ArrayTrack: A free FDA bioinformatics tool to support emerging biomedical research — An update

Joshua Xu,[1] Reagan Kelly,[1] Hong Fang[1] and Weida Tong[2]*

[1]Z-Tech Corporation, an ICF International Company at the National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Rd, Jefferson, AR 72079, USA
[2]Center for Bioinformatics, Division of Systems Biology, National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Rd, Jefferson, AR 72079, USA
**Correspondence to*: E-mail: Zhihua.xu@fda.hhs.gov

Date received (in revised form): 28th May 2010

Abstract

ArrayTrack™ is a Food and Drug Administration (FDA) bioinformatics tool that has been widely adopted by the research community for genomics studies. It provides an integrated environment for microarray data management, analysis and interpretation. Most of its functionality for statistical, pathway and gene ontology analysis can also be applied independently to data generated by other molecular technologies. ArrayTrack has been undergoing active development and enhancement since its inception in 2001. This review summarises its key functionalities, with emphasis on the most recent extensions in support of the evolving needs of FDA's research programmes. ArrayTrack has added capability to manage, analyse and interpret proteomics and metabolomics data after quantification of peptides and metabolites abundance, respectively. Annotation information about single nucleotide polymorphisms and quantitative trait loci has been integrated to support genetics-related studies. Other extensions have been added to manage and analyse genomics data related to bacterial food-borne pathogens.

Keywords: *Microarray, bioinformatics, omics integration, SNP, food-borne pathogens*

Introduction

Microarray technology has been widely used in diverse fields, including pharmacology, toxicology and clinical medicine.[1–5] An integrated bioinformatics infrastructure to facilitate data management and analysis is vital fully to realise the potential impacts of genomics on research and public health. The Food and Drug Administration (FDA) has developed the genomics tool, ArrayTrack™, which has a rich set of functionalities to manage, analyse and interpret genomic data, with a focus on microarray data.[6,7] Most functionality is applicable to biomarker development in biomedical research and personalised medicine using other molecular technologies.

ArrayTrack became the FDA's genomics tool to support the Voluntary Genomics Data Submission (VGDS) programme[8] in early 2004. VGDS provides a novel mechanism outside normal regulatory interactions for sponsors and the FDA to develop expertise, tools and processes appropriate for regulatory interpretation of pharmacogenomics data. In addition to its broad use within the FDA in various research and regulation-related programmes, ArrayTrack is also freely available to the scientific community. Users can gain access to ArrayTrack either through the website (http://www.fda.gov/ArrayTrack) or by

requesting media for local installation. The ArrayTrack user base has grown steadily, and the tool has been adopted by several government agencies (eg the Environmental Protection Agency, Centers for Disease Control and Prevention and National Institutes of Health), academic institutions and private companies. Figure 1 presents an overview of the user base growth trend since 2004.

Over the years, ArrayTrack has been presented and reviewed in several publications.[6,7,9–12] This update will, following a brief overview of its key functionalities, focus on its most recent extensions. The user manual and tutorials are available from the ArrayTrack website (http://www.fda.gov/ArrayTrack). At the time of this writing, ArrayTrack version 3.5, which includes the added capabilities described herein, has just been released.

A brief overview

ArrayTrack was designed as a client-server system that provides: (i) a database server hosting microarray data, related study data and a collection of genomic functional information (eg gene annotations, pathways, gene ontology); (ii) a client application that integrates visualisation and analysis of microarray data with functional information; and (iii) a hierarchically modular structure that provides extensibility for other types of 'omics' data and data from other emerging molecular technologies (eg genome-wide association studies).

ArrayTrack comprises three major integrated components: (i) MicroarrayDB, a database that stores essential data associated with a microarray experiment, including raw gene expression data and information on samples, treatments and phenotypic observations; (ii) LIB, a series of libraries that contain functional information (eg gene annotations, protein function and pathways) from public sources; and (iii) TOOL, a set of algorithms that provide analysis capabilities for data visualisation, normalisation, significance analysis, clustering and classification. A user-friendly interface enables the selection of an analysis method from TOOL, applying the method to selected microarray data stored in MicroarrayDB and linking analysis results directly to associated functional annotations in LIB. Some key functionalities and their applications are highlighted below and the full list of functions is available on the ArrayTrack website.

- Data import: SimpleTox, an internally developed Clinical Data Interchange Standard

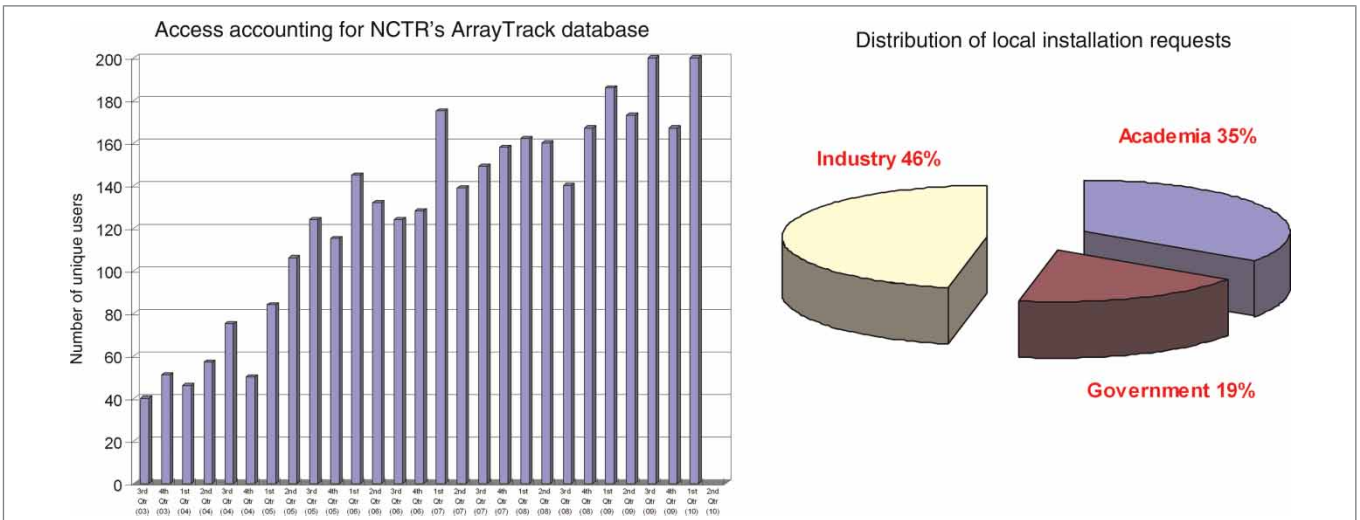

Figure 1. ArrayTrack user base statistics. The bar chart on the left represents the quarterly numbers of unique users who have accessed the ArrayTrack database hosted at the FDA's National Center for Toxicological Research (NCTR), excluding FDA users. The bar chart shows the steady growth of web users of ArrayTrack. The pie chart on the right displays the distribution among institutional type for 150 local ArrayTrack installations since 2004. Many industrial users come from pharmaceutical companies. The 'Academia' category includes public research-oriented organisations.

Consortium/Standard for Exchange of Non-clinical Data (CDISC/SEND)[13]-compliant format, allows essential experiment summaries to be uploaded in a spreadsheet format. This consistent and easy-to-use format facilitates cross-study analysis.

- Class comparison: ArrayTrack offers a rich set of methods for determining a list of differentially expressed genes (DEGs) by comparing two groups (eg treated versus control), such as the simple t-test, the Volcano plot and more advanced statistical approaches, such as false discovery rate (FDR) control and significance analysis of microarrays (SAM). In addition, the analysis of variance (ANOVA) is available for multi-group comparison.
- Class discovery: Principal component analysis (PCA), hierarchical cluster analysis (HCA) and K-means clustering are available for unsupervised class discovery and pattern identification. Such methods are powerful ways to investigate the grouping of samples in terms of their similarities in gene expression.
- Class prediction: For predictive model development, ArrayTrack offers supervised learning methods such as K-nearest neighbour (KNN), linear discriminant analysis (LDA) and support vector machines (SVMs). Model building is an important step towards applying microarray technology to diagnosis, prognosis and treatment outcome prediction.

Importantly, most ArrayTrack functionalities are not limited to microarray data. For example, the information from ArrayTrack's extensive collection of libraries can be applied directly to data from non-microarray technologies, results collected from literature, or other sources (Figure 2). In addition, most statistical, classification and visualisation tools have been implemented as generic tools and can be readily used to analyse data where dependent and independent data are from other research fields.

In summary, ArrayTrack provides a broad range of functionalities, allowing integration of functional information about genes, proteins, pathways and gene ontology[7,14] with data from microarrays and other molecular technologies. Such integrative analysis capability furthers ArrayTrack's ability to support systems biology. ArrayTrack's diverse capabilities have been augmented recently, as described in the next section.

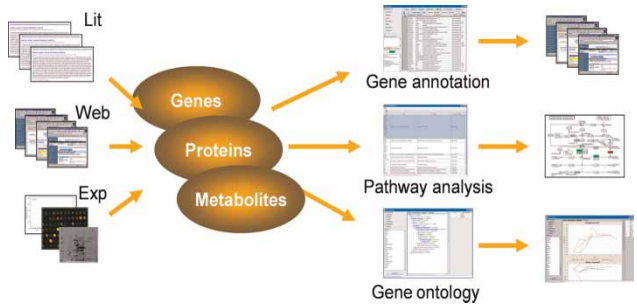

Figure 2. A workflow for data interpretation using the ArrayTrack library collection. The process starts with a list of genes, proteins and/or pathways. The lists can be either generated from in-house experiments (eg microarray or proteomics/metabolomics experiments) or collected from the literature, internet, or communication with peers. The user can easily map the lists independently to different ArrayTrack libraries (eg gene, pathway library) or in an integrative manner for biological interpretation.

New development

ArrayTrack development has continued with the addition of new utilities to address the growing and changing needs of FDA's research programmes. New features facilitate the management of pre-processed proteomics and metabolomics data. The single nucleotide polymorphism (SNP) and quantitative trait locus (QTL) libraries have been integrated to support pathway analysis and data mining for SNP-related studies. Extensive enhancements have also been made to manage and analyse the genetic profiling data related to bacterial food-borne pathogens. These enhancements are depicted in Figure 3 in relation to ArrayTrack's core functionality.

Support for proteomics and metabolomics data

Proteomics and metabolomics have grown steadily in importance in biomedical research, in parallel with microarrays. The integration of tri-omics data

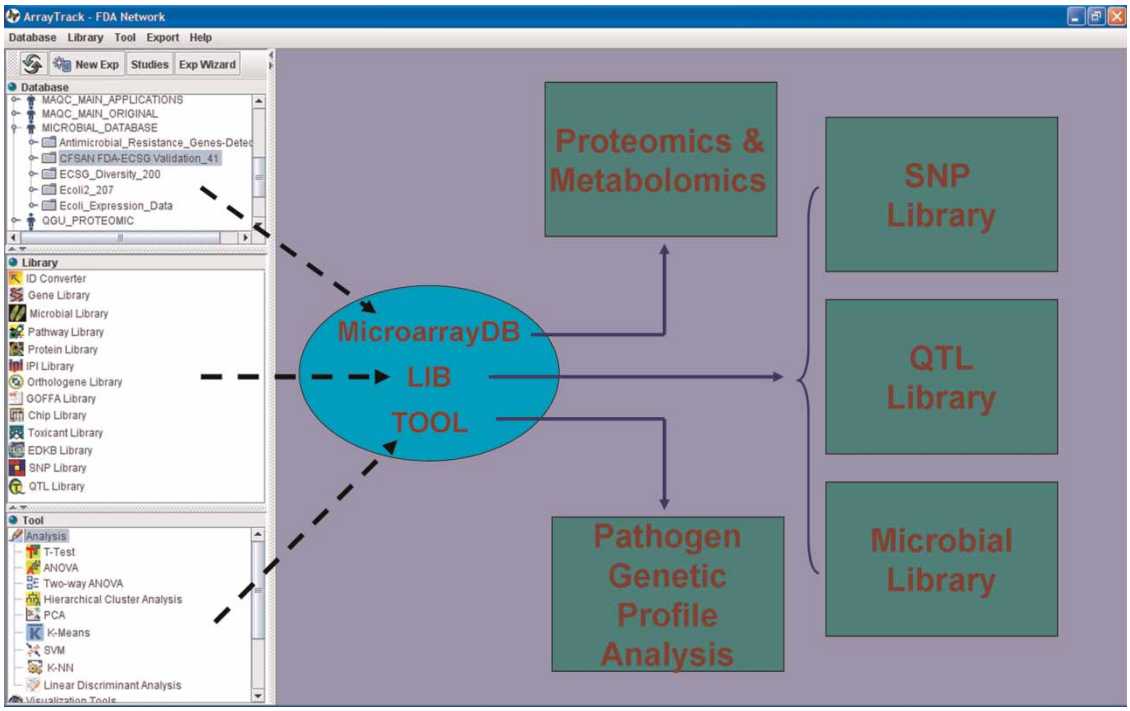

Figure 3. Recent enhancements to ArrayTrack.

(ie genomics, proteomics and metabolomics data) has been a primary goal in systems biology for drug development and safety evaluation. To support this line of research and to review this type of data submitted by sponsors to the FDA through the VGDS — or, as it has been renamed, the Voluntary Exploratory Data Submissions (VXDS) program — ArrayTrack was previously modified to accommodate lists of proteins and metabolites. Additionally, a new systems biology function, called CommonPathway,[10] was added that enables the examination of common pathways and functional categories (eg gene ontology terms) shared by different data types.

ArrayTrack is now capable of analysing data from any mass spectrometry platform once the raw data have been processed for detection and quantification of peptides or metabolites — an important step in the data analysis workflow.[15,16] New tools have been created to simplify the handling of proteomics data from the two most popular database search programs, Mascot and Sequest, for detection of peptides. The tools convert output files from these two programs into ArrayTrack-readable files.

The same interpretation tools used for microarray data in ArrayTrack are extensible for proteomics and metabolomics data. Thus, by linking the results to gene, protein and pathway databases, researchers will be able to contextualise these results in the same way as gene expression experiments. Additionally, a unified interface helps to reduce the 'learning curve' associated with analysing new data types, giving researchers currently working on microarrays an incentive to move towards more integrated approaches that also encompass proteins and metabolites.

SNP and QTL libraries

Recent advances in microarray-based genotyping techniques have enabled researchers rapidly to scan for known SNPs across complete genomes. An efficient data-mining strategy and a set of sophisticated tools are necessary better to understand and utilise the findings from genetic association studies.

One of the focuses in genetic association studies is to relate SNPs to genes and pathways in order to understand the underlying disease mechanisms.

ArrayTrack has already provided a gene–pathway discovery platform. By integrating the SNP library, which contains annotation summary information for SNPs and their mapped relationship to genes, ArrayTrack now provides an integrated SNP–gene–pathway analysis platform for SNP studies.

A QTL is a region of DNA that is associated with a particular phenotypic trait. A common use for QTL data is to identify candidate genes underlying a trait within one or more QTLs. The identification of the SNP–gene–QTL relationship is the basis of tests to determine whether the gene/SNP is associated with the aetiology of a disease in animal models or human studies. The integration of SNP and QTL libraries into ArrayTrack enables dynamic mining of such complex biological interactions.

SNP and QTL libraries[17] have been constructed and incorporated into ArrayTrack. Data from several public repositories were collected in the SNP and QTL libraries and connected to other domain libraries (genes, proteins, metabolites and pathways) in ArrayTrack. Linking the data sets within ArrayTrack allows searching of SNP and QTL data, as well as their relationships to other biological molecules. The SNP library includes approximately 15 million human SNPs and their annotations, while the QTL library contains publicly available QTL data associated with specific phenotypes identified in mice, rats and humans. Case studies demonstrating the utility of these libraries have been reported.[17,18]

Support for microbial pathogen microarray data

Food-borne pathogens are a leading cause of illness in the USA. High-throughput microarray technology provides an effective way to identify, characterise and obtain a nearly complete snapshot of the genetic traits of bacterial strains, such as their pathogenicity, virulence or antimicrobial resistance. Such genome-wide insight is necessary for accurate identification and discrimination of pathogens that may contaminate the food supply.

ArrayTrack has been extended to support microbial genomics research using microarrays.[19] ArrayTrack's libraries have been populated with bioinformatics data relating to bacterial pathogen species from the public domain. Data processing and visualisation tools have been enhanced with customised options to facilitate analysis of genetic profiling microarray data. Specifically, three new functions have been developed and are particularly effective for analysis of these microarray data: flag-based hierarchical clustering analysis (HCA), a flag concordance (FC) heat map and flag indicators in the mixed scatter plot (where 'flag' refers to a gene's presence or absence call). These functions are particularly relevant and effective for the identification and characterisation of bacterial pathogens using microarray genetic profiling data. The enhancements are displayed in Figure 4. For example, the Microbial Library (Figure 4C) is the newest addition to ArrayTrack's collection of libraries. Currently, it holds 270,000 gene records from a total of 84 bacterial strains: 30 *Escherichia coli*, 39 *Salmonella enterica*, ten *Shigella spp.* and five *Vibrio spp.* Thus, as a starting point, the Microbial Library is focused on these four bacterial genera that are common food-borne pathogens. ArrayTrack also holds microbial pathway information from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) for over 50 of these strains[20] and gene ontology information for the *E. coli* K12 substrain MG1655. The gene annotations and sequences for the Microbial Library were downloaded from the National Center for Biotechnology Information (NCBI) website.

Summary

The recent enhancement to ArrayTrack has broadened its capability to support proteomics, metabolomics and SNP-related studies. As an integrated omics data analysis environment, ArrayTrack allows meta-analyses and integration of results from multiple studies with pathway and functional annotations. By providing powerful but easy-to-use utilities, ArrayTrack is positioned to assist in making integrated, contextualised analyses more common, which, in turn, will help to harness genetic knowledge to improve the protection of public health.

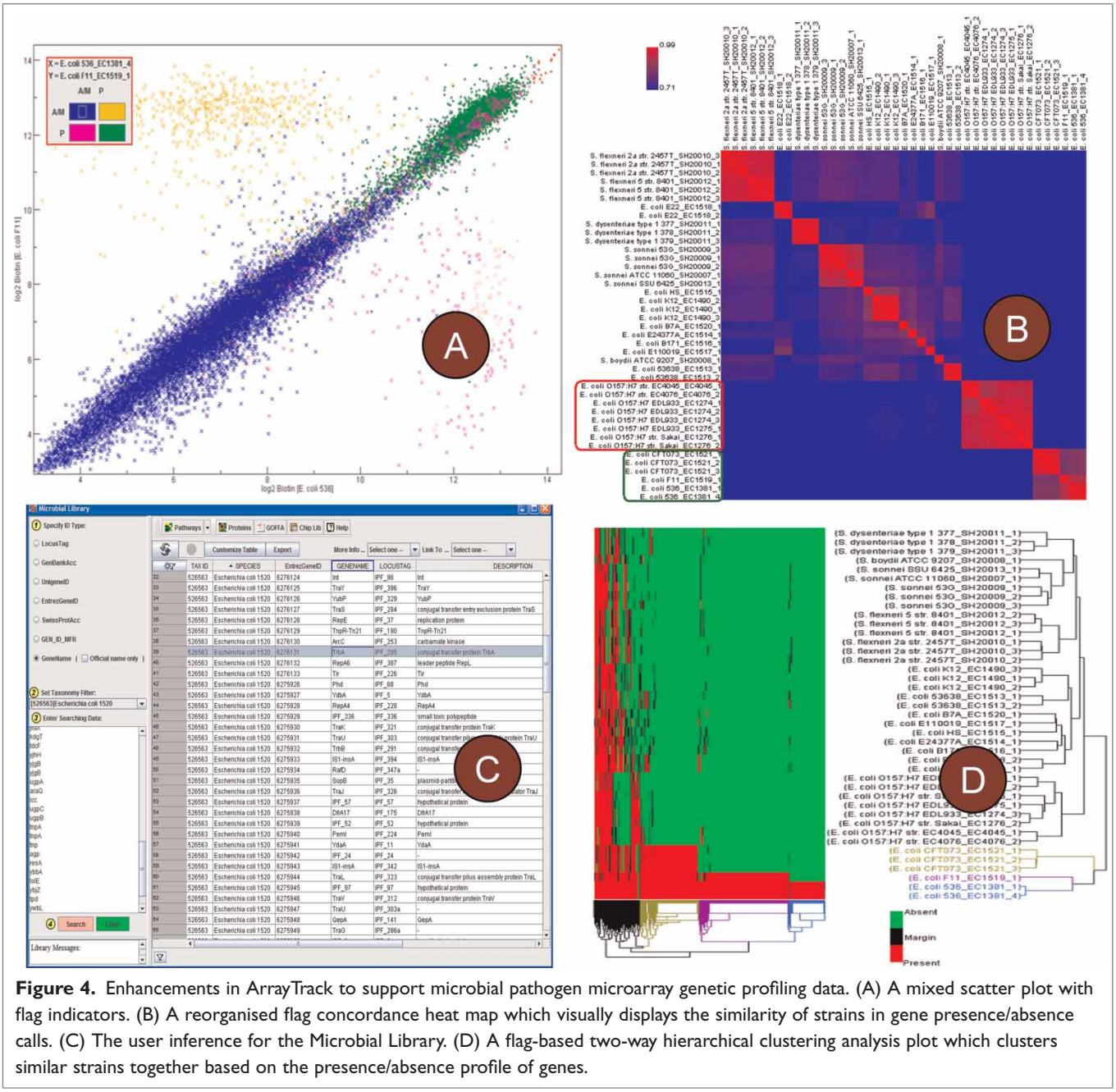

Figure 4. Enhancements in ArrayTrack to support microbial pathogen microarray genetic profiling data. (A) A mixed scatter plot with flag indicators. (B) A reorganised flag concordance heat map which visually displays the similarity of strains in gene presence/absence calls. (C) The user inference for the Microbial Library. (D) A flag-based two-way hierarchical clustering analysis plot which clusters similar strains together based on the presence/absence profile of genes.

Disclaimer

The views presented in this paper do not necessarily reflect those of the US FDA.

Acknowledgments

The authors would like to thank all current and former members of the ArrayTrack development team for their dedication in building up ArrayTrack into an invaluable research tool. The authors would also like to thank all ArrayTrack users for their feedback and support.

References

1. van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D. *et al.* (2002), 'Gene expression profiling predicts clinical outcome of breast cancer', *Nature* Vol. 415, No. 6871, pp. 530–536.
2. Gunther, E.C., Stone, D.J., Gerwein, R.W., Bento, P. *et al.* (2003), 'Prediction of clinical drug efficacy by classification of drug-induced

genomic expression profiles in vitro', *Proc. Natl. Acad. Sci. USA* Vol. 100, No. 16, pp. 9608–9613.
3. Kaushik, N., Fear, D., Richards, S.C.M., McDermott, C.R. *et al.* (2005), 'Gene expression in peripheral blood mononuclear cells from patients with chronic fatigue syndrome', *J. Clin. Pathol.* Vol. 58, pp. 826–832.
4. Bushel, P.R., Heinloth, A.N., Li, J., Huang, L. *et al.* (2007), 'Blood gene expression signatures predict exposure levels', *Proc. Natl. Acad. Sci. USA*, Vol. 104, No. 46, pp. 18211–18216.
5. Oberthuer, A., Warnat, P., Kahlert, Y., Westermann, F. *et al.* (2007), 'Classification of neuroblastoma patients by published gene-expression markers reveals a low sensitivity for unfavorable courses of MYCN non-amplified disease', *Cancer Lett.* Vol. 250, No. 2, pp. 250–267.
6. Tong, W., Cao, X., Harris, S., Sun, H. *et al.* (2003), 'ArrayTrack — Supporting toxicogenomic research at the U.S. Food Drug Administration National Center for Toxicological Research', *Environ. Health Perspect.* Vol. 111, No. 15, pp. 1819–1826.
7. Harris, S.C., Fang, H., Su, Z., Chen, M. *et al.* (2009), 'FDA bioinformatics tool for public use — ArrayTrack™', *Regulatory Research Perspectives* Vol. 8, No. 1, pp. 1–25.
8. Frueh, F.W. (2006), 'Impact of microarray data quality on genomic data submissions to the FDA', *Nat. Biotechnol.* Vol. 24, No. 9, pp. 1105–1107.
9. Fang, H., Harris, S.C., Su, Z., Chen, M. *et al.* (2009), 'ArrayTrack: An FDA and public genomic tool', *Methods Mol. Biol.* Vol. 563, No. 3, pp. 379–398.
10. Fang, H., Perkins, R. and Tong, W. (2007), 'Omics data integration: A systems approach view', *American Drug Discovery* Vol. 2, pp. 49–52.
11. Tong, W., Harris, S.C., Fang, H., Shi, L. *et al.* (2007), 'An integrated bioinformatics infrastructure essential for advancing pharmacogenomics and personalized medicine in the context of the FDA's Critical Path Initiative', *Drug Discovery Today: Technologies* Vol. 4, No. 1, pp. 3–8.
12. Tong, W., Harris, S., Cao, X., Fang, H. *et al.* (2004), 'Development of public toxicogenomics software for microarray data management and analysis', *Mutat. Res.* Vol. 549, pp. 241–253.
13. CDISC, T. *Clinical Data Interchange Standard Consortium (CDISC): CDISC Inc.*, 15907 Two Rivers Cove, Austin, Texas 78717. Available at http://www.cdisc.org/index.html, 2007.
14. Sun, H., Fang, H., Chen, T., Perkins, R. *et al.* (2006), 'GOFFA: Gene Ontology For Functional Analysis — A FDA gene ontology tool for analysis of genomic and proteomic data', *BMC Bioinformatics* Vol. 7 (Suppl. 2), p. S23.
15. Domon, B. and Aebersold, R. (2006), 'Challenges and opportunities in proteomics data analysis', *Mol. Cell. Proteomics* Vol. 5, No. 10, pp. 1921–1926.
16. Spratlin, J.L., Serkova, N.J. and Eckhardt, S.G. (2009), 'Clinical applications of metabolomics in oncology: A review', *Clin. Cancer Res.* Vol. 15, No. 2, pp. 431–440.
17. Xu, J. *et al.* (2010), 'Two new ArrayTrack™ libraries for personalized biomedical research', *BMC Bioinformatics*, In press.
18. Wise, C. and Kaput, J. (2009), 'A strategy for analyzing gene-nutrient interactions in Type 2 diabetes', *J. Diabetes Sci. Technol.* Vol. 3, No. 4, pp. 710–721.
19. Fang, H. *et al.* (2010), 'An FDA bioinformatics tool for microbial genomics research on molecular characterization of bacterial foodborne pathogens using microarrays', *BMC Bioinformatics*, In press.
20. Kanehisa, M. (2002), 'The KEGG database', *Novartis Found. Symp.* Vol. 247, pp. 91–101.