Regulatory Research Perspectives

Impact on Public Health

Volume 8, Issue 1, April 2009

DHHS/FDA/Jefferson Labs, National Center for Toxicological Research, Jefferson, Arkansas 72079

FDA Bioinformatics Tool for Public Use—ArrayTrack™

Stephen C. Harris[1], Hong Fang[2], Zhenjiang Su[1], Minjun Chen[1], Feng Qian[2], Leming Shi[1], Dhivya Arasappan[2], Weigong Ge[1], Xiaohui Fan[1,3], Huixiao Hong[1], Joshua Xu[2], Steven Turner[2], Michelle E. Bishop[1], Martin Jackson[2], Roger Perkins[1], Weida Tong[1*]

[1]Division of Systems Toxicology, FDA's National Center for Toxicological Research, 3900 NCTR Road, Jefferson, AR 72079

[2]Z-Tech Corporation, an ICF International Company at FDA/NCTR, 3900 NCTR Road, Jefferson, AR 72079

[3]Pharmaceutical Informatics Institute, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310027, China

* Corresponding author: Weida Tong (Weida.Tong@fda.hhs.gov)

Disclaimer: The views presented in this article do not necessarily reflect those of the U.S. Food and Drug Administration.

Abstract: Data from pharmaco-, toxico-, and nutrigenomic technologies are crucial for advancing medical-product development and personalizing nutrition and medicine. An integrated bioinformatics infrastructure to facilitate the data review is crucial to realize the benefits of genomics to the public health. An array of FDA efforts on genomics powered by integrated bioinformatics has been taking place within five FDA product centers (CBER, CFSAN, CDER, CDRH, CVM), and among multiple stakeholders in the public domain through collaborations. To facilitate this process, an integrated FDA-bioinformatics tool, called ArrayTrack™, developed at FDA's National Center for Toxicological Research (FDA/NCTR), is being refined as a review tool for managing, analyzing, and interpreting this exploratory data (i.e., genomic, proteomic, and metabolomic data) from both clinical and nonclinical data submissions. ArrayTrack™ stores a full range of information related to DNA microarrays and clinical and nonclinical studies, as well as the digested data derived from proteomics and metabonomics experiments. In addition, ArrayTrack™ provides a rich collection of functional information about genes, proteins, and pathways drawn from various public biological databases for facilitating data interpretation. Many data analysis and visualization tools are available with ArrayTrack™ for individual platform-data analysis, multiple omics-data integration, and integrated analysis of omics data with study data. Importantly, gene-expression data, functional information, and analysis methods are fully integrated so that the data analysis and interpretation process is simplified and enhanced. Using ArrayTrack™, users can select an analysis method from the ArrayTrack™ tool box, apply the method to selected microarray data, and the analysis results can be directly linked to individual gene, pathway, and Gene Ontology analysis. ArrayTrack™ is publicly available online (https://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/index.htm), and the prospective user can also request a local installation version by contacting the authors.

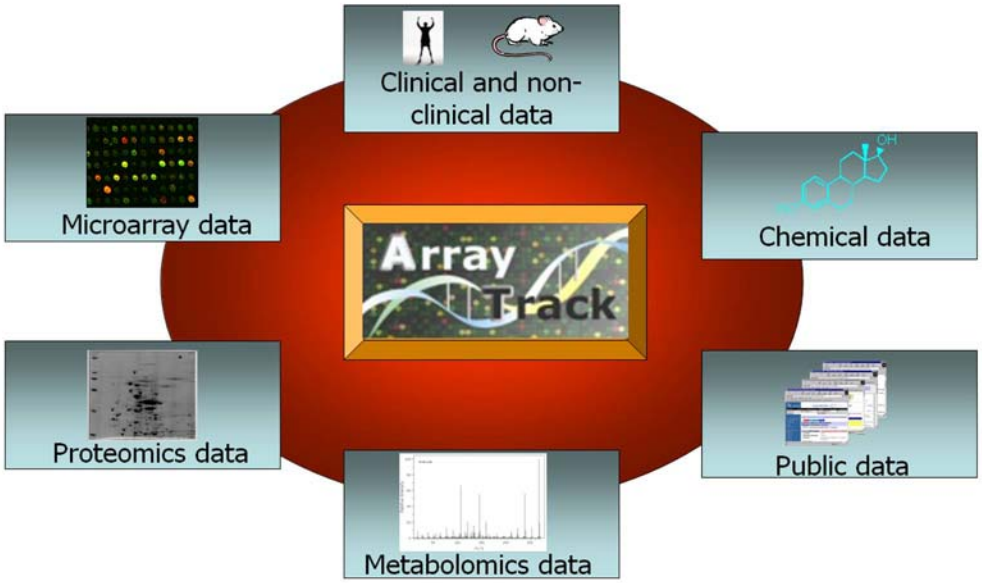


Figure 1. ArrayTrack™ functionality. ArrayTrack™ stores information involved in each step of the microarray experiment, including the raw and normalized gene-expression data, experiment protocol, gene list, and etc. The final results from proteomics and metabonomics experiments, such as protein lists and metabolite lists, can also be managed by ArrayTrack™. In addition, ArrayTrack™ can store both clinical and nonclinical study data in accordance with the CDISC/SEND standard. A chemoinformatics component is available in ArrayTrack™ for chemical structure repository, wherein several structure-searching engines permit identification of chemicals with a predefined similarity. ArrayTrack™ also generates a repository for the public annotation data, such as GenBank, Gene Ontology, KEGG, and etc.

ArrayTrack™–A brief overview

Genomics, proteomics, and metabonomics (collectively called omics), along with other emerging methodologies, e.g., high-density genotyping for Genome Wide Association Studies (GWAS), contribute to our understanding of disease and health. The broad application of omics technologies in drug discovery and development poses a challenge to both sponsors and regulatory agencies. The National Center for Toxicological Research (NCTR) of the U.S. Food and Drug Administration (FDA) has developed an integrated, bioinformatics system meeting the challenge related to these advanced high-throughput and/or high-content genomic assays, with emphasis on DNA microarrays [1]. ArrayTrack™ was originally conceived and developed to provide a one-stop bioinformatics solution for DNA microarray experiments, a capability now extended to integrated analysis of multiple omics expression profiles, such as proteomics and metabonomics (Figure 1).

Over seven years in development as of this writing, ArrayTrack™ has had an increasing and demonstrable impact on FDA programs, of which the Voluntary Genomics Data Submission (VGDS) program [2] and the MicroArray Quality Control (MAQC) project [3] are notable examples. The program roles and demands have, in turn, led to identification and implementation of new capabilities and functionalities.

VGDS is a novel-data submission mechanism within FDA. Through VGDS, the regulated sponsor can interact with FDA by submitting genomic data on a volun-

(Continued on page 3)

(Continued from page 2)
tary basis. ArrayTrack™ became the FDA genomic tool to support VGDS in early 2004. All VGDS DNA microarray data received since 2004 has exclusively been from Affymetrix Gene-Chip® technology. Accordingly, significant ArrayTrack™ development has been oriented to improve Gene-Chip® data handling and analysis. New functionality includes (1) direct loading of CEL files into ArrayTrack™; (2) choice of converting probe-level data to any or all of the probe-set level data types, including MAS 5, RMA, DChip, and PLIER; (3) data filtering based on the presence/absence call; (4) mapping the Affymetrix ID to other types of gene IDs (e.g., Entrez Gene ID), protein IDs (e.g., Swiss-Prot Accession number), different array platform IDs (e.g., Agilent ID); and (5) providing annotation information (e.g., pathways, functions) for all Affymetrix chips.

A primary goal in VGDS is to better understand how the regulated sponsors reach biological conclusions from genomics data, a process requiring reproducing the sponsors' analysis methods. Reanalysis, together with reviewing PGx/TGx studies in literature, enabled delineation of many issues regarding GWAS and expression data, including (1) Array quality–what degree of experiment quality and individual array-platform technical performance should be deemed achievable and adequate? (2) Data analysis issues–what results can be anticipated from different algorithms and approaches, and its corollary: can consensus be reached for a baseline approach to microarray data analysis? and (3) Cross-platform issues–what consistency can be expected among different microarray experimental platforms?

Addressing the above issues were major motivators for initiating the MAQC program in 2005 [3]. MAQC is FDA-led but has a large collaborative community spanning public, private, and academic communities. MAQC Phase I used six different commercial and one institutionally developed microarray platforms, a scope requiring significant expansion of ArrayTrack™ functionalities to manage data. As a result, a generalized data-management scheme was implemented that can handle data from most, if not all, commercial-array platforms. Since most commercial-array types are preloaded in ArrayTrack™ (available from ChipLib in ArrayTrack™), a cross-chip comparison can be carried out to assess commonality and differences between chips provided by the same company (e.g., Affymetrix), as well as between chips provided by different companies (e.g., Affymetrix versus Agilent).

As depicted in Figure 2, VGDS and MAQC emphasize interaction and collaboration with private industry and the research community with the stated objective of moving toward consensus on best practices for microarray data management, analysis, and interpretation. The programs are similarly geared toward advancing the science and consensus. The lessons learned from both VGDS and MAQC are paving the way for development of a Best-Practice Guidance Document for future voluntary, as well as regular submissions of PGx data to FDA. Recently, such a best-practice document draft, a companion document to "Guidance for Industry—Pharmacogenomic Data Submission" was released for comments [4]. ArrayTrack™ both supports VGDS and MAQC and benefits from the programs, contributing to an evermore powerful and versatile FDA-integrated bioinformatics infrastructure to support data management, analysis, and interpretation. Synchronizing ArrayTrack™ development with VGDS and MAQC will assure that the platform meets agency needs to routinely employ PGx/TGx data in regulatory review and decision making, when that time arrives.

ArrayTrack™ development initially focused on management, analysis, and interpretation for DNA microarray data. By the end of 2006, however, the VGDS program had seen proteomics

(Continued on page 4)

Figure 2. A schematic presentation about the integrated nature of an array of pharmacogenomic effort at FDA. (1) the FDA genomic software, ArrayTrack™; (2) the FDA Voluntary Genomics Data Submission (VGDS); (3) the MicroArray Quality Control (MAQC) project; and (4) the best practices presented in the draft companion document to "FDA Guidance for Industry: Pharmacogenomic Data Submissions." VGDS and MAQC are program mechanisms allowing FDA interaction in a collaborative environment with the private sector and research community, respectively. Both programs are aimed at gaining consensus on analysis methods for and valid applications of recently advanced molecular technologies in drug development and regulation. The collective lessons learned from both programs formed the basis to develop the companion document. ArrayTrack™ provides primary support to VGDS and MAQC, thereby continuing its evolution to be the software vehicle that translates best practices into routine application for regulatory review and decision making in FDA.

(Continued from page 3)
and metabonomics data appearing as voluntary submissions. ArrayTrack™ was subsequently modified to accommodate significant lists of proteins and metabolites, and a new systems biology function, CommonPathway, was added that enables examination of common pathways and functional categories (e.g., Gene Ontology terms) shared by different data types (see section with ArrayTrack™ Use Cases).

A VGDS submission normally comes with a large amount of both clinical and nonclinical information. To manage these traditional data types, a general mechanism for handling study data was implemented in ArrayTrack™ using the Study Data Tabulation Model (SDTM) for nonclinical data and clinical data standards suggested by the Clinical Data Interchange Standard Consortium (CDISC) [5]. Additionally, functions were developed to facilitate interpretation of multiple data types (nonclinical, clinical, and omics) in the context of phenotypic anchoring, which, in turn, enabled identification of possible molecular mechanisms related to phenotype (see section on ArrayTrack™ Use Cases).

ArrayTrack™ has been a key genomic tool for the VGDS program and genomic submission in FDA. By now, over 100 FDA reviewers and scientists have attended ArrayTrack™ training. However, the need to make the tool publicly available to the research community was identi-

(Continued on page 5)

(Continued from page 4)
fied early on and has been a continuing priority throughout the planning and development phases of ArrayTrack™. As with VGDS, feedback from the wide-user community has reciprocally benefited ArrayTrack™ by linking its development to emerging common practices and providing validations of functions and usefulness. ArrayTrack™ was made openly available to the public in 2003, whereby users can gain access either through the FDA website [1] or by requesting media for local installation, which would then normally entail local provision of backend database support with ORACLE.

In addition to its broad use within FDA in various regulatory-driven programs, ArrayTrack™ is also freely available to the entire scientific community. The ArrayTrack™ user base has steadily grown, and the tool has been adopted by several government agencies (e.g., EPA, CDC, and NIH), academic institutions, and private sector companies. At this writing, ArrayTrack™ version 3.4 can be accessed through https://edkb.fda.gov/webstart/arraytrack. (https://weblaunch.nctr.fda.gov/jnlp/arraytrack for FDA users). The full-user manual, quick-start manual, and tutorial are available from the ArrayTrack™ website: https://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack™/.

ArrayTrack™ Core Components

The following criteria were considered at ArrayTrack™'s inception and remain salient during continuing development: (1) A rich collection of gene, protein, and pathway functional information to provide context in data interpretation; (2) A software environment that automatically integrates gene-expression data with functional information and visual and analytic tools for efficient and effective data analysis and interpre-

(Continued on page 6)

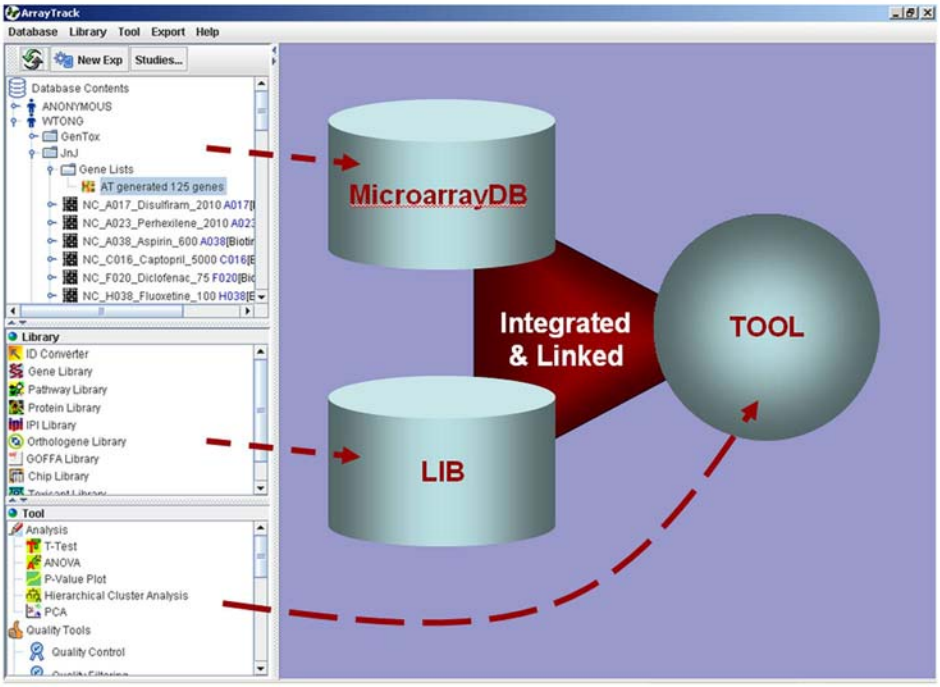


Figure 3. ArrayTrack™ core components. The software consists of three integrated components that are organized as three panels on the left side of the interface: (1) MicroarrayDB captures toxicogenomic data associated with a microarray experiment; (2) TOOL provides data visualization and analysis capabilities; and (3) LIB contains annotated information on genes, proteins, and pathways.

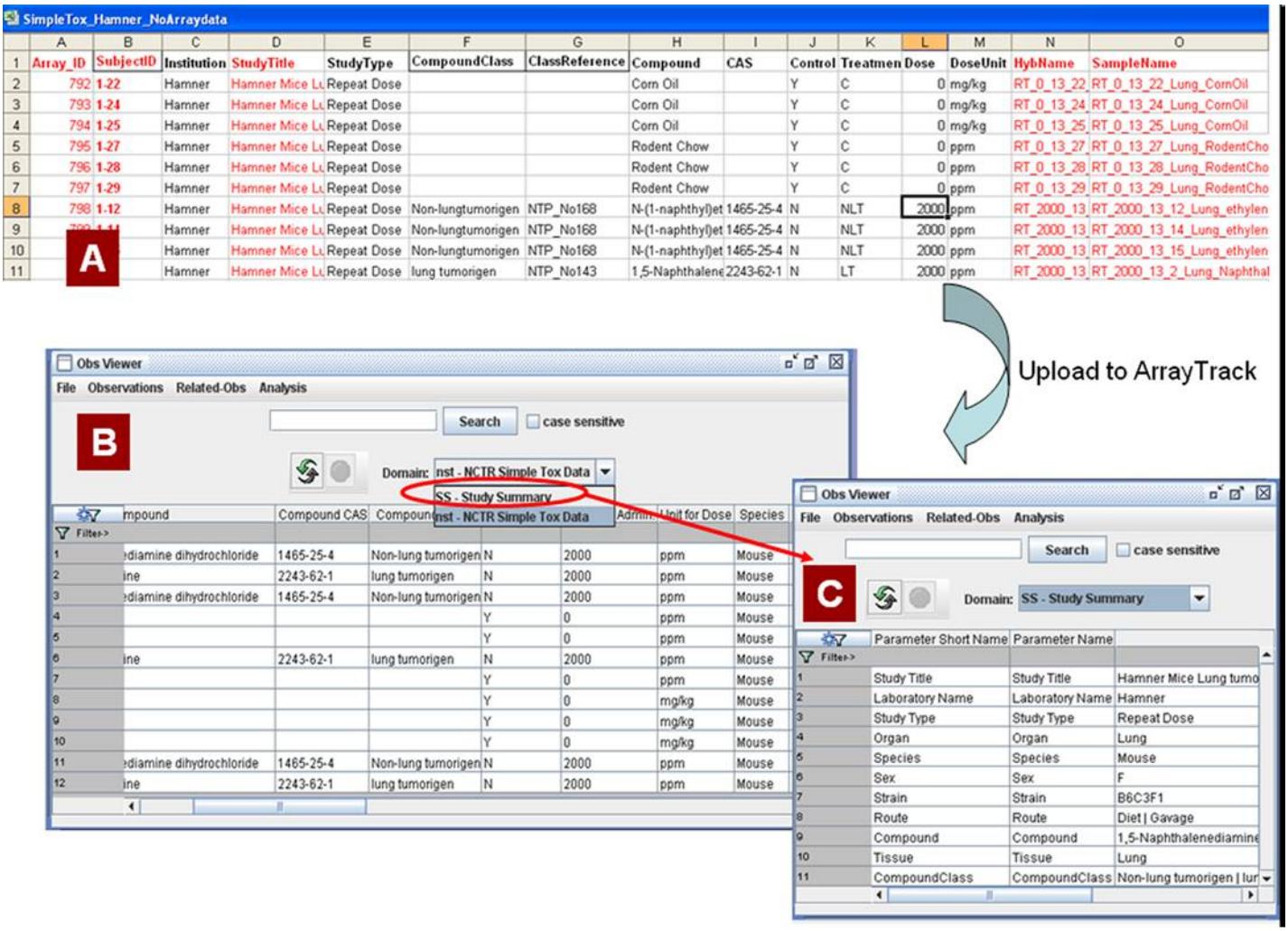


Figure 4. Data uploading using SimpleTox. The SimpleTox is a tabular format, such as an Excel® spreadsheet (A), that can be directly used to input both array and study data into ArrayTrack™. Data can be readily viewed as a spreadsheet (B) or a summary table (C) in ArrayTrack™.

(Continued from page 5)

tation; (3) Ability to cross-link gene expression and conventional toxicological data for phenotypic-driven exploration of underlying mechanisms of toxicity; and (4) modularization for easy extensibility to other types of omics data (e.g., proteomic and metabonomic data) to enable systems toxicology research.

Consequently, ArrayTrack™ comprises three major integrated components (Figure 3): (1) MicroarrayDB that stores essential data associated with a microarray experiment, including raw-gene expression data and information on samples, treatments, and phenotypic observations; (2) TOOL that provides analysis capabilities for data visualization, normalization, significance analysis, clustering, and classification; and (3) LIB that contains information (e.g., gene annotation, protein function, and pathways) from public repositories. 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 the analysis results can be directly linked to associated-functional annotations in the LIB.

(Continued on page 7)

(Continued from page 6)

The key functionalities associated with these three components are discussed below, and the full list of functions is available on the ArrayTrack™ Website [6].

MicroarrayDB

Different from many commercial and public microarray data analysis tools such as Spotfire and Partek®, ArrayTrack™ contains a database structure to store both microarray and study data. This is important because a microarray experiment involves multiple steps, and the data in each step need to be appropriately managed, annotated, and, most importantly, stored in an appropriate data structure for ready access. This enables efficient and reliable access for subsequent data analysis, normally done by a multi-disciplinary group of scientists. The DB structure allows for periodic reexamination of the data in light of continual evolution of gene annotation information in the public domain. Furthermore, reanalysis is likely to be needed as new or more accepted analytic methods evolve, a process much more easily carried out with a well-managed and annotated dataset.

Microarray information, along with study data, can be input through two primary submission formats, batch uploading and SimpleTox format (Figure 4). Both batch uploading and SimpleTox allow a larger number of arrays to be input in batch mode. Input schemas and rationales are as follows: First, we have observed that most biologists tend to organize the data using an Excel® spreadsheet, where rows correspond to array IDs and columns correspond to experiment parameters. Accordingly, both submission formats directly accept such spreadsheet formats (i.e., Excel® or tab delimited). Secondly, to ensure that essential information related to gene expression and study data are being managed in a consistent way for cross-study analysis, the MIAME (Minimum Information About a Microarray Experiment) and SEND (Standard for Exchange of Nonclinical Data) standards are enforced as the column headers for preparing the spreadsheet. The major difference between the batch uploading and SimpleTox is that the latter provides a flexible mechanism that can be used to manage a large variety of data from literature for comparative analysis of multiple studies, which could also ultimately serve as a means for knowledge-base development.

In addition to inputting the raw-gene expression data, a user can also upload any list of genes, proteins, and metabolites into ArrayTrack™. Such lists can be generated outside of ArrayTrack™, such as those calculated using a customized, statistical method or simply assembled from the literature or other knowledge sources. This function is useful in many ways. First, any statistical analysis tool implemented in ArrayTrack™ provides the option to be applied only to a specified-gene list such that, for example, the grouping of the treated samples across different-time points and doses can be examined using a cluster analysis based on a preloaded gene list. Secondly, the preloaded gene list can be directly compared with the gene list generated using the ArrayTrack™ tool for comparative analysis. In VGDS, for example, significant genes chosen by the ArrayTrack™ tool are often compared with the list provided by the sponsor to assess the commonalities and differences in biological interpretation. Thirdly, if the lists of genes, proteins, and metabolites from a multi-omic experiment are input independently into ArrayTrack™, the common pathways and/or functional categories shared by the three lists can be examined (see section on ArrayTrack™ Use Cases).

LIB

Efficient and effective data interpretation is crucial to a microarray experiment, and this demands relevant knowledge for gene annotations, protein

(Continued on page 8)

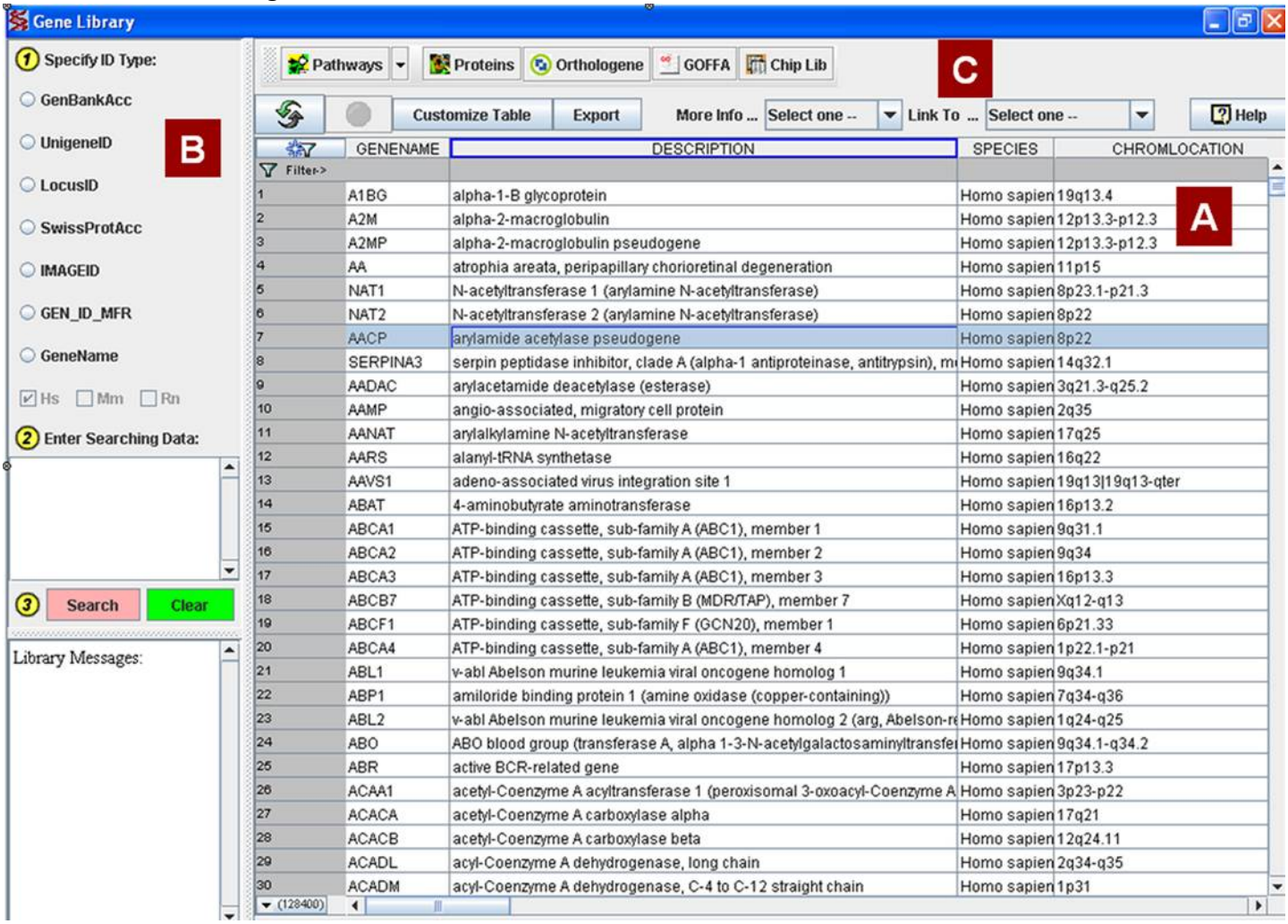


Figure 5. The general layout of ArrayTrack™ LIB. The main part (A) is an Excel®-like spreadsheet, where each row is associated with a gene (or protein, pathway, chemical, SNP) while each column presents a particular functional annotation, such as chromosomal location, pathways, and functional assignments (molecular function, biological process, and cellular component) defined by GO. Users can explore the content in the library using the query mechanism on the left side of the spreadsheet (B). On the top side of the spreadsheet (C), several functions are available to provide detailed information about the selected rows in the spreadsheet, link the findings to the external resources (e.g., GenBank, SWISS-Prot™, etc.), map to other libraries in ArrayTrack™, as well as pathway analysis tools such as KEGG, GeneGo MetaCore™ and IPA, and Gene Ontology.

(Continued from page 7)

functions, and pathways to be readily available and integrated with the data analysis process. The libraries component of ArrayTrack™ makes such biological information readily available in an easy to use format. Each library contains content-specific information that is organized in such a way that it is not only convenient for interpretation of omics data but also useful for other genomic research. Each library has a common look-and-feel. As depicted in Figure 5, the main part of a library is an Excel®-like spreadsheet where each row is associated with an entity of interest, which can be a gene, a protein, a chemical, a pathway, etc., depending on the content of a library. Each column presents particular information for each entity in the row, such as functional annota-

(Continued on page 9)

(Continued from page 8)

tion, chromosomal location, pathways, etc. The query function is on the left side of the spreadsheet, where the user can quickly identify the functional information for a set of significant genes derived from an analysis by searching the library. In addition, a set of functions available on the top of the spreadsheet allows the information in a library to be mapped to other libraries in ArrayTrack™ or to external resources such as GeneGo MetaCore™, IPA (Ingenuity Pathway Analysis), and etc.

ArrayTrack™ contains libraries that partially mirror the contents of GenBank, SWISS-PROT™, LocusLink, Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO), and others. We extract the functional information from these databases to construct several enriched libraries, such as GeneLib, ProteinLib, and PathwayLib, which, as the names suggest, concentrate functional information on genes, proteins, and pathways, respectively [1]. ChipLib contains all the functional information for the probes on a chip provided by the array manufacturers. Since understanding the function and biological characteristics of the probes (genes) present on a mi-

(Continued on page 10)

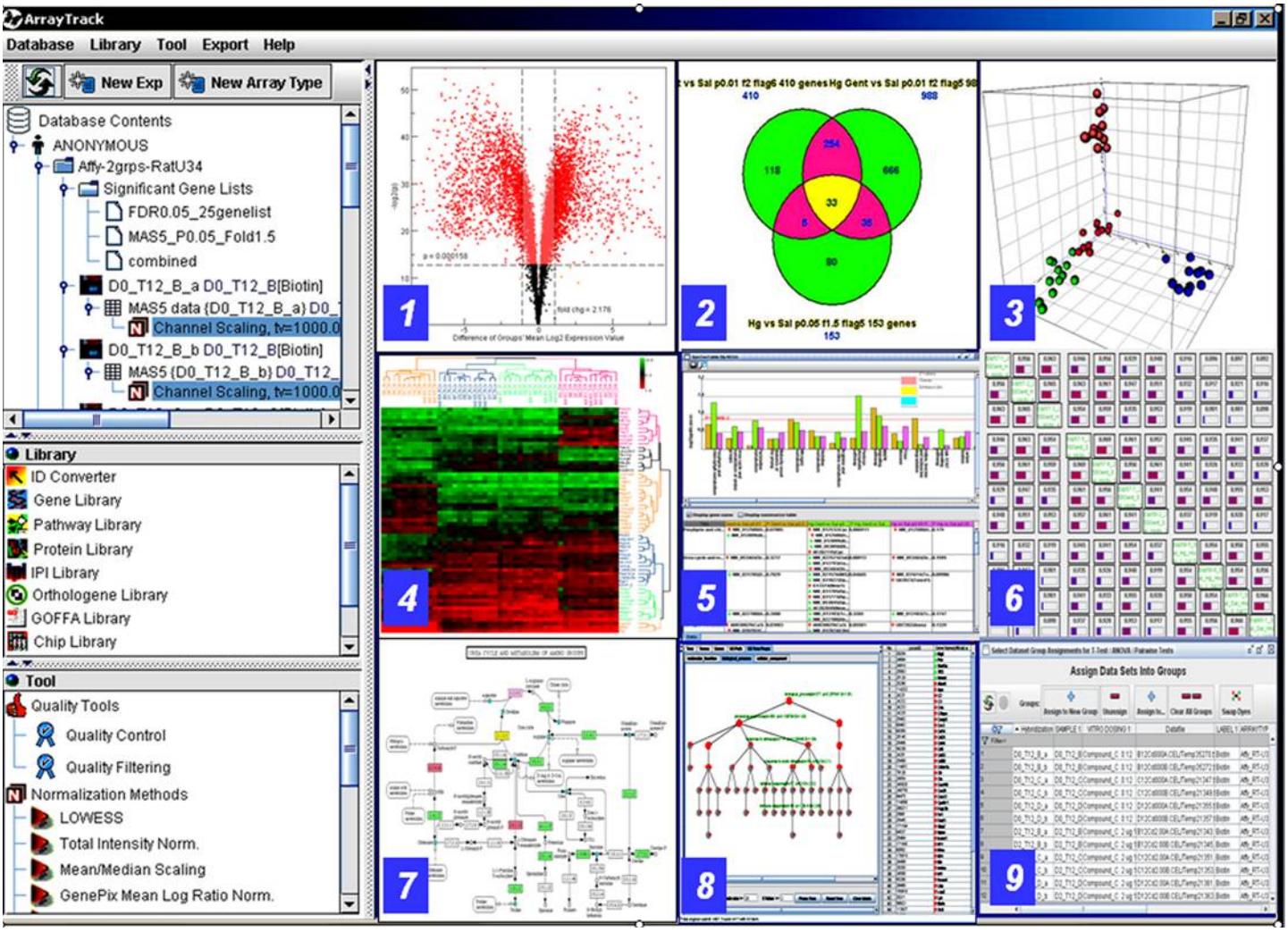


Figure 6. The overview of ArrayTrack™ analysis capability: (1) Volcano Plot; (2) Venn Diagram; (3) Principal Component Analysis; (4) Hierarchical Cluster Analysis; (5) CommonPathway; (6) Correlation Analysis; (7) Pathway Analysis; (8) Gene Ontology tool (GOFFA); and (9) K-Mean clustering.

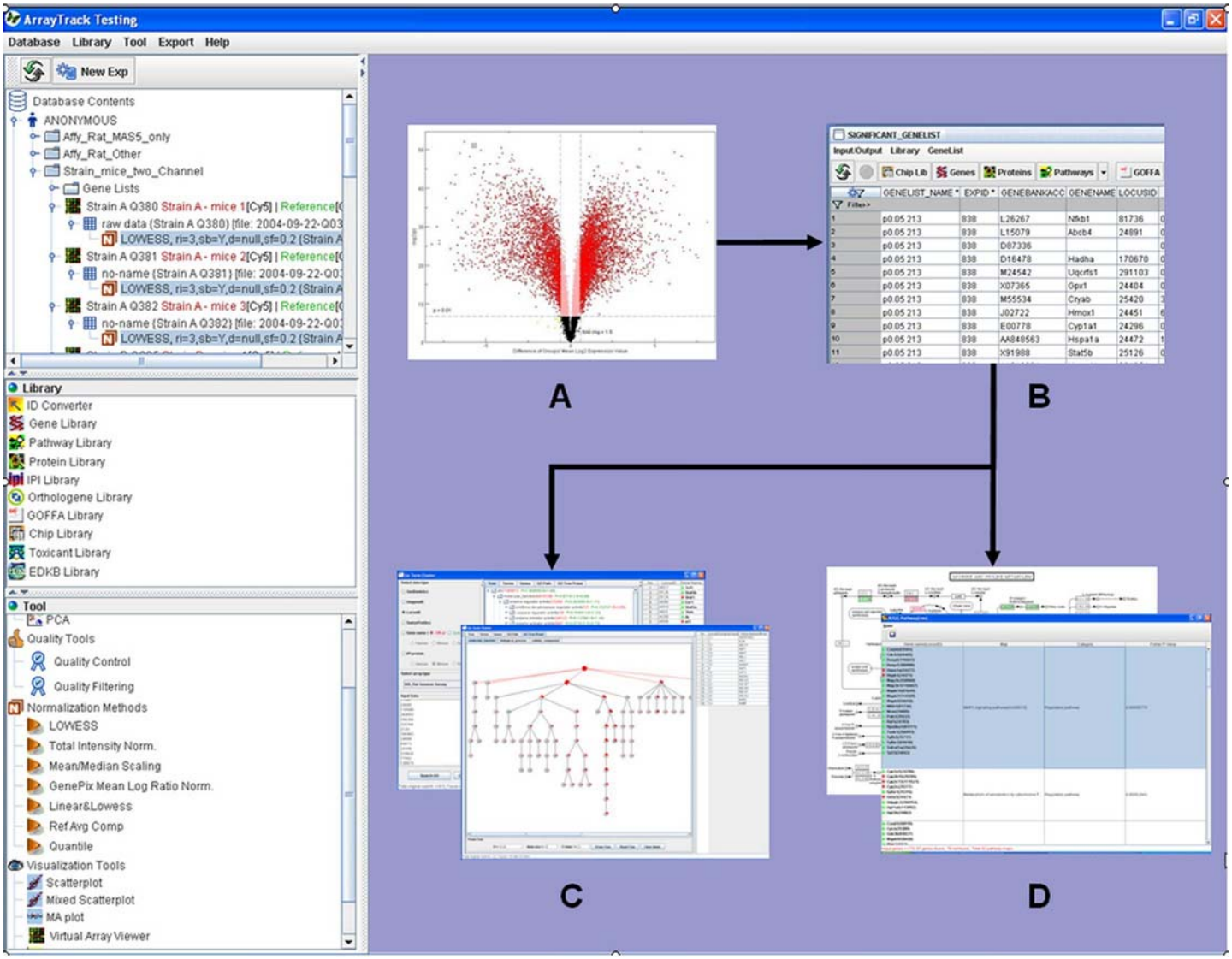


Figure 7. A typical workflow using ArrayTrack™ to identify differentially expressed genes (DEGs) distinguishing treatment and control groups, followed by pathway and Gene Ontology (GO) analyses. (A) DEGs are identified using the Volcano Plot or other means in ArrayTrack™. DEGs can also be identified using other commercial or public tools and uploaded into ArrayTrack™; (B) DEGs are summarized in a table format and can be readily linked to ArrayTrack™ library functions for biological interpretation; (C) Significantly altered KEGG pathways are identified based on DEGs; and (D) DEGs are submitted to Gene Ontology For Functional Analysis (GOFFA) tool in ArrayTrack™ to identify GO terms associated with significantly altered gene expression.

(Continued from page 9)

croarray could be essential for interpretation of microarray results, genes present on the array are also directly linked with other libraries for facilitating biological interpretation of experiment results.

TOOL

A single microarray experiment can produce a large amount of data and a formidable analytical undertaking. Normally the immensity of the data and the analysis effort grows in propor-

(Continued on page 11)

(Continued from page 10)
tion with the complexity of the experiment, directly increasing with the number of technical and biological replicates, and number of time and dose points. The ability to search, filter, and apply mathematical and statistical operations and graphically visualize data quickly with an intuitive-user interface facilitates the laborious process. Microarray-data analysis normally starts with data normalization and quality control, followed by class comparison, class discovery, and class prediction. At this time, ArrayTrack™ provides many analysis capabilities (Figure 6).

ArrayTrack™ provides several normalization methods to convert the probe-level data to probe-set level data for the Affymetrix GeneChip®, including MAS 5, RMA, DChip, and PLIER. The raw-gene expression data from other array platforms can be processed using several global-normalization approaches, such as total intensity normalization [7], log-ratio mean-scale normalization [8], and LOWESS normalization.

One of the most common data-analysis approaches used on DNA-microarray data is determining a list of genes that are differentially expressed by comparing, for example, the treated group with the control group and then using this subset of differentially expressed genes (DEGs) for biological interpretation. Over the years, a number of methods have been proposed to identify DEGs. ArrayTrack™ offers many such methods, such as the simple T-test, ANOVA, the Volcano Plot, and more advanced statistical approaches, such as False Discovery Rate (FDR) and Significance Analysis of Microarrays (SAM)[9].

Two commonly employed tools for class discovery and pattern identification, Principal Component Analysis (PCA) and Cluster Analysis, are available. PCA generates the linear combination of the genes, named principal components, using a mathematical transformation. The algorithm ensures that the first principal component explains the maximal amount of variance of the data. The second principal component explains the maximal-remaining variance in the data subject to being orthogonal to the first principal component, and so on, such that all principal components taken together explain all the variance of the original data. The PCA plot of the first three principal components, which usually explains the majority of variance in the data, is used to inspect the inter-sample and intergene relationships. ArrayTrack™ offers both 2-D and 3-D views of the PCA results, along with the loading tables.

ArrayTrack™ also provides two-cluster analysis methods, a two-way Hierarchical Cluster Analysis (HCA) and K-Means clustering, to investigate the grouping of samples in terms of their similarities in gene-expression profiles, as well as the grouping of genes in terms of their similarity of samples. The primary purpose of two-way HCA analysis is to present data in such a manner that genes with similar expression levels across samples are clustered together, along one axis, while the samples with similar gene-expression patterns are grouped together along another axis. Since the genes in the same cluster are likely to share similar functions, this analysis could reveal the relationships of molecular functions between phenotypes. In contrast, K-Means clustering is mainly used to assess the gene-expression profiles across different-experimental conditions defined in the experiment design.

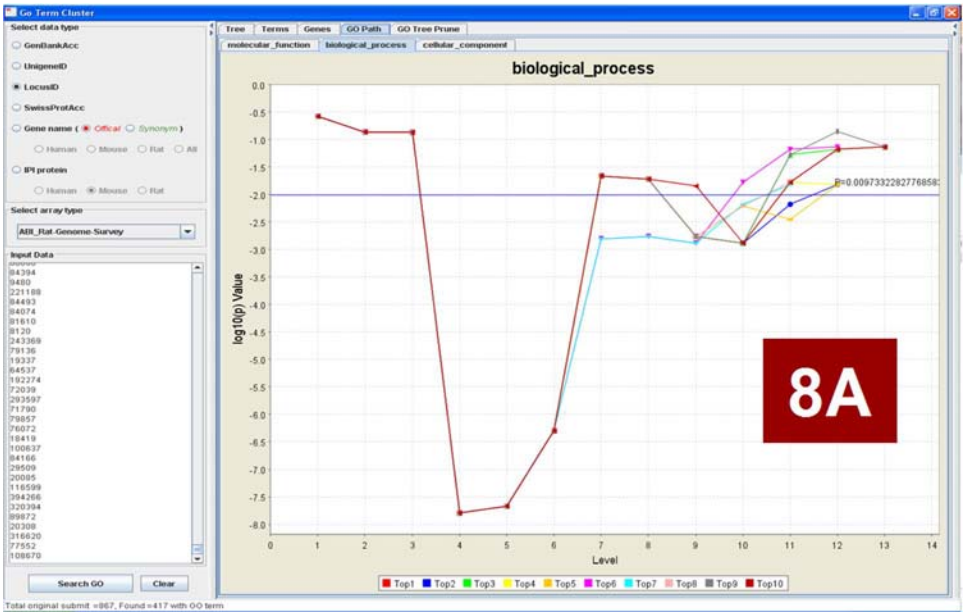
ArrayTrack™ Use Cases

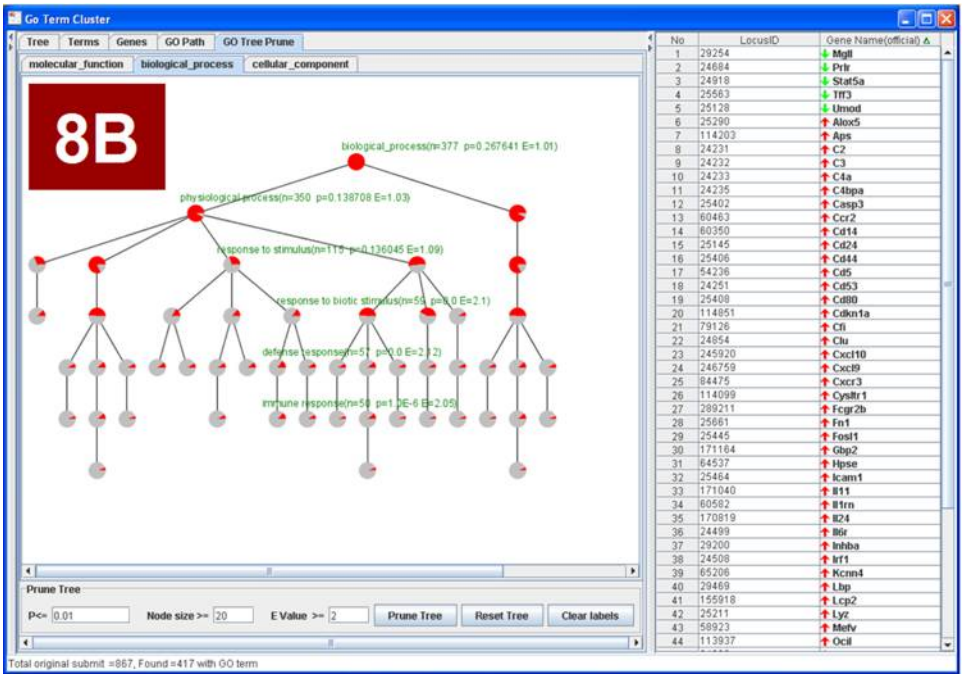
Four examples are provided to illustrate the utility of ArrayTrack™ in addressing the bioinformatics challenges in the FDA VGDS program and research:

Case 1 - A Common Workflow

Drug X was being evaluated for treatment of cancer in a Phase II clinical trial with 100 cancer pa-

(Continued on page 12)





(Continued from page 11)

tients. Before treatment, samples of peripheral-blood mononuclear cells were obtained from individual patients and gene expression in peripheral-blood mononuclear cells measured with Affymetrix microarrays. Treatment benefit was observed for 80 patients but not for the rest. The purpose of this study was to identify a testable hypothesis to explain the treatment outcome. Thus, the analysis required identification of DEGs by comparing patients responsive to treatment with Drug X with those who were not, followed by an interpretation of the biological significance of the comparison.

Figure 7 depicts a prototypical workflow in ArrayTrack™ to carry out the required bioinformatics (i.e., data management,

(Continued on page 13)

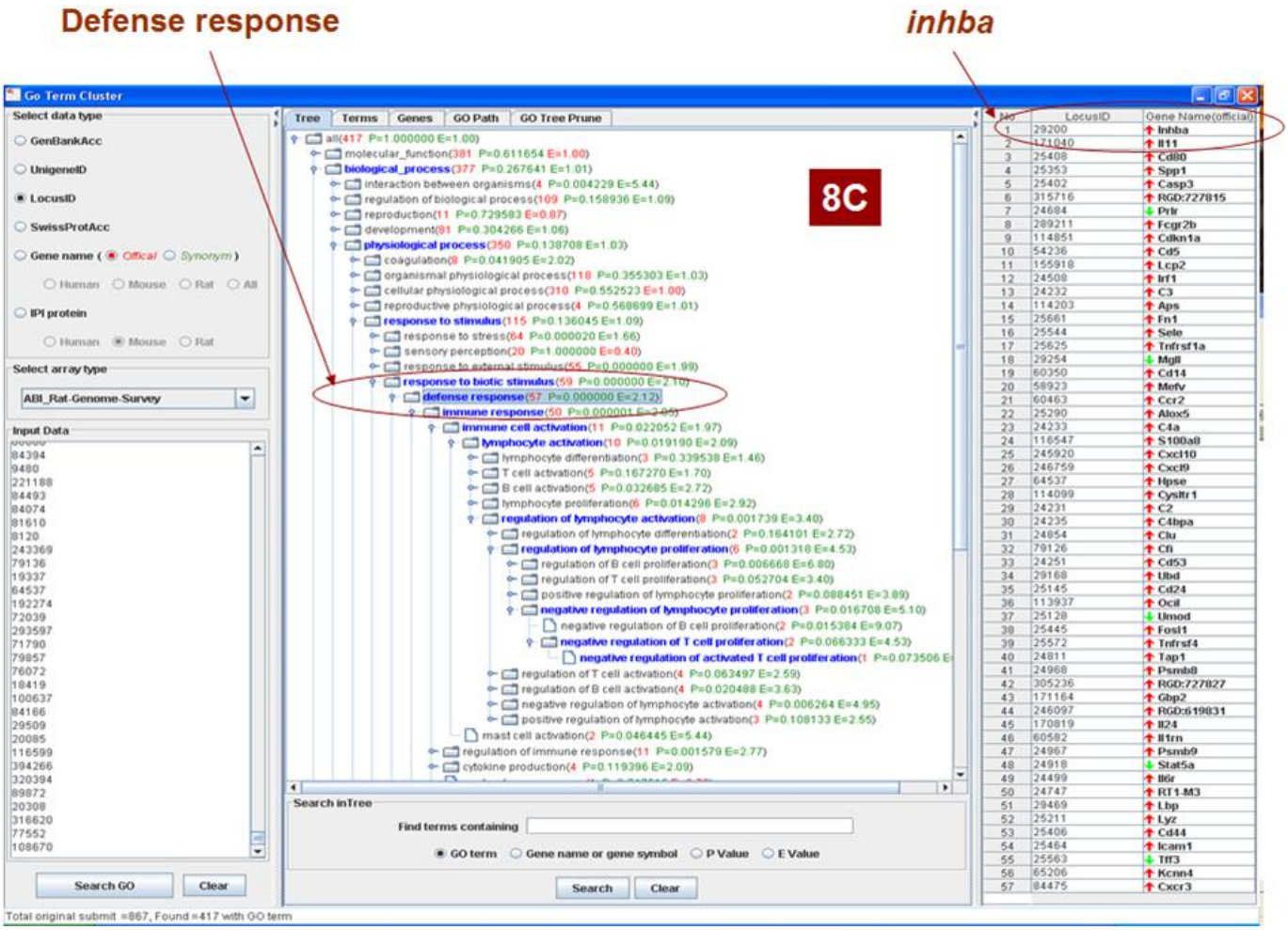


Figure 8. In GOFFA, lists of genes or proteins from an experiment are analyzed by five functional modules: Tree View, Terms View, Genes View, GO Path, and GO TreePrune. (A) GO Path identified the significant GO term based on its path. The most significant 10 paths are graphically displayed, and a color key for the top 10 paths is located beneath the plot. Clicking either a circle in a path in the plot or its corresponding color key launches a Tree View (C) with the selected path highlighted in blue. (B) GO TreePrune display allows the user to filter out nodes and thus reduce the complexity of a tree by specifying the p- and E-value, as well as the user-defined number of genes in the end node.

(Continued from page 12)

analysis, and interpretation), all of which can be done in the single ArrayTrack™-software platform, precluding the need for cumbersome import and export of data between software. ArrayTrack™ was designed *a priori* to provide such a one-stop solution. Using ArrayTrack™, the user can select an analysis method from the TOOL and apply the method to selected omics data stored in DB. The analysis results can then be linked directly to pathways, Gene Ontology database, and other functional information stored in LIB. To further facilitate data interpretation, ArrayTrack™ also provides the capability to directly link the analysis results to external public data repositories, such as OMIM®, UniGene, Chromosomal Map, GeneCard, and etc. Finally, the power and flexibility of ArrayTrack™ is furthered by its interface to, or integration with, many commercial and public software systems, including IPA, GeneGO MetaCore™, PathArt, JMP® Ge-

(Continued on page 14)

(Continued from page 13)
nomics, and the R package.

Case 2 - Gene Ontology analysis using GOFFA

Gene Ontology (GO), which characterizes and categorizes the functions of genes and their products according to biological processes, molecular functions, and cellular components, has played an increasingly important role in interpretation of data from high-throughput genomics and proteomics technologies. A FDA GO tool, Gene Ontology for Functional Analysis (GOFFA), was implemented in ArrayTrack™. With GOFFA, the user can dynamically incorporate ArrayTrack™ analysis functions with the GO data in the context of biological interpretation of gene-expression data.

GOFFA first ranks GO terms in the order of prevalence for a list of selected genes or proteins, and then it allows the user to interactively select GO terms according to their significance and specific biological complexity within the hierarchical structure. GOFFA provides five interactive functions (Tree View, Terms View, Genes View, GO Path, and GO TreePrune) to analyze the GO data. Among the five functions, GO Path and GO TreePrune are unique. The GO Path ranks the GOFFA Tree Paths based on statistical analysis. The GO TreePrune provides a visualization of a reduced GO-term set based on the user's statistical cut-offs. Therefore, GOFFA can provide an intuitive depiction of the most likely relevant biological functions.

A dataset from a toxicogenomics study was used to demonstrate the utility of GOFFA. In this study, the renal toxicity and carcinogenicity associated with the treatment of aristolochic acid (AA) in rats was studied using DNA microarray [10]. The DEG list was determined in ArrayTrack™ and then directly passed to GOFFA for functional analysis. Of 1176 identified genes, 417 genes had GO information for analysis [11]. The GOFFA results are summarized in Figure 8.

The statistics based on a combination of Fisher's Exact Test ($p<0.05$) and Relevant Enrichment Factor ($E>2$) identified 52-enriched GO terms in the GO biological process. The majority of the terms are related to four functional categories: induction of apoptosis, defense response, response to stress, and amino-acid metabolism. These four functional categories reflect the known biological and pharmacological responses of the kidney to the AA treatment [12]. Out of these four functional categories, GO Path ranked "defense response" as an important mechanism associated with the AA treatment (Figure 8A), and similar results were obtained from GO TreePrune as well (Figure 8B). This finding is consistent with the general understanding that defense response, which includes immune response, is a complex, defensive-network response of a tissue to toxins and carcinogens (such as AA). Figure 8C gives the GO Path results in the Tree window, where the majority of genes involved in the defense response are up-regulated to oppose damage by AA. For example, the *inhba* gene (first gene in the right panel) is a growth factor with a 4.1-fold increase in expression in the kidney. This is a tumor-suppressor gene, and it produces a protein that increases arrest in the G1 phase of tumor cells [13]. Therefore, its induction inhibits tumorigenesis in a kidney treated with AA.

Case 3 - Analysis of microarray gene expression data with conventional toxicological endpoints

A number of drugs were recently removed postmarket due to liver toxicity. In fact, hepatotoxicity is recognized as such a significant problem that its study is prevalent in both public- and private-research communities. The VGDS program has observed considerable effort by sponsors to identify relevant preclinical biomarkers for drug-induced liver toxicity.

This example used DNA microarrays to identify a set of genes

(Continued on page 15)

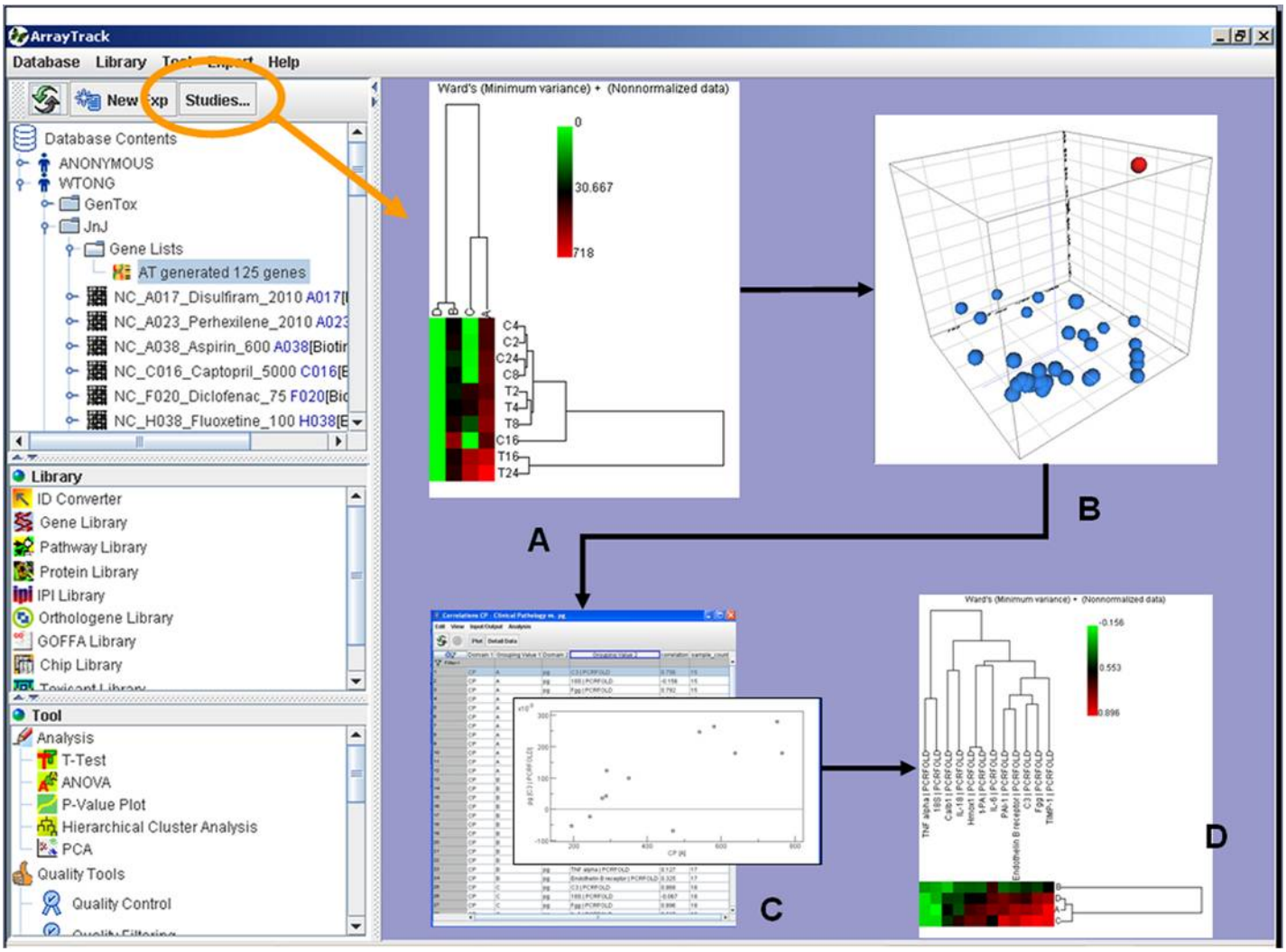


Figure 9. A typical data-analysis procedure and results for Example Study 3 correlating gene expressions at multiple time points with conventional toxicological endpoints. (A) Hierarchical Cluster Analysis is used to assess the ability of clinical pathology to distinguish treatment and control groups. (B) Principal Component Analysis of the clinical pathology data enable an anomalous outlier in the control group to be identified. (C) The DEGs at each time point are correlated with each corresponding set of clinical pathology data. The correlation coefficients are summarized in a table format, and each correlation can also be displayed in a pair-wise plot. (D) The correlation results between the clinical pathology data and gene-expression data are summarized in a heat map, where each cell represents a specific pair (a clinical pathology observation and a gene) in the correlation analysis with magnitude of correlation represented with color (red for the positive correlation and green for the negative correlation).

(Continued from page 14)

with differential expression correlating with clinical pathology parameters associated with, and thus possibly biomarkers for, hepatotoxicity. Specifically, rats were treated with a single high-dose of Drug Y and sacrificed at days 2, 4, 8, 16, and 24. Each time point contained five treated animals along with five matched controls. The liver samples were collected for both treated rats and controls at each time point and analyzed by using Affymetrix microarrays and clinical pathology.

This example required integrating conventional toxicological

(Continued on page 16)

(Continued from page 15)
endpoints with gene-expression data in such a way that phenotype-anchored toxicogenomic analysis could be performed. ArrayTrack™ enables such analyses because a "study domain" is definable based on SDTM developed by CDISC [5]. Using SDTM, ArrayTrack™ is able to concurrently manage disparate clinical and nonclinical data types together with PGx and other biomarker data. Moreover, various statistical analyses at the toxicological level, gene expression level, or in combination, can be conducted.

In this study example (Figure 9), the first step to identify relevant biomarker genes was determining whether the clinical-pathology data contained sufficient biological information to distinguish time points, as well as to distinguish between control and treatment groups. As illustrated in Figure 9A, HCA based on four clinical pathology parameters clearly separated all treatment groups but not the control group sacrificed on day 16. Further analysis using PCA indicated that one of the five control animals had anomalous clinical pathology (Figure 9B) and should be considered for removal before differential expression analysis. Next, the DEGs at each time point were identified, and these genes were correlated with each type of the clinical pathology data (Figure 9C). Genes that showed the highest positive or negative correlations (Figure 9D) with any of the measured clinical-pathology data were identified for further validation as potential biomarkers.

Case 4 - Omics data integration

Integration of gene, protein, and metabolite information for identifying potential biomarkers through perturbed pathways or function is another type of application encountered in the VGDS program. The rationale is that, in the absence of data integration, markers (whether genes, proteins, and metabolites) derived from an individual omics platform are just lists providing but a single level of biological information and subject to Type 1 errors. In contrast, integrating multiple omics data types provides richer elucidation of biological contexts, such as the perturbed functions, signaling pathways, transcription-factor mechanisms of action, gene regulatory networks, and post-translational modifications, among many others. Where differentially expressed genes, proteins, and metabolites implicate the same biological context, there is a qualitative enhancement of both validity and reliability[14].

In this example study, a VGDS submission proposed development of a testable hypothesis for the underlying mechanisms of a disease. The differentially expressed genes, proteins, and metabolites between disease and the disease-free patients were generated from DNA microarrays, proteomics, and metabolomics platforms, respectively. The hypothesis was that pathways common to significant gene, protein, and metabolite lists are more likely to be disease-relevant pathways than pathways identified by a single-significance list.

The CommonPathway function in ArrayTrack™ was used to identify the common pathways or functions shared by a combination of genes/proteins/metabolites differentially expressed between disease and disease-free groups. Figure 10 depicts a typical ArrayTrack™ workflow for required analyses. Once differentially expressed genes, proteins, and metabolites were independently identified from corresponding data, each profile was independently mapped to the pathways to determine which pathways were significantly altered for each data type. The separate pathway lists from the gene, protein, and metabolite profiles were then compared in a Venn Diagram to determine the commonly altered pathways. The statistical significance of each pathway was estimated using Fisher's Exact Test. Details about each significant pathway were also displayed with its differentially expressed genes, proteins,

(Continued on page 17)

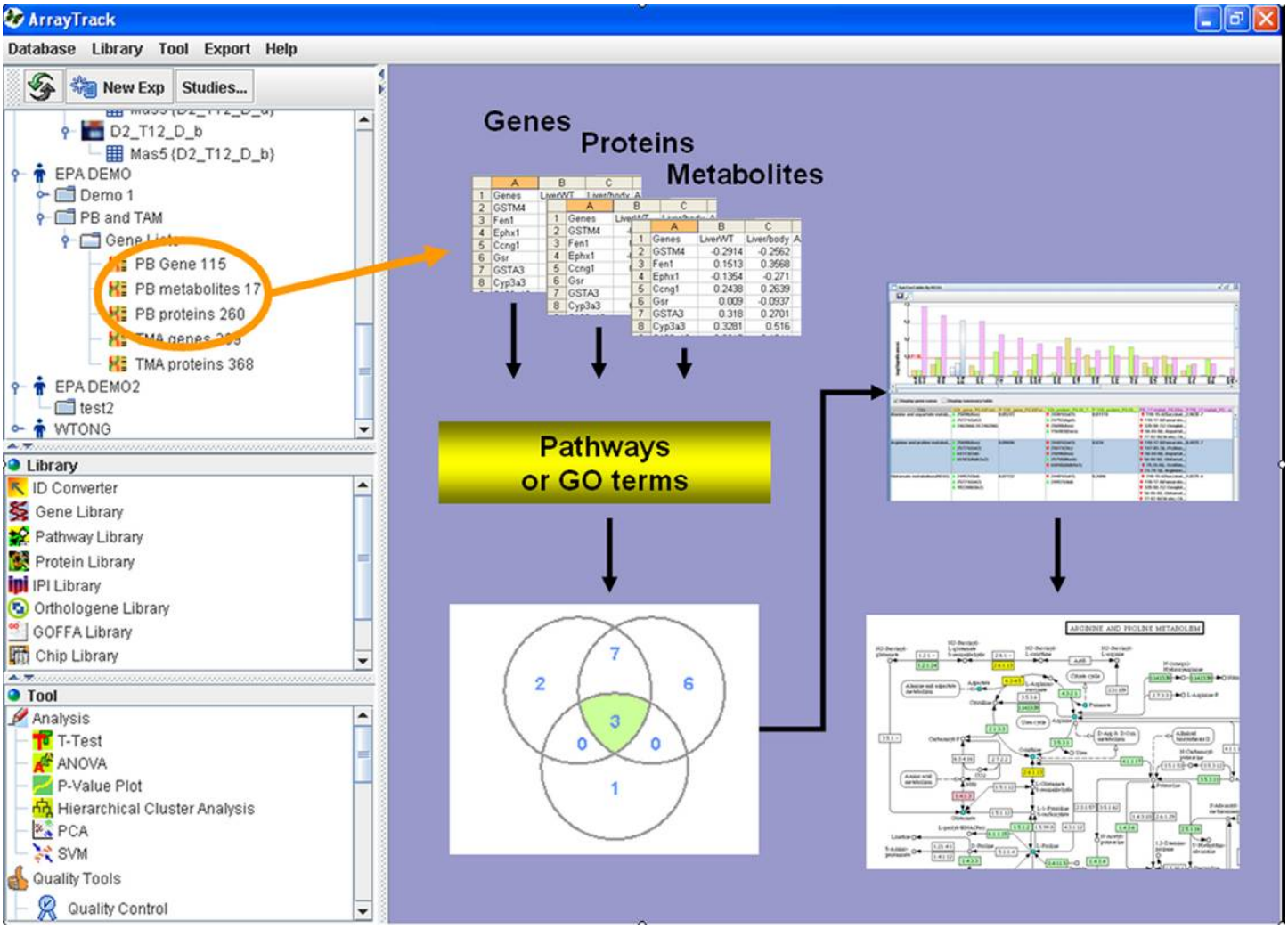


Figure 10. An illustration of the omics data integration workflow in ArrayTrack™. First, differentially expressed genes, proteins, and metabolites are generated or uploaded/stored in ArrayTrack™. Then genes, proteins, and metabolites are each independently mapped to pathways or GO terms, which are considered to also be significantly altered. Altered pathways or GO terms common between data types are next identified using a Venn Diagram. The statistical significance of each common pathways or GO terms is estimated and displayed in a bar chart or spreadsheet. For each common pathway, the detailed pathway map can be viewed with the differentially expressed genes, proteins, and metabolites highlighted in different colors.

(Continued from page 16)
and metabolites highlighted in different colors. The same process can be applied to GO data in order to identify commonly altered GO terms (i.e., gene functions).

Summary

New high-throughput molecular technologies play an increasingly important role in both basic research and in drug discovery and development, and widespread anticipation indicates that this trend will continue.

FDA has gained experience in analyzing new omics data through the VGDS program. The management, analysis, and interpretation of these data constitute a formidable effort for regulatory review. An efficient and integrated bioinformatics

(Continued on page 18)

(Continued from page 17)
infrastructure, within the agency, is therefore essential to review and understand how sponsors reach their biological conclusions, to enable effective interactions with sponsors, and to ensure the incorporation of PGx data into regulatory processes.

ArrayTrack™ continues to undergo constant refinement and enhancement based on the feedback and needs of reviewers. Because ArrayTrack™ has been provided freely to the public, improvements have also been made based on feedback obtained from outside the agency, including academic, pharmaceutical, and other government-agency users. For example, one function recently added to ArrayTrack™ allows for the development of predictive signatures (classifiers) for use in diagnosis, prognosis, and treatment selection relevant to personalized medicine.

ArrayTrack™ has become an integral tool for the analysis and interpretation of genomic and other biomarker data at FDA. The fact that ArrayTrack™ is developed internally, within FDA, has facilitated the integration of enhancements and updates. Several examples illustrate the successful application of ArrayTrack™ in the review of voluntary and nonvoluntary data submissions. With this, ArrayTrack™ and the notion of an integrated, flexible, and robust bioinformatics infrastructure has become a cornerstone of FDA's Critical Path Initiative that is aimed at helping to move medicine from a population-based to a more individually based practice.

References

1. Tong, W., Cao, X., Harris, S., Sun, H., Fang, H., Fuscoe, J., Harris, A., Hong, H., Xie, Q., Perkins, R. *et al 2003.* ArrayTrack–supporting toxicogenomic research at the U.S. Food and Drug Administration, National Center for Toxicological Research. *Environ Health Perspect* 111 (15):1819-1826.
2. Frueh, F. W. 2006. Impact of microarray data quality on genomic data submissions to the FDA. *Nat Biotechnol* 24(9):1105-1107.
3. Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y. *et al* 2006. The MicroArray Quality Control (MAQC) Project shows inter– and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 24 (9):1151-1161.
4. VGDS H. August 2007. Guidance for Industry: Pharmacogenomic data submissions - Companion Guidance: Department of Health and Human Services (HHS), Food and Drug Administration (FDA). http://www.fda.gov/cder/guidance/7735dft.pdf.
5. CDISC T. 2007. Clinical Data Interchange Standard Consortium (CDISC): CDISC Inc., 15907 Two Rivers Cove, Austin, Texas 78717. http://www.cdiscorg/indexhtml.
6. Tong, W., Harris, S., Cao, X., Fang, H., Shi, L., Sun, H., Fuscoe, J., Harris, A., Hong, H., Xie, Q., *et al* 2004. Development of public toxicogenomics software for microarray data management and analysis. *Mutat Res* 549 (1-2):241-253.
7. Quackenbush, J. 2002. Microarray data normalization and transformation. *Nat Genet* 32 Suppl: 496-501.
8. Fielden, M. R., Halgren, R.G., Dere, E., Zacharewski, T. R. 2002. GP3:GenePix post-processing program for automated analysis of raw microarray data. *Bioinformatics* 18(5):771-773.
9. Tusher, V. G., Tibshirani, R., Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 98(9):5116-5121.
10. Chen, T., Guo, L., Zhang, L., Shi, L. M., Fang, H., Sun, Y. M., Fuscoe, J. C., Mei, N. 2006. Gene expression profiles distinguish the carcinogenic effects of aristolochic acid in target (kidney) and non-target (liver) tissues in rats. *Bmc Bioinformatics* 7.

(Continued on page 19)

(Continued from page 18)

11. Sun, H., Fang, H., Chen, T., Perkins, R., Tong, W. 2006. GOFFA: Gene Ontology for Functional Analysis–A FDA Gene Ontology Tool for Analysis of Genomic and Proteomic Data. *BMC Bioinformatics* 7 Suppl 2:S23.
12. Arlt, V. M., Ferluga, D., Stiborova, M., Pfohl-Leszkowicz, A., Vukelic, M., Ceovic, S., Schmeiser, H. H., Cosyns, J. P. 2002. Is aristolochic acid a risk factor for Balkan endemic nephropathy-associated urothelial cancer? *Int J Cancer* 101 (5):500-502.
13. Shav,-Tal, Y., Zipori, D. 2002. The role of activin a in regulation of hemopoiesis. *Stem Cells* 20(6):493-500.
14. Fang, H., Perkins, R., Tong, W. 2007. Omics Integrating Systems Using ArrayTrack and Other Bioinformatics Tools. *American Drug Discovery* 2(4):49-52.

Abbreviations:

BLA—Investigational New Drug
CBER—Center for Biologics Evaluation and Research
CDER—Center for Drug Evaluation and Research
CDISC—Clinical Data Interchange Standard Consortium
CDRH—Center for Devices and Radiological Health
CFSAN—Center for Food Safety and Applied Nutrition
CVM—Center for Veterinary Medicine
DEG— Differentially Expressed Gene
FDA—Food and Drug Administration
GOFFA—Gene Ontology For Functional Analysis
GO—Gene Ontology
HCA —Hierarchical Cluster Analysis
IND—Investigational New Drug
IPA—Ingenuity Pathway Analysis
KEGG—Kyoto Encyclopedia of Genes and Genomes
MAQC—MicroArray Quality Control
MIAME—Minimum Information About a Microarray Experiment
NCTR —National Center for Toxicological Research
NDA—New Drug Application
PCA —Principal Component Analysis
PGx—Pharmacogenomics
TGx—Toxicogenomics
SDTM—Study Data Tabulation Model
VGDS—Voluntary Genomic Data Submission
SEND—Standard for Exchange of Nonclinical Data

Glossary:

ArrayTrack™ is an integrated software system for managing, mining, and interpreting microarray, proteomics, and metabonomics data for systems biology. ArrayTrack™ was developed by NCTR's Center for Toxicoinformatics.

Bioinformatics is the application of information technology to the field of molecular biology. Bioinformatics entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data.

The MicroArray Quality Control (MAQC) project involves six FDA Centers, major providers of microarray platforms and RNA samples, EPA, NIST, academic laboratories, and other stakeholders. The MAQC project aims to establish QC metrics and thresholds for objectively assessing the performance achievable by various microarray platforms and evaluating the advantages and disadvantages of various data analysis methods.

Pharmacogenomics is the branch of pharmacology, which deals with the influence of genetic variation on drug response in patients by correlating gene expression or single-nucleotide polymorphisms with a drug's efficacy or toxicity. Pharmacogenomics is the whole-genome application of Pharmacogenetics, which examines the single-gene interactions with drugs.

Voluntary Genomic Data Submissions (VGDSs) are a novel way to share information with the FDA. Currently, most pharmacogenomic data are of an exploratory or research nature, and FDA regulations do not require that these data be submitted to an IND, or that complete

(Continued on page 20)

(Continued from page 19)
reports be submitted to an NDA or BLA. However, voluntary submissions can benefit both the industry and the FDA in a general way by providing a means for sponsors to ensure that regulatory scientists are familiar with and prepared to appropriately evaluate future genomic submissions.

DNA microarray is a multiplex technology used in molecular biology and medicine. It consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features, each containing picomoles of a specific DNA sequence. This can be a short section of a gene or other DNA element used as probes to hybridize a cDNA or cRNA sample (called target) under high-stringency conditions. Probe-target hybridization is usually detected and quantified by fluorescence-based detection of fluorophore-labeled targets to determine relative abundance of nucleic-acid sequences in the target.

Toxicogenomics is a field that deals with the collection, interpretation, and storage of information about gene and protein activity within particular cells or tissues of an organism in response to toxic substances. Toxicogenomics combines toxicology with genomics or other high-throughput molecular-profiling technologies, such as transcriptomics, proteomics, and metabonomics. Toxicogenomics endeavors to elucidate molecular mechanisms evolved in the expression of toxicity and to derive molecular-expression patterns (i.e., molecular biomarkers) that predict toxicity or the genetic susceptibility to it.

Systems Toxicology is the description of all the toxicological interactions within a living system. Like systems biology, systems toxicology attempts to define the behaviors and relationships of all of the components of a biological system on the premise that global-molecular data can be integrated and modeled computationally.

Database is a structured collection of records or data that are stored in a computer system. The structure is achieved by organizing the data according to a database model. The model in most common use today is the relational model. Other models such as the hierarchical model and the network model use a more explicit representation of relationships.

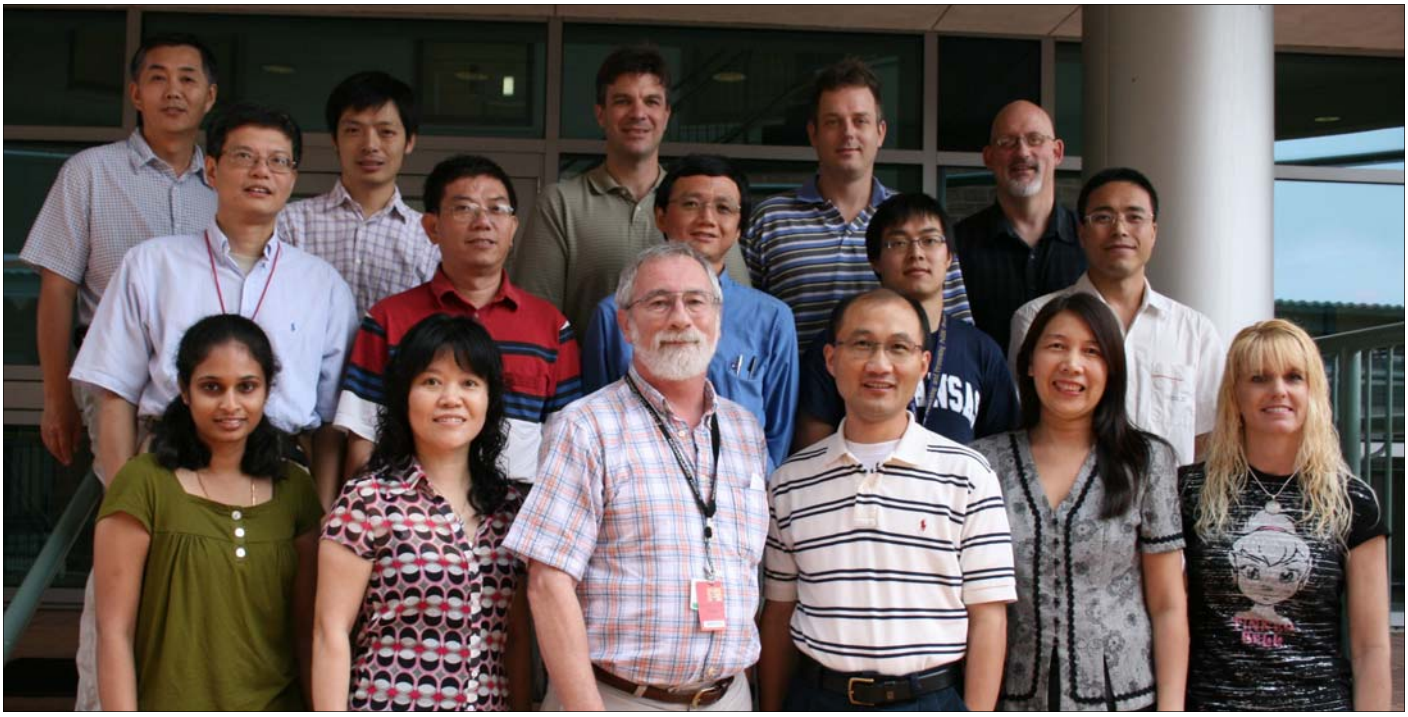
Genomic is the study of an organism's entire genome. The field includes intensive efforts to determine the entire DNA sequence of organisms and fine-scale genetic-mapping efforts. The goal of Genomics is to promote the understanding of the structure, function, and evolution of genomes in all kingdoms of life and the application of genome sciences and technologies to challenging problems in biology, toxicology, and medicine.

The Authors:

Stephen C. Harris received a bachelor's degree in mathematics from Cornell University, in 1992, after which he has gained many years experience of software engineering and systems administration. Currently he is a Principal Software Engineer at the Center for Toxicoinformatics at FDA's NCTR. He has been the principal architect of ArrayTrack™ throughout its design and development. He is also a major contributor to several programs in support of FDA's Critical Paths Initiative, including an electronic-data submission pilot, the Voluntary Genomics Data Submission Program, data standardization using CDISC and SEND for clinical and nonclinical data, respectively, and the associated integration of ArrayTrack™ with the Janus data warehouse.

Hong Fang received her Ph.D. in chemistry from the University of Missouri–St. Louis, in 1995, and worked for two years as a forensic scientist with the St. Louis Police Department. In 1997, she began a two-year postdoctoral assignment at FDA's National Center for Toxicological Research (NCTR) within the Endocrine Disruptor Knowledge Base (EDKB) program. Her EDKB work encompassed design and conduct of assays, as well as using

(Continued on page 21)



Front Row, L–R: Dhivya Arasappan, Feng Qian, Roger Perkins, Joshua Xu, Hong Fang, Michelle E. Bishop
Middle Row, L–R: Weida Tong, Huixiao Hong, Leming Shi, Xiaohui Fan, Zhenjiang Su
Back Row, L–R: Weigong Ge, Minjun Chen, Steven Turner, Stephen C. Harris, Martin Jackson.
Photo/V. B. Taylor

(Continued from page 20)
the data to develop computer-based predictive models for endocrine activity endpoints. In 1999, she joined the NCTR's IT and scientific computing contract as a senior computational scientist, where she currently serves as manager of computational science for Z-Tech, an ICF International Company. She has been particularly involved in the development and support of ArrayTrack™ from its beginning. Her diverse research has involved computational science and integrated bioinformatics approaches applied to chronic-fatigue syndrome, liver cancer, and brain-region differentiation, among others. She has applied various computer-aided drug-design methods to toxicity endpoints involving the endocrine system, cancer, and hepatotoxicity for FDA and Environmental Protection Agency programs. Dr. Fang and her staff collaborate in NCTR research and FDA programs requiring data mining, molecular modeling, computational toxicology, bioinformatics, and associated analytical tools and software development.

Zhenqiang Su is currently a post-doctoral researcher at FDA's National Center for Toxicological Research (NCTR) as one of the main software developers for ArrayTrack™. He received his M.S. in physical chemistry at Chongqing University and his Ph.D. in analytical chemistry at the University of Science and Technology of China. From 1995 to 1999, he worked as a quality control engineer in MCC Shijiu Construction Co., LTD, a company manufacturing metallurgical construction and building materials. From 2002 to 2004, he worked as a software engineer and data analyst in Shenzhen Chipscreen Biosciences LTD, China, where he designed and developed an integrated chemoinformatics and bioinformatics software system to support the company's innovative, chemical genomics-based drug discovery programs. In 2004 and 2005, as a visiting scientist at NCTR, he implemented several data-mining tools, including principal-components analysis and hierarchical clustering in ArrayTrack™. Dr. Su's research has focused on

(Continued on page 22)

(Continued from page 21)
data-mining algorithms, scientific-software development, chemometrics, bioinformatics, and computational chemistry.
Minjun Chen received his Ph.D. in pharmaceutical science from Zhejiang University, Hangzhou, China, in 2003. He worked as an assistant and then associate professor in the School of Pharmacy, Shanghai Jiao Tong University, China, from 2003 to 2006, where he was involved in the studies of mass-based metabonomics and traditional Chinese-herbal medicine's toxicological effects. In 2006, through a postdoctoral fellowship from the Department of Pharmacology at the Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey in Piscataway, New Jersey, he began work as a visiting scientist in the Center for Toxicoinformatics at FDA's National Center for Toxicological Research in Jefferson, Arkansas. At NCTR, he integrated SAM (Significance Analysis of Microarrays) into ArrayTrack™. Dr. Chen is currently working on bioinformatics and a liver-toxicology knowledge base.
Feng Qian received her B.S. in analytical chemistry from Hunan University, China, in 1985, with additional training in information technology at Johns Hopkins University. She is an Oracle-Certified Professional (OCP)/DBA. She worked for Bristol-Myers Squibb as a research chemist focusing on the development of drug candidates. Her work ranged from determining water solubility and partition coefficients of drug candidates using computer-controlled HPLC, to developing quantitative structure-property relationship models for drug candidates, implementing a laboratory information management system and database administration. She joined the NCTR information technology and scientific computing contract in 2003, initially working on a toxicant database and software to calculate chemical-structure descriptors to correlate structure with activity. Since 2004, she has been working for ICF, participating in the development of ArrayTrack™, supporting the MicroArray Quality Control (MAQC) project as the major effort to import huge microarray datasets into the ArrayTrack™ system and deposit them into public databases, such as GEO and ArrayExpress. She is also one of the major staff for ArrayTrack™ customer support. Her work includes writing ArrayTrack™ user manuals, tutorials, training documents, installation media, graphics design, and customer account management. She is also responsible for the design and maintenance of the Center for Toxicoinformatics website.
Leming Shi earned a B.Sc. in analytical chemistry, M.Sc. in chemometrics, and Ph.D. in computational chemistry. He worked as a research associate at Case Western Reserve University (1994-1995), a visiting fellow at the NIH's National Cancer Institute (1995-1997), a computational chemist at R.O.W. Sciences (1997-1999), and a senior scientist at Wyeth and BASF (1999-2001). He was a co-founder and director of informatics at Chipscreen Biosciences Ltd. (2001-2003). Dr. Shi joined the FDA/NCTR in 2003 as a principal investigator in computational chemistry and bioinformatics. He has been leading the MicroArray Quality Control (MAQC) project, an effort with more than 100 participants worldwide aimed at reaching consensus on the "best practices" for the generation, analysis, and application of microarray data in the discovery, development, and review of FDA-regulated products (https://edkb.fda.gov/MAQC/ and http://www.nature.com/nbt/focus/maqc/). A specific focus of his research is to address the technological challenges in predictive and personalized medicine by developing more reliable genomics-based computational models and medical-diagnostic devices for accurately predicting drug efficacy and safety. Dr. Shi is also exploring the potential utilities of stem cells and gene-expression profiles of single cells in evaluating the toxicological effects of FDA-regulated products, including nanoparticles

(Continued on page 23)

(Continued from page 22)
widely used for therapeutic and diagnostic purposes. Dr. Shi has published over 100 peer-reviewed papers and is a co-inventor of nine-issued U.S. and international patents. Two novel drug candidates designed by Dr. Shi and his team for targeting cancer and type 2 diabetes have entered different phases of clinical trials.

Dhivya Arasappan received her B.S. in computer engineering in 2005 from Anna University, India, and her M.S. in Bioinformatics in 2007 from Virginia Commonwealth University. At Virginia Commonwealth University, she worked on developing unique network-analysis metrics and creating a tool for pathway analysis of microarray data. She also interned at Nutrilite, where she performed microarray-data analysis in order to identify genes and novel compounds for the company's sports-nutrition products. She joined the NCTR information technology and scientific computing contract, under Z-Tech, in February 2008 and has been involved in analysis of microarray data and technical writing. She has assisted in ArrayTrack™ documentation, maintenance, and testing, particularly for the locally installed version.

Weigong Ge earned his M.S. in computer science at the University of Bridgeport in 2003. Since then, he has been working at the Center for Toxicoinformatics at FDA's National Center for Toxicological Research. He has been involved in developing, debugging, and testing the Mold^2 system, which calculates molecular descriptors from 1D and 2D structures for drug discovery and toxicity prediction. He has also conducted analyses of proteomics data and microarray data analysis. He is a member of the ArrayTrack™ support team, focusing on extensive testing of the ArrayTrack™ system.

Xiaohui Fan received his Ph.D. in pharmaceutical sciences at Zhejiang University, China, in 2005. He did postdoctoral studies at the NCTR from 2005 to 2007, where his research focused on bioinformatics approaches for analysis of microarray and other omics data. He joined the School of Pharmaceutical Sciences, Zhejiang University, as an associate researcher in 2008, and is currently working on bioinformatics and computational systems biology. He is one of the major players of the MAQC project team at NCTR, especially in developing predictive models using gene expression data.

Huixiao Hong received his Ph.D. in computational chemistry at Nanjing University, China, in 1990. From 1990 to 1992, he completed his postdoctoral fellowship at the Maxwell Institute at Leeds University in the United Kingdom and worked on a computer-assisted organic synthesis system. He was an Associate Professor of Computational Chemistry at Nanjing University from 1992 to 1995, where he investigated computer-assisted molecular design and developed a suite of software for analytical instrumentation simulations. From 1995 to 1998, Dr. Hong worked in the Laboratory of Medicinal Chemistry at the National Cancer Institute at the National Institutes of Health as a Visiting Scientist. He discovered and designed more than 60 potent HIV integrase inhibitors and cyclooxygenase (COX) inhibitors through molecular modeling and pharmacophore searching. He also identified a mononucleotide-binding site in HIV integrase and elucidated the stereoselective binding mode of ibuprofen in COX1 using docking and molecular-dynamics simulation. From 1998 to 2000, he held a Research Scientist position in Sumitomo Chemical Company in Japan. Dr. Hong joined the information technology and scientific computing contract under Z-Tech at FDA's NCTR in 2000, where he conducted diverse analyses and methods development for chemoinformatics and bioinformatics and served as the computational science group manager from 2002. Dr. Hong became an NCTR staff fellow in 2007. His current research interests involve developing methods for analyzing and interpreting genomic, transcriptomic, proteomic, and metabonomic data.

Joshua Xu received his Ph.D. in electrical engineering at Texas

(Continued on page 24)

(Continued from page 23)
A&M University in 1999. His dissertation involved image reconstruction and analysis related to wavelet-based Magnetic Resonance Imaging. Immediately after graduation, he joined the Texas Center for Applied Technology, Texas A&M University, working as a software engineer for the Digital Emergency Medical Services Program, a congressionally funded mobile-telemedicine project that aimed to link emergency personnel in the field with trauma specialists in the hospital through wireless-data communication and medical-device integration. In 2007, he joined the NCTR information technology and scientific computing contract under Z-Tech as a computational scientist. He is currently developing SNPTrack, a bioinformatics software system for supporting FDA's Voluntary eXploratory Data Submission (VXDS) program on genotyping data management, analysis, and interpretation.

Steven Turner received a bachelor's degree in 1991 in computer science and computer engineering technology from the University of Arkansas at Little Rock. In 1994, he joined Acxiom Corporation in Conway, Ark., as a programmer, later becoming a database developer. He joined St. Jude Children's Research Hospital in 2000 as a senior database developer. From 2001 till 2008, he returned to Acxiom in Little Rock, Ark., as a database developer. He is currently employed as a database administrator for NCTR's information technology and scientific computing contract under Z-Tech, an ICF International company. Since joining NCTR, he has played a major role in maintaining the libraries used by the ArrayTrack™ application; this includes rewriting processes using UNIX shell scripting and SQL to add automation, reformatting, and error detection. He is also responsible for the development of the distribution media used for local installation of the ArrayTrack™ database.

Michelle E. Bishop is a support research biologist in the Division of Genetic and Reproductive Toxicology at the NCTR. Ms. Bishop began working at NCTR in 1991 as a student intern in the Division of Chemistry. After receiving her bachelor's in chemistry from the University of Arkansas at Pine Bluff in 1993, she became a full-time employee at NCTR. She is recognized as having established *in vitro* and *in vivo* micronucleus assay techniques at NCTR for the purpose of evaluating drugs or chemicals of interest to the FDA and applying the techniques to assist Principal Investigators at NCTR. She has worked on standardizing, formatting, and importing data into ArrayTrack™ for VGDS submissions. She is currently involved in the National Toxicology Program and National Institute of Child Health and Development projects at NCTR.

Martin Jackson received his bachelor's in computer science from the University of Arkansas-Little Rock in 1989, where he occasionally teaches undergraduate courses. He has worked as a software engineer on the information technology and scientific-computing contract at NCTR, now under Z-Tech, for some 27 years. During this tenure, he has contributed to a large number of scientific-software development efforts spanning across NCTR's research divisions, including a flow-cytometry system, the Endocrine Disruptor Knowledge Base (EDKB) program, the NIDA system, the MBS system, and ArrayTrack™. He is now developing new hardware and software components for future Neurotoxicology operant systems, as well as new software models to facilitate electronic data submissions into FDA's Janus Database.

Roger Perkins received both his bachelor's and master's degrees in nuclear engineering sciences from the University of Florida. He has during his 34-year career either been engaged in the conduct of computational science and engineering or in managing it. He spent a decade in research and development within the U.S. fast-breeder reactor and fusion-reactor programs, including three years as the Chief Nuclear Engineer at INESCO Corp. He joined the program management staff of the National Science Foundation's San Diego

(Continued on page 25)

(Continued from page 24)
Supercomputer Center at its inception in 1984. He later managed a U.S. Air Force computational-science support contract under the National Science Foundation IAG, and from there became director of a U.S. Navy scientific computing and engineering supercomputer center. Mr. Perkins came to NCTR in 1984 as program manager of the information technology and scientific computing contract, a capacity in which he continues to serve Z-Tech, an ICF International Company. The contract integrates IT support for systems, networks, databases, software engineering, biostatistics, animal experimentation, and Good Laboratory Practices, as well as cutting-edge work in bioinformatics and computational biology. During his tenure at NCTR, he has particularly focused on fostering computational-science approaches and capabilities through direct research contributions, project management, and team building.

Dr. Weida Tong is director of the Center for Toxicoinformatics at FDA's National Center for Toxicological Research. He also holds several adjunct academic appointments, including that of associate professor in the Department of Pharmacology, Robert Wood Johnson Medical School, University of Medicine & Dentistry of New Jersey, full-professor in the Department of Computer Science at the University of Arkansas at Little Rock, and assistant professor in the Department of Pharmaceutical Sciences, University of Arkansas for Medical Sciences. The Center for Toxicoinformatics at NCTR is involved in developing bioinformatics systems to support FDA pharmacogenomics-data submission and regulation. Two of the most visible projects from this group are: (1) development of the FDA genomic tool, ArrayTrack™; and (2) leading the effort on the Microarray Quality Control (MAQC) consortium. Dr. Tong is also deeply involved in the FDA Voluntary Genomics Data Submission Program and development of the best-practice document for pharmacogenomics data submission. In addition, Dr. Tong also specializes in the field of computational modeling, chemoinformatics, and QSAR with specific interest in estrogen, androgen, and endocrine disruptors. Dr. Tong has published more than 100 papers and book chapters.

RRP's Research Spotlight

Jim Kaput, Ph.D.

Dr. Jim Kaput joined FDA's National Center for Toxicological Research (NCTR) in November 2007 as Director of the Division of Personalized Nutrition and Medicine. He earned his Ph.D. in the Biochemistry & Cell and Molecular Biology Program, Colorado State University, and performed his postdoctoral fellowship in cell biology at The Rockefeller University under the direction of Günter Blobel (1999 Nobel Laureate in Physiology and Medicine). Dr. Kaput went on to become an Assistant Professor, Laboratory of Cell Biology, The Rockefeller University and, later, Assistant Professor, Department of Biochemistry, University of Illinois, College of Medicine. He also established Northwestern University's Biotechnology Core Laboratory. After two years there, he formed and provided scientific leadership for several private biotechnology- and nutrigenomic-related companies. Dr. Kaput was most recently an As-

(Continued on page 26)

(Continued from page 25)
sistant Professor, Department of Surgery, University of Illinois, Chicago, and Coordinator, Science and Administrative Activities, Section on Cellular and Molecular Biology, Center of Excellence in Nutritional Genomics, University of California, Davis, and Scientific Advisor for International Collaborations with NuGO, the European Nutrigenomics Organization.

Dr. Kaput is well-published in the area of Nutrigenomics and is frequently asked to present his views both nationally and internationally. He has developed extensive course material and electronic delivery options for both literature reviews and educational materials related to nutrition and genetics and their relationship. He is also active in the leadership of several national and international activities focused on nutrigenomics and human genetic diversity. As an example of his international leadership and recognition, Dr. Kaput has just completed his service in South America (Brazil) as a Fulbright Senior Specialist Fellow in Global/Public (nutrigenomics) before accepting his position with NCTR.

NCTR formed the Division of Personalized Nutrition and Medicine (DPNM) in October 2006 with a mandate to develop strategies for individualizing healthcare through personalized nutrition and medicine. The newly formed DPNM is focusing on several parallel initiatives to create infrastructure and novel-research paradigms for developing the path to personalized medicine and nutrition. The key conceptual challenge for personalizing healthcare is that current research strategies are based on population studies. Association studies, whether genetic, nutritional, or nutrigenomic, yield the attributable fraction (AF)— "the proportional reduction in average-disease risk over a specified time interval that would be achieved by eliminating the exposure of interest from the population" while other factors remain unchanged. The AF is usually calculated from population models and is not directly applicable to individuals, because individuals may differ genetically, physiologically, and nutritionally from the population averages. To address this challenge, the division is developing:

▲ A research strategy that merges omic technologies (e.g., genomic, transciptomic, proteomic, metabolomic) with community-based, participatory research strategies. The community-based participatory research's central focus is developing a partnership among researchers and individuals in a community that allows for more in depth lifestyle analyses but also translational research that simultaneously helps improve the health of individuals and communities. DPNM's first effort was to participate in an Obesity Prevention Summer Camp in collaboration with the USDA—Delta Obesity Prevention Research Unit and the Boys, Girls, and Adults Community Development Center in Marvell, Arkansas. This study is linking human-health studies with basic research studies, cell culture, and animal models.

▲ A homeostatic-challenge experiment is being designed to analyze health status in about 100 NCTR employees using the many omic technologies at the Center. Current biomarkers typically focus on diagnosing disease rather than identifying markers for health or for susceptibility to disease. How a healthy individual responds to nutritional challenges (e.g., oral glucose or oral lipid challenges) may predict the path of health for the individual. Investigators from all divisions at NCTR will participate in the design and execution of this study.

▲ A core genomic laboratory for offering genetic-analyses services but also a cutting-edge research unit that will assess existing and novel technologies for implementing personalized nutrition and medicine.

▲ Novel algorithms for classifying individuals into metabolic

(Continued on page 27)

(Continued from page 26)

groups.

Health and the development and progression of disease are produced by interactions between an individual's genetic makeup and environmental exposures. Many experimental designs in the past have ignored one or the other of these influences because of the complexity of environmental factors and the heterogeneity of human populations. The advent of high-throughput technologies, for genetic analyses and ongoing development of statistical tools for analyzing high-dimensional datasets, has led to the realization that biological and environmental complexity may be analyzed for understanding human health. While these technological advances can be used with the reductionist paradigms that have contributed much to our knowledge of health and disease processes, the full-utilization of these technologies requires new experimental strategies that account for genetic differences among individuals and their unique environmental exposures.

The activities of the Division are designed to develop the novel-research strategies, data, and applications to improve personal and public health equitably through personalization of recommendations for nutrient intakes and medical treatments.

Donna Mendrick, Ph.D.

Dr. Donna Mendrick is the Director of the Division of Systems Toxicology at the National Center for Toxicological Research (NCTR), a research arm of the FDA. The division incorporates genomics, proteomics, metabolomics, bioinformatics, spectral modeling, and other approaches to answer the needs of the FDA in terms of drug and food safety and improving the understanding of human disease. Prior to joining FDA, she was a Scientific Fellow and Vice-President of Pharmacogenomics at Gene Logic where she oversaw pharmacogenomics and spearheaded its toxicogenomics effort. For the latter, she formed a pharmaceutical consortium to help guide the development of the program. Before joining Gene Logic, she was a Group Leader in Pharmacology at Human Genome Sciences where she oversaw multiple project teams, toxicity studies, pharmacology studies, etc. Until 1995, Dr. Mendrick was an Assistant Professor of Pathology at Harvard Medical School and Brigham and Women's Hospital where her work focused on biomarkers of renal injury and vascular damage.

Dr. Mendrick has over 25 years of experience in the fields of immunology, pathology, pharmacogenomics, pharmacology, toxicology, and toxicogenomics employing small-molecule drugs, recombinant therapeutic proteins, and monoclonal antibodies. She has served as a speaker and/or a member of the planning committee for many of the FDA co-sponsored workshops on the use of pharmacogenomics in drug and diagnostic kit development.

Dr. Mendrick has published both review articles and opinion pieces on the use of pharmacogenomics to identify biomarkers and co-edited the book titled *Essential Concepts in Toxicogenomics*. She currently is an editorial board member of the *Pharmacogenomics & Personalized Medicine* journal, past president of the National Capital Area Chapter of the Society of Toxicology, and a committee member of the Predictive Toxicology Discussion Group at the New York Academy of Sciences. Dr. Mendrick was on the editorial board of the *Journal of Histochemistry and Cytochemistry* for eight years, a member of the NIH SBIR Immunology Study Section for eight years, and a member of the Board of Direc-

(Continued on page 28)

(Continued from page 27)
tors of the National Kidney Foundation of Massachusetts for four years.

Systems biology is the integration of biologically based systems, chemistry, engineering, and computational science to improve the understanding of molecules, cells, organs, and organisms. The Division of Systems Toxicology (DST) at NCTR is devoted to the use of integrated sciences and evolving approaches to speed innovation in protecting public health to:

- ▲ Improve the safe use of drugs by providing a better understanding of pathogenic processes that lead to cellular and tissue injury as a result of adverse events related to drug therapy and disease progression.

- ▲ Protect the nation's food chain by advancing the detection of bacterial contamination of foods and drinks.

- ▲ Enhance the efficacy of drugs through an understanding of delivery techniques and drug target effects.

- ▲ Improve the performance of medical devices using new computational approaches.

The discovery of, and research on, biomarkers underlies much of the work being performed in DST. Biomarkers have been defined by the Biomarkers Definitions Working Group as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention." A better understanding of biomarkers promises to improve detection of toxicity earlier in the pharmaceutical development pipeline, enhance clinical management of patients, and provide inroads into personalized medicine (usually defined as the right drug for the right individual at the right dose). As we work toward a better understanding of pathogenic processes and the body's response to therapy, it is important to combine multiple disciplines, as this area of science can no longer survive in a siloed environment. Integration of data and computation science is a critical component and will lead to faster discovery of biomarkers and assist in placing such in the appropriate biological context. Equally important is the bridging between lab bench, animal science, and clinical needs to identify and qualify "translational" biomarkers that will provide insight into basic biology and provide actionable information for clinicians and the development of new drugs and medical devices.

Margaret A. Miller, Ph.D., RN

Margaret (Peggy) A. Miller, Ph.D., RN: Dr. Miller joined the FDA's National Center for Toxicological Research (NCTR) in the Office of the Director as the Associate Director for Regulatory Activities in Rockville, Maryland, in October 2007. She not only coordinates research and regulatory activities between the Washington Office and NCTR in Jefferson Arkansas, but she also directs a women's health research group in Arkansas. The research conducted by this group focuses on understanding the molecular basis of drug efficacy and safety and how genetics, sex, diet, and other environmental factors influence these parameters. Specific research projects involve investigating the genetic and epigenetic modulation of cytochrome P-450-metabolizing enzymes and drug transporter molecules and identifying biomarkers of disease.

Dr. Miller received her Ph.D. in endocrinology-reproductive

(Continued on page 29)

(Continued from page 28)
physiology from the University of Wisconsin-Madison in 1981. Her thesis research focused on the hormonal control of the cytochrome P-450 metabolizing enzymes. Following her studies at Wisconsin, she accepted a position as a Postdoctoral Research Associate and Lecturer in the Department of Physiology and Biophysics at the University of Illinois. During this time, she worked with Dr. Benita Katzenellenbogen conducting research on the role of estrogen and other compounds in the control of human breast-cancer cell growth. Her pioneering studies in the area of steroid-receptor molecular biology were unique not only for their scientific merit but for their impact on women's health.

Dr. Miller joined Monsanto Agricultural Company, in 1985, in the Animal Sciences Division where she directed a laboratory responsible for providing analytical support for the clinical trials of bovine and porcine somatotropin. While at Monsanto, she was a member of the Women's Leadership Class, a program specifically designed to encourage women to enter leadership positions.

In 1989, Dr. Miller joined the FDA's Center for Veterinary Medicine (CVM), as a scientific reviewer in the Division of Toxicology and Environmental Sciences. At CVM, she assumed positions of increasing responsibility and eventually managed the Center's Food Safety Program as Deputy Director for Human Food Safety. In this position she was once again one of the few women in leadership positions in the Center and developed programs to support the advancement of women in science and management.

She joined FDA's Office of Women's Health (OWH) in 1999 in the Office of the Commissioner to manage the scientific research program. Through its scientific research program, OWH funds gender-based research that fills critical information gaps needed to ensure FDA-regulated products are safe and effective in both men and women. While in OWH, Dr. Miller started several research initiatives, including studies to investigate pharmaceutical use by pregnant and lactating women and women and cardiovascular disease. The goal of these studies was to improve the safety of FDA-regulated products for women.

In May 2001, Dr. Miller completed her Bachelor's of Nursing degree and in January 2002 became a Registered Nurse. As a nurse, Dr. Miller has gained valuable clinical experience in caring for individuals from many different cultures. This clinical experience not only added a new public-health dimension to her extensive laboratory and regulatory background, but it also provided a direct opportunity to observe the impact of gender disparity in healthcare.

Recently, Dr. Miller was seconded to the World Health Organization (WHO) Headquarters in Geneva, Switzerland (July 2005–July 2007). While at WHO, she had the opportunity for work with several United Nations (UN) agencies. During this time, Dr. Miller, in consultation with experts in social mobilization, developed the "Five keys to Safer Food" training course for women. The course teaches health professionals how to promote the adoption of safe-food handling practices by women in the home. The program was piloted in South Africa in September 2007, in Tunisia in February 2008, and is now being translated into the six official UN languages for distribution worldwide. Dr. Miller continues serving as a WHO consultant.

In addition to her current NCTR position, Dr. Miller has joined the faculty of the George Washington University (Fall 2008) as a professorial lecturer, where she teaches a graduate-level toxicology course.

Jefferson Laboratories of the FDA, Jefferson, Arkansas

Editorial Matters:

DHHS/FDA/Jefferson Labs
National Center for Toxicological Research
3900 NCTR Road, HFT-1
Jefferson, Arkansas 72079-9502
Telephone: (870) 543-7516
Website: https://www.fda.gov/nctr

Address for editorial matters noted above.
Article may be republished without permission. Credit to
Regulatory Research Perspectives as the source is appreciated.

Editor: Thomas Flammang, Ph.D.
E-mail: thomas.flammang@fda.hhs.gov
Managing Editor/Layout/Photography: Virginia B. Taylor
E-mail: virginia.taylor@fda.hhs.gov

Regulatory Research Perspectives: Editorial Board