list and display the highlighted data in KEGG table

In Figure 8-33, the highlighted data in VennDiagram is displayed in the table at right side. The user can highlight any record in KEGG table.

Title	mi/ni	p-value
Sulfur metab...	0.136 1/25	0.136 1/25
Taurine and h...	0.121 1/22	0.121 1/22
Calcium sign...	0.637 3/554	0.637 3/554
Tight junction...	0.872 1/343	0.872 1/343
Inositol phos...	0.63 1/168	0.63 1/168
Huntington's ...	0.073 2/76	0.073 2/76
Pyrimidine m...	0.652 1/178	0.652 1/178
Toll-like rece...	0.065 4/261	0.065 4/261
Arginine and ...	0.493 1/115	0.493 1/115
Apoptosis(KE...	0.004 6/264	0.004 6/264
Tryptophan m...	0.329 2/201	0.329 2/201
Valine, leucin...	0.439 1/98	0.439 1/98
Adherens jun...	0.357 2/214	0.357 2/214
Glutamate m...	0.314 1/64	0.314 1/64
Jak-STAT sig...	0.44 3/415	0.44 3/415
Lysine degra...	0.487 1/113	0.487 1/113
ECM-receptor...	0.759 1/239	0.759 1/239
Circadian rhyt...	0.181 1/34	0.181 1/34
Regulation of ...	0.973 1/595	0.973 1/595
Phosphatidyl...	0.764 1/243	0.764 1/243
Cysteine met	0.186 1/35	0.186 1/35

Figure 8-34: The result

d) Activate VennDiagram from Chip Library. This is very useful for comparing two array types and finding common genes between the two array types. See detail in 4.9 Chip Library (Figure 4-37).

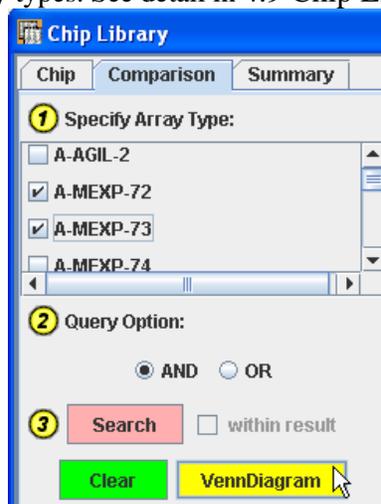


Figure 8-35: Activate VennDiagram from Chip Library comparing 2 or 3 array types

### 8.10.2 Draw VennDiagram by KEGG/PATHART pathway

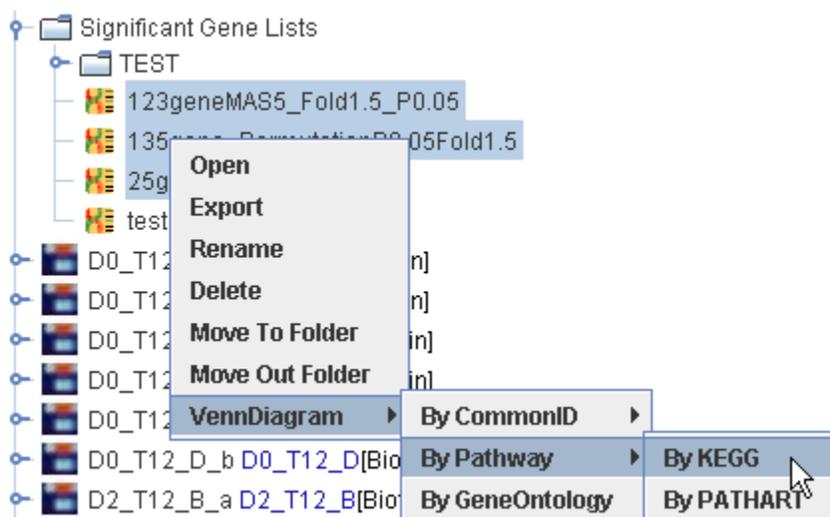


Figure 8-36: Draw VennDiagram matching by Pathway (KEGG)

If the user chooses drawing VennDiagram matching by Pathway, from the result of VennDiagram he can let the highlighted data displaying in KEGG table (see Figure 8-37).

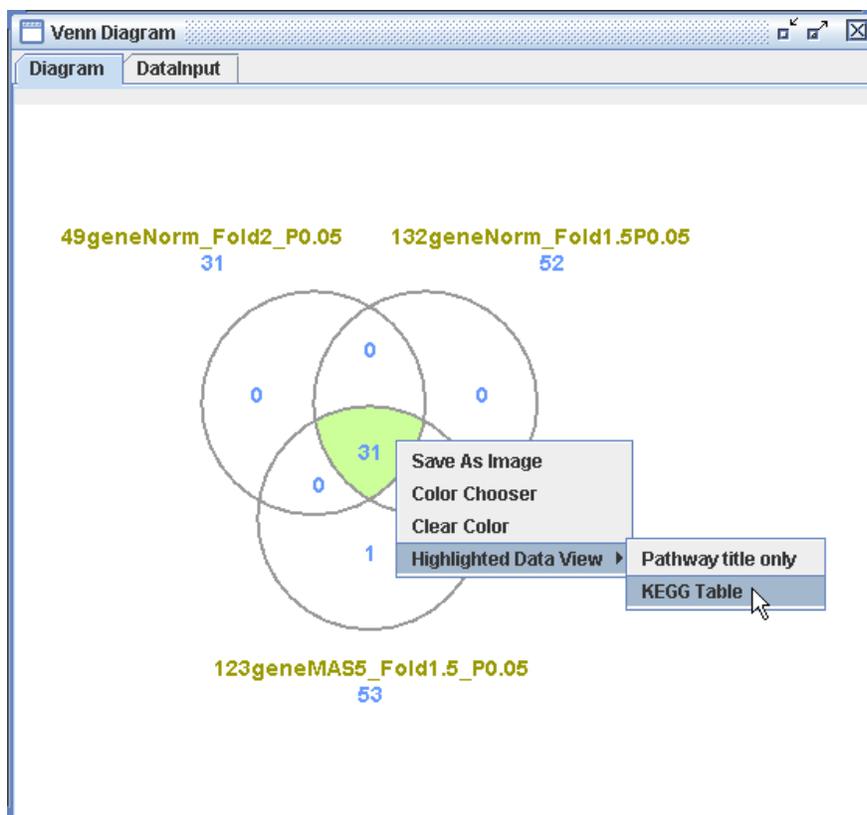


Figure 8-37: the highlighted data can be displayed in KEGG table

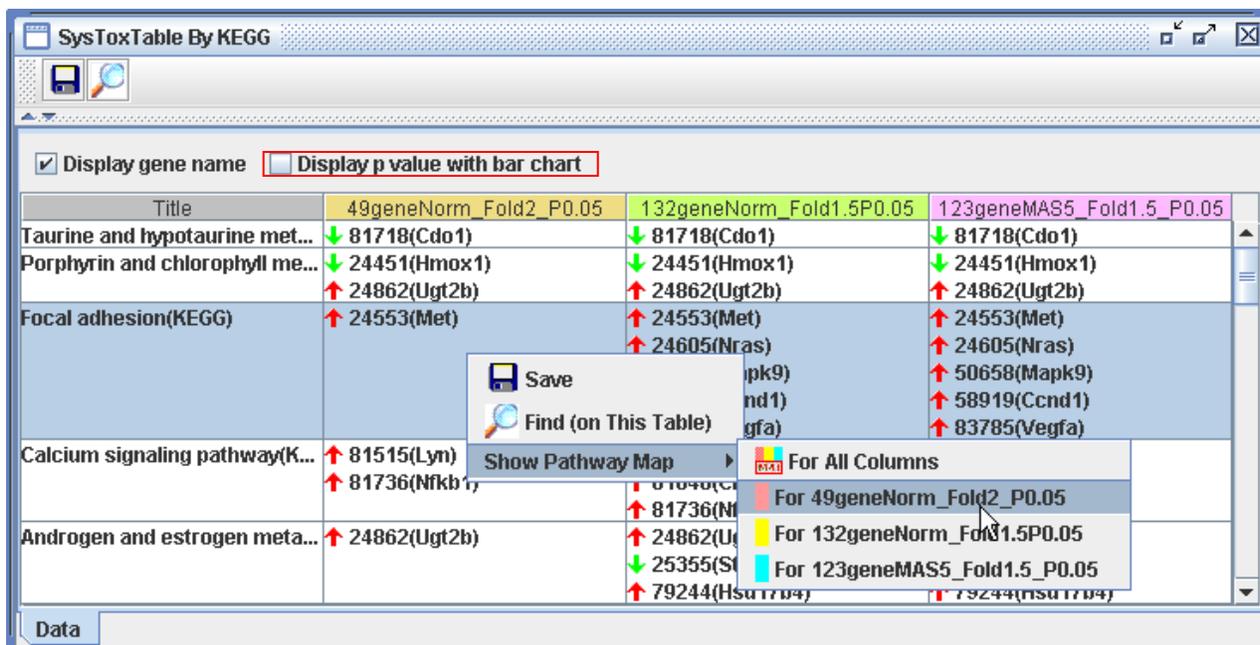


Figure 8-38: the highlighted data displaying in KEGG tables

Figure 8-38 lists the common genes from the three gene lists marked in different colors. The green arrow represents down regulated and red arrow represents up regulated. The user can highlight any record, then right-click -> choose “Show Pathway Map” for any one gene of the gene list. The color block before

the gene list represents that in KEGG any genes from the list that involved in the pathway will be marked in this color. The user can check the box labeled "Display p value with bar chart", and a bar chart for all the common genes will show, see Figure 8-39.

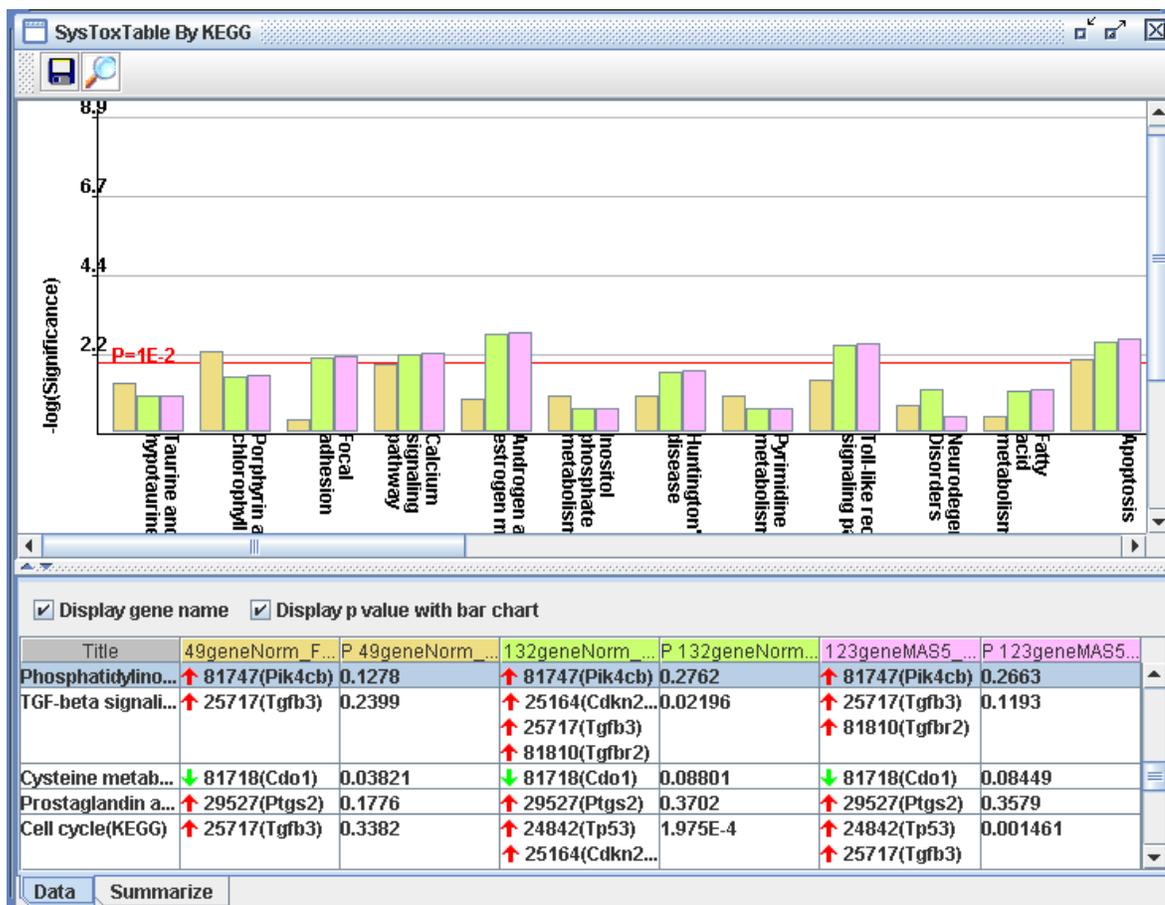


Figure 8-39: Display p value with bar chart for the common genes in Venn Diagram

In Figure 8-39, clicking "Summarize" tab at the bottom will bring the following summarize table:

Title	49geneNorm_Fold2_PO...	mi/ni 49geneNorm_Fold2...	132geneNo...	mi/ni 132gen...	123geneM...	mi/ni 123...
Taurine and hyp...		0.038 1/2		0.088 1/2		0.084 1/2
Porphyrin and c...		0.005 2/6		0.027 2/6		0.024 2/6
Focal adhesion(...		0.436 1/29		0.008 5/29		0.007 5/29
Calcium signali...		0.012 2/9		0.006 3/9		0.005 3/9
Androgen and e...		0.111 1/6		0.002 3/6		0.001 3/6
Inositol phosph...		0.093 1/5		0.206 1/5		0.198 1/5
Huntington's dis...		0.093 1/5		0.018 2/5		0.017 2/5
Pyrimidine meta...		0.093 1/5		0.206 1/5		0.198 1/5
Toll-like recepto...		0.032 2/15		0.003 4/15		0.003 4/15
Neurodegenera...		0.161 1/9		0.058 2/9		0.329 1/9

Figure 8-40: summarize table for the common genes

In Figure 8-40, mi means the number of genes found for pathway i in the input list. And ni means the number of genes found for pathway i in the associated array type. If no array type is specified for input gene list, then ni means number of genes found for pathway i in the whole gene database table.

### 8.10.3 Draw VennDiagram by GeneOntology

The user can also choose drawing VennDiagram matching by GeneOntology, same as matching by Pathway.

#### 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 8-41 is a screenshot of ArrayTrack in which many connected plots and tables are displayed.

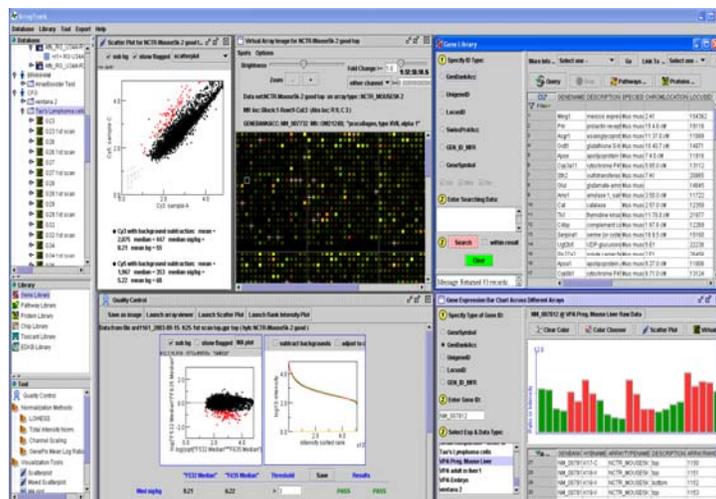


Figure 8-41: A screenshot showing multiple visualization functionalities within ArrayTrack.

## Chapter 9 Working with other Tools

From Tool pull-down menu the user can access some other tools. These miscellaneous tools can not be accessed from tool panel.

### 9.1 Join table (file)

This function is for combining two separate files into one file. Also the file has to be ,txt format. Choosing "File-Manipulation" ->Join Table (File) from Tools pull-down menu, the following window pops up

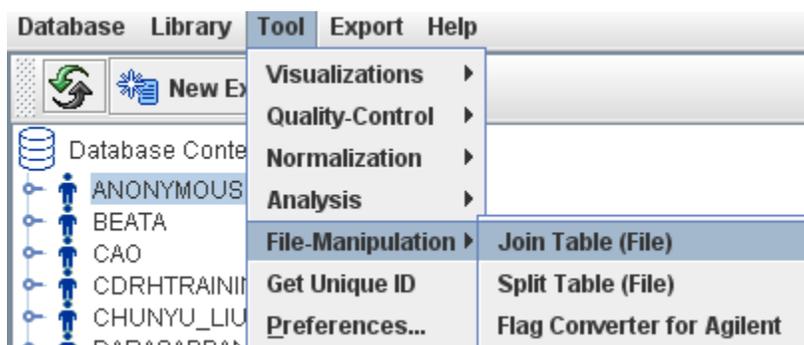


Figure 9-1: File manipulation tool

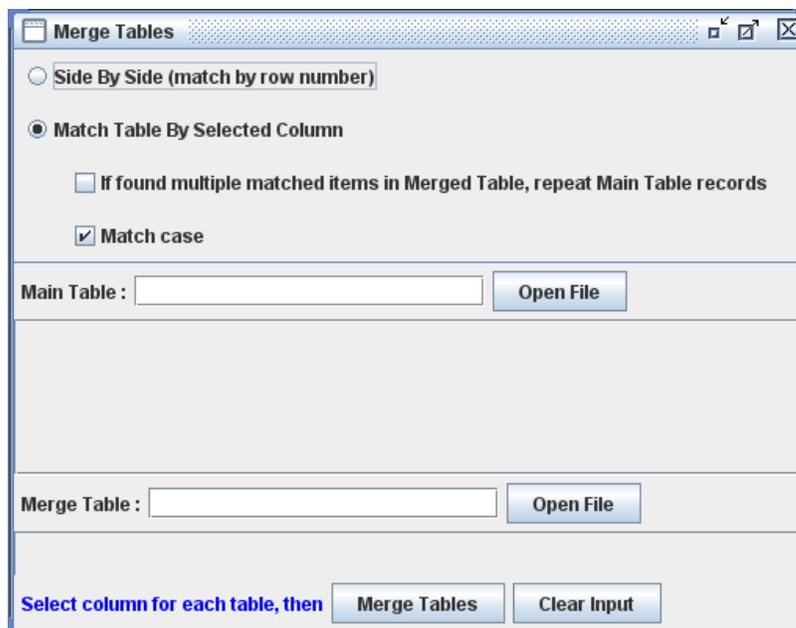


Figure 9-2: Combines files

In Figure 9-2, the user needs to open the two files to be combined and then specify the common ID for joining the two tables. The first table is "Main Table" and second one is "Merged Table". The result of the combined file is dependent on the record of the main table. For example, if the main table has 100 records and the second table has 200 records with 20 records matching the first table. Then the result of the combined table will have 100 records. So the final results will not be more or less than 100 which is the number of the first table.

## 9.2 Split table (files)

When the user needs to do the batch import but has one file that includes the data of all the hybridizations, s/he needs to split the file into separate data file with only one hybridization in each single file. This function is for this purpose.

In Figure 9-1, select “Split File”. The following window will pop up. The user can choose split “By Columns” or “By Rows”.

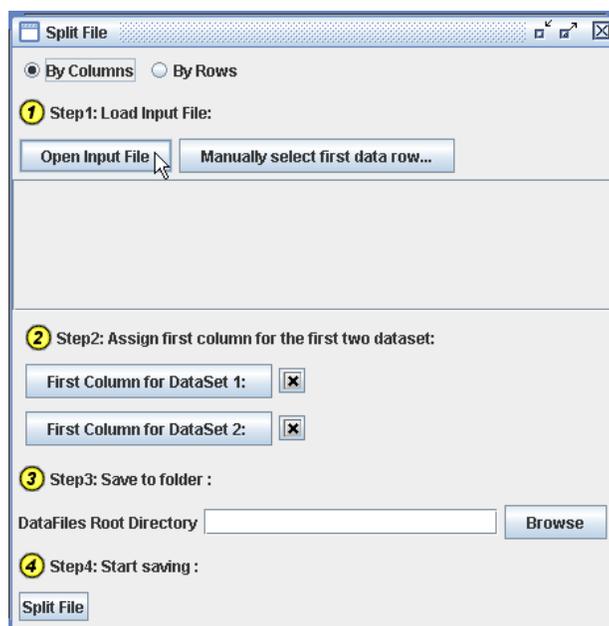


Figure 9-3: Split file

Step 1: Load the file that needs to be split by clicking the button “Open Input File”. The file has to be .txt format. If the user has .csv or .xls file, he needs to save as .txt file first before splitting.

Step 2: Assign first column for the first two dataset, see Figure 9-4. The assigned columns will be marked in yellow color. The user only needs to define the first columns for the first two dataset. Be aware that the name of the first columns for each data set will be the name of the split file, for example in Figure 9-4 the split file name will be A1.txt, B1.txt, ... etc.

Step 3: Specify the location for storing the split files by clicking “Browse” button.

Step 4: Click “Split File” button.

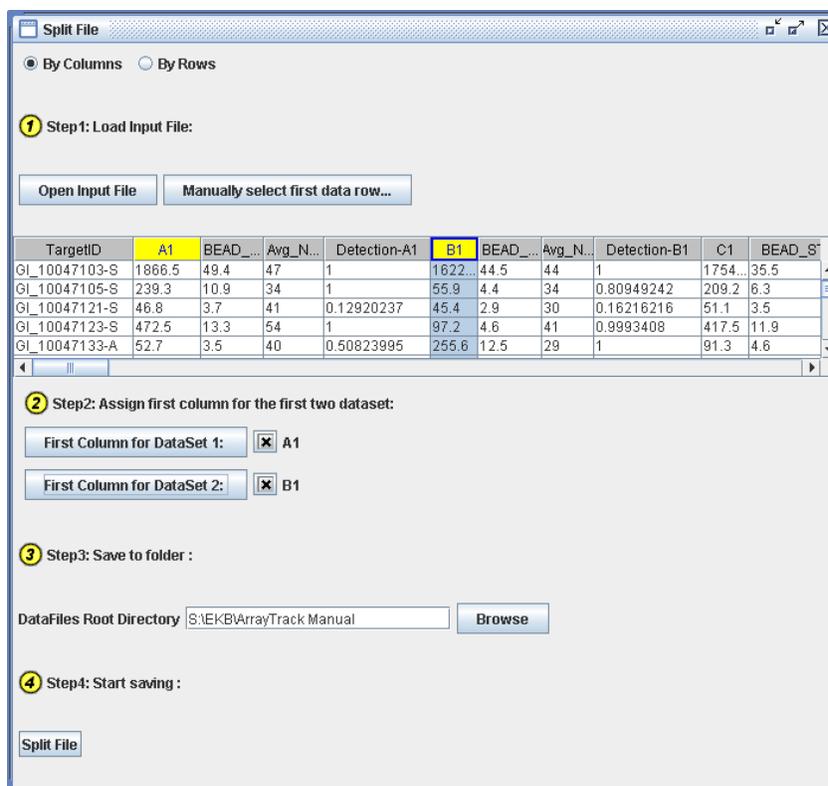


Figure 9-4: Assign columns for the datasets to be split

### 9.3 Get Unique ID

This function is for filtering out the unique IDs from a file that might contain duplicate IDs. Clicking “Get Unique ID” from the Tool pull-down menu a window will pop up (see Figure 9-5A). The user needs to copy the IDs to the empty text box, then click “OK” button. And the result will be displayed (see Figure 9-5B). The column “N” displays the number of times for each ID showing in the original table. Number “1” means that this ID is unique, while “2” means that this ID appears 2 times.

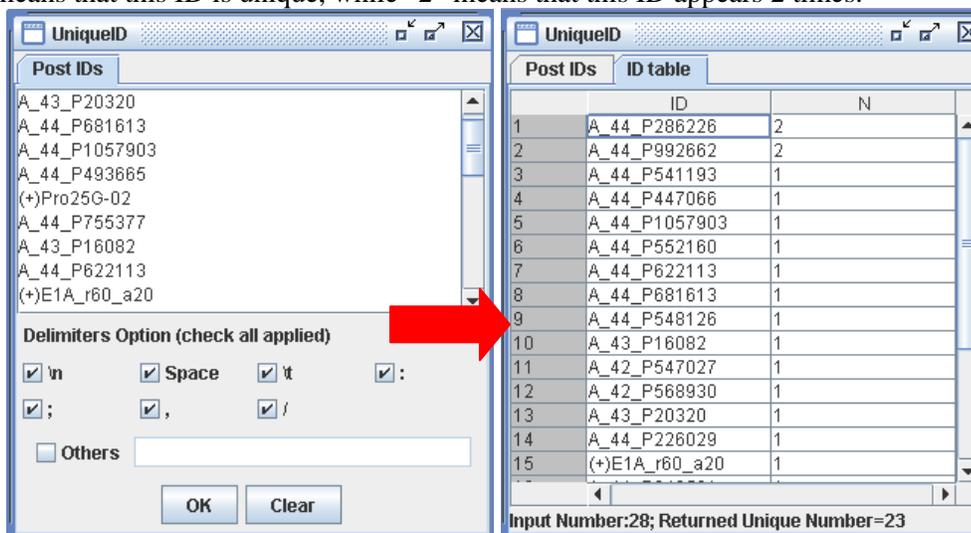


Figure 9-5: Get Unique ID

### 9.4 Convert CEL file to probe set files

The cel files in ArrayTrack database can be converted to probe set files easily by right-clicking the dataset, then choosing “Convert affy cel files to probe sets”. See Figure 9-6.

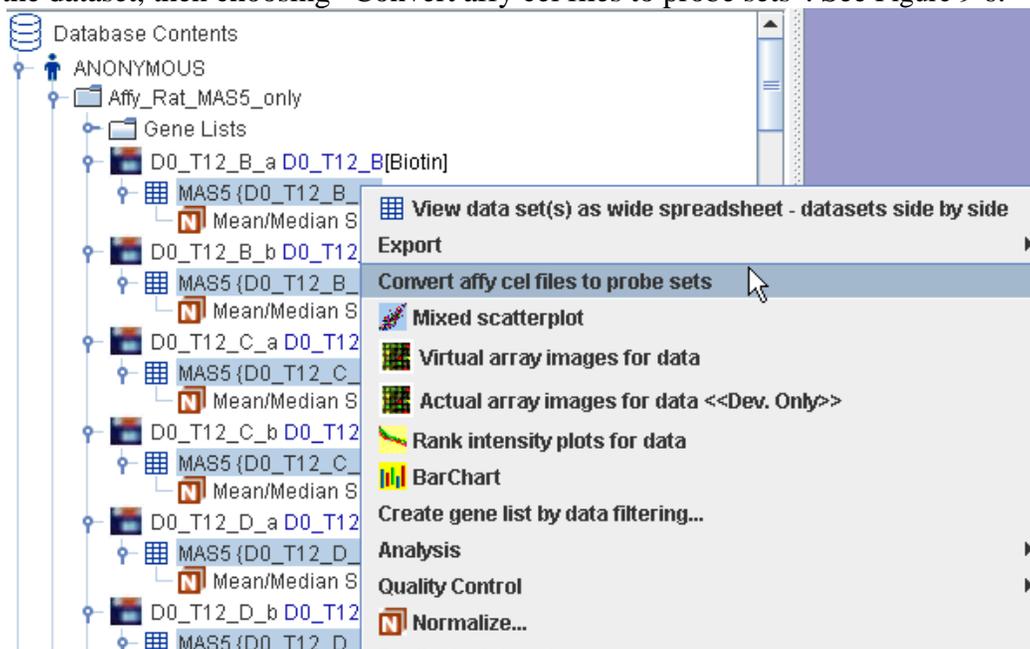


Figure 9-6: convert cel file to probe set files

ArrayTrack provides options of converting to RMA, DChip, Plier and qPlier16. See Figure 9-7. Click OK button, then you will be able to see the converted probe set data in database panel.

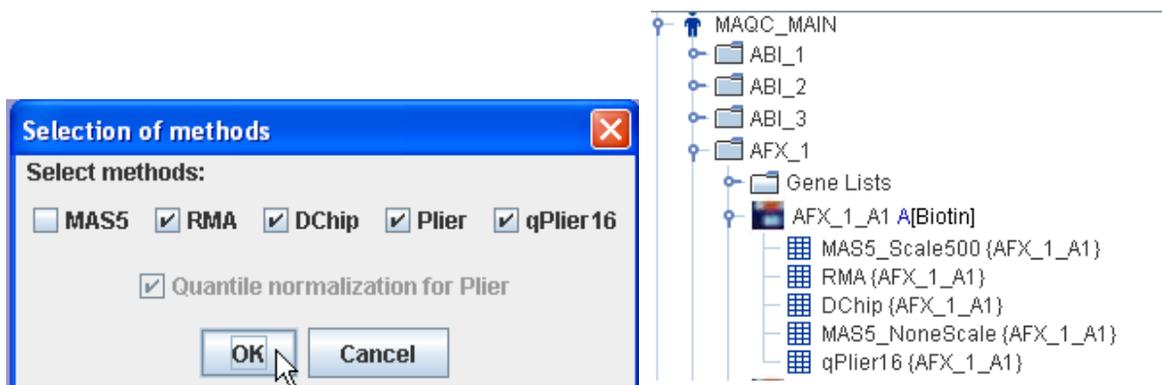


Figure 9-7: options for converting to various probe set files

## Chapter 10 Data Export

### 10.1 How to Access Data Export Functions

Experimental data stored in MicroarrayDB can be conveniently exported to locally-stored data files or ArrayTrack spreadsheets for further analysis purposes. Data export can be accessed from the Export pull-down menu (Figure 10-1), or from right-clicking on selected arrays of the same arrays type (Figure 10-2), where data export options are shown on the top of the list of functions applicable to the selected arrays.

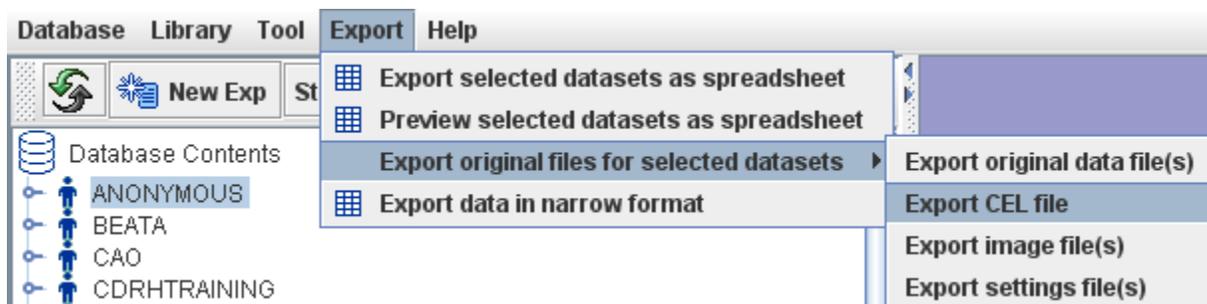


Figure 10-1: Accessing Data Export function from pull-down menu.

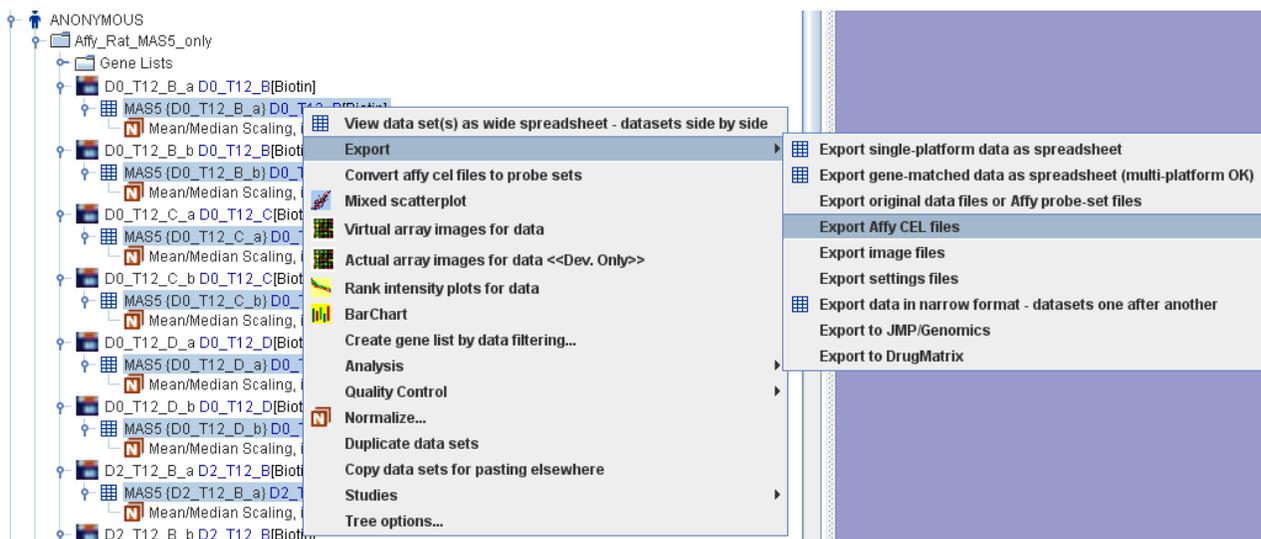


Figure 10-2: Accessing Data Export from right-clicking on selected arrays.

### 10.2 Options for Data Export

If user select the first export option(single-platform data as spreadsheet), the user will see the pop-up Export Options dialog box (Figure 10-3A) 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 five categories: Intensity Data Fields, Dataset Filtering, Gene Filtering, Spot Fields (i.e. annotation information about spots that is to be exported along with intensity data columns) and Dataset Naming. Dataset Filtering and Gene Filtering options allow the user to filter out unwanted data before export. All the other options are self-explaining and will not be discussed here. Options set here will remain effective for Data Export until re-set.

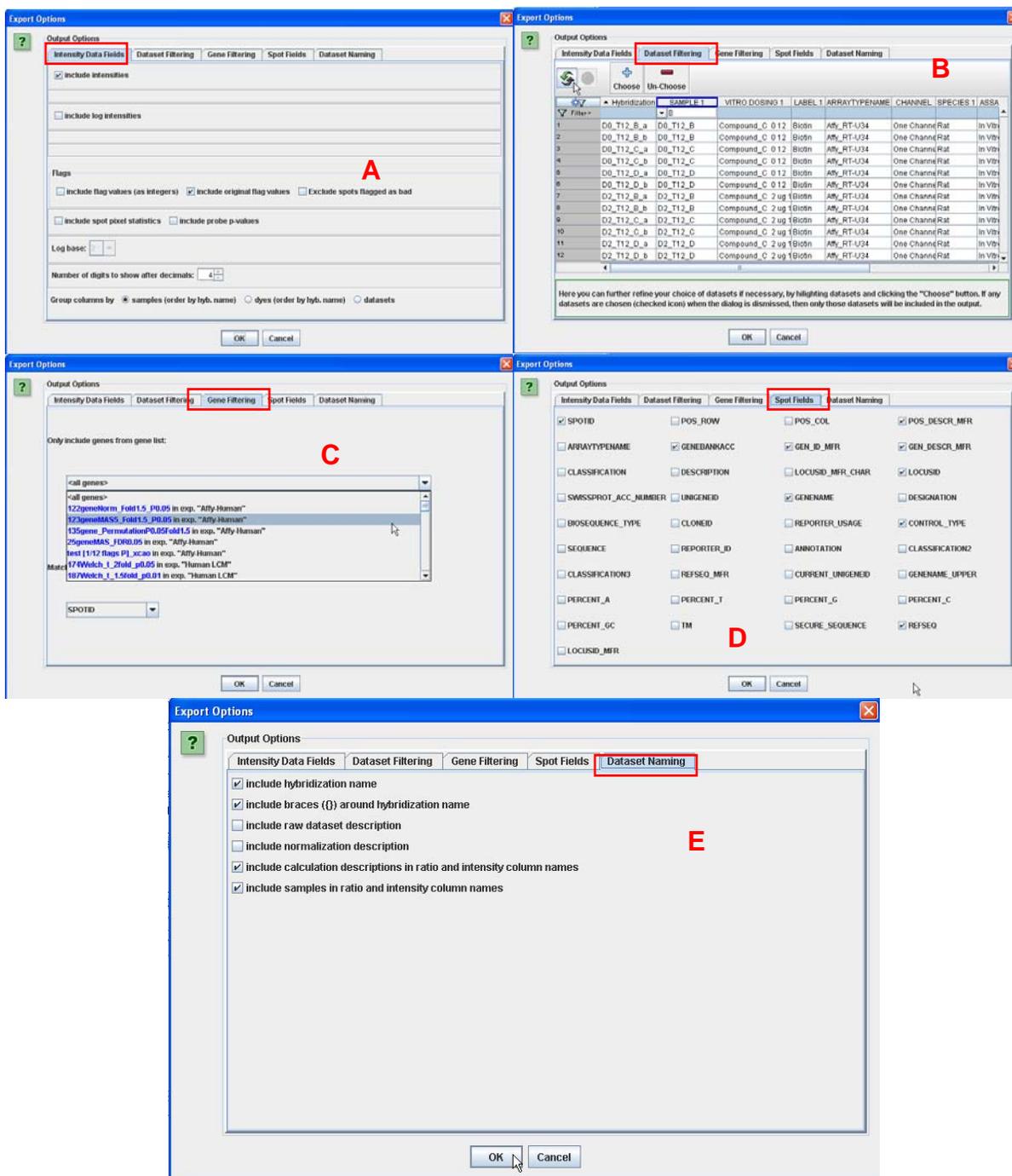


Figure 10-3: Setting Export Options before Data Export. A: Intensity Data Fields options; B: Dataset Naming options; C: Gene Filtering option; D: Spot Fields; E: Dataset Naming

### 10.3 Export selected datasets as spreadsheet

Data for selected arrays can be saved as a local text file in which each row represents a spot (gene) and each column is a particular data 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 10-4).



Figure 10-4: One or more arrays need to be selected before Export selected datasets as spreadsheet.

### 10.4 Export original data files or Affy probe-set files

This export option allows the user to export original data files or Affymetrix probe-set files. Right-clicking the selected data, see Figure 10-5, then choosing “Export” ->”Export original data files or Affy probe-set files” will bring out the window that asks users to choose the location for exported files (Figure 10-6).

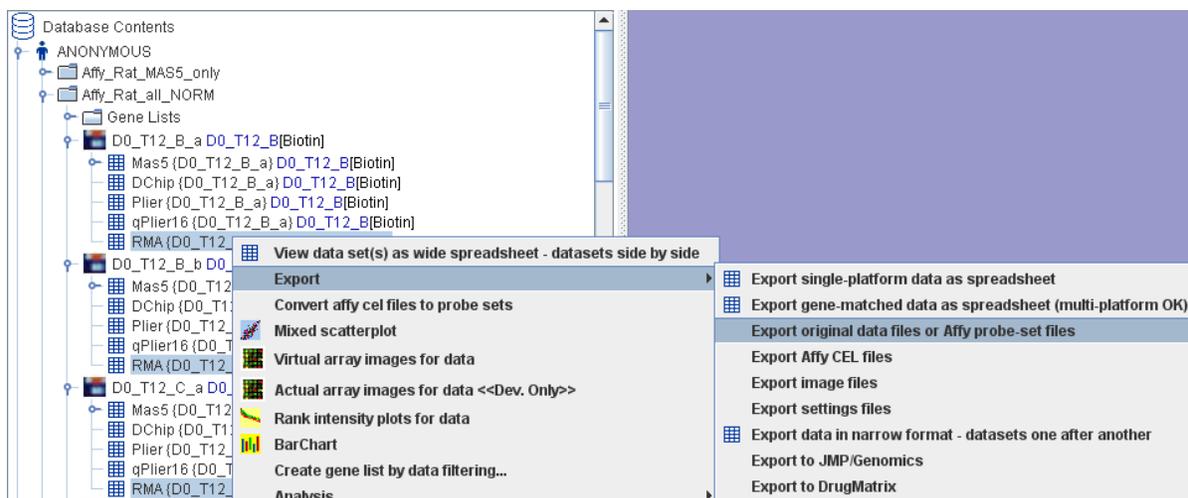


Figure 10-5: export original data files or Affymetrix probe-set files

In Figure 10-6, there are some options for the output file naming: 1) include original file names, 2) include hybridization names, 3) include internal dataset identifiers, 4) include file role (for CelData Setting...). Users can choose these options for their export preference.

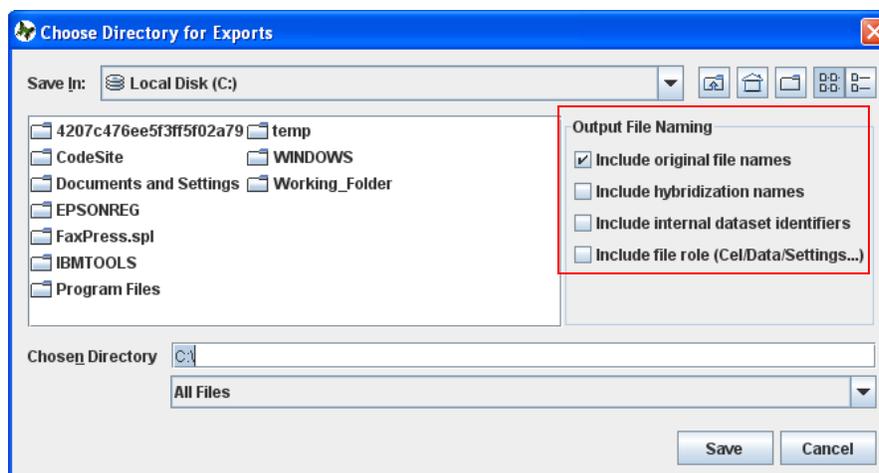


Figure 10-6: choose directory for exports

### 10.5 Export Affy CEL files

To export Affymetrix original CEL file, users can select the MAS5 data imported into ArrayTrack, right-click the select data, then choose “Export” ->”Export Affy CEL files”.

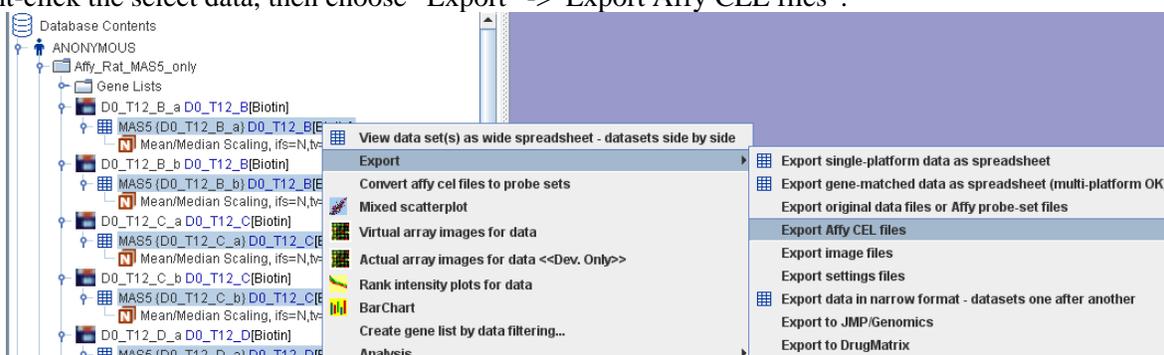


Figure 10-7: export Affymetrix CEL file

### 10.6 Export cross-platform data

ArrayTrack allows users to export datasets with multiple array types, for example Affymetrix data and Agilent data. To do this users first select multiple datasets (hold down Ctrl key), right-click then choose “Export” -> “Export gene-matched data se spreadsheet (multiple-platform OK)”.

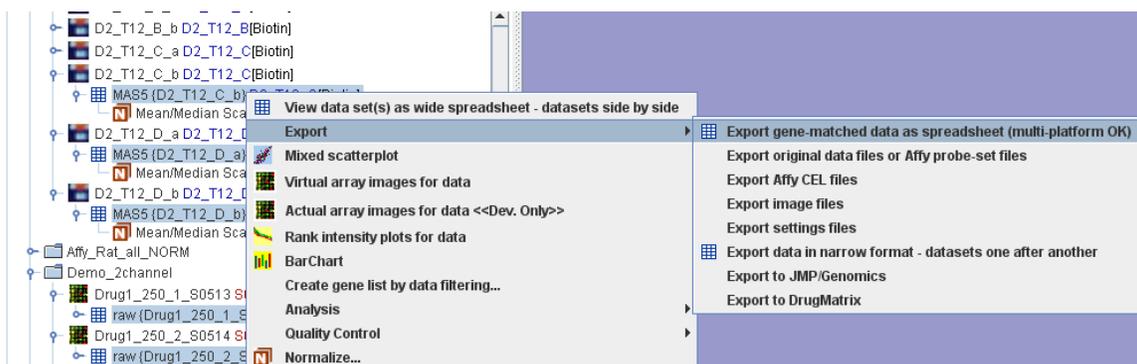


Figure 10-8: export data file with multiple platforms

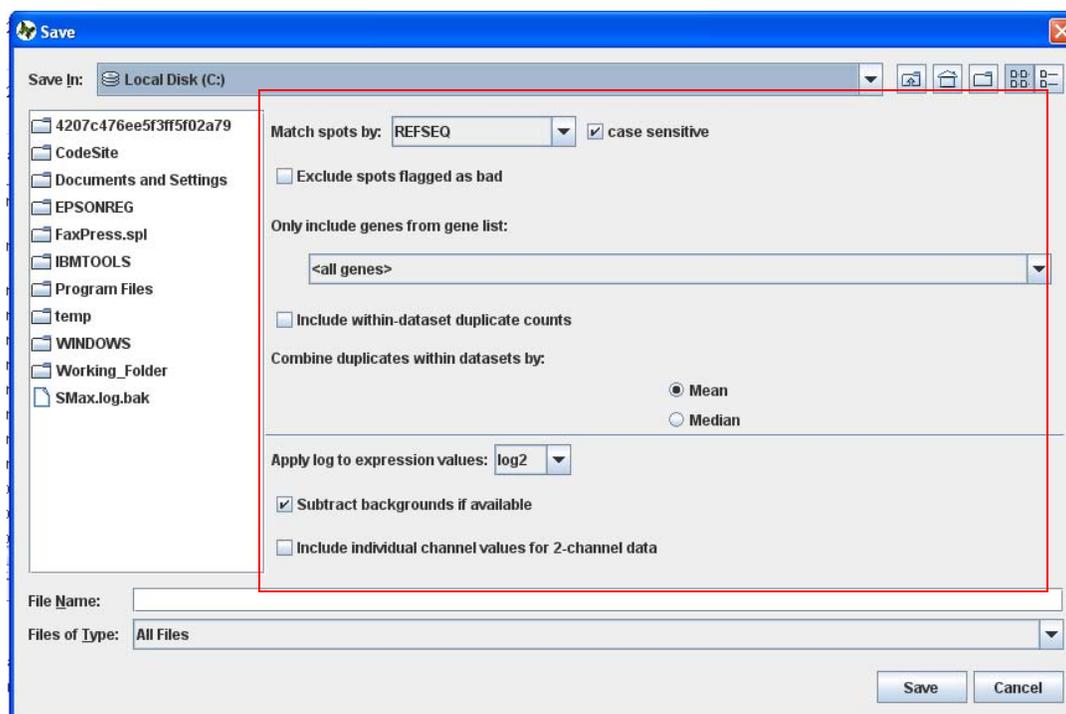


Figure 10-9: options for saving exported data

### 10.7 Export image files and settings files

Users need to select raw data first, then right-click and choose “Export” -> “Export image files” or “Export settings files”. See Figure 10-10.

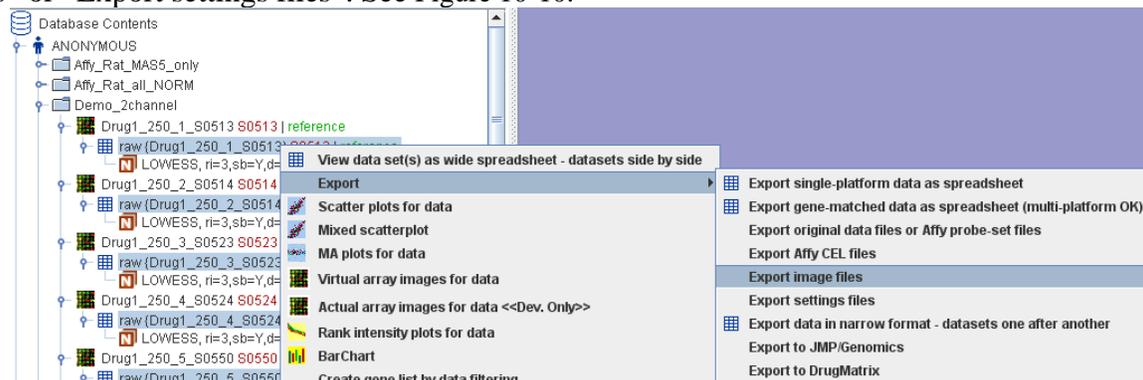


Figure 10-10: export image files or settings files

### 10.8 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.

### 10.9 Export to JMP/Genomics

In Figure 10-2, the user can export the data in JMP/Genomics format and save it to the local drive. The exported file can be directly opened in JMP.

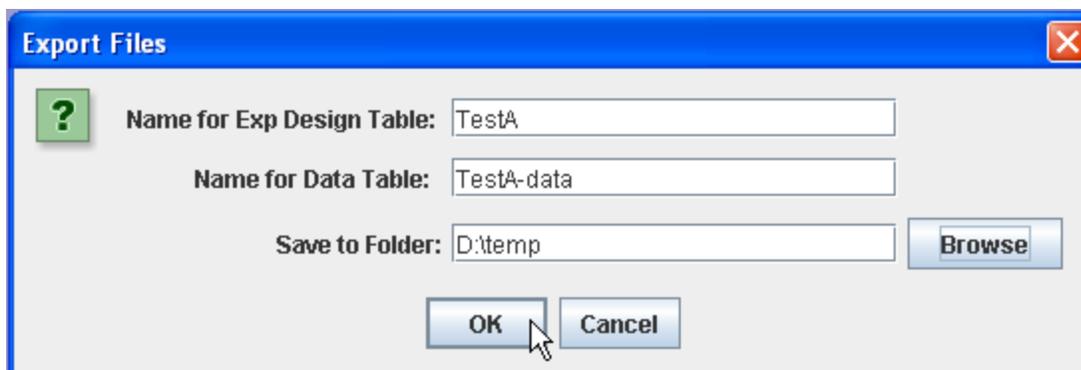


Figure 10-11: Export file compatible with JMP

### 10.10 Export to DrugMatrix

Right now this function is only available for data with array type Affy\_RAE230A, Affy\_RG230\_2 and GEHC\_RAT\_WHOLEGENOME300031. Other datasets with different array types can not be exported to DrugMatrix. If user selects datasets with other array types, he will get the following message:

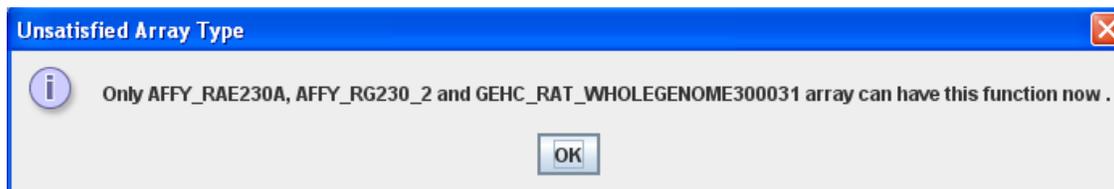


Figure 10-12: export to DrugMatrix is only available for datasets with limited array types

If user select datasets with one of the three array types mentioned above, the following window will pop up and allow user to export the datasets to DrugMatrix.

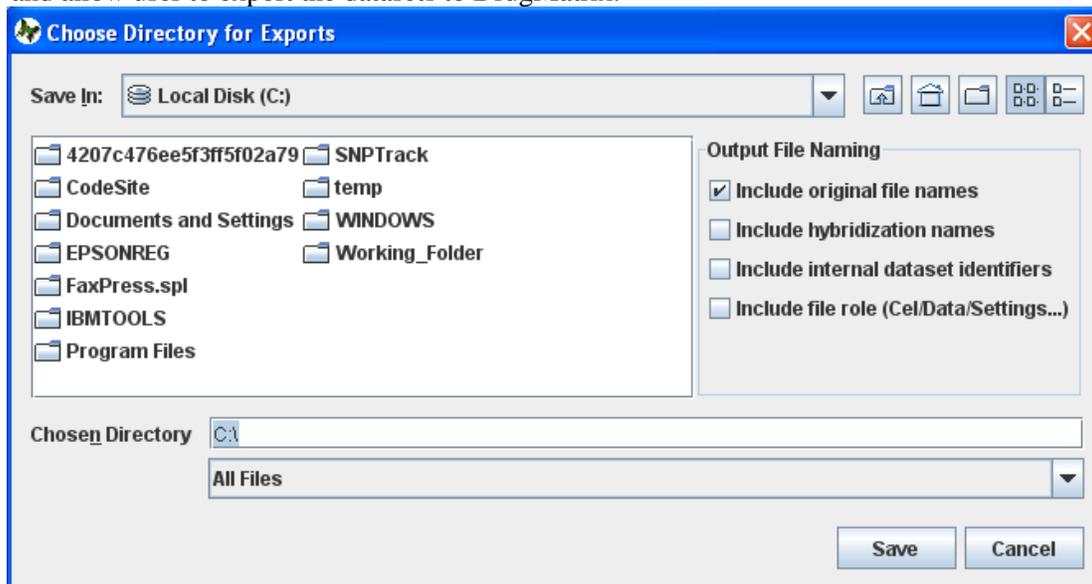


Figure 10-13: export datasets to DrugMatrix

## Appendix 1 Center for Toxicoinformatics of NCTR/FDA

### **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, cheminformatics, 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>.

## Appendix 2 Toxicoinformatics Integrated System TIS

### **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. Error! Reference source not found. 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.

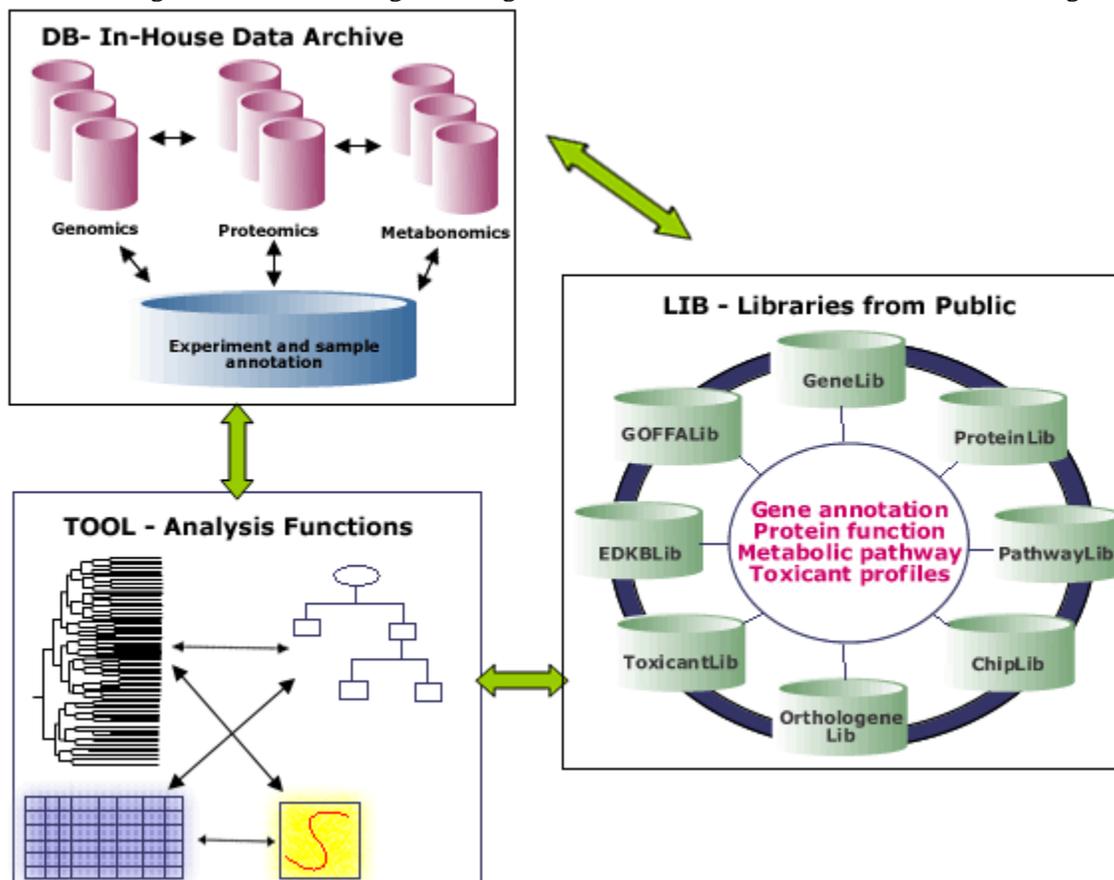


Figure 10-14: An over all system architecture of the Toxicoinformatics Integrated System (TIS).

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.

**System Requirements:** ArrayTrack is a client-server system. To date, the client 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@fda.hhs.gov](mailto:NCTRBioinformaticsSupport@fda.hhs.gov). The application screens are best viewed at 1024x768 (or higher) resolution.

#### Server Requirements:

1. Oracle Database Standard or Enterprise Edition, Version 9i or above.
2. Database storage of 40GB of disk space for a basic installation. (The storage requirements will increase significantly if users wish to use ArrayTrack as the repository for their microarray data.)
3. 2GB or more of memory for the Oracle instance. Refer to Oracle documentation for further information.
4. Oracle import utility (imp) version 10g or higher is required in order to import data into your database. (Windows batch scripts and UNIX shell scripts are included.)

5. A web server (optional). A web server is required when you run ArrayTrack client in prompted mode. It is used to host a java web start (jnlp) file. A web server is not required if run ArrayTrack in unprompted mode. (See "How To Install" section of the ArrayTrack client setup)
6. To have optimal performance, we suggest that you install the ArrayTrack database on a dedicated machine with optimized I/O. Consult with a Database Administrator for details.

#### Client Workstation Requirement:

1. ArrayTrack client needs to be running in Java Runtime Environment (JRE) 5 or above.
2. The database can be installed on the client machine for signal user use.

**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 5.0, if it is not already installed on your machine, by using the provided link to the Sun Microsystems web site in step one (The program will automatically detect whether Java 5.0 has already been 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 5.0 (+) 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](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).`

## 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., Hong Fang, Shi, L., Sun, H., Fuscoe, J., Harris, A., Hong, H., Xie, Q., Perkins, R., and Casciano, D. "Development of Public Toxicoinformatics Software for Microarray Data Management and Analysis." *Mutation Research*, 549:241-253, 2004.
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.
4. Sun, H., Hong Fang, Chen, T., Perkins, R., and Tong W. "GOFFA: Gene Ontology For Functional Analysis- Software for gene ontology-based functional analysis of genomic and proteomic data." *BMC Bioinformatics*, 7(Suppl 2):S23, 2006.
5. Tong W<sup>1</sup> Harris SW, Hong Fang<sup>2</sup>, Shi L, Perkins R, Goodsaid F and Frueh FW "An integrated bioinformatics infrastructure essential for advancing pharmacogenomics and

personalized medicine in the context of the FDA's Critical Path Initiative" *Drug Discovery Today: Technologies*, Volume 4, Issue 1, Autumn 2007, Pages 3-8

### **ArrayTrack Team**

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

Harris, Stephen C.  
Fang, Hong  
Arasappan, Dhivya  
Chen, Minjun  
Ge, Weigong  
Perkins, Roger  
Qian, Feng  
Shi, Leming  
Su, Zhenqiang  
Tong, Weida

### **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, former Director of NCTR/FDA and Dr. William Slikker, Director of NCTR/FDA.

Many early users of ArrayTrack around the world have provided valuable 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

Fax: +1-870-543-7662

E-mail: [NCTRBioinformaticsSupport@fda.hhs.gov](mailto:NCTRBioinformaticsSupport@fda.hhs.gov)  
[Weida.Tong@fda.hhs.gov](mailto>Weida.Tong@fda.hhs.gov), [hong.fang@fda.hhs.gov](mailto:hong.fang@fda.hhs.gov)