Journal of Pharmacogenomics & Pharmacoproteomics

Review Article   **Open Access**

Toxicogenomics Investigation Under the eTOX Project

Olivier Taboureau[1], Anne Hersey[2], Karine Audouze[1], Laurent Gautier[1], Ulrik P. Jacobsen[1], Ruth Akhtar[2], Francis Atkinson[2], John P. Overington[2] and Søren Brunak[1]*

[1]Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet, Building 208, DK-2800 Lyngby, Denmark
[2]EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

Abstract

Attrition of drug candidates during pre-clinical development due to toxicity, especially hepatotoxicity and nephrotoxicity, is an important and continuing problem in the pharmaceutical industry. The reasons for this trend may be multifactorial and there is a need to improve toxicity testing paradigms within the industry. Microarray technologies have the ability to generate massive amounts of gene expression information as an initial step to decipher the molecular mechanisms of toxicologic changes, i.e. toxicogenomics. In the context of the eTOX consortium, one of public private partnership within the framework of the European Innovative Medicines Inititative (IMI), we will discuss here how the integration and analysis of toxicogenomics data can help to understanding the mechanism of toxicity of a compound and so reduce the risk of late-stage failure in pharmaceutical development.

Keywords: Toxicogenomics; Gene expression; Drugs; Rats

Status on Toxicogenomics Studies

The current cost to bring a drug candidate to market is estimated to US $1.8 billion, with an average success rate of 8% [1]. However, there has been a significant decrease in the development of new and effective drugs and one of the most important reasons for attrition was due to clinical side effects and toxicity. Interestingly, since the advent of DNA microarray technology (15 years ago), the field of toxicology started to discuss the great potential of genome-wide expression profiling for toxicity testing: the promise is that the mechanism of action of a chemical at the cellular level, thus the risk of chemical toxicity, can be identified through the transcriptional activity of cells. The keyword toxicogenomics was coined to identify the systematic approach. Moreover, at the molecular level, as the human and rodent genome exhibit more than 90% similarity, toxicogenomics could be of benefit for the extrapolation of toxic effects between species. A similar argument applies for extrapolating *in-vivo* effects from *in-vitro* experiments, although most often different parameters are measured in both experiments [2]. Over the last decade, a number of toxicogenomics studies have been performed taking advantage of the maturity of the microarray technology, and we consider that technology for expression profiling as an indicator at how the concept is gaining adoption. Looking on the number of references mentioning "gene expression" in the PubMed database, we can observe that microarray technology is not applied solely to toxicology but the method allows study of the global transcriptional changes of a given biological system in response to any stress perturbation.

The "toxicogenomics" field was really investigated from 2004 when gene expression experiments of drugs and toxicants started to be publicly available (Figure 1). Toxicogenomics has proven to be useful in toxicology [3,4]. For example in carcinogenicity, gene expression profiling at early time points accurately predicted non-genotoxic carcinogenesis and hepatocarcinogenicity [5,6]. Toxicogenomics was also of relevance to evaluate the potential immunotoxicity of small interfering RNAs (siRNAs) considered for potential therapeutic application [7].

Compounds inducing similar gene expression profiles to known model toxicants can be identified as putatively toxic based on the common mechanisms of response at the molecular level. Nonetheless,

to develop such kind of profiling, access to large and consistent toxicogenomic repositories in conjunction with toxicological outcomes are required. One of the limitations is standardization and consistency across experimental settings, data format and metadata description originating from separate studies. Mostly efforts where a specific scope and rules governing consistency were defined, be it a disease such as cancer or a specific type of cancer, managed to successfully propose specific new knowledge out of the combined data. Within less specific contexts for merging microarray data co-expression of transcripts, giving indications about transcriptional networks in general, is mostly what can be achieved.

Toxicogenomics Initiatives

Among the initiatives in which large toxicogenomics reference data has been generated, DrugMatrix, the Toxicogenomics Projects Japan (TG-GATEs) and PredTox are the largest and most consistent databases that are now available (Table 1).

The DrugMatrix database, established by Iconix Biosciences and recently acquired by the National Toxicology Program, consists of gene expression responses in several tissues including liver, kidney, heart and primary hepatocytes of male Sprague-Dawley rats for over 630 known drugs and toxicants ingested at two or more doses and measured at different time points in triplicate [8,9]. Histopathology, blood chemistry and hematology data are also included with the gene expression data, allowing investigating the relation between the gene expression differentiations and the pathology.

TG-GATEs is a recent collaboration between the National Institute

---

of Health Sciences and 17 pharmaceutical companies in Japan [10]. In this initiative, male Sprague-Dawley rats have been exposed to 131 compounds at three dose levels in single dose experiments where samples were collected at 3, 6, 9 and 24 hours and also repeated dosing experiments where samples were collected up to 29 days. Microarrays have been performed on liver and kidney from *in vivo* experiments as well as *in vitro* hepatocytes from rats and humans. The data include information on histopathology, hematology and clinical chemistry.

In Europe, a collaborative project between pharmaceutical companies, Small and Medium Enterprises (SMEs), and universities, called the 'InnoMed PredTox' project, was performed under the EU Framework Program 6 [11,12]. In this project, 16 proprietary drug candidates that had been discontinued at certain stages of preclinical development due to toxicological findings in liver and/or kidney in 2 to 4-week systemic rat studies were selected for study. Each of the compounds was tested in a 2-week systemic study at a low dose and a high dose using male Wistar rats. For all animals, clinical observations, serum, plasma, blood as well as liver and kidney tissues were collected and analyzed with transcriptomics, proteomics and metabolomics approaches. All the raw data are available at the BioInvestigationIndex (BII) site (www.ebi.ac.uk/bioinvindex/browse_studies.seam).

An analysis of the compounds studied in these three projects shows that there is now publicly available data on 705 distinct compounds. The overlap is shown in Figure 2. Although most of the compounds analyzed are approved drugs by the Federal Drug Administration (FDA), some environmental compounds and natural products have also been studied. Interestingly, a major set of compounds evaluated in TG-GATEs has also been studied in DrugMatrix allowing comparing the reproducibility of the outcome from both studies. In addition, it is possible to integrate gene expression profiles in other tissues (heart,

muscle, bone marrow, spleen, brain and intestine) offering a larger systemic view of the potential toxicity of a molecule.

Data from other microarray experiments involving compound-treatments are also available in ArrayExpress/GEO, but a) the lack of common protocols will make combining and interpreting the data challenging, and b) the chemical indexing of these resources remains a challenge [13].

Data Processing and Analysis

One of the challenges in Toxicogenomics is to find subsets of biomolecules within large genomics data sets that have an obvious meaning. Widely used procedures to analyze transcriptomic data are Bioconductor [14], implemented in R, and DAVID (Database for Annotation, Visualization and Integrated Discovery), a web accessible tool for the interpretation of genome-scale datasets, including those derived from microarrays. DAVID provides exploratory visualization tools that promote discovery through the functional annotation of gene lists [15]. Since, a lot of computational biological methods have been developed and reported for prioritizing candidate genes [16]. Enrichment analyses to a gene/protein network and integration of pathway-level analyses have become important tools for the interpretation of data from transcriptomics [17,18]. For example, a tool like GenMAPP integrates several biological pathways relevant for rat and human toxicity [19]. PINTA is another web resource for the prioritization of candidate genes based on the differential expression of their neighborhood in a genome-wide protein-protein interaction network [20]. Finally, a Predictive Power Estimation Algorithm (PPEA) has been developed to facilitate genomic biomarker discovery for predictive toxicity and drug responses [21].

The application of toxicogenomics as a predictive tool for chemical risk assessment has been under evaluation by the toxicology
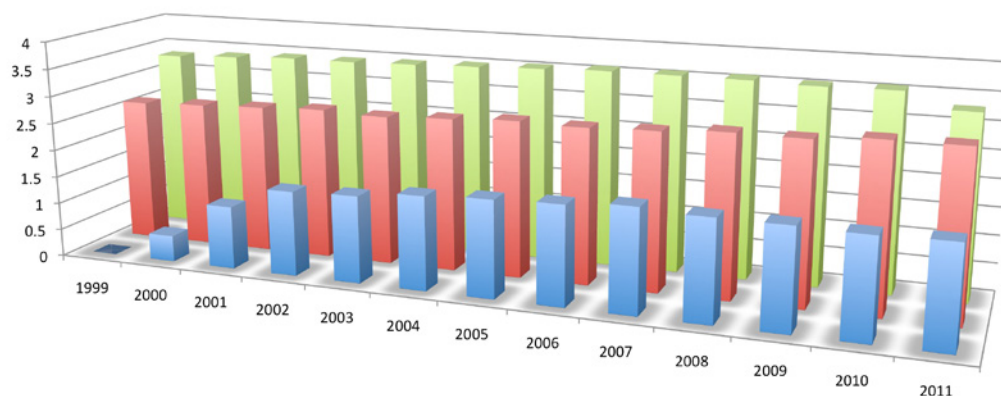

Figure 1: Number of papers in PubMed from 1999 to 2011 mentioning in title or abstract "toxicogenomics" in blue, "toxicology" in red and "gene expression" in green. The Y axis is converted into log (number of paper) for a better visualization.

Dataset	Compounds	Animals	Doses	Dosing	Sampling
Drug Matrix	657	Male S-D rats*	2	daily dosing up to 5 days	6 hrs and 1,3,5 days
TG-Gates	131	Male S-D rats*	3	daily dosing up to 29 days	3, 6, 9, 24 hrs and 4,8,15,29 days
PredTox	16	Male W Rats**	2	daily dosing up to 14 days	1,3,14 days

* S-D: Sprague-Dawley
**W: Wistar rat strains

Table 1: The 3 largest toxicogenomics initiatives.

Figure 2: Overlap of compounds analyzed in the three large toxicogenomics projects.
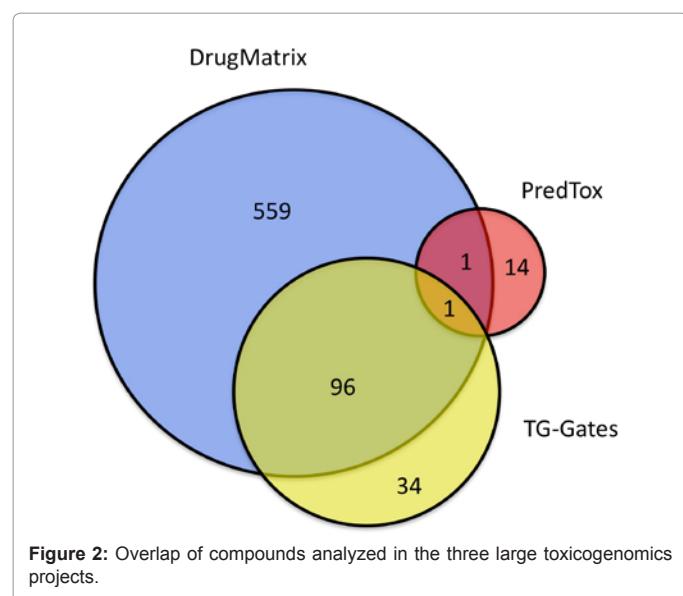
community for more than a decade following the MicroArray Quality Control (MAQC) guidance [22]. Recently, assessment of the level of inter- and intra- laboratory reproducibility between three independent laboratories on benzo[a]pyrene and the human hepatoma cell line HepG2 show that the current generation of microarray technology with a standard *in vitro* experimental design can produce robust and reproducible results [23].

eTOX Project

In 2010, a public-private partnership within the framework of the European Innovative Medicines Inititative (IMI), called eTOX, has been launched to gather relevant toxicity data from public sources (e.g. scientific literatures and non-proprietary databases) and to develop innovative strategies for the *in silico* prediction of the in vivo toxicities of drugs.

The final product should be implemented into a database with integrated and customizable software tools. More information is available at this link: http://www.etoxproject.eu/.

An interesting aspect of this project is the participation of 13 pharmaceutical companies who agreed to provide high quality *in vivo* toxicity data for a pull of compounds, which can be shared among the consortium. The primary focus of their data collections is the systemic toxicity studies (1-4 week repeated dose studies in rats, dogs and other non-rodents). The data collection is intended to be extended to longer studies, as well as to pharmacokinetic/toxicokinetic studies, *in vivo* safety pharmacology or gene expression data sets, during the course of the five years project. Currently, around 2000 reports are planned to be released to the consortium and more than 100 reports processed by contract research organizations (CROs) have been completed.

In parallel to data collection, strategies to improve the toxicity prediction are also currently investigated under eTOX. Development and application of large-scale Quantitative Structure-Activity Relationship (QSAR) models for the prediction of toxicity outcomes and off-target pharmacology such as the hERG K+ channel, metabolites enzymes and drug transporters and pure chemistry-related toxicology (e.g. cationic amphiphilic drug and phospholipidosis, mutagenicity) will be performed and analyzed. Development of multi-scales and
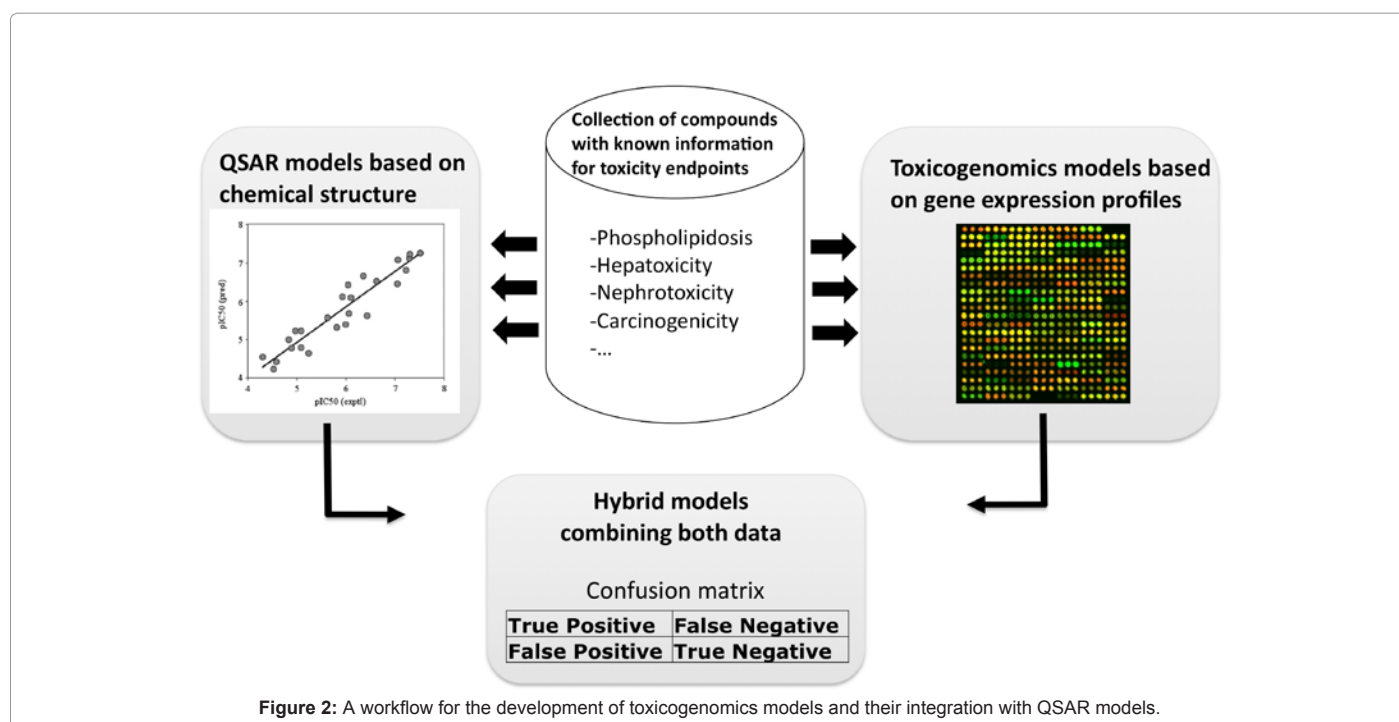
multi-levels modeling techniques to outperform the possible complex relationships existing between biomolecular processes and resulting toxicity outcomes will be explored.

The last objective is to understand the molecular mechanisms associated to *in vivo* toxicicty. To reach this goal incorporation of omics data and cross-omics mapping is intended. As a start, it is planned, to integrate the large toxicogenomics datasets, previously mentioned, in a curated database accessible to the public and then to analyze *in-vivo* (and *in-vitro*) toxicological profiles. To our knowledge, comparison of the data studied in TG-GATEs and DrugMatrix has not been yet performed and an in-depth comparison will be one of the aims in the eTOX project to evaluate the reproducibility of the outcome. Taking advantage of the recent advances in this area, we will assess the transcriptomics data with the integration of biological pathways and gene enrichment and we will try to address the variations of the toxicological events observed in different species and evaluate the translation of toxicity findings across species. One of the challenges will be to predict the potential *in vivo* toxicological profile of a drug and to capture the underlying mechanistic events associated with toxicity. Within the eTOX consortium we will explore the possibility of integrating toxicogenomics data analysis with the more classical QSAR modeling. Recently, it was reported that models from toxicogenomics data on non-genotoxic hepatocarcinogenicity outperformed QSAR model on the same set of compounds [24]. In addition, Low Y. et al. showed that hybrid models combining both chemical descriptors and gene expression profiles could be useful for the interpretation of drug-induced hepatotoxicity [25]. Therefore, in parallel to the QSAR models that will be developed by the eTOX project for several toxicity endpoints (phospholipidosis, hepatotoxicity, nephrotoxicity, carcinogenicity, mutagenicity), toxicogenomics models will be investigated with the same set of available data and combined when possible into a hybrid model. A workflow is depicted in Figure 3. Depending on the biological information associated with a compound, it should be possible to predict the potential toxicity (or non toxicity) in different ways. For example, if a transcriptomic experiment has been performed on a chemical without toxicological information, the potential toxicity can be suggested on the basis of similar gene expression profiles to those of known toxicants. Additionally, prediction can be made through QSAR approaches using molecular descriptors. Using QSAR models in combination with the gene expression profiles will result in hybrid models more predictive and with better interpretation than simple QSAR models.

Ultimately, based on structural similarity or structural alerts, mechanism of action can be proposed, although the accuracy of such procedure is still not optimal. It is expected with the integration of unpublished data from pharmaceutical companies to move towards a predictive and reliable modeling of the complex relationships existing between in vivo observations of the toxicity and safety of drug candidates.

Perspectives

The combination of transcriptomics, proteomics and metabolomics with conventional toxicology approaches has been shown to be useful for mechanistic investigations and the identification of putative biomarkers [8]. In addition, with the advance in Next Generation Sequencing (NGS) technologies, it is now possible to decode an entire human transcriptome, making RNA sequencing a feasible way of obtaining global transcriptome information with reduced time and cost [26]. NGS will significantly accelerate genomic research and

Figure 2: A workflow for the development of toxicogenomics models and their integration with QSAR models.

discovery with a potential contribution on personalized medicine [27]. By providing the means to sequence up to few human genomes in a single run with an high-end sequencer such as currently an HiSeq for Illumina or a SOLiD 5500 from Life Technologies, recent machines have opened the door to the systematic exploration of mutations and epigenetic patterns. Recently a comparison of the NGS and microarray technologies on toxicogenomics data was performed on aristolochic acid (a nephrotoxic compound) [28]. Although RNA-seq was more sensitive in detecting genes with low expression levels, similar gene expression patterns were observed for both platforms and encouragingly the biological interpretation was largely consistent between the RNA-Seq and microarray data. However, it is too early to estimate if such data will be available in large enough quantities to be of value to integrate in the eTOX project. In the context of personalized medicine we are only at the beginning and the toxicological component will play an obvious and important role: as much as the efficacy of compounds will vary across individuals according to specific genetic backgrounds the toxicity and side effects of compounds will also vary according to genetics. The best treatment for a given individual will be the drug that offers the best efficacy for the lowest amount of side effects.

Overall, integrating toxicogenomics in combination with other "omics" and other sequencing of mammalian genomes should open the development of new approaches for the understanding of toxicity as it might be affected by genetic variability.

Acknowledgement

References

1. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, et al. (2010) How to improve R&D productivity: the pharmaceutical industry´s grand challenge. Nat Rev Drug Discov 9: 203-214.

2. Stierum R, Heijne W, Kienhuis A, van Ommen B, Groten J (2005) Toxicogenomics concepts and applications to study hepatic effects of food additives and chemicals. Tox Appl Pharmacol 207: 179-188.

3. Afshari CA, Hamadeh HK, Bushel PR (2011) The evolution of bioinformatics in toxicology: advancing toxicogenomics. Toxicol Sci 120: 225-237.

4. Pognan F (2007) Toxicogenomics applied to predictive and exploratory toxicology for the safety assessment of new chemical entities: a long road with deep potholes. Prog Drug Res 64: 219-238.

5. Nie AY, McMillian M, Parker JB, Leone A, Bryant S, et al. (2006) Predictive toxicogenomics approaches reveal underlying molecular mechanisms of nongenotoxic carcinogenicity. Mol Carcinog 45: 914-933.

6. Ellinger-Ziegelbauer H, Gmuender H, Bandenburg A, Ahr HJ (2008) Prediction of carcinogenic potential of rat hepatocarcinogens using toxicogenomics analysis of short-term in vivo studies. Mutat Res 637: 23-39.

7. Jackson AL, Linsley PS (2010) Recognizing and avoiding siRNA oof-target effects for target identification and therapeutic application. Nat Rev Drug Discov 9: 57-67.

8. Ganter B, Tugendreich S, Pearson CI, Ayanoglu E, Baumhueter S, et al. (2005) Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. J Biotechnol 119: 219-244.

9. Natsoulis G, Pearson CI, Gollub J, Eynon B, Ferng J, et al. (2008) The liver pharmacological and xenobiotic gene response repertoire. Mol Syst Biol 4: 175.

10. Uehara T, Ono A, Maruyama T, Kato I, Yamada H, et al. (2010) The Japanese toxicogenomics project: application of toxicogenomics. Mol Nutr Food Res 54: 218-227.

11. Suter L, Schroeder S, Meyer K, Gautier JC, Amberg A, et al. (2011) EU framework 6 project: Predictive toxicology (PredTox) – overview and outcome. Toxicol Appl Pharmacol 252: 73-84.

12. Biotier E, Amberg A, Barbié V, Blichenberg A, Brandenburg A, et al. (2011) A comparative integrated transcript analysis and functional characterization of differential mechanisms for induction of liver hypertrophy in the rat. Toxicol Appl Pharmacol 252: 85-96.

13. Williams-Devane CR, Wolf MA, Richard AM (2009) Toward a public

toxicogenomics capability for supporting predictive toxicology: survey of current resources and chemical indexing of experiments I GEO and ArrayExpress. Toxicol Sci 109: 358-371.

14. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: 80.

15. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44-57.

16. Tranchevent FB, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, et al. (2010) A guide to web tools to prioritize candidate genes. Brief Bioinform 12: 22-32.

17. Riedel RF, Porrello A, Pontzer E, Chenette EJ, Hsu DS, et al. (2008) A genomic approach to identify molecular pathways associated with chemotherapy resistance. Mol Cancer Therap 7: 3141-3149.

18. Kamburov A, Cavill R, Ebbels TM, Herwig R, Keun HC (2011) Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. Bioinformatics 27: 2917-2918.

19. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet 31: 19-20.

20. Nitsch D, Tranchevent LC, Goncalves JP, Vogt JK, Madeira SC, et al. (2011) PINTA: a web server for network-based gene prioritization from expression data. Nucleic Acids Res 39: 334-338.

21. Liu J, Jolly RA, Smith AT, Searfoss GH, Goldstein KM, et al. (2011) Predictive power estimation algorithm (PPEA) – a new algorithm to reduce overfitting for genomic biomarker discovery. PLoS One 6: e24233.

22. Severnini M, Bicciato S, Mangano E, Scarlatti F, Mezzelani A, et al. (2006) Strategies for comparing gene expression profiles from different microarray platforms: Application to a case-control experiment. Anal BioChem 353: 43-56.

23. Scott DJ, Devonshire AS, Adeleye YA, Schutte ME, Rodrigues MR, et al. (2011) Inter- and intra laboratory study to determine the reproductivity of toxicogenomics datasets. Toxicology 290: 50-58.

24. Liu Z, Kelly R, Fang H, Ding D, Tong W (2011) Comparative analysis of predictive models for nongenotoxic hepatocarcinogenicity using both toxicogenenomics and quantitative structure-activity relationships. Chem Res Toxicol 24: 1062-1070.

25. Low Y, Uehara T, Minowa Y, Yamada H, Ohno Y, et al. (2011) Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. Chem Res Toxicol 24: 1251-1262.

26. Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 8: 469-477.

27. Ruden DM (2011) Frontiers in toxicogenomics – the grand challenge: to understand how the genome and epigenome interact with the toxic environment. Frontiers in genetics 2: 1-2.

28. Su Z, Li Z, Chen T, Li QZ, Fang H, et al. (2011) Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. Chem Res Toxicol 24: 1486-1493.