ArrayTrack—Supporting Toxicogenomic Research at the U.S. Food and Drug Administration National Center for Toxicological Research

Weida Tong,[1] Xiaoxi Cao,[2] Stephen Harris,[2] Hongmei Sun,[2] Hong Fang,[2] James Fuscoe,[3] Angela Harris,[4] Huixiao Hong,[2] Qian Xie,[2] Roger Perkins,[2] Leming Shi,[1] and Dan Casciano[5]

[1]Center for Toxicoinformatics, Division of Biometry and Risk Assessment, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, Arkansas, USA; [2]Northrop Grumman Information Technology, Jefferson, Arkansas, USA; [3]Center for Functional Genomics, Division of Reproductive and Genetic Toxicology, [4]Center for Hepatotoxicity, [5]Office of Director, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, Arkansas, USA

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 indicates an unprecedented advancement in the methods of understanding the expression of toxicity at the molecular level. At the National Center for Toxicological Research (NCTR) of the U.S. Food and Drug Administration, core facilities for genomic, proteomic, and metabonomic technologies have been established that use standardized experimental procedures to support centerwide toxicogenomic research. Collectively, these facilities are continuously producing an unprecedented volume of data. NCTR plans to develop a toxicoinformatics integrated system (TIS) for the purpose of fully integrating genomic, proteomic, and metabonomic data with the data in public repositories as well as conventional *in vitro* and *in vivo* toxicology data. The TIS will enable data curation in accordance with standard ontology and provide or interface a rich collection of tools for data analysis and knowledge mining. In this article the design, practical issues, and functions of the TIS are discussed through presenting its prototype version, ArrayTrack, for the management and analysis of DNA microarray data. ArrayTrack is logically constructed of three linked components: *a*) a library (LIB) that mirrors critical data in public databases; *b*) a database (MicroarrayDB) that stores microarray experiment information that is Minimal Information About a Microarray Experiment (MIAME) compliant; and *c*) tools (TOOL) that operate on experimental and public data for knowledge discovery. Using ArrayTrack, we can select an analysis method from the TOOL and apply the method to selected microarray data stored in the MicroarrayDB; the analysis results can be linked directly to gene information in the LIB. *Key words:* bioinformatics, data analysis, database, genomics, infrastructure, MIAME, microarray, systems toxicology, toxicogenomics, toxicology. *Environ Health Perspect* 111:1819–1826 (2003). doi:10.1289/txg.6497 available via *http://dx.doi.org/* [Online 15 September 2003]

While modern toxicology has focused on understanding biological mechanisms involved in the expression of toxicity at the molecular level, a technological revolution has occurred enabling researchers to perform experiments on a scale of unprecedented proportions (Marshall and Hodgson 1998; Ramsay 1998). High-throughput experimentation is producing large amounts of data impossible to analyze without informatics-related support (Bellenson 1999; Spengler 2000). We see a paradigm shift in toxicology research, where hypothesis-driven research is complemented by data-driven experimentation designed to be hypothesis generating (Afshari et al. 1999). Although toxicogenomics, the study of toxicology using high-throughput "omics" technologies (Aardema and MacGregor 2002; Hamadeh et al. 2002; Nuwaysir et al. 1999; Schmidt 2002; Ulrich and Friend 2002), and systems toxicology, the study of toxicology through data integration (Waters et al. 2003), have advanced rapidly and are likely to continue to advance, development of software infrastructures to manage, analyze, and integrate the diverse data has lagged behind. Recently, Waters et al. (2003) proposed a conceptual framework of chemical effects in biological systems [(CEBS) Chemical Effects in Biological Systems knowledge base] to meet the expanding toxicogenomic research needs at the National Center for Toxicogenomics (NCT) (Tennant 2002), including both NCT intramural research and research within the Toxicogenomics Research Consortium (TRC) (Medlin 2002). Both the NCT and the TRC are located at the National Institute of Environmental Health Sciences (NIEHS) in the Research Triangle Park, North Carolina.

Implementing toxicogenomic technologies is a high-priority initiative at the U.S. Food and Drug Administration (U.S. FDA) National Center for Toxicological Research (NCTR). A microarray core facility using validated and standardized protocols has been established. Similar facilities for proteomics and metabonomics are at an advanced stage of development and are preparing for validation of protocols. A toxicoinformatics integrated system (TIS) is concurrently being developed to meet the data management and analysis challenges associated with these efforts. The TIS is designed to aggregate data from toxicogenomic research with traditional toxicological end points and chemical data, along with sequence, gene function, and pathway data in public repositories. Through integration of different data types with analysis capabilities, the TIS will enable extraction of a tailored data set for data interpretation and hypothesis generation and testing.

In this article, the prototype of TIS, ArrayTrack, is presented in the context of meeting the following bioinformatics challenges associated with DNA microarray experiments in toxicology:

- How to manage the massive information associated with a microarray experiment and determine what relevant toxicology-specific experimental information or ontology needs to be acquired for the database.
- What visualization and analysis capabilities are required to efficiently extract knowledge from the microarray data.
- How the microarray experimental data should be linked with data from public databases to make the germane information on gene annotation, protein function, and pathways readily available for data interpretation.

Methods

ArrayTrack contains three integrated components: *a*) MicroarrayDB, which stores essential data associated with a

Address correspondence to W. Tong, Center for Toxicoinformatics, Division of Biometry and Risk Assessment, NCTR, 3900 NCTR Rd., HFT-020, Jefferson, AR 72079 USA. Telephone: (870) 543-7142. Fax: (870) 543-7662. E-mail: wtong@nctr.fda.gov

The authors gratefully acknowledge V. Desai at NCTR, who provided encouragement and insightful suggestions for ArrayTrack.

The authors declare they have no conflict of interest.

Received 3 June 2003; accepted 15 September 2003.

microarray experiment, including information on slide samples, treatment, and experimental results; *b*) TOOL, which provides analysis capabilities for data visualization, normalization, significance analysis, clustering, and classification; and *c*) LIB, which contains information from public repositories (e.g., gene annotation, protein function, and pathways). MicroarrayDB and LIB are used to store in-house experimental results and public data, respectively, whereas TOOL provides various algorithms for data visualization and analysis. At the time of this writing, ArrayTrack is not open-source software but can be accessed through the World Wide Web (http://edkb.fda.gov/webstart/arraytrack/). Prospective users can also acquire the software free of charge by contacting the authors.

Both MicroarrayDB and LIB were developed based on the Oracle relational database management system (Oracle Corp., Redwood Shores, CA). The database structure of MicroarrayDB and LIB was designed to accommodate the essential data associated with a microarray experiment as well as the data from the public repositories on genes, proteins, and pathways (database schema available upon request). The robust design allows data entities (tables of the identical type of data) and their relationships in the databases to be conveniently added and modified to accommodate needs of ever-evolving microarray technology and public databases. The diverse data in MicroarrayDB and LIB are stored in an IBM storage area network (SAN), and backed up daily using TSM (the Tivoli storage manager system).

User interface components providing query analysis and visualization capabilities are programmed in the Java language, ensuring portability to most computer operating systems as well as enabling easy Web deployment. Interfaces have been built for several data exchange formats, including flat text files and Microsoft Office Excel spreadsheets (Microsoft Corp., Redmond, WA). The "data drilling" capabilities were developed to allow the user to lock down and requery the database across other data within the realm of the previous query.

Controlling access to experimental data is a sensitive issue for many organizations and researchers. ArrayTrack allows only the owner of the data and members of groups approved by the owner to access the data to either read or write.

Results

Figure 1 depicts the ArrayTrack comprising three integrated components: *a*) MicroarrayDB, *b*) TOOL, and *c*) LIB. Through a user-friendly interface, the user can select an analysis method from the TOOL, apply the method to selected microarray data stored in the MicroarrayDB, and link the analysis results directly to gene information in the LIB. Additionally, ArrayTrack also allows data to be directly linked with other public databases.

MicroarrayDB

Microarray experimentation is one of the fastest-growing methods used in genomic research and has led to a broad diversity of microarray databases in both the public domain and commercial domains (Gardiner-Garden and Littlejohn 2001). Unfortunately, most if not all of them fail to accommodate the toxicology-related information needed for toxicogenomic studies. Recently, a joint effort between International Life Sciences Institute's Health and Environmental Sciences Institutes (ILSI HESI), NIEHS-NCT; and European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI) resulted in a draft defining the required information for toxicogenomic experiments. Augmenting the original Minimum Information about a Microarray Experiment (MIAME; http://www.mged.org) proposal (Brazma et al. 2001), the MIAME/Tox document outlines the minimum information required for a toxicogenomic experiment to ensure that the results are interpretable and the experiment is replicable.

Our goal is to develop a validated microarray database as a rich resource for cross-experiment and platform comparison to derive toxicity-specific signatures. By validated, we mean that data are stored if and only if they meet prescribed standards for completeness and accuracy as well as conformance to the applicable ontology. MicroarrayDB was designed to support toxicogenomic studies adhering to the MIAME guidelines. Currently, a number of journals, including *Nature*, the *Nature* group of journals, *Cell*, *The Lancet*, *EMBO*, and *Toxicology Pathology*, require an accession number from the public microarray databases developed based on the MIAME guidelines, which must be supplied on or before acceptance of publication (Anonymous 2002; Ball et al. 2002). The following practical issues were specifically discussed for implementing the MIAME guidelines among software developers, bioinformaticians, and toxicologists who

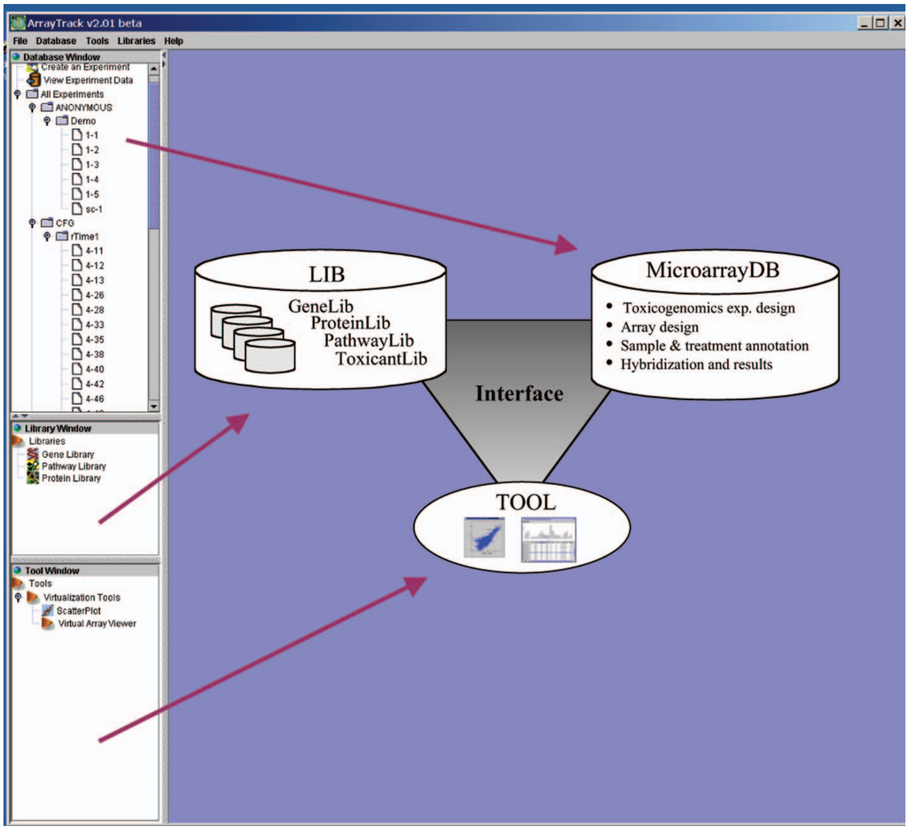


Figure 1. Overall system architecture of ArrayTrack. The software consists of three integrated components: MicroarrayDB captures toxicogenomic data associated with a microarray experiment; TOOL provides data visualization and analysis capabilities; and LIB contains annotated information on genes, proteins, and pathways.

work closely together to understand both the structure of the database and the structure of the data to be stored in the database:

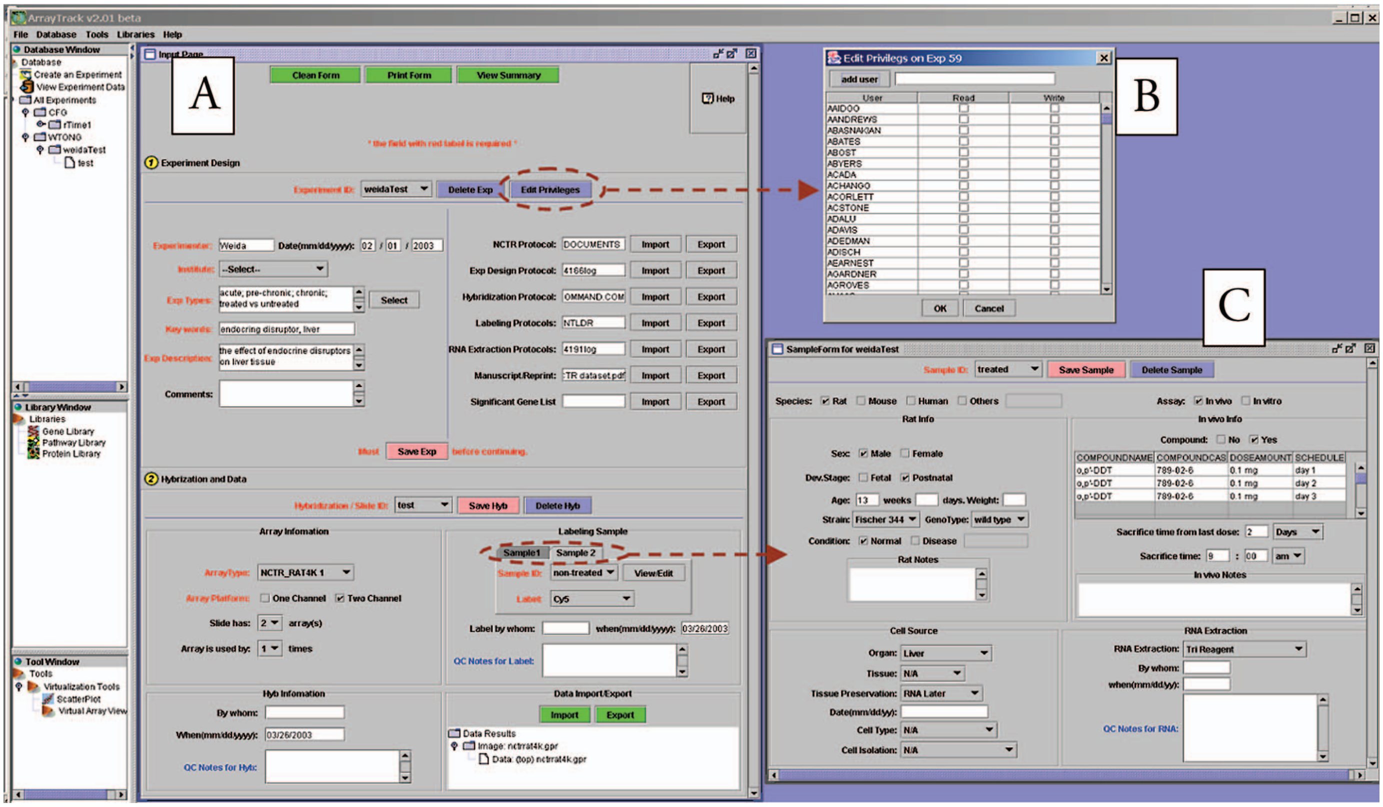
- MIAME versus database: MIAME specifies the content of the information to be available, whereas the database addresses how the content should be managed, and most importantly, queried. In other words, there is a distinction between the way the database handles all available information and a subset that is searchable. Technically, both available and searchable information can be treated in the same way. However, practically, such an approach usually imposes an inevitable burden on the end user to enter all information into the database in a tedious way, which might hinder their participation. Therefore, it is critical to define a balance point that can be accepted by both experimentalists and bioinformaticians.
- Local versus global repository: The MIAME guidelines broadly specify required data with the goal of a truly global repository for public data deposition and data exchange that would evolve as needs change. However, most databases similar to ArrayTrack are intended primarily, at least initially, for local use within an institution. For local institutional use, the extensive MIAME format can be simplified while still retaining essential information for toxicogenomics experiment interpretation and replication.

Thus, the ArrayTrack is MIAME-compliant, with inclusion of additional parameters related to toxicogenomics, using controlled vocabularies. Figure 2 gives the data submission requirements for essential information from both the microarray and toxicology perspectives. Currently, MicroarrayDB contains over 650 array data. We are closely following the current development of MicroArray Gene Expression Markup Language (MAGE-ML) standards (Spellman et al. 2002) that represent microarray data using markup language. We will develop a mean using MAGE-ML—an XML-based data exchange format—to allow data in MicroarrayDB to be communicated with other microarray data repositories such as ArrayExpress (http://www.ebi.ac.uk/arrayexpress; Brazma et al. 2003) and the Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo; Edgar et al. 2002).

LIB

The public domain has a rich and diverse collection of biological databases that greatly facilitates microarray experiment interpretation and associated knowledge discovery (Baxevanis 2003). A comprehensive summary of the major databases can be found in a special issue of *Nucleic Acids Research* (2003). The databases we find most valuable for microarray research are GenBank (http://www.ncbi.nlm.nih.gov/; Benson et al. 2003), SWISS-PROT (http://www.expasy.org/sprot/ and http://www.ebi.ac.uk/swissprot/; Boeckmann et al. 2003); LocusLink (www.ncbi.nlm.nih.gov/LocusLink; Kanehisa 2002); Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.ad.jp/kegg/; Kanehisa 2002); and Gene Ontology (GO; http://www.godatabase.org/dev/database/; Ashburner et al. 2000), each of which emphasizes a different type or aspect of biological information. ArrayTrack's LIB is a compilation of the essential public data to facilitate annotation and interpretation of microarray expression data.

We downloaded a number of public databases to create local mirrored databases that are automatically updated every



① Experiment Design

Experiment ID: weidaTest

Experimenter: Weida Date(mm/dd/yyyy): 02 / 01 / 2003

Institute: --Select--

Exp Types: acute; pre-chronic; chronic; treated vs untreated

Key words: endocring disruptor, liver

Exp Description: the effect of endocrine disruptors on liver tissue

Comments:

NCTR Protocol: DOCUMENTS

Exp Design Protocol: 4166log

Hybridization Protocol: OMMAND.COM

Labeling Protocols: NTLDR

RNA Extraction Protocols: 4191log

Manuscript/Reprint: :TR dataset.pdf

Significant Gene List

Must Save Exp before continuing.

② Hybrization and Data

Hybridization / Slide ID: test

Array Infomation

ArrayType: NCTR_RAT4K 1

Array Platform: □ One Channel ☑ Two Channel

Slide has: 2 array(s)

Array is used by: 1 times

Labeling Sample

Sample1 Sample 2

Sample ID: non-treated

Label: Cy5

Label by whom: when(mm/dd/yyyy): 03/26/2003

QC Notes for Label:

Hyb Infomation

By whom:

When(mm/dd/yyyy): 03/26/2003

QC Notes for Hyb:

Data Import/Export

Data Results
Image: nctrrat4k.gpr
Data: (top) nctrrat4k.gpr

Edit Privilegs on Exp 59

User	Read	Write
AAIDOO	□	□
AANDREWS	□	□
ABASNAKIAN	□	□
ABATES	□	□
ABOST	□	□
ABYERS	□	□
ACADA	□	□
ACHANGO	□	□
ACORLETT	□	□
ACSTONE	□	□
ADALU	□	□
ADAVIS	□	□
ADEDMAN	□	□
ADISCH	□	□
AEARNEST	□	□
AGARDNER	□	□
AGROVES	□	□

SampleForm for weidaTest

Sample ID: treated

Species: ☑ Rat □ Mouse □ Human □ Others

Assay: ☑ In vivo □ In vitro

Rat Info

Sex: ☑ Male □ Female

Dev.Stage: □ Fetal ☑ Postnatal

Age: 13 weeks days. Weight:

Strain: Fischer 344 GenoType: wild type

Condition: ☑ Normal □ Disease

Rat Notes

In vivo Info

Compound: □ No ☑ Yes

COMPOUNDNAME	COMPOUNDCAS	DOSEAMOUNT	SCHEDULE
o,p'-DDT	789-02-6	0.1 mg	day 1
o,p'-DDT	789-02-6	0.1 mg	day 2
o,p'-DDT	789-02-6	0.1 mg	day 3

Sacrifice time from last dose: 2 Days

Sacrifice time: 9 : 00 am

In vivo Notes

Cell Source

Organ: Liver

Tissue: N/A

Tissue Preservation: RNA Later

Date(mm/dd/yy):

Cell Type: N/A

Cell Isolation: N/A

RNA Extraction

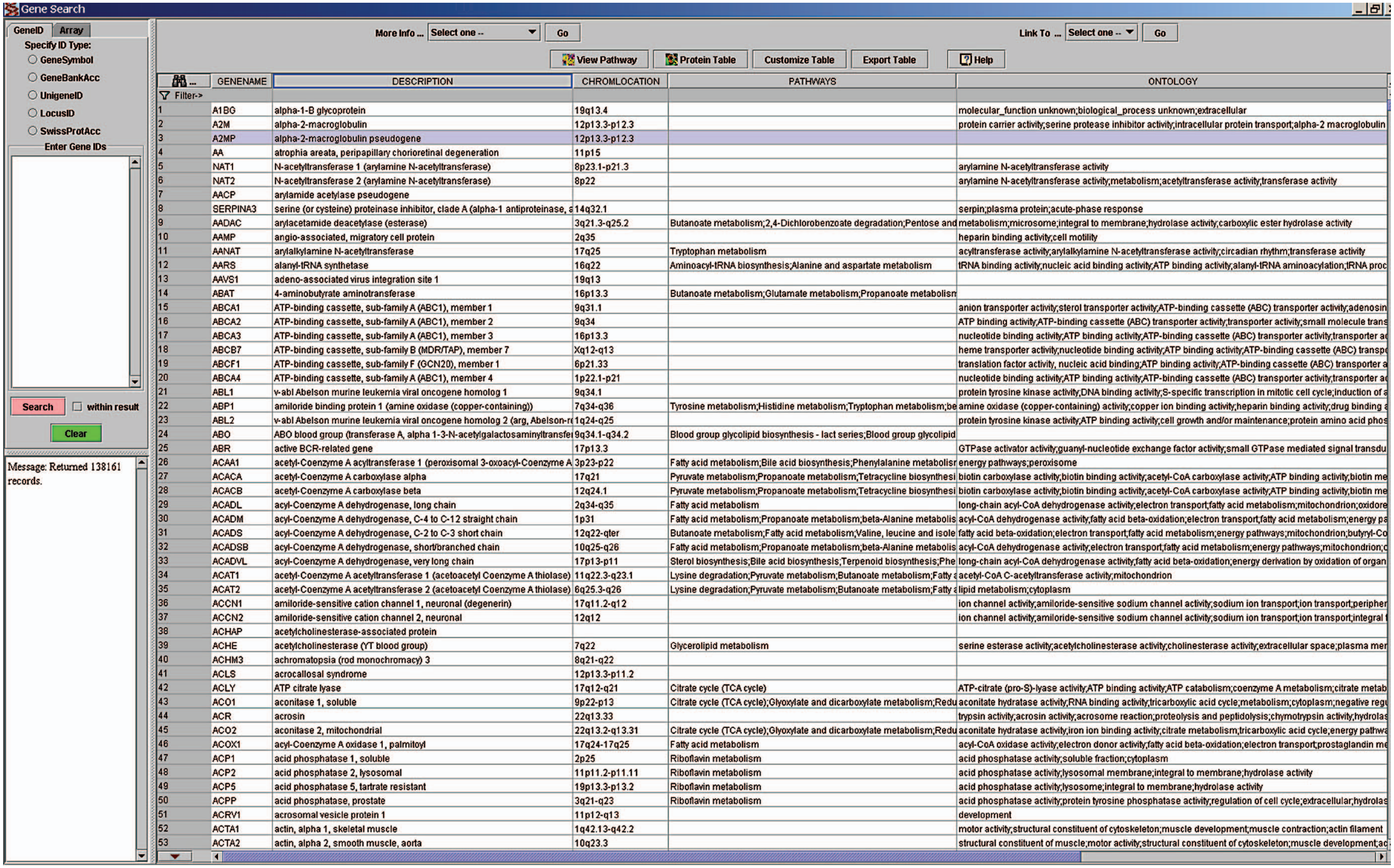
RNA Extraction: Tri Reagent

By whom:

when(mm/dd/yy):

QC Notes for RNA:

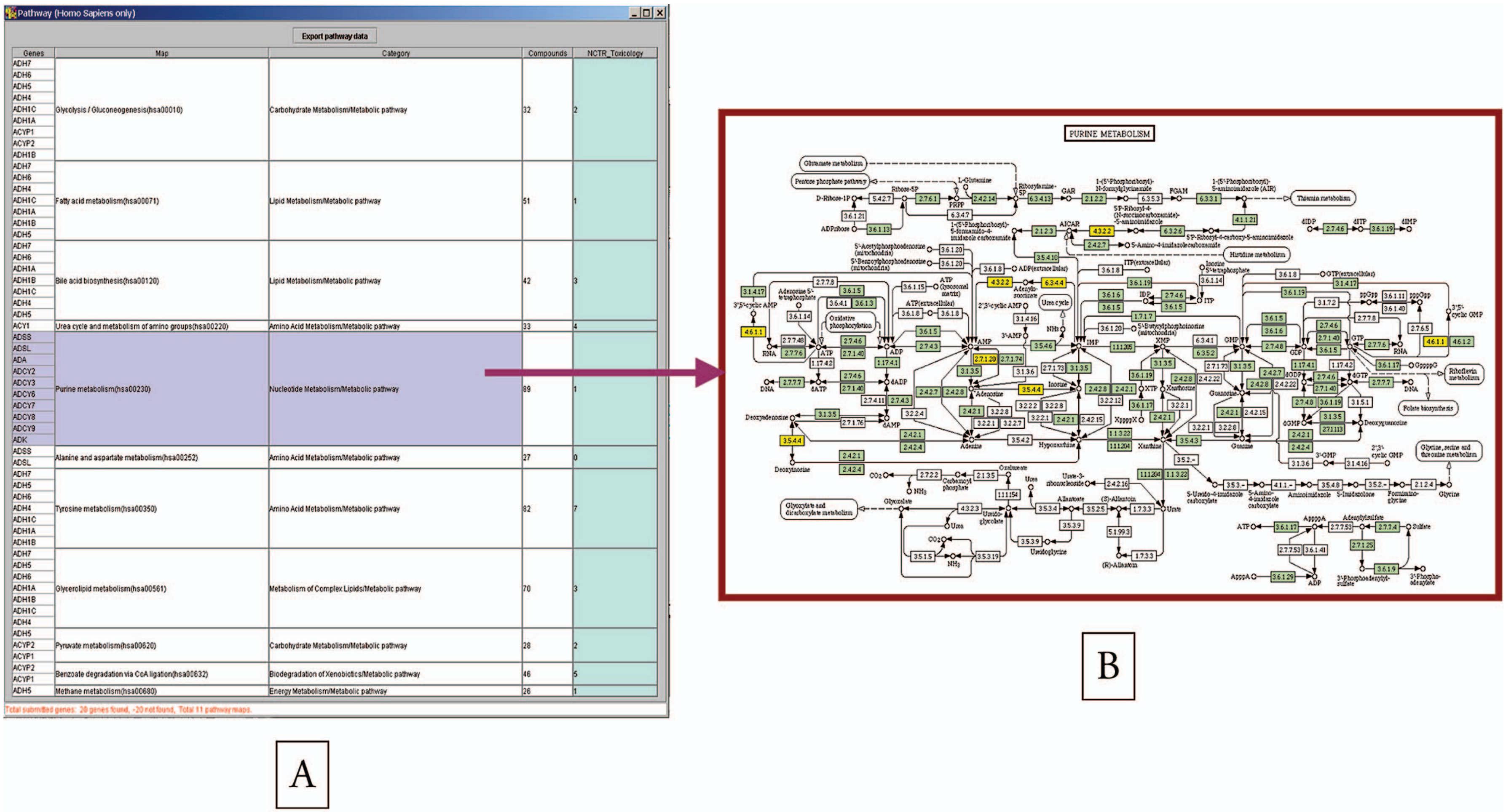
Figure 2. Data submission form. The form (*A*) consists of two sequential components: (1) "Experiment Design" and (2) "Hybridization and Data." The description of an experiment and its associated experimental protocols are input in the section "Experiment Design," where the owner of the experiment can define "read and/or write" privilege (*B*) to share the experiment with collaborators. The content and output of the experiment are input in the section "Hybridization and Data." The experiment details on the sample preparation can be input through this section (*C*).



Message: Returned 138161 records.

	GENENAME	DESCRIPTION	CHROMLOCATION	PATHWAYS	ONTOLOGY
1	A1BG	alpha-1-B glycoprotein	19q13.4		molecular_function unknown;biological_process unknown;extracellular
2	A2M	alpha-2-macroglobulin	12p13.3-p12.3		protein carrier activity;serine protease inhibitor activity;intracellular protein transport;alpha-2 macroglobulin
3	A2MP	alpha-2-macroglobulin pseudogene	12p13.3-p12.3		
4	AA	atrophia areata, peripapillary chorioretinal degeneration	11p15		
5	NAT1	N-acetyltransferase 1 (arylamine N-acetyltransferase)	8p23.1-p21.3		arylamine N-acetyltransferase activity
6	NAT2	N-acetyltransferase 2 (arylamine N-acetyltransferase)	8p22		arylamine N-acetyltransferase activity;metabolism;acetyltransferase activity;transferase activity
7	AACP	arylamide acetylase pseudogene			
8	SERPINA3	serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, a	14q32.1		serpin;plasma protein;acute-phase response
9	AADAC	arylacetamide deacetylase (esterase)	3q21.3-q25.2	Butanoate metabolism;2,4-Dichlorobenzoate degradation;Pentose and	metabolism;microsome;integral to membrane;hydrolase activity;carboxylic ester hydrolase activity
10	AAMP	angio-associated, migratory cell protein	2q35		heparin binding activity;cell motility
11	AANAT	arylalkylamine N-acetyltransferase	17q25	Tryptophan metabolism	acyltransferase activity;arylalkylamine N-acetyltransferase activity;circadian rhythm;transferase activity
12	AARS	alanyl-tRNA synthetase	16q22	Aminoacyl-tRNA biosynthesis;Alanine and aspartate metabolism	tRNA binding activity;nucleic acid binding activity;ATP binding activity;alanyl-tRNA aminoacylation;tRNA proc
13	AAVS1	adeno-associated virus integration site 1	19q13		
14	ABAT	4-aminobutyrate aminotransferase	16p13.3	Butanoate metabolism;Glutamate metabolism;Propanoate metabolism	
15	ABCA1	ATP-binding cassette, sub-family A (ABC1), member 1	9q31.1		anion transporter activity;sterol transporter activity;ATP-binding cassette (ABC) transporter activity;adenosin
16	ABCA2	ATP-binding cassette, sub-family A (ABC1), member 2	9q34		ATP binding activity;ATP-binding cassette (ABC) transporter activity;transporter activity;small molecule trans
17	ABCA3	ATP-binding cassette, sub-family A (ABC1), member 3	16p13.3		nucleotide binding activity;ATP binding activity;ATP-binding cassette (ABC) transporter activity;transporter ac
18	ABCB7	ATP-binding cassette, sub-family B (MDR/TAP), member 7	Xq12-q13		heme transporter activity;nucleotide binding activity;ATP binding activity;ATP-binding cassette (ABC) transpo
19	ABCF1	ATP-binding cassette, sub-family F (GCN20), member 1	6p21.33		translation factor activity, nucleic acid binding;ATP binding activity;ATP-binding cassette (ABC) transporter a
20	ABCA4	ATP-binding cassette, sub-family A (ABC1), member 4	1p22.1-p21		nucleotide binding activity;ATP binding activity;ATP-binding cassette (ABC) transporter activity;transporter ac
21	ABL1	v-abl Abelson murine leukemia viral oncogene homolog 1	9q34.1		protein tyrosine kinase activity;DNA binding activity;S-specific transcription in mitotic cell cycle;induction of a
22	ABP1	amiloride binding protein 1 (amine oxidase (copper-containing))	7q34-q36	Tyrosine metabolism;Histidine metabolism;Tryptophan metabolism;be	amine oxidase (copper-containing) activity;copper ion binding activity;heparin binding activity;drug binding a
23	ABL2	v-abl Abelson murine leukemia viral oncogene homolog 2 (arg, Abelson-re	1q24-q25		protein tyrosine kinase activity;ATP binding activity;cell growth and/or maintenance;protein amino acid phos
24	ABO	ABO blood group (transferase A, alpha 1-3-N-acetylgalactosaminyltransfe	9q34.1-q34.2	Blood group glycolipid biosynthesis - lact series;Blood group glycolipid	
25	ABR	active BCR-related gene	17p13.3		GTPase activator activity;guanyl-nucleotide exchange factor activity;small GTPase mediated signal transdu
26	ACAA1	acetyl-Coenzyme A acyltransferase 1 (peroxisomal 3-oxoacyl-Coenzyme A	3p23-p22	Fatty acid metabolism;Bile acid biosynthesis;Phenylalanine metabolisr	energy pathways;peroxisome
27	ACACA	acetyl-Coenzyme A carboxylase alpha	17q21	Pyruvate metabolism;Propanoate metabolism;Tetracycline biosynthesi	biotin carboxylase activity;biotin binding activity;acetyl-CoA carboxylase activity;ATP binding activity;biotin me
28	ACACB	acetyl-Coenzyme A carboxylase beta	12q24.1	Pyruvate metabolism;Propanoate metabolism;Tetracycline biosynthesi	biotin carboxylase activity;biotin binding activity;acetyl-CoA carboxylase activity;ATP binding activity;biotin me
29	ACADL	acyl-Coenzyme A dehydrogenase, long chain	2q34-q35	Fatty acid metabolism	long-chain acyl-CoA dehydrogenase activity;electron transport;fatty acid metabolism;mitochondrion;oxidore
30	ACADM	acyl-Coenzyme A dehydrogenase, C-4 to C-12 straight chain	1p31	Fatty acid metabolism;Propanoate metabolism;beta-Alanine metabolis	acyl-CoA dehydrogenase activity;fatty acid beta-oxidation;electron transport;fatty acid metabolism;energy pa
31	ACADS	acyl-Coenzyme A dehydrogenase, C-2 to C-3 short chain	12q22-qter	Butanoate metabolism;Fatty acid metabolism;Valine, leucine and isole	fatty acid beta-oxidation;electron transport;fatty acid metabolism;energy pathways;mitochondrion;butyryl-Co
32	ACADSB	acyl-Coenzyme A dehydrogenase, short/branched chain	10q25-q26	Fatty acid metabolism;Propanoate metabolism;beta-Alanine metabolis	acyl-CoA dehydrogenase activity;electron transport;fatty acid metabolism;energy pathways;mitochondrion;c
33	ACADVL	acyl-Coenzyme A dehydrogenase, very long chain	17p13-p11	Sterol biosynthesis;Bile acid biosynthesis;Terpenoid biosynthesis;Phe	long-chain acyl-CoA dehydrogenase activity;fatty acid beta-oxidation;energy derivation by oxidation of organ
34	ACAT1	acetyl-Coenzyme A acetyltransferase 1 (acetoacetyl Coenzyme A thiolase)	11q22.3-q23.1	Lysine degradation;Pyruvate metabolism;Butanoate metabolism;Fatty a	acetyl-CoA C-acetyltransferase activity;mitochondrion
35	ACAT2	acetyl-Coenzyme A acetyltransferase 2 (acetoacetyl Coenzyme A thiolase)	6q25.3-q26	Lysine degradation;Pyruvate metabolism;Butanoate metabolism;Fatty a	lipid metabolism;cytoplasm
36	ACCN1	amiloride-sensitive cation channel 1, neuronal (degenerin)	17q11.2-q12		ion channel activity;amiloride-sensitive sodium channel activity;sodium ion transport;ion transport;peripher
37	ACCN2	amiloride-sensitive cation channel 2, neuronal	12q12		ion channel activity;amiloride-sensitive sodium channel activity;sodium ion transport;ion transport;integral t
38	ACHAP	acetylcholinesterase-associated protein			
39	ACHE	acetylcholinesterase (YT blood group)	7q22	Glycerolipid metabolism	serine esterase activity;acetylcholinesterase activity;cholinesterase activity;extracellular space;plasma mer
40	ACHM3	achromatopsia (rod monochromacy) 3	8q21-q22		
41	ACLS	acrocallosal syndrome	12p13.3-p11.2		
42	ACLY	ATP citrate lyase	17q12-q21	Citrate cycle (TCA cycle)	ATP-citrate (pro-S)-lyase activity;ATP binding activity;ATP catabolism;coenzyme A metabolism;citrate metab
43	ACO1	aconitase 1, soluble	9p22-p13	Citrate cycle (TCA cycle);Glyoxylate and dicarboxylate metabolism;Redu	aconitate hydratase activity;RNA binding activity;tricarboxylic acid cycle;metabolism;cytoplasm;negative regu
44	ACR	acrosin	22q13.33		trypsin activity;acrosin activity;acrosome reaction;proteolysis and peptidolysis;chymotrypsin activity;hydrolas
45	ACO2	aconitase 2, mitochondrial	22q13.2-q13.31	Citrate cycle (TCA cycle);Glyoxylate and dicarboxylate metabolism;Redu	aconitate hydratase activity;iron ion binding activity;citrate metabolism;tricarboxylic acid cycle;energy pathwa
46	ACOX1	acyl-Coenzyme A oxidase 1, palmitoyl	17q24-17q25	Fatty acid metabolism	acyl-CoA oxidase activity;electron donor activity;fatty acid beta-oxidation;electron transport;prostaglandin me
47	ACP1	acid phosphatase 1, soluble	2p25	Riboflavin metabolism	acid phosphatase activity;soluble fraction;cytoplasm
48	ACP2	acid phosphatase 2, lysosomal	11p11.2-p11.11	Riboflavin metabolism	acid phosphatase activity;lysosomal membrane;integral to membrane;hydrolase activity
49	ACP5	acid phosphatase 5, tartrate resistant	19p13.3-p13.2	Riboflavin metabolism	acid phosphatase activity;lysosome;integral to membrane;hydrolase activity
50	ACPP	acid phosphatase, prostate	3q21-q23	Riboflavin metabolism	acid phosphatase activity;protein tyrosine phosphatase activity;regulation of cell cycle;extracellular;hydrolas
51	ACRV1	acrosomal vesicle protein 1	11p12-q13		development
52	ACTA1	actin, alpha 1, skeletal muscle	1q42.13-q42.2		motor activity;structural constituent of cytoskeleton;muscle development;muscle contraction;actin filament
53	ACTA2	actin, alpha 2, smooth muscle, aorta	10q23.3		structural constituent of muscle;motor activity;structural constituent of cytoskeleton;muscle development;ac

Figure 3. GeneLib. The main part of the GeneLib is an Excel-like spreadsheet, where each row is associated with a gene, whereas each column presents a particular functional annotation, such as chromosomal location, pathways, and functional assignments (molecular function, biological process, and cellular component) defined by GO. On the top and left side of the spreadsheet, several functions are available to assist further exploration of the contents in the GeneLib.



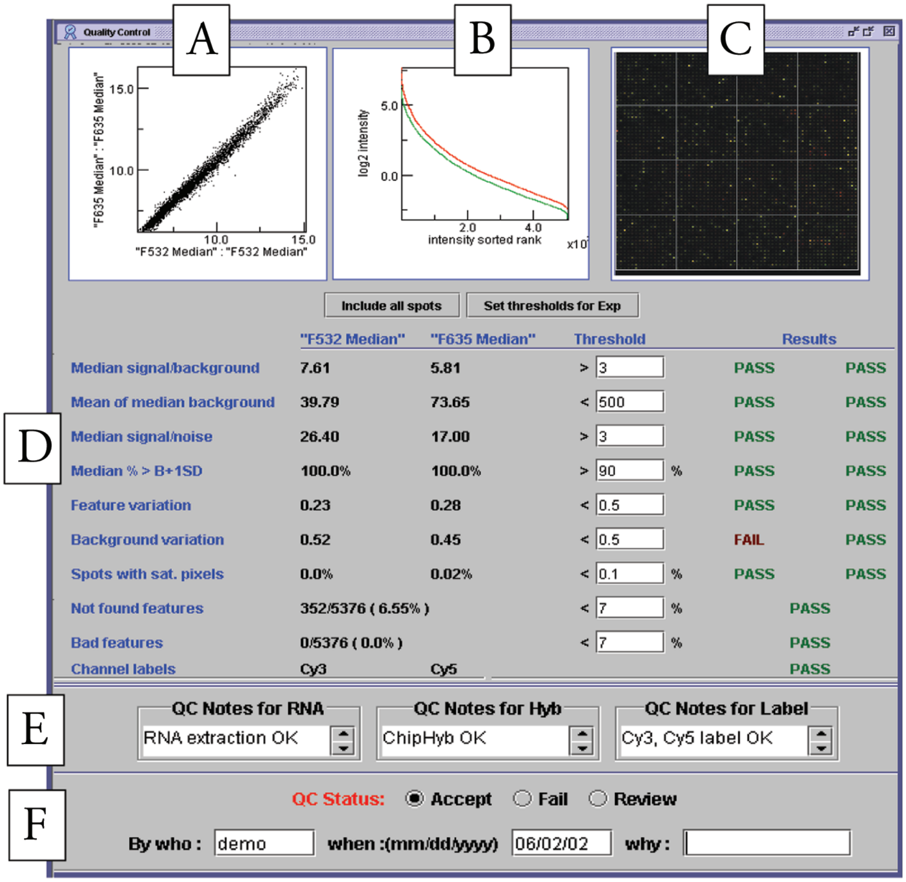
Genes	Map	Category	Compounds	NCTR_Toxicology
ADH7 ADH6 ADH5 ADH4 ADH1C ADH1A ACYP1 ACYP2 ADH1B	Glycolysis / Gluconeogenesis(hsa00010)	Carbohydrate Metabolism/Metabolic pathway	32	2
ADH7 ADH6 ADH4 ADH1C ADH1A ADH1B ADH5	Fatty acid metabolism(hsa00071)	Lipid Metabolism/Metabolic pathway	51	1
ADH7 ADH6 ADH1A ADH1B ADH1C ADH4 ADH5	Bile acid biosynthesis(hsa00120)	Lipid Metabolism/Metabolic pathway	42	3
ACY1	Urea cycle and metabolism of amino groups(hsa00220)	Amino Acid Metabolism/Metabolic pathway	33	4
ADSS ADSL ADA ADCY2 ADCY3 ADCY6 ADCY7 ADCY8 ADCY9 ADK	Purine metabolism(hsa00230)	Nucleotide Metabolism/Metabolic pathway	89	1
ADSS ADSL	Alanine and aspartate metabolism(hsa00252)	Amino Acid Metabolism/Metabolic pathway	27	0
ADH7 ADH5 ADH6 ADH4 ADH1C ADH1A ADH1B	Tyrosine metabolism(hsa00350)	Amino Acid Metabolism/Metabolic pathway	82	7
ADH7 ADH5 ADH6 ADH1A ADH1B ADH1C ADH4	Glycerolipid metabolism(hsa00561)	Metabolism of Complex Lipids/Metabolic pathway	70	3
ADH5 ACYP2 ACYP1	Pyruvate metabolism(hsa00620)	Carbohydrate Metabolism/Metabolic pathway	28	2
ACYP2 ACYP1	Benzoate degradation via CoA ligation(hsa00632)	Biodegradation of Xenobiotics/Metabolic pathway	46	5
ADH5	Methane metabolism(hsa00680)	Energy Metabolism/Metabolic pathway	26	1

Total submitted genes: 20 genes found, -20 not found, Total 11 pathway maps.

Figure 4. The genes were ordered based on their common pathways (*A*). Each pathway was directly linked to the detailed pathway map (*B*) in KEGG, where the genes involved in the pathway (*A*) were highlighted in yellow on the map (*B*) to indicate their roles in the pathway.

2 weeks using scripts. The LIB was the selected aggregation of the information in the mirrored databases that was relevant for interpretation of microarray results. Currently, the LIB comprises three sublibraries, GeneLib, ProteinLib, and PathwayLib, which concentrate public data on genes, proteins, and pathways, respectively. Each contains only the most relevant selected information from UniGene (http://www.ncbi.nlm.nih.gov/UniGene/), LocusLink, SWISS-PROT, KEGG, and GO.

The three libraries (GeneLib, ProteinLib, and PathwayLib) have the same design and functional interface. A screen shot of the GeneLib is displayed in Figure 3. The gene information is displayed in an Excel-like spreadsheet. Each row is associated with a gene, and each column is a particular functional annotation, such as chromosomal location, pathway, or functional assignment (molecular function, biological process, and cellular component) defined by GO (Ashburner et al. 2000). The spreadsheet can be customized by including/excluding a specific functional annotation. The common functions such as sorting, ranking, and querying are available for comparison across the entire gene list. The genes can also be categorized on the basis of their common pathways (Figure 4). In addition, detailed information on each gene is available, including synonym, sequence, chromosomal map, and reference. Information for genes not contained in the GeneLib is readily available by hot link to a wide range of public data repositories.



	"F532 Median"	"F635 Median"	Threshold	Results	
Median signal/background	7.61	5.81	> 3	PASS	PASS
Mean of median background	39.79	73.65	< 500	PASS	PASS
Median signal/noise	26.40	17.00	> 3	PASS	PASS
Median % > B+1SD	100.0%	100.0%	> 90 %	PASS	PASS
Feature variation	0.23	0.28	< 0.5	PASS	PASS
Background variation	0.52	0.45	< 0.5	FAIL	PASS
Spots with sat. pixels	0.0%	0.02%	< 0.1 %	PASS	PASS
Not found features	352/5376 (6.55%)		< 7 %	PASS	
Bad features	0/5376 (0.0%)		< 7 %	PASS	
Channel labels	Cy3	Cy5		PASS	

QC Notes for RNA: RNA extraction OK

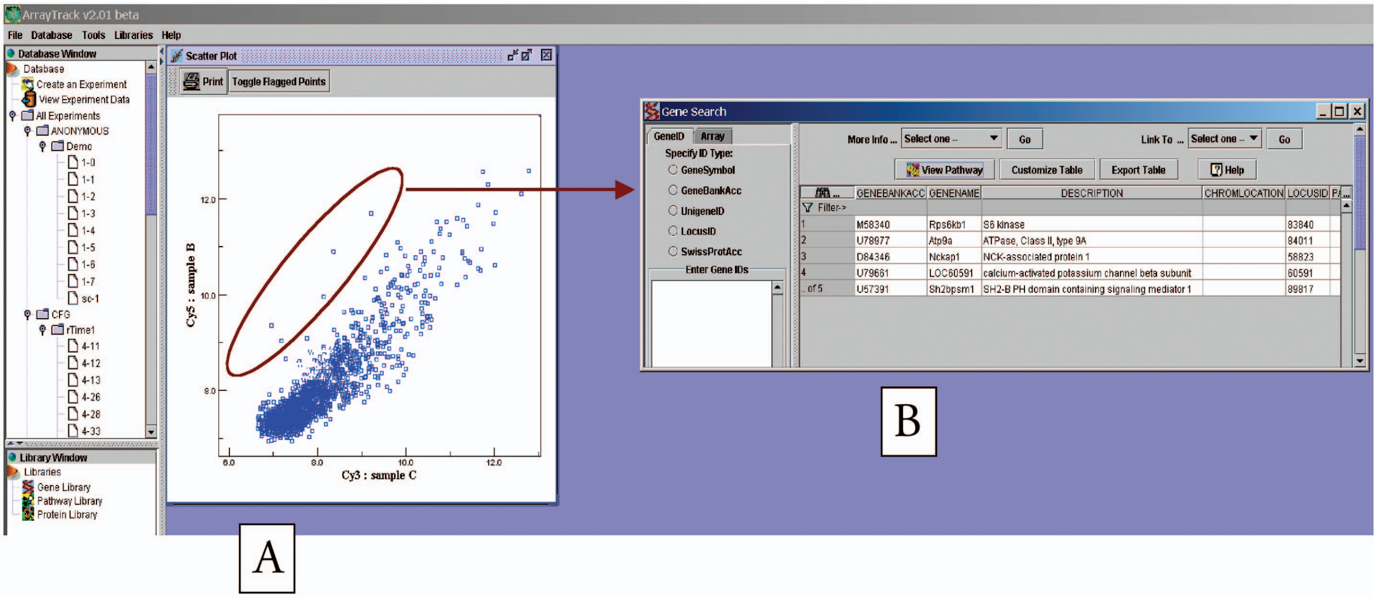
QC Notes for Hyb: ChipHyb OK

QC Notes for Label: Cy3, Cy5 label OK

QC Status: ● Accept ○ Fail ○ Review

By who : demo when :(mm/dd/yyyy) 06/02/02 why :

Figure 5. Quality control interface summarizes most relevant information into one interface to facilitate the process of quality control of a microarray experiment. The tool allows investigator(s) to make a QC/QA decision (*F*) for a slide using the following information: (*A*) the Cy3 versus Cy5 plot; (*B*) the rank intensity plot of Cy3 and Cy5; (*C*) the slide image; (*D*) the summary of various statistics; and (*E*) the experimental annotation.



	GENEBANKACC	GENENAME	DESCRIPTION	CHROMLOCATION	LOCUSID	PA...
Filter->						
1	M58340	Rps6kb1	S6 kinase		83840	
2	U78977	Atp9a	ATPase, Class II, type 9A		84011	
3	D84346	Nckap1	NCK-associated protein 1		58823	
4	U79661	LOC60591	calcium-activated potassium channel beta subunit		60591	
.. of 5	U57391	Sh2bpsm1	SH2-B PH domain containing signaling mediator 1		89817	

Figure 6. ScatterPlot viewer. The function plots gene expression profiles of one sample versus another sample (*A*). The information on any gene circled in the plot (*A*) will be displayed in the GeneLib (*B*).

TOOL

The TOOL was designed to provide a spectrum of algorithmic tools for microarray data visualization, quality control, normalization, significant gene identification, pattern discovery, and class prediction.

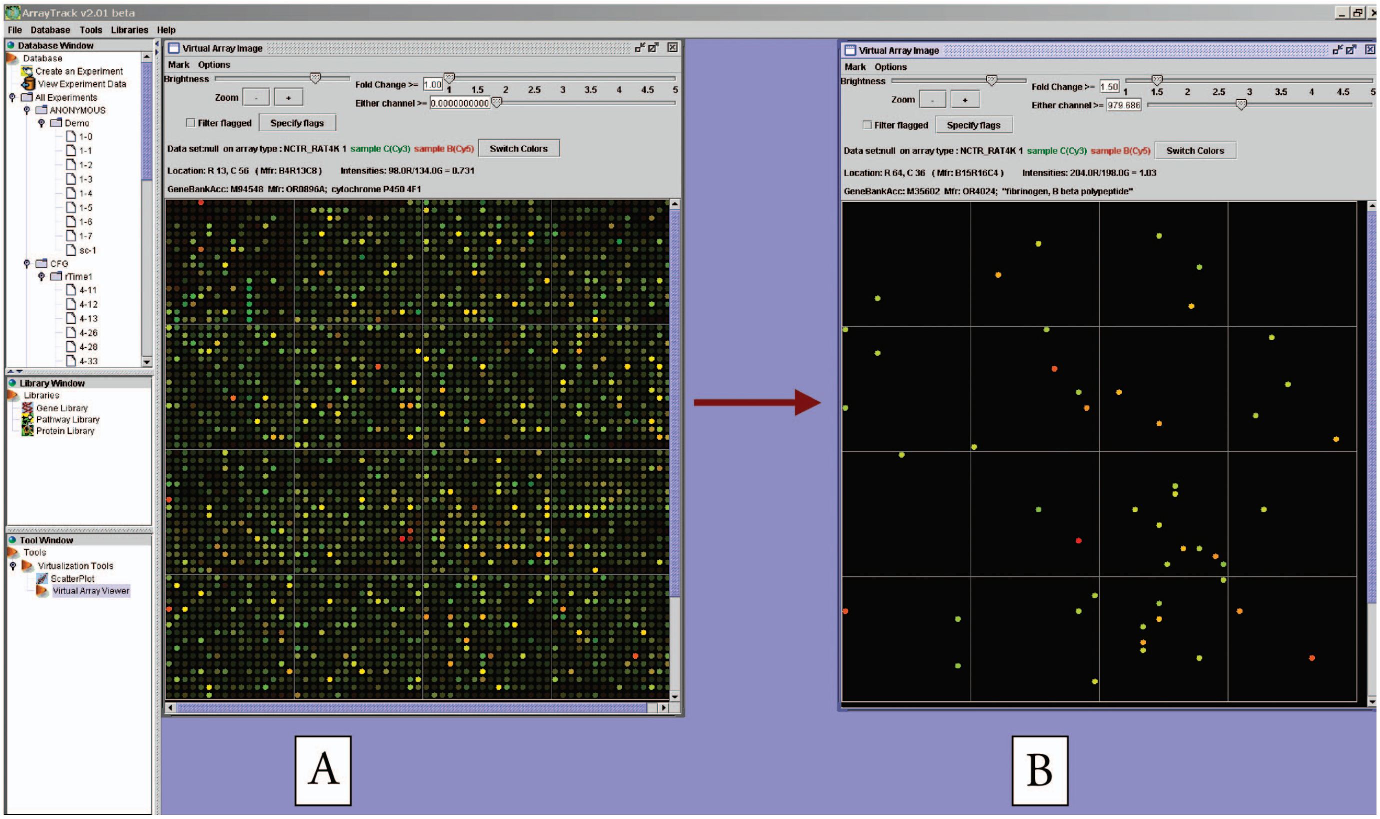
A quality assurance/quality control tool was developed to assist quality control of slide array results (Figure 5). The tool summarizes most relevant information into one interface to facilitate the process of quality control. The user can determine the quality of individual microarray results through visualizing data, applying statistical measures, and viewing experimental annotation. Statistical measures are provided to assess the quality of a hybridization result based on the raw expression data, including signal-to-noise ratio, the percentage of nonhybridized spots, etc. The experimental annotations associated with the processes of hybridization, RNA extraction, and labeling are also available to the end user. Additionally, a scatterplot of Cy3 versus Cy5, together with the original image, is available for visual inspection for quality control purposes.

Two data visualization methods are currently provided—ScatterPlot Viewer and VirtualImage viewer. The ScatterPlot viewer plots gene expression profiles of one sample versus another sample (Figure 6), whereas the VirtualImage viewer displays expression pattern in an array image format (Figure 7). Both functions permit visual identification of significant genes and hyperlink directly from the graph to additional detailed library information on any particular gene.

Discussion

The GeneLib, ProteinLib, and PathwayLib components of ArrayTrack contain general but essential information for functional genomics research. These libraries also provide a basis for linking and integrating various omics data. For example, lists of genes, proteins, and metabolites derived from various omics platforms could be cross-linked based on their common identifiers through these three libraries. An additional library, ToxicantLib, is being developed for ArrayTrack and will similarly provide linkage between toxicological data and the different types of omics data. The ToxicantLib contains the chemical name and structure together with toxicological end points. Through the similarity comparison of the chemical structure of a toxicant with the structures of the metabolites in the PathwayLib, we might be able to examine the toxicity effect of a particular toxicant at the molecular level. The first toxicological data in ToxicantLib are data from our endocrine disruptor knowledge base (EDKB; http://edkb.fda.gov/; Tong et al. 2002) and the carcinogenicity potency database (CPDB) (Gold and Zeiger 1997). Other specific toxicology libraries will be added in the near future, including LiverLib (gene/protein associated with liver toxicity) and SNPsLib (containing information on single-nucleotide polymorphism).

Development of commercial software for visualizing and analyzing microarray data is currently an area of vigorous effort by bioinformatics-oriented companies. Representative software providers for microarray data analysis include Spotfire, Silicon Genetics, BioDiscovery , and Partek. Similarly, a diversity of software is available in the public domain, some of which can be accessed through a website at Stanford University (http://genome-www.stanford.edu/). Collectively, commercial and public software



A: Fold Change >= 1.00; Either channel >= 0.0000000000; □ Filter flagged

Data set:null on array type : NCTR_RAT4K 1 sample C(Cy3) sample B(Cy5)

Location: R 13, C 56 (Mfr: B4R13C8) Intensities: 98.0R/134.0G = 0.731

GeneBankAcc: M94548 Mfr: OR0896A; cytochrome P450 4F1

B: Fold Change >= 1.50; Either channel >= 979.686; □ Filter flagged

Data set:null on array type : NCTR_RAT4K 1 sample C(Cy3) sample B(Cy5)

Location: R 64, C 36 (Mfr: B15R16C4) Intensities: 204.0R/198.0G = 1.03

GeneBankAcc: M35602 Mfr: OR4024; "fibrinogen, B beta polypeptide"

Figure 7. VirtualArray viewer. The function shows a reconstruction of the original array image from the expression data derived from the array (*A*). This virtual array image provides a visual representation of data in the format of the original image. There are several functions on the top of the image that allow browsing the contents of the array, identifying significant spots and information about their corresponding genes. For example, there are two sliding controls for filtering out unwanted spots. The upper sliding control is used to eliminate spots whose expression fold change is less than the predefined criteria. The other sliding control is used to eliminate spots for which the intensity of both Cy3 and Cy5 channels falls below the selected threshold. The resulting image (*B*) contains only those genes that meet both ratio and intensity criteria. Those genes can be directly linked to the GeneLib.

provide many redundant capabilities, though particular software may have unique features or other attributes or familiarity that appeal to end users. Consequently, we have developed and will develop more interfaces (as part of TOOL) to provide interoperability between ArrayTrack and other analysis software. ArrayTrack includes some tools common to other bioinformatics software, but future development will focus on novel analysis approaches and tools for toxicology-specific problems. For example, we developed a novel class prediction method, heterogeneous decision forest (Tong et al. 2003), that could be useful for omics data analysis generally and for development of predictive models in particular.

ArrayTrack has been developed and programmed in a modular manner and uses a Java library such that the code is readily extensible for other omics data, such as proteomics and metabonomics, as well as for conventional toxicology data. The extended system, supporting the diversity of data types, is the TIS, which will be under further development and evolution for several more years. The ultimate goal is for TIS to serve as a general, broad repository for diverse data sources (e.g., omics, toxicology, and chemical structure data), supporting broad data mining and meta-analysis activities as well as development of robust and validated predictive systems. The TIS will facilitate scientific discovery and productivity via effective management of diverse data and knowledge and by integration of toxicological information at different levels of biological complexity. Through cross-linking gene, protein, and pathway information available in public databases, and experimental data from multiple experiments, protocols, and labs, systems toxicology will allow a fuller understanding of toxicological mechanisms.

References

Anonymous. 2002. Microarray standards at last [Letter]. Nature 419:323.

Aardema MJ, MacGregor JT. 2002. Toxicology and genetic toxicology in the new era of "toxicogenomics": impact of "-omics" technologies. Mutat Res 499:13–25.

Afshari CA, Nuwaysir EF, Barrett JC. 1999. Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation. Cancer Res 59:4759–4760.

Anonymous. 2002. Microarray standards at last [Letter]. Nature 419:323.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25:25–29.

Ball CA, Sherlock G, Parkinson H, Rocca-Sera P, Brooksbank C, Causton HC, et al. 2002. An open letter to the scientific journals [Letter]. Science 298(5593):539; Bioinformatics 18(11):1409; Lancet 360:1019.

Baxevanis AD. 2003. The molecular biology database collection: 2003 update. Nucleic Acids Res 31:1–12.

Bellenson JL. 1999. Expression data and the bioinformatics challenges. In: DNA Microarrays: A Practical Approach (Schena M, ed). Oxford, UK:Oxford University Press.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2003. GenBank. Nucleic Acids Res 31:23–27.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. 2003. The SWISS-PROT protein knowledgebase and its supplement TREMBL in 2003. Nucleic Acids Res 31:365–370.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29:365–371.

Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, et al. 2003. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 31:68–71.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207–210.

Gardiner-Garden M, Littlejohn TG. 2001. A comparison of microarray databases. Brief Bioinform 2:143–158.

Gold LS, Zeiger E. 1997. Handbook of Carcinogenic Potency and Genotoxicity Databases. Boca Raton, FL:CRC Press.

Hamadeh HK, Amin RP, Paules RS, Afshari CA. 2002. An overview of toxicogenomics. Curr Issues Mol Biol 4:45-56.

Kanehisa M. 2002. The KEGG database. Novartis Found Symp 247:91–101.

Marshall A, Hodgson J. 1998. DNA chips: an array of possibilities. Nat Biotechnol 16:27–31.

Medlin J. 2002. Toxicogenomics research consortium sails into uncharted waters. Environ Health Perspect 110:A744–A746.

Nucleic Acids Research. 2003. 2003 database issue. Nucleic Acids Res 31(1):1–516.

Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. 1999. Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog 24:153–159.

Ramsay G. 1998. DNA chips: state-of-the art. Nat Biotechnol 16:40–44.

Schmidt CW. 2002. Toxicogenomics: an emerging discipline. Environ Health Perspect 110:A750–A755.

Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, et al. 2002. A. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 3(9):R46.

Spengler SJ. 2000. Techview: computers and biology. Bioinformatics in the information age. Science 287:1221, 1223.

Tennant RW. 2002. The national center for toxicogenomics: using new technologies to inform mechanistic toxicology. Environ Health Perspect 110:A8–A10.

Tong W, Hong H, Fang H, Xie Q, Perkins R. 2003. Decision forest: combining the predictions of multiple independent decision tree model. J Chem Info Comput Sci 43:525–531.

Tong W, Perkins R, Fang H, Hong H, Xie Q, Branham SW, et al. 2002. Development of quantitative structure-activity relationships (QSARs) and their use for priority setting in the testing strategy of endocrine disruptors. Regul Res Perspect 1:1–16.

Ulrich R, Friend SH. 2002. Toxicogenomics and drug discovery: will new technologies help us produce better drugs? Nat Rev Drug Discov 1:84–88.

Waters M, Boorman G, Bushel P, Cunningham M, Irwin R, Merrick A, et al. 2003. Systems toxicology and the chemical effects in biological systems (CEBS) knowledge base. Environ Health Perspect 111:811–824.