COT DISCUSSION PAPER: USE OF TOXICOGENOMICS IN TOXICOLOGY – DESIGN, ANALYSIS AND STATISTICAL ISSUES

COVER PAPER

Review background

1. In 2001, the COT/COM/COC held a Joint Symposium[1] to discuss issues relating to the use of genomics and proteomics in toxicology. A joint statement outlining the conclusions reached was subsequently published on the Committees websites [http://cot.food.gov.uk/pdfs/JointCOT-COM-COCStatement.PDF]. The Committees agreed to periodically review the literature to consider whether the conclusions made in 2001 needed revising. A series of COT discussion papers[2] followed resulting in a joint statement published in 2004 that updated the previous conclusions reached from the 2001 Joint Symposium [http://cot.food.gov.uk/pdfs/cotstatementtoxicogen0410.pdf].

2. The following general conclusions were made by the Committees as published in the 2004 Joint Statement:

a) *We recognise the rapid development in toxicogenomic methods (transcriptomics, proteomics and metabonomics) in toxicological hazard identification and characterisations since 2001.*
b) *We confirm that these techniques may serve as adjuncts to conventional toxicology studies. There is a need to provide appropriate data from studies on gene expression, protein levels and metabolite changes in order to provide sufficient information on toxicologically relevant pathways.*
c) *However, we consider that further research and validation is required before these techniques can be considered for routine regulatory toxicological risk assessment[3]. At present toxicogenomic approaches can provide valuable supportive data on mechanisms of target organ toxicity which can aid in the risk assessment process.*
d) *There is a need for further refinement and optimisation of methods used, approaches to data interpretation and evaluation using statistical and bioinformatics methods and development of appropriate publically accessible databases.*
e) *We note the need for generic guidance on the most suitable methods for statistical evaluation of different types of toxicogenomic data.*

[1] A full write up of the meeting was published by Barlow et al (2003). See references section.
[2] COT discussion papers: TOX/2003/08; TOX/2004/02; TOX/2004/26 and TOX/2004/27.
[3] A review paper by Battershill (2005) published in the Human & Experimental Toxicology provided a regulatory perspective of issues relating to proposed TGX applications, a possible approach to integrating TGX and conventional data into risk assessments, highlighting potential areas for future consideration.

3. The 2004 Joint Statement also includes conclusions reached by individual committees, and those published by the COT (following COT discussions at its February and September 2004 meetings) can be grouped under the following broad categories/themes[4]:

a) Study design/reproducibility, evaluation and statistical analysis of raw data
b) Pattern recognition, phenotypic anchorage and systems biology;
c) Target organ (time course/reversal), prediction (NOAEL), validation, in-house screening, regulatory submission;
d) Variation in transcriptomics, array design
e) High density vs. low density array design
f) Proteomic methods
g) Metabonomics/metabolomics, metabolite pattern changes, tracjectomes
h) Toxicogenomics integration into risk assessment
i) Animal use
j) Epidemiology and toxicogenomics
k) Database management and bioinformatics

Review aims, objectives and layout

4. This discussion paper (TOX/2010/18) focuses on issues represented by the above categories a), d) and e) as they relate to transcriptomic studies, i.e. the design, data and statistical evaluation of transcriptomic analyses. A literature search was conducted to retrieve useful reviews and original studies that report on relevant developments which can be used to update the previous COT conclusions. Studies derived from key organisations working in the field (e.g. ISLI-HESI, US EPA Toxcast Programme, NIEHS, and MGED Society), provide additional sources of authoritative information and were also considered in this review. In addition, discussions and conclusions arising from the COT 21st Century Toxicology Workshop held in February 2009 were used. This paper therefore provides an update to the previous statement by reviewing issues currently impacting on the design and reproducibility of transcriptomic (TRSX)-based toxicology studies, and the analysis and statistical evaluation of the resultant raw data.

5. The review uses the following COT 2004 Joint Statement conclusions, (and related previous discussion papers) to structure the layout of this paper.

a. *There had been improvements in the design and reproducibility of studies, and approaches to the analysis of raw data and statistical approaches to evaluation and identification of toxicologically relevant patterns for gene changes.*

[4] Conclusions represented by categories b), c), h), i) and j) principally address issues of relevance to regulatory risk assessment.

d. Regarding transcriptomic methods it was agreed that there were a considerable number of sources of variance which might affect the results of studies. The COT confirmed that for the present it was necessary to confirm key gene changes independently such as by quantitative PCR2 analysis of mRNA. The design of experiments (e.g. pooling of samples), reproducibility of replicate mRNA analyses, the approach to assessment of background changes, use of different fluorometric methods to assess gene expression changes, use of housekeeping genes, variation between laboratories regarding analysis of mRNAs in particular the use of different platforms, and validation of the genes incorporated into microarrays were all examples of the potential sources of variation in transcriptomic analyses.

e. There are few comparative data on the use of high density cDNA microarrays (e.g. with thousands of genes) and low density cDNA arrays (with small numbers of genes targeted for a limited number of toxic mechanisms). In general high density arrays are comparatively of greater difficulty and expense to develop and the evaluation and interpretation of data is complex. Low density arrays are cheaper, easier to evaluate, but may miss novel mechanisms and have limited coverage of genes.

6. Section 1 discusses issues relating to the design of transcriptomic-based toxicology studies (and includes consideration of microarray platform density); Section 2 considers the approaches used to analyse raw transcriptomic data while Section 3 focuses on the statistical approaches used to identify and evaluate toxicologically relevant gene expression changes. Issues relating to quality control and sources of variation of transcriptomic-based analyses (which permeate all levels of a TGX study i.e. sample preparation, data generation and data analysis stages) are addressed as separate topics in Sections 4 and 5 respectively. Issues relating to the reproducibility of TGX data will be discussed in a subsequent paper at the next COT meeting in September 2010.

7. Annex 1 contains details of the literature search strategy used. Narrative summaries/abstracts of selected key reviews and original studies are provided in Annex 3. Members should note that TOX/2010/18 is not based on a comprehensive systematic review of the literature due to the extensive amount of work published in the field rendering such an approach unfeasible. TOX/2010/18 is a discussion paper that uses selected review papers, original studies and work published from key organisations to address issues previously raised and reveal current developments and associated challenges. In doing so, it is hoped this will further advance the integration of toxicogenomics into toxicology.

EXECUTIVE SUMMARY

I. Design of transcriptomic-based toxicology studies

1. Issues relevant to considering the design of transcriptomic-based toxicology studies are those relating to the generation of accurate, precise and reliable data, and the approaches undertaken to minimise the effect of factors altering the ability to answer the question of interest as clearly and efficiently as possible. Deciding a prior objective is considered particularly important as it impacts on the subsequent design. These experimental objectives have been categorised by NAS (2007ab) as either class comparison, prediction or discovery (paragraphs 4-6). Selecting appropriate animal species or cell lines constitutes a key consideration to minimise false positives and maximise true positives (paragraph 7).

2. Reflecting interest in using species from lower phyla, the zebra-fish has been investigated as a potential alternative to mammalian models in several studies (paragraphs 8-9).

3. Undertaking preliminary calculations of appropriate sample size is essential to ensure studies have sufficient power to detect regulated genes at acceptable cost. Various factors must be considered and several methods are available to calculate the most appropriate size including Wilks' lamda score F-test, and the Microarray Power Atlas reported by Page et al (2006) (paragraphs 10-15).

4. Deciding dose and time points for studies represents a significant design consideration. The value of multiple dose groups and appropriate time course experiments is described in relation to elucidating the shape of the dose response curve at low levels of exposure and establishing relationships with downstream changes, respectively (paragraphs 16-20).

5. The sampling of RNA must be carefully considered to avoid introducing bias and various approaches are described i.e. in-vivo approaches using whole or regional tissue, as well as the use of laser capture microscopic techniques to extract single cells (paragraphs 22-31). Studies either support or criticise the use of in-vitro approaches in gene expression profiling and these are described in paragraphs 33-36. Other issues addressed include the use of peripheral blood (paragraphs 29-31) and information arising from activities of the International Life Sciences Institute (ILSI) Health and Environmental Sciences Institute (HESI) and the European Centre for the Validation of Alternative Methods (ECVAM) regarding the use of in-vitro toxicogenomic (TGX) studies (paragraphs 37-38). The conclusions from the COT workshop on 21st Century Toxicology are summarised in relation to the application of large scale non- in-vivo TGX studies in chemical risk assessment (paragraphs 40-46) with further information of two FP7 Programme projects incorporating TGX methods (paragraphs 47-49).

6. Various design issues relating to microarray technology are noted including the advantages and disadvantages of the two main microarray

platforms used (paragraphs 50-57); platform density (paragraph 58); the significance of microarray gene target selection to characterise a toxic response (paragraph 59); and the advantages and disadvantages of using single or double channel microarrays (paragraphs 60-62). The overall limitations of the technology in relation to its inability to detect gene expression changes at levels of environmental exposure , the lack of correlation between platforms and, more importantly, the lack of correlation with downstream (proteomic) changes (paragraphs 63-68) are also considered. A summary of the use of alternatives to microarray technology as reviewed by Gant (2007) is provided, with particular attention on the mRNA translation assay (paragraphs 69-73). The remaining design issues discussed include the need to consider the type of hybridisation approach to fit the experimental objective e.g. direct, loop or reference design (paragraph 74), the various sources of bias (paragraphs 75-79) and the approaches used to minimise their effects e.g. blocking and randomisation (paragraph 80), and the use of replicates (paragraphs 81-85). Finally, whether or not to pool samples and other recommendations are provided in paragraphs 86-88.

II. Low level data analysis of raw transcriptomic data

7. The analysis of transcriptomic data involves two distinct stages. The first stage involves analysis of the raw transcriptomic data and is referred to as data processing, pre-processing or low-level data analysis. Data processing principally aims to correct data sets for sources of variability arising from random and systematic error during the experimental procedure. Various factors must be considered, which include the choice of scanning approach (which can optimise data acquisition) (paragraphs 93-95); the image analysis protocol for either spotted cDNA or oligonucleotide microarray; (paragraphs 96-102); and the various methods available to sort poor, i.e. uninformative, data from potentially useful data e.g. via visual representation or using quality metrics (such as Percent present) and Pearson's correlation (paragraphs 103-105).

8. Transforming the data to correct for background noise represents a critical preprocessing step and can be done either via log transformation (paragraphs 106-107) or normalisation, for which various methods and approaches are available which largely depend on platform type used (paragraphs 108-116). Filtering represents a vital pre-processing step to remove unreliable data prior to analysis and is discussed in paragraph 118.

III. High-level data analysis (statistical and computational approaches)

9. The second stage of data analysis (also referred to as high-level data analysis) refers to the statistical and computational manipulation of transformed data. There are two key objectives (i) to identify significant gene expression changes (hypothesis testing) and (ii) evaluate toxicologically relevant patterns (data mining).

10. Hypothesis testing requires that investigators first decide on the level of gene analysis to perform i.e. whether to make single or multiple gene

comparisons (paragraphs 123-126). Other considerations include the approach used to adjust for multiple testing e.g. calculating the family wise error rate (FWER) or controlling the false discovery rate (FDR) (paragraphs 137-139) and whether to use threshold or statistical based approaches to select differentially-expressed genes (DEGs) (paragraphs 127-136). The outcome of hypothesis testing is a list of genes, changes in which are associated with the condition being tested. It should be noted that the outcome of such studies is often more hypothesis generation than hypothesis testing.

11. The approaches used in data mining depend on the experimental hypothesis/ type of study e.g. class prediction or class discovery, and involve either supervised or unsupervised approaches (or both). The statistical/computational methods used in class prediction are typically supervised and involve the application of a classifier (the type used depending on the level of gene analysis) e.g. K-Nearest Neighbour for individual gene analysis and Support Vector Machines for multiset analysis (paragraphs 140-145). Class discovery studies use unsupervised methods such as principal component analysis (PCA) for individual gene analysis (paragraphs 146-147), or the various clustering algorithms for multiple gene analysis (paragraphs 148-156). The various software packages used in data mining are briefly summarised in paragraphs 158-159.

12. Approaches used to validate the observed differentially regulated genes are discussed in relation to quantifying gene expression via real time RT-PCR (paragraphs 162-169), or interpreting their biological significance via post-analytical approaches (paragraph 170). These include structural gene annotation (paragraphs 171-173) followed by the identification of pathways and networks overrepresented in a given gene list via pathway and network analysis (paragraphs 174-185). The quality of databases storing TGX data impacts on data interpretation and various issues relating to management of the databases (paragraphs 186-190) and their role in supporting data comparison (paragraphs 191-193) and standardisation of TGX protocols (paragraphs 194-198) are discussed.

IV. Quality control and sources of variation

13. The final two sections of the paper discuss issues of relevance to both design and analysis of TGX data. The application of quality control (QC) measures before hybridisation (paragraphs 200-207) or after hybridisation (paragraphs 208-222) is considered. Pre-hybridisation QC measures relate to the various approaches used to assess RNA quality and target preparation. Post hybridisation measures can be based on either assessment of individual spots (which links to image analysis data quality assessments) (paragraphs 208-210), within individual chips/hybridisations (based on a selection of spots e.g. housekeeping genes or spike in controls (paragraphs 211-216), or between chips e.g. by comparing intensities and expression ratios across chips (paragraphs 217-220). Issues relating to the validation of TGX platforms are discussed in paragraph 223-224. Finally, the different sources of variation in TGX data are described (paragraphs 225-227) including the approaches

that have been used to identify and characterise them (paragraphs 228-233) and the implications for reproducibility (paragraph 234).

14. Members are reminded that the COT secretariat plans to discuss issues relating to the reproducibility of TGX data as a separate discussion paper at the next COT meeting in September 2010. This will include consideration of factors affecting reproducibility (i.e. specific aspects of TGX design and analysis that either enhance or impair reproducibility); comparative studies (e.g. cross platform correlation studies); the MicroArray Quality Control (MAQC) Project (evaluation of inter and intra-platform reproducibility); and the findings from inter-laboratory studies.

QUESTIONS FOR THE COMMITTEE

15. Members are asked to comment on the updated information in relation to the design, analysis and statistics of TGX studies.

CONTENTS

SECTION 1: DESIGN OF TRANSCRIPTOMIC-BASED TOXICOLOGY STUDIES

INTRODUCTION

1. As with any experimental investigation, design issues of transcriptomic-based toxicology studies principally relate to those aspects necessary to ensure the generation of accurate, precise and reliable data, and approaches undertaken to minimise the effect of factors altering the ability to answer the question of interest as clearly and efficiently as possible. These issues are extensively discussed in the published literature, and the most commonly raised topics are outlined below.

2. Various approaches are undertaken to minimise/avoid introducing bias in the data and improve the power of studies to detect genes differentially expressed between treated and control samples. These include the use of replication, randomisation and blocking. The identification and estimation of sources of variability and the approaches used to reduce their effects are considered separately in section 5.

3. Members may wish to note that Dr Richard S. Paules, who heads the Environmental Stress and Cancer Group and the NIEHS Microarray Group within the Laboratory of Toxicology and Pharmacology recently gave an award lecture at the 2010 SOT Conference in Salt Lake City, Utah, U.S. discussing TGX at the NIEHS and how it is impacting on toxicology. Paules noted (in a summary abstract) that although technical problems associated with signals, the bioinformatic determination of significant changes and reliability across platforms and different users have been addressed, there is still a critical need to address the more complex and significant issues of the appropriate experimental design and interpreting the vast amounts of information produced in TGX studies.

A. EXPERIMENTAL OBJECTIVES

4. Outlining the main goal of the study in advance is routinely stated in the literature as a prerequisite to producing well designed investigations (Lee et al 2005). This is because the design must reflect the objective and the practical constraints of the experiment being done, so that the most appropriate procedures and methods are selected and applied to answer the research question. Typical constraints include sample numbers and availability and cost which determine the number of slides used and ultimately the reliability of the data produced. Research objectives in TGX microarray studies typically aim to delineate mechanisms, classify toxicants, predict toxic endpoints and identify biomarkers.

5. The National Research Council (NRC) Committee on the Validation of Toxicogenomic Technologies held a workshop in 2005 to consider current practices and advances. A summary was published and various fundamental topics were covered including the categorisation of experimental objectives

into class comparison, class prediction and class discovery types (NAS, 2007a).

6. The term class comparison is used to describe studies that seek to identify a list of genes that are differentially expressed among predefined classes of samples e.g. for comparing control and test samples exposed to a particular toxicant. Such approaches can be used to identify mechanisms of action of particular toxicants. In class prediction studies the objective is to develop a method capable of predicting whether a sample belongs to a particular predefined class by way of gene expression data. It therefore has a similar setup to class comparison in that there is a prespecified group. Such an approach requires use of predefined classes to develop the predictive method i.e. samples treated with a particular toxicant to produce a known gene expression profile. In class discovery studies, there are no predefined classes, which rather are constructed during the course of data analysis. The researchers are interested in finding some sort of structure in the data set. Either genes or samples can be classified.

B. ANIMAL SPECIES

(i) Key issues

7. The type of animal species used in *in-vivo* based transcript profiling toxicology studies constitutes a key design consideration (Lee et al 2005). Species differences in xenobiotic metabolising enzymes require that the most appropriate species is selected to minimise false positives and maximise true positives. However, TGX-based approaches could potentially overcome the species-specific differences that typically complicate safety/risk assessment, particularly in view of the findings of a study that sought to determine whether comprehensive gene expression data from rat in-vivo/in-vitro and human in-vitro systems could explain the well-known species-specific difference in the toxicity of coumarin (Uehara et al., 2008). The authors reported that the overall responsiveness of genes identified (relating to glutathione metabolism and oxidative stress) were, as expected, much higher in rats than in humans. Mammalian animal models, such as rodents, are typically used although the use of non-mammalian alternatives such as the zebrafish (*Danio rerio*) is the subject of increasing debate. The use of zebrafish embryo models in toxicogenomics is particularly pertinent to regulatory risk assessment initiatives seeking to find alternatives to reduce the number of higher-phyla animals currently used in toxicity testing, which is discussed further in this section.

(ii) Zebrafish (*Danio rerio*)

8. The use of zebrafish in toxicity assessment was discussed in a symposium organised as part of the British Toxicology Society (BTS) Annual Meeting in 2008. Issues raised included the similarity of zebrafish tissues to their mammalian counterparts and the applicability of zebrafish as a model organism, although their use as a frontloaded screen rather than a replacement for current regulatory models was emphasised. The session also

discussed the methodologies being developed to produce a high throughput approach to screening using zebrafish larvae, and an example was provided of a study demonstrating how gene expression profiling of zebrafish exposed to TCDD enabled identification of the Ahr2 gene as a mediator of TCDD toxicity.

9. A scan of the published literature identified two useful reviews and several studies using zebrafish models in transcriptome analysis[5]. Ju et al (2007) discussed the use of gene expression profiling in relation to its application to aquatic model research. Key issues noted included the lack of Affymetrix chips bearing probes (gene targets) for zebrafish due to the limited use of the species rendering probe synthesis for zebrafish genechips uneconomical; spotted arrays are therefore typically used in zebrafish transcriptome analyses. A review by Scholz et al (2008) presented examples of the use of zebrafish embryos to study the effect of chemicals on mRNA (and protein) patterns and the potential implications of differential expression for toxicity. The authors considered zebrafish embryos excellent models for studies aimed at understanding toxic mechanisms and identifying possible adverse and chronic effects. Lam et al (2008) performed expression-based chemogenomics[6] on adult zebrafish (using PAHs and oestrogenic compounds). Knowledge-based data mining of human homologs of zebrafish genes revealed highly conserved chemical-induced biological response/effects, health risks and novel biological insights associated with the aryl hydrocarbon receptor and oestrogen receptor, from which relevance to humans could be inferred. This led them to conclude that zebrafish were in a strategic position to bridge the gap between cell-based and rodent models in chemogenomics research and applications. Wahl et al (2008) analysed the effects of an abundant polybrominated diphenyl ether (PBDE) congener on AhR activity and signalling and noted changes in gene expression and toxicity similar to those with known AhR agonists. Usenko et al (2008) used zebrafish embryos as a model organism to confirm the potential of the nanomaterial fullerene C60to elicit oxidative stress responses. Evidence for the applicability of zebrafish to integrative toxicogenomic approaches was provided in a study by De Wit et al (2008) who exposed two adult zebrafish populations for 14 days to 0.75 and 1.5 uM TBBPA (a frequently used HPV brominated flame retardant) and employed a combined transcriptomic and proteomic approach to evaluate molecular hepatotoxic effects. Gene expression findings (confirmed by RT-PCR) enabled the authors to hypothesize several working mechanisms of TBBPA thereby demonstrating the potential of a combined genomic and proteomic approach to generating detailed mechanistic toxicological information. Finally, given the limited knowledge on the action of manmade chemicals on vertebrate development, Yang et al (2007) exposed zebrafish embryos to a range of environmental toxicants to determine whether distinct chemicals would induce specific transcriptional profiles. A barcode-like response was observed in what has been considered the most comprehensive report on zebrafish embryo toxicogenomics to date, in which

[5] Most zebrafish studies relate to its application to ecotoxicology and developmental issues.
[6] Chemogenomics is defined as the study of genomic responses to chemical compounds.

11 distinct compound and stage specific transcription patterns were identified for embryos exposed for 24 h at different time frames (Scholz et al 2008).

C. SAMPLE SIZE

10. Various definitions of sample size exist in the published literature e.g. sample size has been used to refer to biological replicates, and to the total number of microarray slides or individuals (Lee et al., 2005). Choosing the optimal number of biological replicates is a critical step in the design of a TGX experiment as the larger the size the more reliable the results, although the more expensive it becomes. Therefore, calculating the most appropriate sample size is essential – the most appropriate being a size that maximises the scientific information at minimal cost i.e. the smallest sample size that still provides sufficient power to recognise genes regulated at a specified level, while controlling the false discovery rate (FDR) at an acceptable level (Elashoff et al 2008).

11. Calculation of sample size is a complex process as it requires consideration of: the magnitude of the variability of the population (gene expression levels); magnitude of effect (i.e. expression changes that are biologically meaningful or desirable to detect); acceptable false positive/discovery rate (FDR) (e.g. 0.05); and the desired power to detect differences (e.g. to detect only the 10% most-regulated genes requires a power = 0.1) (Lee et al 2005; Ahmed, 2006a; Page et al 2006; Jorstad et al 2007). This information is frequently derived from previous pilot studies performed by the research team or from similar data in the literature. However, it is also possible that relevant information is not available, for example the minimum change in the magnitude of expression of a given gene that is biologically meaningful.

12. Several statistical/ computational based methods are in use, for example power analysis can be used to estimate the minimum number of array samples for two colour multiclass discrimination (Ahmed, 2006a). The procedure employs the Wilks' lambda score (F-test) together with the leave-one-out cross validation to measure the proportion of variance, and Fischer discriminant analysis (FDA) to find linear combinations of discriminatory genes that characterise/separate samples. Sample size formulas are also available for class comparison studies which take into account how the relative sources of variation impact on the sample size requirement and how the design decisions (i.e. pooling, technical replicates and dye swaps) influence the costs associated with the arrays (Ahmed, 2006a; NAS, 2007a). For class prediction several guidelines and methodologies have been suggested, while size calculation for class discovery studies is considered more problematic (NAS, 2007a).

13. Wei et al (2004) reported that available studies in general used sample sizes that were too small to detect a 2-fold change with 90% probability and a p-value of 0.01 in humans. For experimental animals, Lee et al (2005) suggest that practically up to 10 inbred mice per group are required for treatment with toxic agent, while in humans a much larger number is needed

e.g. 85 individuals (Lampe et al., 2004). The reason why more subjects are needed for studies that involve humans (or any other outbred population) cf. studies that employ samples from an inbred population is because humans typically exhibit larger variability than those seen in experimental animals or cell cultures due to genetic influence on gene expression. Ahmed (2006a) reports that approximately 5 times as many humans are required relative to mouse samples to detect the same magnitude of change with the same statistical power at the same significance level for cDNA array data. To detect a twofold change with 90% probability and a p-value of 0.01 in humans requires at lease 20 samples in the 75% least-variable genes. This is much larger than the number of samples commonly employed in such studies.

14. Page et al (2006) developed the *Microarray PowerAtlas,* considered to be a valuable resource for estimating required sample sizes. The atlas enables investigators to build on previous studies that have similar experimental characteristics and also allows researchers to upload their own pilot data to derive power and sample size estimates. At the time of publication, estimates in the *PowerAtlas* were based on 632 experiments from Gene Expression Omnibus (GEO). These are regularly updated with new datasets from GEO and other databases. The authors comment that the use of *PowerAtlas* not only prevents investigators using too many samples in a group (which is cost inefficient) but also helps by eliminating costs arising from experiments that have too few replicates to have sufficient power to yield good results.

15. In a previous meeting, Members commented on the potential value of TGX in reducing the number of animals used in toxicological tests. Initiatives to reduce animal numbers and the potential role of TGX are discussed further in paragraphs 38-48. NB. Such issues are not driven solely by financial concerns but also by humane considerations and the biological relevance of toxicity testing in animals.

D. DOSE AND TIME-POINTS

(i) Dose dependent analysis

16. The dose level/treatment regime used in TGX studies constitutes another key design consideration, particularly with regard to their potential application in risk assessment. Incorporating multiple dose groups enables toxicity thresholds to be determined (Lee et al 2005). Multiple doses in the low dose range may enable detection of small effects (if they exist) at low levels of exposure i.e. lower than the lowest observed effect level for traditional toxicity endpoints, which thereby makes it possible to address questions about the shape of the dose-response curve at these lower exposure levels. Studies using inappropriate doses can limit the value of their data, as the gene expression changes detected in those that use only high doses may relate mainly to overt cytotoxic mechanisms. Conversely, too low a dose may result in no discernable toxicologically relevant gene signatures being detected, an outcome that might be considered to be cost ineffective. However, this very much depends on the objective of the study.

17. Clearly, the use of multiple dose levels can yield more reliable and relevant data but such experiments are neither simple nor inexpensive. Dose response gene expression studies that integrate other analyses (such as histopathology or proteomics) should incorporate not only multiple dose levels but also multiple time points to accommodate any downstream related changes. Furthermore, the number of animals at each dose and time point must be sufficient to provide reasonable statistical power. Andersen et al (2008) provide a good illustration of the use of global gene expression analysis to support conclusions about dose-dependent transitions[7] in toxic responses, in this case the response of the rodent nasal epithelium to formaldehyde (tumourigenic). The study included both time (6h inhalation per day, 5 days/week for up to 3 weeks) and dose (0-15 ppm) dimensions and the two lowest doses (0.7 and 2 ppm) did not produce discernable effects at the histological or macroscopic level at any time point. Furthermore, gene expression data provided confirmation that the responses were not linear at low doses.

18. Burgoon & Zacharewski (2008) developed an automated application called ToxResponse Modeler that can be used to analyse any large dose response data set, such as that generated in a TGX study. The ToxResponse Modeler uses an automated process capable of large scale modelling and model selection to streamline analyses and point of departure calculations across hundreds of responses. The authors propose that the application could be used to assist in the ranking and prioritisation of compounds that warrant further investigation and development.

(ii) Temporal analysis

19. Temporal analyses of gene expression changes provides further insight into the key biochemical mechanisms and processes associated with toxicant exposure (particularly chronic exposures) (Lee et al 2005). Incorporating a time course into an experiment enables researchers to determine not only genes involved in early/adaptive responses but those involved in the elicitation and progression of adverse effects. The use of inappropriate time points can preclude association with other downstream changes thereby hindering data interpretation and phenotypic anchorage of toxicologically relevant gene expression changes. Naciff et al (2007) evaluated the temporal response of the uterus to an oestrogenic stimulus (17 alpha-ethinyl estradiol – EE) and detected expression changes in families of genes responsible for eliciting each stage of the oestrus cycle prior to appearance of cellular and morphological changes in the tissue. Genes that control cell division and suppress apoptosis were expressed a few hours before the onset of measurable cell proliferation. Toxicologically relevant time- and dose-dependent gene expression profiles were identified in a study by Kwon et al (2008) who exposed mice to multiple doses of a bile duct-damaging chemical (4,4'-methylene dianiline – MDA) (10 or 100 mg/kg b.w.)

[7] Dose-dependent transitions are described by Daston (2008) as inflection points in the dose-response curve that occur when the concentration of the exogenous agent is sufficiently high to alter normal physiological function.

killed 6, 24 and 72 hours after treatment. Serum chemistry, histopathological and transcript profiling analyses were performed. Resultant bile duct cell injury followed by regeneration was associated with up and down regulation of various functionally defined and undefined genes, verified by RT-PCR. The authors were able to hypothesize that the chemokine-mediated Th1 pathway was involved in the inflammatory process, and Wnt/beta-catenin signalling pathways were responsible for the repair of the MDA-injured liver.

20. Morgan et al (2004) noted a branch of mathematics known as Fourier Analysis of Time Series that considers the interactive/changing (dynamic) nature of gene expression changes via time series experiments. It is thought such an approach provides a better understanding of a system's structure/behaviour in response to chemical exposure ultimately leading to improved study design and data that are more reliable.

E. RNA SAMPLING

21. The procedures used and tissues sourced to derive RNA for transcriptomic analysis must be carefully considered to avoid introducing bias into the results. Investigators conducting in-vivo studies must weigh the advantages against the limitations of sampling the whole organ, region or specific single cells from a chosen tissue. In-vitro studies, however, while not faced with tissue-handling considerations have drawbacks of their own.

(i) In-vivo approaches

1. Whole organ, regional tissues

22. Although potentially any organ can be used as a source of RNA, to date, the liver has been used most often for gene expression profiling experiments due to its involvement in the biotransformation of compounds to their toxic metabolites, ease of sampling and removal, and the high quality RNA yielded from liver tissue preparations (Irwin et al 2004). However, the drawback of using the whole liver (as with any whole organ) arises from the fact that any treatment response represents an average of all cells/locations (Morgan et al 2004). Spatial and regional issues abound due to the existence of zonal/lobular gene expression differences (i.e. genes coding for metabolic enzymes exhibiting a gradient of expression), and differential sensitivity to a toxicant's effects due to the mixed cell population[8] of the liver. Furthermore, there are also temporal issues due to the liver's dynamic transcriptome activity. Other organs[9] exhibiting similar zonation issues include the kidney.

23. Irwin et al (2004) suggests possible approaches to addressing zonation issues with liver samples, and highlights potential flaws in paracetamol studies using intra-peritoneal injection as a route of administration. These

[8] Cell types include hepatocytes, endothelial cells, Kuppfer cells, Ito cells (hepatic stellate cells) and biliary epithelial cells. NB. Sixty per cent of nuclei in the rat liver are hepatocytes yet all cells in a sample will contribute to the analysis after homogenisation).

[9] In fact, the liver (in comparison to other organs) is one of the more homogeneous tissues. Almost all other tissues have zonal problems.

authors have proposed that IP administration bathes the liver in solution of the compound resulting in a non-uniform exposure of liver lobes, producing inaccurate/ misleading gene expression patterns that do not reflect what would occur from oral administration.

24. RNA sourced from particular regions of a tissue would clearly yield more accurate data than whole organ based analyses and approaches such as laser microscopic dissection (or laser capture microscopy (LCM)) have been developed to isolate specific tissue regions. Plummer et al (2007) used LCM to isolate cells from either the interstitium or seminiferous cords of the fetal rat testes. LCM use requires that tissues first undergo staining to visualise cell boundaries and allow for sufficient structural orientation. However, this usually results in a considerable reduction in RNA content of dissected specimens. To circumvent this, Ruetze et al (2010) developed a modified hematoxylin/eosin staining protocol that allows concurrent visualisation of important structures and subsequent isolation of sufficient RNA for use in linear amplification and quantitative analyses.

2. Single cells

25. The ideal method of choice would be to conduct transcriptome analyses using specific cells of interest isolated from toxicant exposed tissues. Single cell gene expression profiling is increasingly being used as an approach to generating treatment responses that are specific to a particular cell/location.

26. Tietjen et al (2003) provides an early study example of attempts to monitor expression profiles of individual cells. Single mature and progenitor cells of the highly heterogeneous mammalian olfactory system were collected in an effort to generate mechanistic data on neuronal differentiation and diversification. Transcriptome data was generated following PCR amplification of synthesised cDNA. Retrospective PCR and Southern blot analysis was used to determine the identity and developmental stage of the cells. The authors identified a wealth of transcriptional differences between different cells and were able to define signal pathways expressed by individual progenitors at precise developmental stages.

27. Plummer et al (2007) used LCM to isolate specific cells (Leydig and Sertoli cells) from microscopic regions of sectioned fetal rat testes to determine in which compartment dibutyl phthalate (DBP) induced gene expression changes occurred. The procedure involved marking out the area to be dissected (with only microscopically clearly definable regions outlined) and cutting using a PALM UV laser microbeam. Samples were then pooled and RNA extracted for microarray analysis. Plummer et al (2009) were able to identify and localise DBP regulated genes to the Leydig cells and hypothesise a possible mechanism of action.

28. Roach et al (2009) exposed primary hepatoctyes to oxidative stress over multiple timepoints and dynamically monitored the responses for each cell by developing a microwell cytometry platform consisting of a

microfabricated in-vitro device with high density arrays of cell-sized microwells and custom software for automated image processing and data analysis. The cells were labelled with fluorescent probes that were sensitive to mitochondrial membrane potential and free radical generation. The authors found that the cytometry platform was able to provide a detailed picture of the heterogeneity present in cell responses to oxidative stress and concluded that it was a particularly useful tool for delineating issues with heterogeneity in cell populations .

3. Peripheral blood

29. The use of peripheral blood (PB) as a tissue source of RNA offers significant advantages to TGX research. As a non-invasive surrogate for inaccessible tissue, PB offers translational benefits to clinical settings; the rationale being that circulating blood might reflect physiological and pathological events occurring in different tissues in the body. Various techniques are used to prepare PB for gene expression analysis such as PAXgene (isolates RNA from whole blood), QIAamp (selectively lyses eythrocytes prior to RNA isolation), Ficoll-Hypaque (separates peripheral blood mononuclear cells prior to RNA isolation) (Thompson & Hackett, 2008).

30. A microarray study by Debey et al (2004) assessed whether blood isolation factors could affect gene expression analysis and thereby potentially bias the results. RNA was isolated from blood taken from human volunteers with a 20-24h delay in processing. The authors observed expression of hypoxia-related gene signatures and noted that these and change in the expression of other genes prevented the assessment of gene signatures of inter-individual variation. Gene expression patterns were also found to be dependent on the cell type chosen and RNA technique used. The authors recommended that TGX studies should aim to reduce the time to RNA isolation and consider the cell type prior to conducting the study.

31. Aside from the obvious limitations of using surrogate tissues, the predominance of globin mRNA from the reticulocyte population in total blood is reported to reduce the sensitivity (of detecting changes) when PB is used as an RNA source (Thompson & Hackett, 2008). A possible redress is the removal of blood components or the use of fractionated blood (for peripheral blood mononuclear cells which are the most transcriptionally active cell population in blood) although this can interfere with microarray data and potentially bias the results.

4. Archival tissue

32. Archiving tissue is considered to be particularly useful to enable subsequent gene expression studies. To preserve tissue for later RNA isolation, tissue fixative and processing methods are employed e.g. liquid nitrogen immersion, which has been reported to compromise RNA integrity (Thompson & Hackett, 2008). However, a study by Sumida et al (2007) provides evidence that tissues and RNA quality are well preserved even after

freezer storage for up to 2.5 years. Furthermore, LCM can be used to extract RNA from formalin-fixed paraffin embedded tissue although there is concern that this can mute lower fold changes in expression.

(ii) *In-vitro* approaches

1. Advantages and disadvantages

33. Aside from being cheaper than in-vivo testing and more readily automated (and thus faster), in-vitro based analyses provide opportunities for improved understanding of a transcriptomic response to toxicants. This is because of the complexity and detail of the studies that can be performed. For example, Morgan et al (2002) used cDNA microarrays to examine chemically induced alterations of gene expression to detect one or more selected mechanisms of toxicity in HepG2 cells exposed to a diverse group of toxicants. Gene expression correlated with morphological and biochemical indicators of toxicity and there was good correlation between biochemical measures of oxidative stress and transcriptional measures. However, there is concern that in-vitro studies are limited by the fact that the data do not directly compare to the results obtained in vivo (at least not on a gene-to-gene comparison basis). Indeed, questions about their relevance to in-vivo toxicity limit their use in the regulatory decision making process without additional supporting data.

34. A study conducted by Boess et al (2003) highlighted the problems of using in-vitro data. The authors characterised and compared the microarray gene expression of several in-vitro systems (i.e. rat cell lines, primary hepatocytes in conventional monolayer or sandwich culture and liver slices) with gene expression of whole liver tissue. Their study findings led them to conclude that primary in-vitro systems result in pronounced gene expression changes related to adaptation and de-differentiation. However, two studies identified in the published literature counter the concerns over the relevance of in-vitro to in-vivo-based analyses.

35. Suzuki et al (2008) examined the feasibility of screening for hepatotoxicity by an in-vitro gene expression analysis using rat primary hepatocytes and Affymetrix arrays. Hepatocytes were exposed to for 6 or 24 h to eight drugs with different mechanisms of hepatoxicity (i.e. paracetamol, cyclophosphamide, clofibrate, chlorpromazine, lithocholic acid, cisplatin, diclofenac and disulfiram) at one third of the cytotoxic concentration (TC_{50}). The types of transcriptional changes observed were generally consistent with previously reported in-vivo data to the extent that the authors were able to conclude that in-vitro gene expression analysis of hepatocytes provides a useful tool for evaluating the toxicological profile of drugs and in screening for the direct toxicity of drugs against hepatocytes.

36. Elferink et al (2008) analysed the effect of different model hepatotoxins (liposaccharide, paracetamol, carbon tetrachloride and gliotoxin) at the gene expression level in rat liver both in-vivo and in-vitro (via a precision cut liver slice model whereby all liver cell types are present in their natural

architecture). The authors consider this model particularly useful owing to the multi-cellular nature of toxicant-induced effects that involve hepatocytes and other cell types. The authors found that the in-vitro profiles of gene expression could predict the toxicity and pathology observed in vivo and concluded that the rat liver slice system could be used as an appropriate tool for the prediction of multi-cellular liver toxicity.

37. The HESI Committee on the Application of Genomics to Mechanism-based Risk Assessment conducted a cross-sector international online survey to assess the state of TGX and identify real and potential barriers to progress (Pettit et al 2010). From the 112 respondents, it appeared that in-vitro models (cell lines and primary cultures) were favoured more than whole organ or in-vivo approaches. It was suggested that this was due to the fact that their inherent simplicity renders data interpretation more straightforward. Also, given the cost of generating samples from in-vitro systems is substantially less than the costs required to generate samples from in-vivo systems, running inexpensive experiments to generate samples for use on expensive platforms is generally considered reasonable.

38. The European Centre for the Validation of Alternative Methods (ECVAM)[10] aims to promote the scientific and regulatory acceptance of non-animal tests (i.e. in-vitro, in-silico tests) through research, test development and validation, and a database service (http://ecvam.jrc.ec.europa.eu/). ECVAM considers TGX approaches as second generation of alternatives that can be applied in areas where no satisfactory alternatives exist e.g. carcinogenicity, endocrine disruption and chronic toxicity. Through workshops, task-force meetings and special symposia, experts review current status on specific topics and make recommendations about the best way to integrate in-vitro tests and alternative methods into the regulatory process. The first workshop on the topic, entitled 'Validation of TGX-Based Test Systems' held in Italy in 2003 (jointly organised by the Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM) and National Toxicology Program (NTP) Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM)), discussed and defined principles applicable to the validation of TGX platforms and specific toxicological test methods that incorporate TGX technologies (Corvi et al 2006). The validation of TGX platforms is discussed further in section 4.

39. A recent attempt to integrate in-vitro toxicology into regulatory risk assessment is provided by Muellner et al (2010) who conducted a focussed TGX analysis of the regulated disinfection by-product (DBP) bromoacetic acid (BAA)[11]. The authors aimed to determine how BAA regulates expression of genes involved in DNA damage/repair and toxic response in non transformed human cells. Transcriptome profiles for 168 genes with 30 min and 4h exposure times that did not produce cytotoxicity were generated, and the levels of 25 transcripts were significantly modulated following 30 min BAA

[10] ECVAM is part of the EC Joint Research Centre, Institute for Health & Consumer Protection

[11] DBPs arise from the application of current water disinfection methods and produce toxicologically relevant effects and thus represent an important class of environmentally hazardous chemicals with potential long term health implications (Muellner et al., 2010)

treatment (16 up/ 9 down). The majority of transcripts with altered profiles were from genes involved in DNA repair, especially repair of double stranded DNA breaks. The authors concluded that their study was the first TGX study in a nontransformed human cell of a regulated drinking water disinfection by-product, and implicated double strand DNA breaks as a consequence of BAA exposure.

2. Twenty-First Century Toxicology

40. Hazard screening typically involves in-vivo and in-vitro tests . The critical question is whether TGX can improve hazard screening by making these tests faster, more comprehensive, less reliant on higher order animals and more predictive and accurate without being prohibitively expensive (NAS, 2007b).

41. The COT held a workshop in February 2009 entitled *'21st Century Toxicology'* to discuss issues emerging from the 2007 report *'Toxicity Testing in the 21st Century: A Vision'* published by the National Research Council of the US National Academies (NRC, 2007). The report documented a 10-20 year strategy to develop and validate toxicological protocols that produced better science and reduced animal testing. The reasoning behind this initiative centres on the fact that although in-vivo testing covers many biological processes which in-vitro models cannot emulate, there are drawbacks and these relate to concern over the ethics of the scale of animal experimentation, the relevance of animal to human extrapolations, the cost of animal studies, and their inability to identify idiosyncratic human responses. The COT workshop noted the establishment of international efforts to meet these aims, there being a particular focus on predicting human in vivo responses to exposures of substances assessed under REACH and similar initiatives. It is envisioned by the NRC and others that this will be achieved via the development and validation of novel methods that can predict hazards, determine mechanisms and integrate data e.g. in-vitro (biochemical and cell-based assays), in-vivo (lower order organisms e.g. zebrafish, *C. elegans*) and computational models of biological systems. TGX was identified as a key approach to addressing the following issues[12]:

a). Prioritisation and prediction of environmental toxicity

42. This issue relates to the thousands of substances in use/produced for use for which there are inadequate toxicological data. Prioritising the safety assessment of substances that raise most potential concern is considered the most appropriate approach given the sheer volume of candidate chemicals to be assessed and the fact current in-vivo approaches are untenable.

[12] Other issues discussed in the workshop included the metabolic profiling strategies for characterisation of toxic mechanisms and a tiered approach for the use of non-testing methods in the regulatory assessment of chemicals. These are not addressed in this paper, as they do not specifically relate to transcriptomic profiling analyses.

43. In the United States, the Environment Protection Agency (EPA) ToxCast Program plans to do this via the application of computational toxicology, defined as “the integration of modern computing and information technology with molecular biology to improve Agency prioritisation of data requirements and risk assessment of chemicals”. Kavlock et al (2008) notes that the main difference between computational and traditional toxicology is scale with regards to the number of chemicals studied, the breadth of endpoints and pathways covered, the levels of biological organisation examined, the range of exposure conditions considered and the coverage of life stages, genders and species. ToxCast therefore involves conducting research to characterise hazards via high throughput assays that measure the impact of substances on various endpoints and then combining these results with a priori information to identify compounds likely to present greatest hazard/risk. The ToxCast Program also aims to predict hazards by screening for potential toxic in-vivo effects, without testing in animals, via predictive modelling. This would involve developing predictive models via a training set of substances well-characterised in-vivo (e.g. pesticides) and running poorly characterised or uncharacterised compounds through the models for comparisons to see if any similarities exist that are predictive of possible in-vivo effects. However, further validation of the training set is required with plans to include a further 300 or more data-rich chemicals and examination of animal to human extrapolations using known human toxicants. Nanomaterials would be used as a pilot before passing any other poorly characterised or uncharacterised substances through the models. It is envisioned that this will result in fewer animals being used in the process and a significant reduction in the assessment period.

44. Computational toxicology is being led on a global level by a collaboration between the secretariats of the International Programme on Chemical Safety (IPCS) and the Organisation for Economic Co-operation and Development (OECD). This collaboration arose from the need to prioritise co-ordination and exchange of information, given the rate of change in TGX and the scarcity of resources. Workshops organised in 2003 (in Germany) and 2004 (in Japan) provided a focus to work on various joint objectives including surveying available omics tools. A meeting in 2007 for the IPCS/OECD advisory group (open to representatives of other sectors) provided an open forum on computational toxicology.

b) The EU FP6/7 and other Europe-wide projects contributing to the vision

45. Questions over the accuracy, specificity and relevance of current carcinogenicity assessment methods and concerns about their cost, speed and use of animals have led to calls from EU REACH regulation and elsewhere for alternative approaches. TGX forms the basis of these alternative tests, particularly in relation to two initiatives: (a) the Carcinogenomics Project, launched under the 6th Framework Programme and presented as an in-vitro alternative to rodent bioassays; and (b) the Children’s Environment & Health Action Plan for Europe, which is applying omics based biomarkers to analyse perinatal exposures to carcinogenic agents following concern over environmental health effects in European children.

46. Toxicogenomic approaches were also applied in the FP6 Programme InnoMed PredTox which aimed to identify early in development drugs that would result in unacceptable toxicity in chronic bioassays and thereby provide substantial cost savings and reduce animal use. The plan involved exposing animals (short term) to failed drug candidates (with known chronic liver/kidney toxicity) and generating microarray data of gene expression profiles to predict future chronic toxicity. Such an approach is considered valuable as it provides an indication of chronic effects in short term study. The FP7 Programme Predict-IV extends the aims of PredTox to improve prediction of drug toxicity to accelerate the drug development process and reduce failure rates in later stages of development. The work involves a multidisciplinary approach that incorporates TGX methods.

47. The Innovative Medicines Initiatives (IMI)[13] (initially launched under the FP6 for Research) represents one of the first Joint Technology Initiatives introduced under the 7th Framework programme to realise Public-Private Partnerships at the European research level. In view of several challenges within the European biopharmaceutical sector (i.e. insufficient R&D investment, complex technologies and the fragmented nature of research in Europe) IMI's overall goal is to reinvigorate the biopharmaceutical sector in Europe, and in particular, overcome difficulties in predicting safety and efficacy, poor knowledge management, and gaps in education and training. IMI hope to achieve this by pooling competencies and resources from the public and private domain (via a unique collaboration between competitor pharmaceutical companies) and produced a research agenda that sets forth recommendations to overcome the above research bottlenecks in the drug development process. Fifteen IMI projects are currently ongoing (from the first call for proposals in 2008) and the following two Norvartis Pharma co-ordinated projects represent TGX-based efforts.

48. The MARCAR project/consortium[14] (an acronym for BioMARkers and molecular tumor classification for non-genotoxic CARcinogenesis) seeks to address the lack of reliable tools for predicting which compounds have a potential for later cancer development. The MARCAR project will apply mechanism-based (TGX) approaches to establish reliable biomarkers for the early prediction of potential for non-genotoxic carcinogenesis and to improve the molecular classification of tumors that can be caused by non-genotoxic carcinogenesis.

49. The eTOX Project[15] aims to develop a computer-based database and novel software tools to enable better in silico prediction of the toxicological profiles of new compounds in the early stages of drug development. This follows from the lack of a comprehensive computer toxicology database and a lack of integrative tools capable of exploiting such a database. The eTOX database, therefore, aims to be the largest toxicology repository for data

[13] IMI website: http://imi.europa.eu/index_en.html
[14] MARCAR is managed by University of Dundee,
[15] eTOX is managed by the private non-profit independent Spanish organisation Fundacio IMiM

derived from a range of toxicology-related disciplines. eTOX plans to combine public data and historical data from 14 pharmaceutical companies and device management tools for large databases to make computer based predictions of the potential toxicity profiles of new compounds based on their chemical structure.

F. MICROARRAY

50. Microarray design issues comprise a key subject area and mostly relate to the different platforms available and the experimental approach used in TGX studies.

(i) Microarray platforms

51. The type of microarray platform used is a key design consideration for any TGX microarray study (Lee et al 2005). Evaluation of the literature shows that to date two main types of DNA microarrays are widely used in TGX research: spotted and Affymetrix (high density) gene chip microarrays, although other platforms are increasingly being used e.g. Illumina (as noted below).

1. Spotted microarrays

52. Spotted[16] microarrays use either cDNA or oligonucleotides as their gene targets, which confer specific advantages and disadvantages over high density arrays (Lee et al 2005; Ju et al 2007). Gene targets are represented by a single cDNA [or oligonucleotide] clone spotted on the array (Morris et al 2006), and contain between 10-20K in one microarray slide (Yang & Speed, 2002). Notable advantages include the cost (spotted arrays are considerably cheaper per array than high density chips), and the fact that spotted arrays are customised which allows the researcher to determine the microarray content. cDNA microarrays are also not as sensitive to variation in sequence polymorphisms as occurs with short oligonucleotide arrays (Ahmed, 2006a). However, spotted microarrays are limited by their long set-up time, the variable amounts of DNA that can be placed on spots, their increased susceptiblility to contamination and high background levels. Investigators may use spotted oligonucleotide microarrays (or long[17] oligonucleotide microarrays (LOM)) in preference to their cDNA-based counterparts (or amplicon[18] arrays), following reports of better correlation between LOM expression data and quantitative real-time PCR (Ju et al 2007). LOM data is also noted to have greater concordance with Affymetrix data. However, given that amplicon arrays use longer gene targets, this enables more stringent washing conditions which results in stronger signals and less background.

[16] The word "spotted" refers to the process by which sequences of DNA are attached to a glass slide or other surface (Kerr & Churchill, 2001).
[17] Note. These targets are longer than the oligonucleotides used in Affymetrix arrays and are based on expressed sequence tags (ESTs) (Ju et al 2007).
[18] Due to the PCR amplification of the cDNA fragments (Ju et al 2007)

2. Oligonucleotide microarrays

53. In high density oligonucleotide arrays, gene targets are represented by "probes" or short sequences of nucleotides from the target gene sequence (Morris et al 2006). Affymetrix Inc. is the largest producer of oligonucelotide arrays (GeneChips), other types typically used include Agilent and Nimblegen. A single Affymetrix array slide contains between 200-500K probes (Yang & Speed, 2002). These oligonucleotides are noted for providing better characterised gene targets than spotted cDNA arrays (as the concentration and sequences are known and probe pairs are used to ensure specificity (Lee et al 2005). Furthermore, the fact that Affymetrix chips can be automated produces fewer problems with samples.

54. GeneChips contain multiple probes for each gene aka 'probeset' as an attempt to average out the natural variability among probes wrt their binding to matrix. The probes (based on sequence information contained in GenBank, Unigene, RefSeq databases) are 25 bases long and classed as either Perfect Match (PM) probes i.e. the target sequence or mismatch (MM) probes that are identical to the PM probes but their middle position base (13) is substituted by its complimentary base (these corresponding MM probes are for normalisation purposes). A probe pair comprises one PM and one MM, with a single probe pair scanning a particular sequence of a gene. A probe set comprises 11-20 related probe pairs for a target RNA.

55. Another microarray platform documented in the literature is the nylon cDNA array (Atlas array), considered to be an early form of microarray that uses radioactive rather than fluorescent labelling. Atlas arrays are thought to increase sensitivity (NRC, 2007) but they are limited by their low density and their high false positive rate (which arises from bleeding of highly overexpressed genes into adjacent targets on the membrane autoradiography (Irwin et al 2004)).

56. The Illumina Microarray (aka BeadArray) is becoming a popular microarray platform due to its cost effectiveness and accuracy (Kuhn et al 2004;). The BeadArray is based on randomly arranged beads, and each bead binds many identical copies of a gene-specific probe sequence; each type of bead having on average 30 randomly positioned replicates. Such a design is thought to contribute to the BeadArray's enhanced measurement precision and reliability (as it yields higher confidence calls and more robust estimations compared to other microarray types). However, its unique design does make preprocessing and quality control steps significantly different from other types of microarrays, such that most Illumina-based analyses incorporate preprocessing methods originally designed for Affymetrix microarrays. Du et al (2008), however, describe the development and implementation of 'lumi', a Bioconducter package especially designed to process Illumina microarray data.

57. Overall, the range of choices available for microarray platforms is a limitation as it contributes to variation in data (Yauk & Berndt 2007). Other challenges identified relate to technical problems arising with gene annotation,

programmes generating gene lists and the target's location on the chip, although these issues could possibly be resolved by considering the biological plausibility of results and using quantitative PCR to confirm a subset of genes identified (Irwin et al 2004).

3. Platform density

58. This issue was identified in the 2004 Joint statement (See COT conclusion (e)) and relates to the advantages and disadvantages of using high or low density microarray platforms and the lack of studies comparing the type of data generated in both. Concern was raised that the expense of using high density microarray platforms resulted in studies using less microarrays thereby compromising the quality of study design. This issue has however been resolved particularly in view of the findings of cross platform studies led by the HESI Nephrotoxicity Working Group, that examined the ability of different microarray platforms (e.g. custom (spotted) cDNA microarray vs. high density oligonucleotide Affymetrix microarray) to identify gene expression changes in kidneys of rats treated with cisplatin (Thompson et al 2004). A set of 93 differentially expressed genes associated with cisplatin-induced renal injury were identified on the cDNA array of which 48 could be identified as differentially expressed on the Affymetrix platform. The authors suggested that these findings demonstrate that gene profiles linked to specific types of tissue injury or mechanisms of toxicity (and identified in well-performed replicated microarray experiments) may be extrapolatable across platform technologies. Members are advised that the issue of cross-platform comparisons will be further addressed in a separate COT discussion paper on reproducibility.

(ii) Gene target selection for microarray

59. The type and number of genes to measure in a microarray that will ultimately help characterise a toxic response constitutes another key design issue (Lee et al 2005). Genes fall into many categories (e.g. xenobiotic metabolism, DNA repair, etc) and ready-made online resources of categorised genes are available to investigators conducting knowledge-based microarray experiments i.e. where and *a priori* hypothesis exists in relation to the toxicant mechanism of action. Notable online resources include GeneCards, Kyoto Encylcopedia of Genes and Genomes (KEGG), Toxicogenomics Research Consortium (TRC) and Chemical Effects in Biological Systems (CEBs). The first experiment is usually an exploratory one employing comprehensive arrays that include as many genes as possible in order to generate hypotheses (Ahmed, 2006a). These are followed by focussed arrays to test the hypotheses and elucidate mechanisms.

(iii) Single vs. double-channel microarray approaches

60. TGX studies use one of two microarray approaches: the single (one-colour) fluor/channel approach or the double (two colour) fluor/channel approach (Irwin 2004; Hayes & Bradfield 2005). The single channel approach involves hybridising control and experimental samples on separate

microarrays. Affymetrix arrays are inherently single channel, though some associated analysis tools facilitate pairwise comparisons (Slonim & Yanai, 2009). NB. Agilent and NimbleGen can be run using either one or two channels. For oligonucleotide arrays, labelling is achieved via use of an antisense copy of RNA with biotinylated nucleotides and after hybridisation, gene expression is measured by treating the chip with streptavidin-labelled with phycoerthyrin dye and scanned followed by calculation of hybridisation intensities.

61. In contrast, the more commonly used double channel approach involves hybridising control and experimental samples against the same array. Such an approach provides additional options for experimental design, that is it allows control for some technical issues by allowing a direct comparison in one hybridisation, where the second channel can either be used as part of the experiment or as an external control (Ahmed 2006a; Slonim & Yanai, 2009). cDNA arrays typically involve two channels. For cDNA (spotted) (and Agilent, Nimblegen), labelling is achieved by use of cy3 and cy5 fluorophores (fluorescent tags) with different excitation and emission spectra, from which the expression ratio for both samples at the same location is calculated and the data presented as a heat map. A useful schematic of the different labelling procedures used in cDNA spotted and high density oligonucleotide arrays are provided in Lee et al (2005). Also, Repsilber & Ziegler (2005) describe the experimental steps for a typical two colour microarray gene expression experiment particularly in relation to the cDNA spotted and Agilent platforms that use the two colour approach.

62. Experimental design and analysis is generally considered more straightforward with one-colour than two colour microarrays which can limit options for downstream analysis (Ahmed, 2006a). Slonim & Yanai (2009) cite the findings of a paper by Patterson et al (2006), which compared single and two channel methods on three platforms that allow for both options. It was suggested that two-channel arrays may have greater sensitivity while single channel arrays were more accurate for estimating fold changes. It was further suggested that single channel arrays allowed for more flexibility in the analysis and were better geared towards estimating raw transcript abundance, partly due to the lack of competition between samples for the same probes. A potential drawback of single channel arrays is the apparent increase in the number of arrays needed (and hence cost). However, responses to a cross-sector international online HESI Survey question about the type of high-density[19] microarray technologies used revealed that 46% of the respondents used single channel high-density oligonucleotide arrays (Affymetrix arrays) in 2007, compared to only 16.7% of respondents using cDNA microarray, a trend which rarely changed since 2005 (Pettit et al., 2010).

[19]Pettit et al (2010) describe high density microarrays as those generating more than 6000 data points

(iv) Limitations of microarray technology

63. Clearly, the main application of microarray technology is to determine mRNA levels from a large number of genes simultaneously thereby providing an indirect measure of global gene expression. However, several review papers identify various challenges associated with microarray gene expression detection, in particular, the inability of microarrays to detect chemical-induced TGX expression changes at environmental exposure levels (Lee et al 2005). There is concern that the wide variation in baseline gene expression levels amongst individuals and experimental animals may hinder the detection of (small) changes induced by environmental chemical exposure.

64. In the past criticisms levied against the use of microarrays to measure gene expression have included the fact that: the genes surveyed are limited to those included in the microarray (NAS, 2007b); cross hybridisation between similar sequences restricts microarrays to using non-repetitive fractions of genomes which can complicate analysis of related genes; a reliance on PCR-based amplification of biomaterial can introduce bias into samples; and the fact that since microarray design requires a priori knowledge of the genome this presents a problem for incomplete/incorrectly annotated genomes (Hurd & Nelson (2009).

65. The Microarray Quality Control (MAQC) Project led by the US Food & Drug Administration (FDA) involving 137 participants from 51 organisations was initiated to address concerns about the reliability of microarray technology, performance and data analysis issues (Shi et al 2006). These concerns relate principally to the fact that studies using different microarray platforms to analyse identical RNA samples are obtaining dissimilar or contradictory results. NB. Reproducibility issues will be further discussed in a subsequent COT paper.

66. In functional genomics, the assessment of mRNA profiles is considered to provide a measure of the practical[20] functional genome of any cell i.e. DNA that encodes proteins actually functioning within a cell. However, a more general (and perhaps more critical) concern relating to the use of microarray technology, is the lack of evidence that the transcribed mRNA undergoes further translation en route to protein synthesis (Gant 2007). Attempts have been made to address the latter issue (through proteomic approaches), although these are reportedly hampered by limitations associated with, for example 2D gel resolution.

67. Maier et al (2009) reviewed the available literature on the correlation of mRNA and protein abundances in cells and found that most of the published literature focussed on yeast species with very few studies conducted in bacteria and mammalian cells. The authors noted that protein and mRNA abundances do not appear to follow a normal distribution and hence their

[20] The practical functional genome is part of the theoretical functional genome i.e. DNA that codes for RNA message that can in theory be translated into a protein (comprises only 1-2% of genome) (EPA. 2004).

correlation is best described using Spearman rank coefficient (r_s) rather than the Pearson correlation coefficient (r_p). A weakly positive correlation was revealed in a study using two hematopoietic mouse cell lines $r_p = 0.59$, 425 mRNA-protein pairs (Tian et al., 2004). Yu et al (2007) developed the Protein Abundance and mRNA Expression (PARE) web tool which allows for rapid assessment of mRNA-protein correlation for complex samples with data-sets for rat and mice provided. The review authors concluded that the available literature failed to provide strong evidence of mRNA-protein correlation and suggested possible factors responsible for the quantitative differences between the transcriptome and translatome. These include technical factors such as experimental noise and error and post-transcriptional and post-translational factors e.g. RNA secondary structure, Shine Dalgarno sequence differences, regulatory proteins and sRNAs, codon bias and codon adaptation index, ribosomal density and occupancy and protein half-lives.

68. The IPCS/OECD secretariats conducted two surveys to review the available TGX tools in OECD member countries. The first survey was conducted in preparation for the 2004 Workshop on ecotoxicogenomics. The second survey (which was led by Japan) was conducted (and published) in 2009 to follow up on the current approaches used given the rate of acceleration in the field's development. A questionnaire[21] was used to gather information on available TGX methods and queried the type of mammalian effects analysed and respondents' experiences of using omics tools for the evaluation of chemicals. The following eight countries participated in the survey: Japan, Korea, The Netherlands, Switzerland, US, UK, France and Germany. The results (shown in Table 1) of the survey showed that during the three years since the first 2004 survey, the application of omics technologies to toxicology drastically changed. In 2004, nearly 80% of studies were transcriptome analysis but in 2007 the ratio had decreased to 55% largely due to the introduction of the emerging omics technologies, metabolomics and proteomics (in particular, the number of applications in metabolomics increased from 2 to 11). The report concluded on the importance of periodical surveillance of the omics technologies in order to evaluate progress in this field.

Table 1. Comparison of the IPCS/OECD survey results obtained in 2004 and 2007 (Reproduced)

	2004	**2007**
Countries	5	8
Total studies	42	62
Transcriptomics	33	32
Proteomics	7	19
Metabolomics	2	11
Published studies	10	32

[21]The questionnaire was distributed in July 2006 and collected information between August and Nov 2006

(v) Alternative applications of microarray technology

69. Given the above challenges associated with the significance of mRNA level measurements, alternative applications of microarray technology have been developed, and a review by Gant (2007) suggests that these novel techniques could provide more extensive information on the molecular changes that characterise a toxic response. These techniques provide information of events either upstream of mRNA synthesis (e.g. array comparative genomic hybridisation (ArrayCGH) or downstream of mRNA synthesis (e.g. the mRNA translation assay). Table 2 summarises the application, value and challenges of each of these methods (see end of section 1).

70. The mRNA translation assay may prove to be a key gene expression assessment tool in view of its ability to bridge the transcription vs. translation dilemma. The assay works by exploiting the differential densities of different mRNA fractions. The less dense (light) monosomal fraction bears ribosomal RNA only. In contrast, the denser (heavy) polysomal fraction includes attached ribosomes, which are considered a measure of active translation (in which the number of bound ribosomes is proportional to the protein amount produced). The fractions are separated by running them on a polysomal gradient and visualised via a UV tracer. The mRNA content of each layer is then measured and the layers compared via microarray analysis.

71. The use of density RNA fractionation with microarrays in toxicology is limited. Two studies identified have used the assay to further characterise gene expression changes following a toxic insult. Mazan-Mamczarz et al (2005) exposed human cells to UV light and monitored mRNA distribution along polysome gradients with each layer analysed via spotted cDNA array analysis. The authors were able to identify and verify translationally induced/repressed mRNAs, and concluded that the mRNA translation assay provides key information in relation to which genes are ultimately expressed by determining degrees of translational engagement. The mRNA translation assay was also used in a study by Shenton et al (2006) which analysed the regulation of protein synthesis after oxidative stress in yeast exposed to hydrogen peroxide. The authors were able to conclude that translational control is a key component of a cell's response to oxidative stress.

72. Gant (2007) also suggests that the pattern of microRNA (miRNA expression profiling) could be used to identify specific toxicities, although to date there has been no application of this technology in toxicology. MicroRNAs are single stranded RNA molecules that control gene (mRNA) translation. They are transcribed from polycistronic regions of the genome via RNA polymerase II and III to produce immature transcripts. These transcripts are processed in the nucleus/cytoplasm to produce mature miRNA (21-23 nucleotides long). MiRNA regulates translation by interacting with a multiprotein complex called RNA-inducing silencing complex (RISC) which essentially represses translation. MiRNAs store mRNAs in P-bodies within the cytoplasm which are later retrieved for translation. Such actions can increase

protein levels without new transcription. The potential regulation of miRNA expression by chemicals is thereby significant since any alteration in miRNA expression could alter the cell protein complement and a cell's subsequent response to chemical exposure. MiRNA profiling is fraught with technical challenges associated with the short nature of mature miRNA species. Use of the RNA tailing method for labelling and use of modified targets on microarrays (aka locked nucleic acid nucleotides) for hybridisation are suggested as ways forward (Castoldi et al 2006).

73. Finally, Gant (2007) also discusses the significance of epigenetic modifications in inducing transmissible genomic changes and their assessment via microarray and immunoprecipitation methods. The epigenome serves as an interface between the dynamic environment and the inherited static genome and given the potential impact epigenetic mechanisms could have on the toxic action of xenobiotics, incorporating epigenetics in the assessment of the safety of chemicals may become a standard requirement (Szyf, 2007). This subject was recently discussed at a 2009 ILSI HESI Workshop designed to evaluate and enhance the scientific knowledge base regarding epigenetics and its role in disease (Goodman et al (2010). Through several breakout groups of cross sector representatives the workshop addressed the issue of what needs to be known prior to thinking about incorporating an epigenetic evaluation into safety assessment. It was concluded that a great deal still needs to be learned before an epigenetic evaluation can be rationally incorporate into safety assessment.

G. TYPES OF HYBRIDISATION APPROACH

74. Hayes & Bradfield (2005) describe the three different types of hybridisation approaches available to two-channel microarray studies to help achieve particular study objectives (hence hybridisation approach is often referred to as the experimental design of a microarray study i.e. microarray experiment type). The direct design is considered the most sensitive and involves hybridising the control sample directly against the test sample. This approach is particularly useful for studies seeking to determine the identity of gene targets i.e. what genes are expressed at a particular time, and thus requires comparing each time point to time-matched controls. Loop designs involve hybridising biological replicates against each other i.e. the control sample from one animal hybridised against the control sample of another (presumably at different timepoints and a comparison of the range of treatment groups). This approach is used for studies seeking to determine the temporal nature of a gene target i.e. how gene expression changes over time, and requires comparing each time point against preceding and subsequent time points. Although better information is provided on how time influences a target gene, such an approach reduces the sensitivity of detecting any changes. The final approach, the reference design, involves hybridising the control and test sample against a common reference sample. The reference design is particularly useful for studies seeking to determine both the identity of a target gene and the temporal nature of changes in its expression i.e. what gene is differentially expressed and how does its expression change over

time.. However, the reference design is reported to be limited by reducing the study's statistical power.

H. SOURCES OF BIAS

75. The multitude of measurements and estimates made in microarray studies increases the likelihood of multiple systematic errors arising. Systematic errors are problematic because they introduce bias in the data i.e. a systematic directional distortion of measured gene expression values or other related data from the true actual value. Such errors result in consistently inaccurate readings/results irrespective of the number of repeat measurements made. The development of approaches to better identify and characterise sources of bias in TGX studies and thus correct-for or limit their effect comprises a significant research priority.

76. Bias can be categorised as systematic, selection or confounding. Systematic bias relates to bias of a measurement system or estimation method. This can therefore arise in any quantification step employed in a microarray experiment.

77. Selection biases in TGX studies can arise from errors made in, for example, the selection of gene targets for microarray assessment, or from the tissue region selected for RNA extraction (in which the latter may be particularly sensitive to a particular toxicant resulting in a biased gene expression profile) (Thompson & Hackett, 2008). NB. Issues relating to RNA sampling are further discussed elsewhere in this paper.

78. Confounding bias arises from factors that obscure the real effect of exposure to a particular agent, and a review by Thompson & Hackett (2008) describes how circadian rhythm regulation, vehicle anaesthesia, human variation and the two-colour labelling step contribute to the production of biased gene expression data. Circadian rhythm regulation is an important design consideration for TGX studies. For example, hepatic gene expression varies during the day and since tissue collection can also vary during the day circadian rhythm can confound the TGX data obtained. Boorman et al (2005) evaluated temporal hepatic gene expression in untreated rats and reported differential expression in day/night comparisons. The study noted periodically expressed genes in liver samples collected at multiple times during the day and also observed circadian genes in rat livers demonstrating rhythms of gene expression, thereby concluding that prominent circadian rhythm gene expression exists in the rat. Takishima et al (2006) found that the corn-oil vehicle used in a single bolus or repeat dose TGX study modulated fatty acid metabolism genes. Human variation is also reported to have a confounding effect on differential gene expression as demonstrated in a study exploring inter-individual and temporal variation in gene expression patterns (Whitney et al 2003). Finally the two-colour labelling microarray approach can introduce bias due to the differential rates at which dyes can be incorporated into the sample or the differences in quantum efficiencies between the two dyes as well as the differential sensitivities of Cy5 and Cy3 dyes to quenching, photobleaching and degradation.

79. As in toxicology in general, other confounders include the nutritional status of animals used in in-vivo studies, which presents a particular challenge due to fed and fasted animals responding differently, for example to treatment with paracetamol, thereby requiring that treated and control groups are matched with respect to their nutritional status (Irwin et al 2004). A related potential confounder arises when treatment affects food and water consumption rate resulting in reduced body weight which can influence the pattern of gene expression in the liver and other tissues.

80. Several design and statistical approaches are used to respectively minimise and correct for the sources of bias discussed above. However before biases can be remedied they must first be detected and quantified. To reduce the likelihood of bias arising in a microarray study investigators can incorporate various experimental design approaches such as matching, e.g. the use of time-matched controls to minimise bias due to differences in diurnally expressed genes (Thompson & Hackett, 2008). Randomisation is considered one of three fundamental experimental design approaches and is used to reduce the likelihood of systematic biases caused by selection or assignment (Chen et al 2004). It is suggested that randomisation be applied throughout a whole experiment i.e. randomising biological samples to a particular treatment and randomising measurements, etc. Blocking and replication represent the two other fundamental experimental design concepts. Blocking is described as an approach used to increase the precision of any estimates made in a study and involves arranging a TGX experiment into a smaller subset of homogenous experimental units to enable the study to be conducted at different times and locations (Chen et al 2004). Replication is a particularly significant design feature for three main reasons. The number of replicates incorporated into a microarray study not only determines the type (quality) of statistical methods that can be used to analyse the data, but also the more replicates a study uses the greater its ability to detect small differences in gene expression and distinguish differential gene expression from noise (Chen et al 2004). Replication can be incorporated at every level of data generation e.g. wrt a sample – use replicate no. of animals/tissues/cell types; array – use replicate no. of arrays; spot – use replicate no. of spots of the same gene, which is more common (Lee et al 2005). However, different types of replicates are used depending on the aim of the experiment as described below.

I. REPLICATES

81. Two types of replicates are commonly used: technical and biological.

(i) Technical replicates

82. Technical replicates describe replications in which the mRNA sample is derived from one single source (Chen et al 2004; Irwin et al 2004; Lee et al 2005). These replicates are typically incorporated into a study to reduce experimental variabilities i.e. data variation in measurements (and thereby ensure reproducibility) or when an individual response/gene expression profile

is desired. Technical replicates can help identify data variation in measurements because using the same sample renders it more likely that any variation in data arises from technical procedures, enabling the assessment of experimental variation. Subsequent data analysis would thereby utilise a statistical test based on technical replicates.

83. A technical replicate commonly used in two-colour microarray channel studies is the dye reversal/flip design. This approach aims to compensate for dye bias and involves using two microarrays, in which the hybridisation of the first microarray is based on labelling the test and control sample with Cy5 and Cy3 dye respectively and for the second microarray switching the dye orientation between the samples. Such an approach is reported to help reveal systematic bias in labelling reaction/fluorescence yield. However, Rosenzweig et al (2004) considers the above approach impractical and inefficient because it requires additional microarrays and reagents which can be costly to end users. The authors propose an alternative approach that involves incorporating split-control microarrays within a set of concurrently processed hybridisations, which specifically measures dye bias and maintains experimental accuracy and technical precision.

84. Fujita et al (2009) highlight the lack of attention generally paid to checking the technical replicates (i.e. to ensure that the error of measure is small enough to be of no concern) and provide an interpretable and objective way to ensure the technical replicates quality.

(ii) Biological replicates

85. A limitation of studies that use only technical replicates is their inability to provide information on average response, which can be obtained via the use of biological replicates. Biological replicates describe replications in which the mRNA sample is derived from more than one discrete biological sample , and for this reason are also referred to as the sample size (Lee et al 2005) - as it often refers to the number of animals used per treatment/control group (Irwin et al 2004). Biological replicates are typically used in studies seeking to obtain an estimate of the variability about the average response, with subsequent data analyses utilising statistical tests based on a biological replicate design. This design approach enables investigators to determine the extent to which individual responses vary between treated and control groups. A study investigating variability in gene expression data also evaluated the effect of introducing replication to data consistency and reliability (Lee et al., 2000). After combining data from all replicates, the authors found that fewer genes were incorrectly classified as having altered expression, which led them to suggest that pooling replicate data provides a more reliable analysis. The authors further concluded that experiments should be designed using a minimum of 3 biological replicates which should help reduce misclassification rates.

J. POOLING

86. RNA samples are sometimes pooled prior to labelling and hybridisation to either avoid the need for RNA amplification, in cases where individual RNA samples are insufficient, or to reduce costs arising from using multiple arrays. Sample pooling is performed in order to reduce the effects of biological variation, without having to measure multiple individual samples. This is because differences due to individual variations will be minimised, making substantive differences easier to detect (Ahmed, 2006a). Pooling is considered worthwhile only when samples are cheap relative to microarrays (NAS, 2007a). Pooling different biological samples can enable an average response to be obtained, although at the expense of obtaining information on variation in response between samples. Therefore, if samples are valuable (e.g. human samples) then pooling is not recommended due to loss of information on individual samples (i.e. lack of individual variability data).

87. Members previously raised concern that pooling to save money can result in poor quality study designs due to too loss of information on biological variability and the subsequent statistical implications. Ahmed (2006a) assessed studies examining the implications of pooling in detecting differential gene expression. It was noted that pooling should only be recommended when there is insufficient RNA from each individual sample to perform an analysis. The review also identifies a study by Zhang & Gant (2005) which provided formulae to estimate conditions under which a pooled design is preferred versus a nonpooled design (taking into account unit costs of microarray platform and biological subject).

88. Members also previously commented on the failure of published studies to clarify whether samples were pooled (and the type of replicate used). Jafari & Azuaje (2006) reviewed hundreds of MEDLINE-indexed papers involving gene expression data analysis published between 2003 and 2005 on the basis of their reporting practices/ standards. A total of 293 studies were identified that used microarray transcript profiling methods and of these only 33 studies reported their pooling procedures. The authors concluded that studies published at the time lack key information required for properly assessing their design quality and potential impact and suggested the need for more rigorous reporting of important experimental factors such as statistical power and sample size as well as the correct description and justification of statistical methods applied. The authors further concluded that their study highlights the importance of defining a minimum set of information required for reporting on statistical design and analysis of expression data.

K. RECOMMENDED STUDY DESIGNS

89. Adherence to recommended study designs not only facilitates quality assurance and peer-review processes but also the reproducibility of the results. Elashoff (2008) recommends that a TGX study should be designed with 3-6 animals per control/dose group with a single RNA sample run per animal, with animals exposed to several doses to enable comprehensive dose

response analysis (wrt overall toxicity and on an individual gene/pathway level basis). As there is still uncertainty over the most appropriate study timepoint to use a range of 3-4 timepoints is suggested. Ahmed (2006a) suggests that the following guidelines should be considered when designing a microarray experiment: use of technical replication and repeated measurements to ensure effective precision (when variability of measurement is greater than the variability between experimental units); pooling samples when biological variability between individual samples is large and units not too costly; use of cDNA microarrays when minimal or no information is known about the study organism; use of dye swapping and looping to balance dyes and samples and using at least a single round of RNA amplification when the starting amount of RNA is limiting to ensure that there is enough material for analysis.

Table 2. Summary of some other microarray technology approaches (Gant, 2007)

Method	ArrayCGH	Epigenetic analysis	ChIP analysis	Transcription rate analysis	mRNA translation assay
Description	Array comparative genomic hybridisation	Study of inheritable gene expression changes arising in the absence of changes to DNA sequence	Chromosome immunoprecipitation analysis	Measure of rate of gene transcription e.g. following chemical exposure	Technique used to determine whether mRNA is translated and if this occurs differentially following chemical exposure.
Application	To identify chemicals that may act by causing deletions/ amplifications in genome (chromosomal changes in genome) - Characterising cells and animal strains for testing purposes - Drug efficacy/safety evaluation	To identify/further characterise chemicals that may act (regulate gene expression) by causing epigenetic modifications (e.g. DNA methylation)	Map localisation of modified histones/transcription factors on the genome (Collas 2009) To evaluate chemicals that may act (affect gene expression) by regulating binding of transcription factors to promoter regions of genes??	To obtain quantitative information about the relative rates of transcription in different genes from isolated nuclei Alternative indirect measure of mRNA expression	To provide a more accurate measure of gene expression
Protocol	(For two-colour channel microarray): - Extract genomic DNA (gDNA) - Label control/test sample with different fluorescent dyes - Hybridise probes onto same microarray - Scan image - Determine ratio of fluorescence dye (where > test DNA ➔ amplification; < test DNA ➔ deletion) - Produce chromosome map	- Microarray gene target: print spots on slides containing target sequences from gene promoter regions - Sample preparation: Fragment DNA; immunoprecipitation of methylated fragments (fragments identified using ab raised against 5-methylcytosine) - Identify differential methylation: via microarray analysis (by compare differences in DNA sequence between immunoprecipitated fractions)	Similar methodology as used in epigenetic analysis except - ab against transcription factor of interest (not 5-methylcytosine) - DNA fragment is crosslinked to transcription factor prior to immunoprecipitation	(For microarray approach): - Isolate nuclei from test and control sample - Incorporate labelled nucleotide in RNA (during transcription) - Isolate RNA (as via nuclear run on assay) - Hybridise on microarray - Detect differential gene transcription	- Inhibit mRNA translation - Separtae monosomal from polysomal mRNA - Visualise two fractions - Conduct microarray analysis

Value	Over other cytogenetic methods include - higher resolution and throughput shorter turn around time (no need for cell culture); high reproducibility; robustness; precise mapping of aberrations (Shinawi & Cheung 2008)	Aids understanding of how drugs/chemicals affect DNA methylation patterns and subsequent gene expression changes leading to toxicity. NB. Thought to account for differences in susceptibility and resistance	Generates mechanistic data	- Provides a direct measure of the activity of genes (since hybridisation assays only give a measure of how much RNA is present (steady-state level). Hence allows changes in transcription rate to be measured - Aids mechanistic understanding of toxicants - Compliments ChIP analysis	Provide a more complete account of gene expression changes cf. mRNA assessment
Limitations/ challenges	Inability to identify balanced rearrangements (e.g. translocations/inversions) and detect polyploidy Area of genome assayed dependent on microarray targets present. (Shinawi & Cheung 2008)	- Current methods analyse only one type of epigenetic modification at a time (Cipriany et al 2010) - Requires substantial input material (Cipriany et al 2010) - Need to conduct research to determine the role of epigenetic modifications in chemically mediated toxicities	- Cumbersome procedure (Collas 2009) - Requires large number of cells (Collas 2009) - Lack of microarray targets with suitable promoter fragments (NB. Companies developing appropriate microarrays)	- Presence of artefacts from the nuclear isolation steps - Time consuming - Requires use of radioactivity and large number of cells - Limited data set - Methodology not fully established	(Melamed et al 2009) - Sedimentation in polysomal gradient may not always be due to ribosomes (Melamed et al 2009) - Include large cellular complexes and their associated mRNAs (Melamed et al 2009) - No of ribosomes bound to the mRNA does not always correlate with its translational status(Melamed et al 2009) - Uses fractionated mRNA so must be conducted with care - Limited use of density RNA fractionation with microarrays in toxicology

SECTION 2. APPROACHES USED TO ANALYSE RAW TRANSCRIPTOMIC DATA

INTRODUCTION

90. The analysis of toxicogenomic data was previously discussed at the February 2009 COT Workshop on 21st Century Toxicology (published as a statement at http://cot.food.gov.uk/cotstatements/cotstatementsyrs/cotstatements2009/cot200903). A presentation by Dr Cliff Elcombe entitled, 'Toxicogenomic tools for chemical safety assessment' summarised approaches used to analyse transcriptomic data and grouped them into three main categories: data preprocessing, data analysis and data validation and interpretation (the latter including bioinformatics[22] and pathway analyses). Several reviews (e.g. Butte et al 2002; Boes & Neuhauser., 2005; Ittrich., 2005; Rahnenfuhrer, 2005a Ahmed, 2006b) and book chapters (NAS, 2007b; Durinck, 2008) discuss the analysis of raw transcriptomic data with different perspectives on what constitutes data analysis. For example, Butte (2002) and NAS (2007b) describe data analysis on the basis of the type of analytical methods used i.e. unsupervised vs. supervised, while other papers consider data analysis in terms of strata i.e. with reference to oligonucleotide /Affymetrix GeneChip microarrays. Low-level data analysis applies to methods (background correction, normalisation, PM adjustment and summarisation) used to calculate the expression values of probe sets from scanned image values (pixels) of probes (Boes & Neuhauser., 2005). These methods survey how to correct for typical biases in microarray gene expression data (Repsilber et al 2005). High-level data analysis, therefore, refers to the methods applied to transformed raw data. This section discusses the steps involved in low-level data analysis. Steps represented by high-level data analysis comprise largely of statistical and computational manipulation of *transformed* data to identify gene expression changes and evaluate toxicologically relevant patterns and are therefore further discussed in Section 3 on statistical analysis.

A. DATA PROCESSING

91. Slonim & Yanai (2009) note that the task of analysing microarray data is typically more time consuming than the laboratory protocols required to generate the results. This is in part due to the pre-processing or data preparation stage, which principally focuses on assessing the quality of data and ensuring that all samples are comparable for further analysis. Data preparation/pre-processing aims to correct data sets for sources of variability arising from random and systematic error during experimental procedure (Magglioli et al 2006; Suarez et al 2009) and a schematic of the data processing stage is provided in Annex 2.

[22] The US, EPA defines bioinformatics as data acquisition and processing technologies that store and analyse data generated from omic technologies (EPA, 2004)

92. Data processing essentially involves scanning the array slide; saving the image/text data in a database; discarding poor quality images and filtering extremes; normalising qualified data (Ju et al 2007). Although scanning and image analysis can be considered to represent post-hybridisation (data acquisition) steps, the application of analytical approaches to sort and process microarray data for subsequent high-level data analysis justifies their inclusion in this section.

(i) Scanning

93. Microarray slides are scanned to detect hybridised spots via the use of lasers that excite the dyes with the resultant image saved on a computer. In two channel microarray approaches the amounts of bound targets are quantified by recording fluorescence signals resulting in a Tagged Image File Format (TIFF). Selection of appropriate scanning setting is particularly important as it can affect data acquisition (Williams & Thomson, 2010). Various options have been proposed, in particular multi-scanning noted for its ability to reduce quantification error and minimise the effects of saturation.

94. Skibbe et al (2006) developed a scanning approach that extends the dynamic data range by acquiring multiple scans of different intensities. The authors used multiple scan and linear regression approaches to identify and compare the sets of genes that exhibit statistically significant differential expression. Data were separately analysed from each of three scan intensities (low, medium and high). In the multiple scan approach only one third of DEGs were shared among the three scanning intensities, and each scan intensity identified unique sets of DEGs (all verified via qRT-PCR). The authors found that signal intensities (average) of DEGs were highest for low intensity scans and lowest for high intensity scans and suggested that the low intensity scans should be used to detect expression of high signal genes and high intensity scans should be used to detect expression of low signal genes. The authors concluded that the multiple scan approach effectively identifies a subset of statistically significant genes that linear regression approaches are unable to identify.

95. Once an image has been scanned, all the data are fixed regardless of image quality (Rahnenfuhrer, 2005a). Poor quality images automatically lead to a decrease in power of subsequent statistical analyses.

(ii) Image Analysis

96. Repsilber et al (2005) consider image analysis[23] to be the first step of the statistical analysis of microarray experiments and outline two key aims: (1) to identify spots belonging to a single transcript, (2) to quantity the signal intensity of spots identified. The output of image analysis is the assignment of a signal intensity for each gene. To achieve this, the background must be corrected for, as the signal observed (at this stage) comprises the true

[23] Commercial image acquisition programmes are available e.g. Automated Microarray Image Analysis (AMIA) software (Yauk & Berndt 2007).

(foreground) signal (from the specific hybridisation of interest) and the background signal (due to non-specific hybridisation and/or contamination) (Suarez et al 2009). The standard approach for background correction is to estimate background intensity[24] and subtract this from the spot intensity (and an alternative approach would be to examine image plots of log intensities). Qin et al (2004) notes that this can substantially change data and questions whether background subtraction from spot intensity measurements improves accuracy.

97. Procedures conducted from this point onwards depend on the type of microarray platform used i.e. whether the array is a spotted cDNA microarray or oligonucleotide microarray. Spotted cDNA microarrays are able to generate either one- or two channel data and contain a single probe for each target RNA (Suarez et al 2009). For two-channel cDNA microarray approaches, after hybridisation the two differently colour-labelled biological samples are scanned separately and the relative expression is determined by comparing intensities (Butte, 2002). Spotted cDNA microarrays therefore report differences in gene expression between two samples. This contrasts with the oligonucleotide microarray (e.g. Affymetrix GeneChip), which typically deploys one channel approaches and reports absolute expression levels (Boes & Neuhauser, 2005).

98. AMIA software uses three steps to assign a signal intensity to a gene (in two channel analyses) as described for both platforms below.

1. Image analysis for spotted (cDNA) microarrays

99. For spotted cDNA microarrays, after obtaining TIFF files, the next step is to derive a ratio of fluorescence measurements for green and red dyes, which represents the relative abundance of corresponding mRNA. This involves generating intensity values for red dye, green dye and their ratio via the process of gridding, segmentation and intensity extraction.

100. Gridding localises areas in an image that belong to a spot i.e. identifies regions on the slide containing single spots by assigning co-ordinates (Rahnenfuhrer, 2005a; Suarez et al 2009). The spot and its background represent the target area (patch). Segmentation describes the process of partitioning the target area of every spot into two distinct regions: the spot area containing the signal of interest from the background area surrounding the spot (Repsilber et al 2005; Suarez et al 2009). Segmentation is used to differentiate the pixels within a spot containing region into the foreground (true signal) and background (Rahnenfuhrer, 2005a). Intensity/information extraction uses the pixels of an image to calculate 'summarising' values for foreground and background intensities (Suarez et al 2009).

101. Ideally all spots would be separated by the same distance and have a circular shape however Rahnenfuhrer (2005a) notes that the available

[24] Background intensity is liable to increase from dust, fibres, fingerprints, auto fluorescence of coated glass, hybridisation problems, or residual effects from inadequate washing (Suarez et al 2009)

technology does not make this possible. The causes of various errors in image analysis include variable spot size and shape, artefacts caused by printing process and hybridisation technique, scanner sensitivity and experimental quality (Suarez et al 2009). It is important to rectify these errors and several algorithms for this are used that fall into two method groups: spatial methods attempt to capture the shape of a spot by fixing a circle with a constant diameter to all the spots in the image; while distribution methods apply a threshold value using Mann-Whitney test to classify pixels as either foreground or background depending on whether their value is greater or less than the threshold (Rahnenfuhrer, 2005a).

2. Image analysis for oligonucleotide microarrays

102. For oligonucleotide microarrays a slightly different procedure is used. After scanning, gridding is performed using a set of probes present on the borders and the middle of the array. MM oligonucleotide probes are used to calculate cross-hybridisation and local background signals. MM probe intensities are then subtracted form the intensities of the corresponding PM probes. If the MM value is less than the PM value it is possible to estimate background intensity. Sometimes MM probes have intensities higher than their corresponding PM probes resulting in negative expression values.

(iii) Data quality assessment/data sorting

103. Data quality assessment represents a key pre-processing step in microarray TGX analysis, and essentially separates poor data from potentially useful data (Morgan et al 2004). Various methods exist to assess the quality of the microarray experiments and graphical representations of array data can help quickly identify bad arrays and the need for normalisation (Durinck, 2008). For spotted cDNA microarrays, hybridisation problems are often identified by plotting images of foreground and background intensities for each channel used, and for Affymetrix Genechips the images of the PM and MM values are plotted. The quality of a particular hybridisation can also be assessed by either identifying the number of spots above background, in which low numbers would suggest repeating the hybridisation (Durinck, 2008), or by visualising the differences between arrays which can be done in two ways: plotting boxplots[25] of raw intensities grouped per array or correlation heatmaps. For heatmaps, correlations between hybridisations are expected to be high while for the boxplot approach, failing arrays show up as outliers (the pattern also reveals the necessity for between array normalisation). Either way, hybridisations failing these quality assessments would suggest the array is discarded, redone or given lower priority in subsequent analysis.

104. Elashoff (2008) provides an alternative description of the data quality assessments used and considers quality metrics and correlation as the two main approaches. Quality Metrics employs various methods to detect variation in data quality. These include: percent present (PP) or PP calls (PPC) –

[25] Also known as box-and-whisker diagram or plot. Boxplots are a convenient way of graphically depicting groups of numerical data. They can identify outliers and are non-parametric

considered the most informative measure of gene expression quality, which measures the percentage of genes present (expressed) in a sample as a fraction of (genes deemed present / total no of genes present on chip) although PP is limited by the fact that the value depends on the type of chip and sample used; the threshold benchmark approach – metrics that fall short of a predetermined value fail; consistency – failing metric values lie outside the norm within a study; and balance – which compares the distribution of metric values between study groups. Elashoff (2008) describes Pearson's correlation as an example of a correlation measure used to visualise differences between arrays. Pearson's correlation measures the similarity of expression log values between pairs of samples and uses the entire set of genes to derive correlation values[26] (or average) for each sample relative to another within the same study, which are used to produce a correlation matrix (heatmap). By taking the mean (expression value) for each sample a low value for a sample (i.e. < 0.9) would suggest that the expression profile differs from others in the study i.e. sample is of low quality compared to others in the study.

105. Other methods used in the data quality assessment of microarray data include the 5'3' ratio for specific control genes, which measures RNA degradation; scale factor, which involves scaling unnormalised gene expression mean values; specifically for Affymetrix GeneChips the MM > PM, which provides a measure of chip quality by ensuring that PM probe pairs > MM probe pairs (Elashoff (2008); Thompson & Hackett, 2008)

(iv) Standard Transformations

106. Typically background corrected data are subject to standard transformations to make them more suitable for statistical and biological analysis between biological samples (Irwin et al 2004). Log transformation and normalisation are the two most commonly described approaches in the literature.

1. Log transformation

107. Log transformation is typically required to address the limitations associated with evaluating cDNA microarray data. Spotted cDNA microarray data are usually evaluated by looking at ratios e.g. the ratio between two conditions on the same array, to provide a measure of gene expression changes. However, the drawback of such an approach is the different way in which up and down regulated genes are treated. For example, genes upregulated by a factor of 2 have an expression ratio of 2 and genes downregulated by a factor of 2 have an expression ratio of 0.5. Visualisation in a graph results in upregulated genes having a much wider range than down regulated ones (i.e. a positively skewed graph). Log transformation of expression ratios treats numbers and their reciprocals symmetrically (i.e. numerically equilibrates similar magnitudes of increases and decreases) whereby a $\log_2(1)=0$, $\log_2(2)=1$, $\log_2(1/2)= -1$. It therefore transforms positively

[26] Typical average correlation values are ≥0.95

skewed data into a more symmetrical distribution around 0 and graphically treats up and down regulated genes in a similar fashion. Log transformation is typically performed after background subtraction but before normalisation

2. Normalisation

108. Repsilber et al (2005) considers normalisation as the starting point of the so-called high-level analyses as it allows informative comparison of expression intensity values, without which it would be impossible to identify differentially expressed genes (Boes & Neuhauser, 2005). Normalisation is applied to raw microarray data to correct for (miminise) sources of technical (systematic) variation i.e. unequal quantities of starting RNA, differences in labelling or detection efficiencies between dyes which can lead to systematic biases in measured gene expression levels (Lee et al., 2005, Ahmed, 2006b). Minimising the amount of non-biological variation makes it possible to focus on the real biological changes during data analysis, however this represents a big challenge. Thus, normalisation becomes necessary particularly when dealing with experiments involving multiple arrays (Suarez et al 2009). Normalisation is also conducted to help eliminate questionable measurements and adjust measured intensities thereby aiding comparisons between samples and subsequent selection of DEGs between samples.

109. Various normalisation approaches and methods are available and have a profound effect on the expression levels and consequently on the detection of differentially expressed genes (Boes & Neuhauser, 2005). Some methods combine information from all arrays while other methods first determine a baseline array and normalise other arrays on the basis of the baseline array values (which presents issues over deciding on which array should be the baseline array). Normalisation can be done either between channels or slides/experiments. Visualisation of the raw data is an essential part of choosing a normalisation method and estimating the effectiveness of normalisation (Slonim & Yanai 2009), however, there is no general consensus on which is the best method to use although the microarray platform type influences the approach used.

a) Normalisation for spotted cDNA microarrays

110. For two-channel spotted cDNA microarrays, normalisation can be applied either within a single slide (involves normalising the log ratios of red and green channel intensities separately for each slide); between pair of slides (for dye swap experiments); or among multiple slides (involves adjusting for scale differences between slides) (Ittrich, 2005, Ahmed, 2006b). NB. To normalise single channel arrays all slides of the experiment are incorporated, with log ratios computed afterwards (Ittrich, 2005).

111. Which genes on which to base a normalisation method presents a further consideration. The simplest approach is a global average (all genes) which uses the majority of spots on a chip for applying robust normalisation procedures (Ittrich, 2005, Ahmed, 2006b). However, this is only applied on the condition that there is a large number of spots on a chip, similar number of

up/down regulated genes can be assumed, only a relatively small proportion of genes are expected to vary significantly in expression between two samples or there is symmetry in expression levels of regulated genes. Ideally, normalisation methods are based on a set of genes assumed to be non-differentially expressed between samples in the experiment. When a large proportion of genes on probes on a chip are expected to be differentially expressed between conditions it then becomes necessary to include a set of controls that function as a subset of genes with a constant expression to which the complete dataset can be normalised. One approach is to use a combination of housekeeping genes believed to have a constant expression across a variety of conditions. However, if there are too few genes or the intensities do not cover the whole range of different intensity levels this may preclude use of complex normalisation methods like intensity dependent normalisation. Furthermore, genes may not be constitutively expressed at constant levels (Ittrich, 2005). The use of exogenous universal control genes (aka spiked controls) as defined by their synthetic DNA sequences or DNA sequences from a different organism provides an alternative approach. NB. Also referred to as spike-in normalisation (Slonim & Yanai, 2009). These controls are spotted onto the array and also included in the two different samples (discussed further in section 4). Other approaches include the use of genomic DNA, and a microarray sample pool, which describes a set of controls analogous to genomic DNA but lacks non-coding regions (Ittrich, 2005). MSP (microarray sample pool) is derived by pooling PCR-amplified ESTs of all spots on an array, diluting and then spotting on the array in each print tip group.

112. Another consideration is how to perform the normalisation and two types of approach are commonly documented in the published literature: total intensity (global) and intensity-dependent normalisation. Total intensity normalisation makes a number of assumptions, one being that the total hybridisation intensities summed over all elements in the arrays are the same for each sample i.e. that dye intensities within a slide are related by a constant factor, so the intensity of each spot can be scaled by that factor. However, the fact that stronger signals dominate the summation is a drawback of this approach, although variations in the method have been developed to address this and include the median method, trimmed mean (trims 5% of highest and lowest extreme values and then globally normalises the data using the mean method) and the global intensity method (specifically for oligonucleotide arrays). Intensity-dependent normalisation can be linear or non-linear. Ahmed (2006b) states that intensity-dependent linear regression is noted for its ability to normalise intensity dependent dye bias arising in 2 colour channel microarrrays (which occurs when fluorescent dyes Cy3 and Cy5 emit unequal light resulting in low correlation of signals between Cy dyes (Ju et al., 2007)). Intensity-dependent linear regression involves firstly making a visual display of the data distribution to visualise how the intensity and dye bias relate via an MA plot (i.e. a scatter plot of log ratios vs. log intensities) that enables intensity specific artefacts to be revealed. Normalisation is then performed by applying a statistical regression method known as locally weighted scatterplot smoothing (LOWESS) (Kepler et al 2002).. Ahmed (2006b) also notes that a

colour normalisation method is used to eliminate data artefacts introduced by dyes.

113. After using normalisation to remove or minimise the systematic variations, gene expression matrix tables are generated in which rows represent genes and columns represent various samples e.g. experimental conditions or tissues.

b) Normalisation for oligonucleotide microarrays

114. For oligonucleotide microarrays, several linear and non-linear normalisation approaches are available. Boes & Neuhauser (2005) refer to the linear scaling approach as a method that uses a baseline array, in which the first array or the one in the middle of the dataset can be specified as the baseline. Slonim & Yanai (2009) refer to this approach as mean-signal or "scaling" and consider it to be the simplest normalisation method since each microarray's expression level is adjusted against the same level. Scaling, although minimal, avoids over-normalisation and is thought particularly worthwhile if the samples to be compared are expected to have similar average levels (e.g. they come from the same tissues and developmental stage, or have similar mRNA quality, etc). Non-linear methods such as cyclic-loess, contrast based method and quantile normalisation are described as complete data methods because they make use of data from all arrays in order to form the normalisation relationship i.e. they normalise without specifying a baseline array. For example quantile normalisation equalises the distribution of probe signals over all n arrays, so that after normalisation all arrays have the same distribution of probe intensities such that the most highly expressed value is set to be the same across arrays, as is the next most highly expressed and so on (Slonim & Yanai, 2009). Boes & Neuhauser (2005) recommend this method but emphasise that this cannot be defined as a gold standard, and indeed that it is not possible to define one, as no single method could ever be suitable for all circumstances.

(c) Limitations of normalisation

115. A significant drawback of normalisation is the possibility that it may modify the data by reducing both technical and biological variation. It is considered particularly important that biological variation is not reduced otherwise this may affect the outcome of significance testing and elevate false discovery rates. Therefore, it is best to minimise the amount of normalisation by having a good experimental design.

116. Which normalisation method to choose depends on the design of the microarray used (platform) and how much the resulting data set changes when the normalisation method is applied. The global method is recommended when the dataset is generated from large microarrays containing thousands of gene sequences reflecting a broad range of cellular activities. Normalisation methods based on housekeeping genes are recommended when the dataset is generated from a focussed array. Boes & Neuhauser (2005)) define a good normalisation method as one that considers

precision, accuracy, practicability (computing time) and the impact on significance testing and false discovery rate. Ittrich (2005) notes that the usefulness of different normalisation procedures can be compared by assessing both the correction of bias (improved accuracy) and the improvement of variance (precision). However, normalisation is considered a trade-off[27] between bias and variance as both parameters cannot be optimised simultaneously.

117. Various software tools are freely available as either a desktop package e.g. BRB-ArrayTools or via the Internet e.g. SNOMAD (Ahmed, 2006b) for normalisation.

(v) Filtering

118. Filtering is considered to be a vital preprocessing step to remove unreliable data prior to analysis (Yauk & Berndt (2007). Filtering removes genes that are either not expressed over all samples or show little variation across sample types. Those probes with low intensities have values near or below the noise level of the assay and therefore represent questionable results and so filtering them out improves the reproducibility of subsequent gene lists, reducing the number of genes that have to be tested for differential gene expression (Ahmed, 2006b, Durinck 2008). Filtering is typically applied prior to detection of differentially expressed genes (Durinck 2008), as a threshold for the variance of the gene across chips, an arbitrary threshold for a test statistic based on for e.g. t-test, or on excluding a certain percentage of the genes (Ahmed, 2006b). However, because the methods used are arbitrary, they should be used conservatively to filter out only the least differentially expressed genes (Ahmed, 2006b).

[27] It is generally accepted that as the model complexity of a procedure increases, variance tends to increase and squared bias decreases; the opposite behaviour occurs as the model complexity is decreased (Ittrich, 2005).

SECTION 3. STATISTICAL APPROACHES USED TO IDENTIFY/ EVALUATE TOXICOLOGICALLY RELEVANT GENE EXPRESSION CHANGES

INTRODUCTION

119. Members previously noted that the statistical analysis of microarray data is conducted on both the gene level (to test hypotheses, estimate the size of differences in gene expression and explore data) and on the pathway level (to determine which genes within a pathway correlate and to establish which pathways are differentially regulated). Indeed, statistical methods are employed throughout both high- and low-level data analysis, and are used in the former:

a) to test hypotheses and thereby identify toxicologically relevant gene changes (i.e. detect differentially expressed genes (DEGs)), and;
b) to evaluate these toxicologically relevant gene changes by detecting patterns in data (data-mining) and interpreting them by classifying the functional dependency of these genes (Suarez et al., 2009).

120. This section discusses the various statistical approaches used in the high-level data analysis of toxicogenomic (transcriptomic) data and approaches to validate/interpret toxicologically relevant gene changes.

A. DATA ANALYSIS

121. High-level data analysis comprises the second part of the analysis of raw (albeit transformed) toxicogenomic data. Its two key objectives are to: (i) identify gene changes by testing hypotheses of interest (e.g. class comparison, class prediction and class discovery), generating a gene list of differentially expressed candidate genes; (ii) extract patterns/trends in the data, i.e. pattern recognition by applying data mining methods, graphically presenting gene groupings that could be further investigated (Repsilber et al 2005). Which data analysis strategy to adopt depends on the purpose of the microarray experiment and the extent of the user's knowledge of the biology of the system being studied. Various methods are available to determine either approach, and these have been categorised into one of two of the following groups: supervised[28] learning methods (that use information about the various samples being analysed in a supervised fashion and are applied to both class comparison and class prediction studies); and unsupervised learning methods (that characterise components of [i.e. find relationships within] a dataset without prior information about the sample – also referred to as exploratory analysis and applied to class discovery studies) (Butte et al 2002, Ahmed, 2006b; NAS, 2007b). It is worth noting that although class comparison, prediction and discovery studies each have theirs own set of statistical methods, these methods can overlap.

122. A schematic of the data analysis stage is provided in Annex 2

[28] Most statistical approaches are supervised (NAS 2007)

(i) Hypothesis testing

123. The key aim of testing hypotheses of microarray experiments is to identify differentially expressed genes i.e. whether or not the functional group shows altered expression. To achieve this, investigators must first ensure that the experiment is powered to answer the hypothesis tested (which often requires using a smaller number of probe sets/functional groups of genes) (Ahmed, 2006b). Investigators must decide on the level of gene analysis to perform (i.e. whether to make single or multiple gene comparisons), the approach used to adjust for multiple testing and select DEGs. The outcome of hypothesis testing is a list of genes that are believed to be regulated by the condition being tested, although sometimes the outcome should be considered more as hypothesis generation. These issues and approaches are described for class comparison studies in the following text.

1. Single or multiple gene level comparison

124. Class comparison studies seek to find genes with expression levels that are significantly different between groups of samples and investigators are required to choose whether to base the analysis on single gene or multiple gene comparisons. Single gene (pairwise) approaches compare microarrays one pair at a time, and examine each gene or transcript individually to find genes that [by themselves] have statistically significant differences in expression between samples with different phenotypes or characteristics (Slonim & Yanai 2009). Once these genes have been identified they will undergo further examination to see if they are over-represented in specific functions or pathways (See Section 3.2). It is thought that single gene comparisons may be more appropriate for studying a biological process that is poorly understood (as it allows hitherto unexpected genes and gene sets to be implicated) although the fact that they could miss trends existing between measurements is a significant drawback (Butte, 2006; Slonim & Yanai 2009).

125. Multiple gene (or gene set) comparisons identify groups (sets) of functionally related genes ahead of time and test whether these gene sets (as a group) show differential expression (also referred to as Gene Set Analysis - GSA) (Slonim & Yanai 2009). GSA is considered a powerful alternative to pairwise comparisons as it can detect subtle changes in gene expression that individual gene expression analysis may miss. GSA also combines identification of differential expression and functional interpretation into a single step, however, it is limited by the need to identify the appropriate gene sets ahead of time.

126. Hayes & Bradfield (2005) summarise the approach used to identifying sets of co-ordinately regulated (overrepresented) genes, which involves assuming all genes are regulated independently and look for genes that deviate from this. Any gene profiles that correlate can then be examined for co-regulation and associated biology.

2. Selecting differentially expressed genes

127. Threshold and statistical-based approaches are typically used to select DEGs.

a) Threshold based approaches

128. In single gene comparisons, threshold approaches have used cut-offs such as fold change to select DEGs. Such studies end up using fewer microarrays as only one or two microarrays are required per experimental condition (Ahmed, 2006b). However, the drawbacks of using this approach are well documented in the published literature; the main limitations being the arbitrary nature of fold changes (in which the spot intensity is used to determine how reliable fold increases and decreases are) thereby questioning its sensitivity and reliability (Irwin et al 2004). Furthermore, the fact that fold-change cannot address the reproducibility of absolute differences or provide a level of confidence about statistical significance of microarray data renders its use a dubious one (the exception being if the microarray study is used solely as a preliminary or coarse screen (Ahmed, 2006b)). The use of statistical significance as a criterion to generating a gene list (P-value) is reported to be a possible solution. Statistical approaches provide the most reliable and unbiased way of selecting DEGs, enabling the precise measurement of genes exhibiting even a small fold increase or decrease in expression (into which many important genes fall) (Irwin et al 2004). Indeed, Morgan et al (2004) consider that the statistical level of change is more relevant than fold change and suggest that use of fold-change cut off approaches should be avoided.

b) Statistical based approaches

129. Various types of statistical methods are available to detect DEGs. These 'supervised approaches' report the probability of the observed test score occurring by chance under the null hypothesis that there is no difference in expression related to the phenotype being studied (Butte, 2002; Slonim & Yanai, 2009). The chosen method depends largely on the level of gene analysis employed.

130. For single gene comparisons, the available statistical tests are described as being either parametric, non-parametric or Bayesian based. Parametric tests make assumptions about the normality of data. Examples include the t-test statistics (paired t-test, Welch t-test) and ANOVA. These tests look for differences in the average expression level between groups (Irwin et al 2004; Morgan et al 2004; Slonim & Yanai, 2009). However, since the assumptions regarding normality are often inappropriate the reported P-values are more appropriately used as a guide to prioritise genes rather than accurate probabilities. ANOVA is used to determine the statistical significance of increases and decreases in gene expression and provides a solid statistical basis[29] (based on p-values) for identifying DEGs (Irwin et al 2004). ANOVA F-test has been used with the One–Vs–All (OVA) test to identify genes that

[29] But based on same assumptions re: variance distribution as t-tests

significantly varied (changed in expression) in treatment groups (Tsai et al 2005).

131. Non-parametric tests make no assumptions as to the distribution of the data and various tests are used. These include: the Wilcoxon-signed rank test/Mann-Whitney U Test – an alternative to parametric t tests (although its use is limited by its reduced power to detect important differential expressions); and the significance analysis of microarrays (SAM) – that can also be used to correct for multiple experiments by utilising the false discovery rate (FDR) concept to assist in determining a cut-off after performing adjusted t-tests (although this can become computationally intensive) (Ahmed, 2006b; Zhou et al 2009).

132. Bayesian methods are used to order microarray expression (and are also used in error measurements and quality control).

133. Multiple gene comparisons may use the Bonferroni correction method to adjust for multiple significance testing (although it is mainly used to adjust for large false positive rates (see below) (Lee et al 2005).

134. Butte (2002) notes that use of the above statistical approaches requires consideration of the following factors to help rank genes that are most significantly different: absolute expression level (i.e. is expression high or low, since low-level expression is often (but not always) associated with less reliable measurements and poor reproducibility); subtractive degree of change between groups (i.e. the difference in expression level between samples); fold change between groups (i.e. ratio of difference in expression level between samples); and the reproducibility of measurement (i.e. do similar samples produce same levels of expression).

135. Volcano plots are also used in studies to graphically represent DEGs and help in selection, and involve comparing the size of the fold change to the statistical significance level.

136. A key limitation noted for statistical based approaches to identifying DEGs is that the final result can be dependent on the algorithm used (Ju et al 2007). Therefore, it is recommended that investigators use different analytical methods with the same data sets to determine which best suit the experimental design.

3. Adjusting for multiple testing

137. A key stage in analysing microarray data and indeed a critical statistical issue in class comparison studies is adjusting the data for multiple testing. The COT previously reported the problems associated with multivariate analysis[30], notably the n poor, p rich dilemma, in which testing a large number

[30] Multivariate analysis is defined as the simultaneous analysis of multiple variables (gene expressions). NB. Univariate analysis (e.g. pairwise comparisons) is concerned with Type I (i.e. the probability of rejecting the null hypothesis when it is true) and Type II (i.e. the probability of accepting the null hypothesis when it is false) errors (Suarez et al 2009).

(thousands) of variables (genes or ‘p’), with comparatively only a few number (hundreds) of samples (experiments or ‘n’) analysed, increases the likelihood of false positives. This typically results in the over-fitting of data onto statistical models, greatly inflating the number of significant results (when in fact most are false positives (FP)). Such high dimensionality presents a significant challenge for statistical methods as it makes the visualisation of samples difficult and thereby limits the exploratory potential of the data. Various approaches are available to account for this. However, most of these methods involve specific assumptions and characteristics that the experimenter should be aware of before choosing to apply them. The Family Wise Error Rate (FWER) defined as the probability of yielding one or more FP out of all hypothesis tested is commonly used, although the associated low power of the FWER method may cause many potentially interesting genes to be missed. The most preferred approach to adjust the high dimensionality of multivariate analysis is to control the false discovery rate (FDR).

a) Controlling the False Discovery Rate

138. The FDR is defined as the probability that any particular significant finding is a false positive (Slonim & Yanai, 2009). Ahmed (2006b) further defines FDR as the fraction of truly unchanged genes that appear as FP or false negatives (FN) i.e. the rate at which significant features are truly null (whereby a FDR of 5% means that on average, 5% of the genes found to be significantly differentially expressed are not i.e. are FP). Controlling this is a common approach to balance FP and FN (Zhou et al 2009)

139. Methods used to control FDR include the use of the Bonferroni correction (which adjusts the false positive rate according to study objective (Lee et al 2005)), although it has been considered too conservative and inappropriate for when many thousands of genes are being compared as it can result in many FN (Ahmed 2006b)); permutation-based methods; and the Benjamini-Hochberg step down method. These methods calculate a ‘q’ value, which is similar to the p value as a measure of significance and offer a reasonable combination of statistical rigor and power. However, Vlaanderen et al (2010) notes that the strongest safeguard against FP results is provided by replication of initial findings in follow-up studies.

(ii) Data mining

140. Data mining aims to find patterns in data and the approaches used depend on experimental hypothesis/ type of study e.g. class prediction or class discovery which adopt either supervised or unsupervised approaches (or both) (Ahmed, 2006b). Rahnenfuhrer (2005b) notes that these computational (i.e. bioinformatical) methods help to classify samples and genes. Sample classification approaches aim to find groups of samples that have similar gene expression patterns while gene classification approaches aim to identify groups of genes with similar expression values in different samples (based on the biological assumption that functionally related genes exhibit similar expression patterns under different conditions).

1. Pattern recognition in class prediction studies

141. Class prediction studies typically aim to classify unknown samples into predefined groups based on the expression levels of key genes. These studies also attempt to predict the toxicological class of an unknown toxicant based on gene expression signatures of samples and therefore constitute key approaches in predictive toxicology. Both objectives are achieved via the application of a classifier (i.e. supervised learning method) to gene signatures of training set samples, which generates a mathematical model for predicting the toxicological class of the unknown sample (Magglioli et al 2006). A range of classifiers are available, and the type used depends on whether single or multiple genes are being analysed (Hayes & Bradfield 2005). For individual gene analyses, Bayesian probability, Linear Discriminant Analysis (LDA), and Genetic algorithm (GA)/K-Nearest Neighbours (KNN) have been used, while for multiple (geneset) analyses Support Vector Machines (SVM) are commonly used as well as decision trees and neural networks. Evaluating the classifier-generated models using individual samples from the training set comprises the final validation step of a class prediction study, including estimating the model's success rates to predict toxicological class of unknowns.

142. Thomas et al (2001) applied a probabilistic approach based on Bayesian statistics to classify toxicants based on their mRNA transcript profile effects. Male C57BL/6J mice were exposed to 24 known (model) toxicants that fell into 5 toxicological classes i.e. non-coplanar PCBs, peroxisome proliferators, inflammatory agents, hypoxia inducing agents and aryl hydrocarbon receptor agonists. Following microarray analysis of gene expression changes in liver, the authors were able to classify toxicants with up to 70 % accuracy using the total gene set of 1200 transcripts, and were also able to identify and use a diagnostic set of 12 transcripts to predict with 100 % accuracy. The authors concluded that the use of classifiers not only provides huge cost savings but identifying a diagnostic gene set renders large arrays unnecessary for classification purposes. However, the authors acknowledge caveats within the study including how the toxicological categories selected primarily reflect the model compounds that toxicologists had extensively studied at the time, which represented only a small percentage of the chemicals in commerce to date. The authors consider their study as an early step towards accurately classifying toxic chemicals according to their transcript expression profiles.

143. Although, clustering techniques are generally used as unsupervised approaches (see below) the nearest neighbour (NN) clustering algorithm is commonly used in a supervised fashion to find genes whose patterns match a designated query pattern (Butte, 2002). The K-NN algorithm is very simple to use and understand and its accuracy as a classifier can be improved via the use of specific noise reduction techniques. The algorithm first compares the similarity of expression patterns of measured genes (i.e. a test set) with an ideal gene pattern (training set) (by calculating distance of these expression patterns from training set) and ranks them according to their similarity with the ideal gene pattern to decide on similarity of mechanism of toxicity. Limitations

associated with this algorithm include the fact that it is sensitive to irrelevant or redundant features (since all features contribute to the similarity). Furthermore, K-NN may be outperformed by other techniques such as SVM.

144. Steiner et al (2004) used two different SVM algorithms to produce predictive models to determine whether biological samples from rats treated with various compounds could be classified into different classes of hepatotoxicants based on gene expression profiles. Recursive feature elimination was also used to enhance the ability of SVM to create sets of informative genes. The authors were able to predict toxicity as well as mode of toxicity to discriminate hepatotoxic from non-hepatoxic compounds and correct toxicant class. Steiner et al (2004) also investigated the effect of strain differences for classification and generated a SVM algorithm using Wistar rat data to see whether it could correctly classify individual animals from Sprague Dawley rats. They found that the predictive model built on transcripts from Wistar strain could successfully classify profiles from Sprague-Dawley strain.

145. Various limitations associated with the use of these algorithms and potential ways forward have been reported. Magglioli et al (2006) notes that given the numbers of classes and chemicals within each class will increase, the validation of a robust method that can incorporate and accurately predict toxicity using much larger data sets, represents a significant obstacle which hopefully future algorithms may address (Magglioli et al., 2006). Furthermore, to address the tendency of these algorithms to create models that over-fit the data thereby making it difficult to predict a future dataset using the same prediction rule, Mayo et al (2006) recommend splitting data into training set and validation set.

2. Pattern recognition in class discovery studies

146. Class discovery studies use unsupervised learning methods to visualise gene expression similarity (Magglioli et al (2006). However, as with other analytical approaches, the chosen method depends largely on whether single or multiple genes are being analysed. Single gene based analyses use principal component analysis (PCA) to find genes with interesting properties without looking for an a priori pattern. In multiple gene analyses, unsupervised methods such as clustering analysis and self-organising maps are used to find groups of genes/samples with similar patterns of gene expression. Unsupervised methods are also used on a network level to find interactions between genes. Examples include Boolean, Bayesian and Relevance networks, which are discussed further below. Class discovery is considered a subjective approach as the results tend to be influenced by the clustering algorithm and similarity metrics selected (Magglioli et al 2006).

a) Principal Component Analysis (PCA)

147. PCA is a multi-purpose mathematical approach that serves as both a visualisation and analytical technique, although its former use is considered more valuable (Butte, 2002). In class discovery studies, PCA is used to describe the variation seen in a multiple value/ expression data set (i.e. to

enable a visual assessment of the similarities and differences between samples) and thereby determine whether samples/genes can be grouped (Hayes & Bradfield 2005). This is achieved by summarising (reducing) the dimensionality of the data and graphically representing the reduced data to identify clusters and detect outliers (Thompson & Hackett, 2008). Reduction is accomplished by identifying directions (also known as principal components) along which the variation of data is maximal i.e. by linearly combining the components (transcripts) so that the first components represent the greatest amount of variability (Mei et al 2009). By using only a few components (e.g. the first three) each sample or gene can be represented by relatively few numbers of points in a multidimensional space instead of by values for thousands of variables (Ringner, 2008). PCA is particularly advantageous because it is able to reduce the dimensionality of data while retaining most of the variation in the data set. However, it is limited by its inability to describe how best to separate groups of genes or samples and the need to know whether genes have been centred (normalised?) before analysis which is not always documented (Butte, 2002)

b) Cluster analysis

148. Cluster analysis is a commonly used unsupervised technique to visually determine patterns in large data sets (Irwin et al (2004); Hayes & Bradfield (2005); Ju et al (2007)). Cluster analysis helps identify relevant biological structure[31] in the data and can be applied either to samples to identify those sharing similar gene expression profiles, or to genes to identify those that behave similarly across various experimental conditions, which may indicate a possible relationship i.e. belonging to the same biological pathway (Ahmed, 2006b).

149. Several essential steps are performed in cluster analysis, which include (i) calculating the Euclidean distance from heatmap; (ii) applying average linkage algorithm to distances between genes and plotting as a dendogram (iii) selecting an appropriate clustering method (iv) performing correlation tests.

150. Heat maps are grids of coloured cells where each colour represents a gene expression value in the sample i.e. a matrix of gene expression values where genes are represented by rows and columns represent samples (Rahnenfuhrer (2005b); Ahmed, 2006b). Heat maps are considered particularly useful as they provide an overall view of the expression profile. Conventionally, red denotes increased expression, green denotes decreased expression and black denotes intermediate expression (Ahmed, 2006b). Although heat maps are used to visually determine patterns in data sets, they also provide a useful starting point for cluster analysis as they help calculate similarity or dissimilarity values used in constructing a hierarchical tree dendogram.

151. Similarity or dissimilarity measures (aka metrics) indicate the degree of similarity between genes. They are calculated from heatmaps and used by clustering methods to build groups of genes with similar patterns of

[31] Cluster analysis also identifies structure caused by systematic biases in the data (Mayo et al 2006).

expression, resulting in, for example, the construction of dendograms (Rahnenfuhrer, 2005b). An appropriate dissimilarity measure must be chosen so that an appropriate analytical technique can be applied. A common metric used is the Euclidean distance, which ranks similarities of gene expression profiles by treating each gene as a point in a multidimensional space (where the X and Y axis represent different samples and the axis co-ordinates denote the amount of gene expression per sample) (Butte, 2002). Use of this metric is limited by the fact that it can miss correlations of measurements (if measurements are not normalised) and genes negatively associated with each other (e.g. those associated with tumour suppressor genes). Pearson's correlation coefficient (r) provides an alternative metric and measures the strength of association between genes and is calculated from distances of each point from the linear regression line (aka residuals). However, r is particularly sensitive to outliers and involves various assumptions (i.e. normal distribution, which may not apply to datasets arising from oligonucleotide microarrays, and linear gene interactions). This can be rectified by replacing measurements with ranks (via calculation of rank correlation coefficients) (Butte, 2002).

152. Several types of clustering methods (algorithms) are available and most expression analysis tool kits include some clustering or visualisation tools (Slonim & Yanai, 2009). Clustering tools are often accompanied with non-distance dimension reduction-based methods such as PCA and multidimensional scaling, to facilitate visualisation and provide new smaller sets of independent dimensions (which contain most of the information from the original data) by projecting data onto a lower dimensional space (Rahnenfuhrer, 2005b; Ahmed, 2006b). The two most common clustering techniques (hierarchical and K-Means clustering) are described below.

153. Hierarchical clustering (HC) is a commonly used unsupervised technique that builds clusters of genes that have similar patterns of expression (Butte, 2002). It uses the similarity of expression to organise data into groups of highly correlated genes (clusters) (Hayes & Bradfield 2005). These algorithms find successive clusters using previously established clusters and are usually either agglomerative (bottom-up) or divisive (top-down). HC is considered a particularly useful clustering algorithm as it produces a dendogram to visualise overall similarities in expression patterns observed in an experiment. Dendograms visualise resultant clusters by representing genes as leaves of a large branching tree. The branches link genes and the branch length indicates level of correlation whereby short branches denote similarity between genes and long branches dissimilarity between genes. HC is also a popular algorithm because it allows users to easily estimate the number and size of expression patterns within a data set (Rahnenfuhrer, 2005b). Furthermore, the fact that each cluster is further divided into subtrees makes it more informative than k-means, for example (Ahmed, 2006b). Disadvantages associated with HC include the loss of information (e.g. negative associations) due to the enforcement of tree structure to data and the lack of a probabilistic foundation to guide decision as to where to cut the dendogram, which can be rectified by using external criteria to choose the number of clusters (Butte, 2002). HC is also considered

unsuitable for finding up and downregulated genes in experiments (Ahmed, 2006b). HC is considered more suitable for clustering genes than samples since the number of genes is usually several magnitudes greater and therefore often only compact subgroups of genes are sought after (Rahnenfuhrer, 2005b).

154. K-Means clustering is a type of partitioning algorithm that determines all clusters at once. Such algorithms seek to minimise the heterogeneity of clusters and maximise their separation for a given number of clusters. K-means is the most widely used partitioning cluster algorithm due to its simplicity/ low computational complexity and speed (Ahmed, 2006b). It organises the data by producing divisions of the data set (bins) that are based on a predetermined number of groups/clusters that are viewable in tabular format (Hayes & Bradfield 2005; Magglioli et al 2006). However, it is limited by its assumption that each gene fits into only one cluster.

155. Zhou et al (2009) report on a newly developed profile-based method for Extracting microarray gene expression Patterns and Identifying co-expressed Genes (EPIG) (first reported by Chou et al 2007). EPIG uses a filtering process to extract biologically informative patterns and co-expressed genes more effectively than other techniques such as CLICK (cluster identification via connectivity kernels). It evaluates the correlations among profiles, the magnitude of variation in gene expression profiles, and profile signal-to-noise ratios without a pre-defined seeding of the patterns.

156. Slonim & Yanai (2009) highlight a significant drawback of using clustering approaches i.e. the possibility of finding predominant patterns in the data that do not correspond to the phenotypic distinction of interest in the experiment. They suggest the use of more directed methods to identify gene expression patterns related to a particular distinction. The fact that every clustering algorithm yields some grouping regardless of the true structure of the data is of concern as this can lead to an overly optimistic interpretation (Rahnenfuhrer, 2005b). Therefore, the deployment of an objective judgment of clustering results to assess the quality is recommended. Slonim & Yanai (2009) recommend the use of different methods which can reveal different patterns. In addition, the use of different algorithms to look at broad patterns in each data set can help rule out correlations with possible confounding variables, such as day effects.

c) Self-organising maps

157. Self-organising maps (SOM) are similar to HC algorithms but use a different approach to survey expression patterns within a data set (Butte, 2002; Ju et al 2007). SOM represent genes as points in a multidimensional space and provide a 2-D visualisation of expression patterns, with reduced computational requirements. However their use is limited by the arbitrary nature of the shapes and issues related to their reproducibility. SOM are also criticised for their potential to miss negative associations.

(iii) Microarray data analysis packages

158. A range of commercial and publicly available software packages are used to analyse TGX data (Slonim & Yanai, 2009). Commercial packages although plentiful are limited by their cost. Furthermore, they also have limited flexibility. However, a number of web-based tools are freely accessible online and include: Gene Set Enrichment Analysis (Broad Institute) used in gene set analysis; DNA-Chip Analyser (dChip) (Harvard University) – a Windows software package for probe-level (e.g. Affymetrix platform) and high level analysis of gene expression microarrays and SNP microarrays; BRB-Array tools that provide various utilities for processing expression data from multiple experiments, visualising data, multidimensional scaling, clustering and classification and prediction of samples; and Pipe, MeSHer[32] and RACE (Ahmed, 2006b; Suarez et al., 2009).

159. Open-source software packages provide their programme source-code freely for use or modification. The most commonly used is the Bioconducter Project, an open-development computational platform that provides tools for the analysis and comprehension of genomic data. Bioconducter is continually updated with the development of new methods and uses R-programming language based statistical software systems such as R/maanova that can be used for data quality checks and visualisation, data transformation, ANOVA model fitting and various statistical tests including cluster analysis (Ahmed 2006b; Suarez et al 2009). Other available open-source software packages include the Java-based TM4 software system developed by the Institute for Genomic Research, MD, USA and BASE, a web-based system developed at Lund University, Sweden (Repsilber et al 2005).

160. Members previously commented on the need for generic guidance on the most suitable methods to evaluate TGX data, given the wide range of analytical approaches available and subsequent lack of standardisation in data analysis methods. Members also highlighted the lack of statistical experts to peer review published studies, which would exacerbate the situation. Findings from the HESI Committee on the Application of Genomics to Mechanism-based Risk Assessment cross-sector international online survey revealed that principal component analysis (PCA), hierarchical clustering, and analysis of variance (ANOVA)/statistical analysis of microarrays (SAM) were the most commonly used statistical (computational) methods for analysing microarray data across all 112 respondents (Petit et al 2010).

161. Suarez et al (2009) considers further development is required in the following areas: methodology for pairwise comparison; preprocessing data (wrt different platforms to collect microarray data, segmentation procedures, normalisation methods, use of mismatched probes); and statistical inferences (wrt different t statistics, methods to control proportion of false positive declarations and problems in controlling correlation among genes and among tissues).

[32] Now integrated into MeV (multi Experiment Viewer)

B. DATA VALIDATION/CONFIRMATION

162. Computer-generated gene lists contains 10s to 1000s of transcripts that have been statistically significantly up/down regulated cf. reference population. After detecting patterns in regulated genes, further analysis is necessary to biologically validate the data. Several levels of validation can be distinguished. The first level requires use of an alternative assay/gene expression technique to determine whether there was a real difference in expression between the samples in a study (and typically involves the numerical verification of expression levels). The second level seeks to determine whether the expression is really affected by the treatment and achieves this via biological replication. Establishing the biological relevance of these gene expression changes comprises the next validation step of TGX data. The objective is to ascertain meaning (i.e. interpret the significance of) these regulated genes e.g. do the changes observed actually mean anything (and typically involves finding common promoter regions and biological relationships between genes) (Butte, 2002).

(i) Numerical verification of regulated gene expression levels

163. Confirmation of microarray data requires complimentary validation on a few genes by a quantitative method. Indeed current publication guidelines require that all microarray results are confirmed by an independent gene expression profiling method. Various approaches can be used such as Northern blotting, ribonuclease protection assays, and *in-situ* hybridisation. However, quantitative real-time reverse transcription polymerase chain reaction (RT-PCR)[33] is generally considered the method of choice (Ahmed, 2006b).

1. Quantitative real-time RT-PCR

164. Real time RT-PCR validates genes whose expression was found to be altered by the microarray analysis. It takes sample mRNA and quantifies the amount present thereby providing a measure of gene expression. Real time RT-PCR is considered to be particularly advantageous because it rigorously quantifies gene expression when mRNA levels are low and automates laborious processes involved in PCR (i.e. data analysis, standard curve generation and copy number generation) by quantifying reaction products for each sample in every cycle, thereby removing the need for user intervention or replicates.

165. It is recommended that real time RT-PCR is performed for all genes of interest and a suitable housekeeping (reference) gene (whose relative expression level can used to normalise the expression of the genes of interest to control for sample to sample variation) (VanGuilder et al 2008). The same housekeeping gene used for the microarray experiment could also be used (if

[33] It is worth noting that there the published literature uses RT-PCR to abbreviate either 'real-time PCR' or 'reverse transcription PCR' which can be confusing. In this review, RT-PCR refers to reverse transcription PCR.

applicable), however mRNA for glyceraldehdye-3-phosphate dehydrogenase, beta-actin, MHC1 and cyclophilin tend to be more commonly used. The design and availability of primer pairs (sets)[34] may impact on the outcome of the gene expression validation step. This is because in-house primer sets are relatively cheap compared to those commercially available. However, it is reported that in house primer sets have potential for yielding non-specific amplification products while commercial primer sets are more reliable as they are designed by experimentally verified computer algorithms and tested in a quality control assay.

166. After the PCR amplification rounds (range 20-40) that use fluorophores that permit measurement of DNA amplification during PCR in real time, two quantification methods are used to determine gene expression. The standard curve method determines the relative level of expression of the genes of interest and housekeeping gene by firstly constructing a standard curve from RNA of known concentration (standards used can be either RNA, purified plasmid ds DNA, in-vitro generated ssDNA, and cDNA sample expressing the target gene). The curve is then used as a reference standard for extrapolating quantitative information for mRNA targets of unknown concentrations, followed by spectrophotometric assessment of the concentration of standards. The comparative c_t method involves comparing c_t values of the sample of interest with a control/calibrator (i.e. non-treated sample or RNA from normal tissue) and normalising the values of both calibrator and samples of interest to an appropriate endogenous housekeeping gene (Ma et al., 2006).

167. The different real time PCR approaches (or chemistries) include TaqMan Probes ®, Molecular Beacons, Scorpions ® and SYBR® Green. These chemistries allow detection of PCR products via the generation of a fluorescent signal. SYBR Green is considered the simplest and most economical format for detecting and quantifying PCR products in real time PCR. The SYBR Green dye binds to double stranded DNA but not to single-stranded DNA and is frequently used in real-time PCR reactions. When bound to double stranded DNA SYBR Green strongly fluoresces. However, users are required to design gene specific primer sets to avoid co-amplification of non-specific secondary products since SYBR Green can detect any double stranded DNA non-specifically.

168. In general, reverse transcriptase PCR methods are limited by the fact that the extra reverse transcription step makes it less quantitative than PCR of DNA. However, northern blotting is not as popular a gene expression validation method, because it requires relatively large amounts of RNA, and provides only qualitative or semi-quantitative information of mRNA levels. Similarly, use of *in-situ* hybridisation is limited by the fact that it provides qualitative rather than quantitative information.

169. Members previously queried whether studies were using alternative platforms to validate microarray data. Petit et al (2010) reports that most (60%)

[34]PCR Primer sets are complimentary to a defined sequence on each of two strands of DNA and extended by DNA polymerase.

HESI Committee survey respondents used RT-PCR to assess expression levels of specific genes and confirm array results. RT-PCR confirmed 60-89% of microarray data in most sectors. It was further suggested that the reliability of TGX technology is currently being miscommunicated at higher levels within organisation structures due to the varying responses among vice presidents and study directors as to the frequency with which confirmatory analyses supports primary microarray findings (e.g. vice presidents considered the confirmation rate to be 10% cf. 50% reported by study directors).

(ii) Data interpretation

170. As previously noted the result of high-level data analysis is the generation of a list of genes (or more precisely probes represented on an array) believed to be significantly regulated by the condition under scrutiny. However, this list is limited by the fact that it does not indicate the pathological or physiological processes that took place in the samples under consideration (Brors, 2005). The challenge, therefore, is to transform this list of DEGs, predictive or co-regulated genes into entities that are more quickly understood in terms of traditional biology. Consequently, an approach to validating the biological significance of regulated gene patterns is to interpret the results in relation to their regulation, function and biological relevance (Butte, 2002; Morgan et al., 2004). Such post analytical work is considered to be a rate limiting step and is represented by the following steps. Firstly, researchers must annotate the probes (genes) by collecting structural information on the probes regulated on the microarray (i.e. associate the probe with biological entities e.g. genes, transcripts or proteins). This is accomplished by surveying databases that provide information for example, on biological sequences to identify genes represented on microarrays. Next, the function of gene products must be defined in a systematic and consistent way i.e. mapping microarray probes onto protein databases providing information on protein functions e.g. via Gene Ontology System, and SwissProt keywords. The preceding steps represent structural and functional components of gene annotation i.e. the assignment of gene level information based on the probeset sequences whereby the name and symbol of the gene is interrogated as well as its general function. Finally, pathways or biological functions that are overrepresented in a given gene list can be identified (pathway analysis). It is therefore essential that gene annotations are regularly updated to enable systems biology modelling.

<u>1. Structural gene annotation</u>

171. The structural annotation of genes with sequence, structure and locus information constitutes a significant challenge due to the fact that genome databases are either incomplete or faulty. Brors (2005) notes that the challenge of correctly identifying the intron-exon structure of genes is further aggravated by the fact that the coding sequence in humans is only three percent, which thereby results in many errors in genome databases. Other annotation issues highlighted in the literature relate to significantly regulated genes not being ascribed names (official gene name); the lack of detailed information e.g. predicted protein domains or gene ontology classification; and

incorrect probeset design (e.g. against wrong DNA strand/species or probe sequences may not be unique for a particular gene) (Butte, 2002). Furthermore, the trajectory of data generation is set to increase given that before 2003 only 6 animals species were completely sequenced and published which rose to 103 animal species from 2003 to date (Van den Berg et al 2010).

172. Biological sequence databases exist to provide a repository for sequence information, further information on sequence domains and cross-references to other databases and literature (Brors, 2005). Sequence information for both nucleic acids and proteins is available in, for example EMBL/GenBank/DDBJ (which represent the same information but in different formats), which enable users to deposit nucleic acid data directly into the database. However, as the responsibility to curate[35]/correct errors lies with the depositor, databases are often not regularly curated which may result in them containing hundreds of redundant entries and inconsistent descriptions. In contrast, protein sequence databases (e.g. SwissProt) are regularly curated. Other databases used to provide structural annotation include various domain databases e.g. Prosite, medical literature databases e.g. Pubmed, protein structure databases e.g. Structural Classification of Proteins (SCOP) and position in genome databases e.g. Locus Link. The NCBI GEO is cited as being the primary repository for structural annotations of most commercial and custom-made microarrays (Van den Berg et al 2010).

173. Despite the above assurances, recent work suggests that gene annotations are still not regularly updated consistently after publication. For instance, Van den Berg et al (2010) found that although the FHCRC Chicken 13K cDNA v.2.0 microarray was developed and structurally annotated in 2004 and published in 2005 in GEO[36], it has only received one updated since 2006, despite the assignment of new and/or corrected structural and functional annotations. However, it is thought that this situation is likely to change with the acquisition of new genomics data and development of annotation tools.

2. Functional gene annotation

174. Bioinformatics provides insight into gene function, interactions, biomarkers, networks and pathways (Ju et al, 2007). Therefore, functional gene annotation consists of attaching biological information to the regulated probesets (i.e. biochemical and biological functional information and the regulatory and interactive associations). Although the same above challenges apply to the databases used to provide functional annotation, more than half the probes on a given array can be mapped onto databases (depending on the organism under investigation) (Brors, 2005). Various databases are available e.g. KEGG for biological pathway information, Transfac for regulatory signal information and Transpath for signal transduction information.

[35] Databases are often supervised by curating scientists (curators) who ensure consistency and non-redundancy of database entries (Brors, 2005).
[36] Gene Expression Omnibus (GEO) is a public functional genomics data repository supporting MIAME-compliant data submissions

a) Network analysis

175. Not only the profile of genes associated with a specific biological process are tested but also functional interactions between genes in a profile (Vlaanderen et al 2010). Various unsupervised learning methods are used to find interactions between genes. These include Relevance networks (RN), Boolean networks, and Bayesian networks (the latter used to search for true genetic regulatory networks i.e. hypotheses that the expression of one gene correlates with expression of another) (Butte, 2002).

176. Rho et al (2008) define a biological network as a composite of nodes (e.g. DNA, mRNAs, proteins and metabolites of cellular systems) and edges (e.g. the interactions between these nodes, which can be between genes, between gene and protein, between chemical and gene, etc). Such networks enable systems (defined as organs, tissues, cells and subcellular compartments) to function and they receive signals from these systems. Network modules describe a particular portion of a biological network activated to execute certain functions to offset perturbations caused by environmental or genetic events.

177. RN is an unsupervised technique that builds networks from genes, phenotype and clinical measurements. RN searches pair of genes that are likely to be co-expressed and compares features with each other by calculating a correlation coefficient (or other similarity measure) and choosing a threshold value whereby only those features with a measure greater than threshold are kept (NB. Can be used as a dial to increase and decrease the number of connections shown). Use of RN is considered particularly advantageous as it enables more than one data type to be represented together e.g. linking systolic blood pressure and the expression of a particular gene. It also allows for the visualisation of a variable number of associations for a particular feature (with nodes representing genes/ phenotypical measurements and edges (between nodes) representing associations), including the visualisation of negative associations. RN is limited by the fact associations at low thresholds are rather complicated.

178. Network approaches have been reported in the literature. Kulkarni et al (2008) propose a mathematical model entitled Toxicologic Prediction Network (TPN) as an approach to correlating gene expression changes caused by drug exposure to chronic toxicity. It is suggested that this approach avoids the difficulty associated with exposure to hepatotoxic drugs and the time consuming nature of animal experiments assessing their chronic toxicological impact.

b) Pathway analysis

179. Pathway analysis is an approach used to phenotypically anchor altered genes into biologically relevant pathways (Elashoff et al 2008). It detects pathways by identifying sets of genes with common characteristics and measures whether pathways are affected by compound either via an equation ([No of genes in pathway regulated by compound/ No of genes in pathway]/

[No of genes not in pathway that are regulated by compound/ No of genes not in pathway]) – which is limited by the fact that it assumes genes act independently; or by using a measure that accounts for the correlation between genes within a pathway. The procedure typically involves loading accession numbers and log-fold change values of DEGs into pathway analysis software e.g. Ingenuity Pathway Analysis (IPA) which then overlays/ integrates set of DEGs onto a knowledge database which provides a classification of gene products into molecular functions, biological proocesses and cellular components (Mei et al 2009). Pathway analysis can provide illustrations of linked transcripts but this is only speculative and further confirmatory investigations are needed. Therefore, once these data have been integrated, regulatory networks, functional analysis and canonical pathways altered in response to the chemical treatments can be explored. NB. Adjustment for multiple testing may be desirable although it is not easy to find appropriate method. Although there are currently large gaps in knowledge of biological pathways, each new study helps build the knowledge base (Vlaanderen et al 2010).

c) Ontological approaches

180. The Gene Ontology Database, conceived by the Gene Ontology Consortium describes a series of integrated publicly available tools that facilitate annotation by providing information on pathways, biological mechanisms and molecular functions. This ongoing Gene Ontology project provides ontologies of defined terms representing descriptions of gene products i.e. the non-overlapping domains of molecular and cellular biology (Ahmed, 2006b). This includes three independent categories for biological process (i.e. operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms), molecular function (i.e. the elemental activities of a gene product at the molecular level, such as binding or catalysis) and cellular components (i.e. the parts of a cell or its extracellular environment). Gene sets identified in microarray experiments as DEGs are tested for their association with a profile in the GO library (Vlaanderen et al 2010). The GO project not only enables inferences to be made of biological roles but it also thereby provides a means of transferring annotation from one organism to another.

181. Other ontological tools include Onto-Tools an annotation and integrated web accessible data mining suite (e.g. Onto-Express and Onto-Compare) that integrates data from sequence, gene, protein and annotation databases designed and implemented by the Intelligent Systems and Bioinformatics Laboratory (IBSL), Wayne State University, Detroit, USA. Onto-Express (OE) automatically translates lists of DEGs into functional profiles, which thereby help reveal biological mechanisms characterising effect of the treatment under study. OE also constructs function profiles for each of the GO categories and provides information on statistical significance of each pathway and categories used in the profiles to help users distinguish which mechanisms are significantly affected from those due to chance. Onto-Compare was developed to enable access to the biological significance of microarray data. Standards and Ontologies for Functional Genomics (SOFG) Mouse Ontology

Resources provides standards and ontologies for functional genomics (for mouse).

182. Members previously highlighted that interpreting the functional significance of gene expression changes comprised a critical area. The approach used in functional genomic analysis was recently considered in a study by Van den Berg et al (2010) that sought to demonstrate the impact of structural and functional re-annotation on systems biology modelling of functional genomics data. The authors chose to structurally and functionally re-annotated a chicken functional genomics dataset as it exemplified rapidly evolving annotations. The authors re-analysed a previously published differentially expressed mRNA experimental dataset generated using the FHCRC Chicken 13K cDNA v.2.0 microarray (previously used as a tool in cancer research) and quantified the impact of re-annotation on (a) the array by comparing the quality of new annotations with that of prior annotations, and (b) on systems biology modelling. Their findings led to the conclusion that re-annotation should be the standard first step when analysing functional genomics data as it not only provides more structural and functional-annotations but also improves the power of functional genomics modelling. Furthermore, the authors considered that re-annotation can result in a different knowledge outcome derived from previous published research findings. This approach was considered especially valuable for those species in which data and resources are rapidly expanding (including those for which genomic sequence information is only recently available).

183. Wu et al (2009) comments that although the breadth of available gene annotation resources benefits the TGX community, there is yet to be a single resource that completely describes everything a researcher might want to know about a gene's function. Many researchers ultimately visit different sites for each gene of interest to get as complete a picture as possible of gene function, which is considered highly inefficient and cumbersome for end users (since user interfaces vary dramatically to the extent that researchers must learn and remember how to navigate each site). A further drawback arises from the fact that new online resources are continually being developed and thus staying abreast of these tools and evaluating their utility is a time-consuming and recurring task. Consequently, Wu et al (2009) developed BioGPS, a centralised gene annotation portal for aggregating distributed gene annotation resources. By embracing the principle of community intelligence, and enabling any user to easily and directly contribute to the BioGPS platform, the developers hope the BioGPS portal will overcome the above bottlenecks in functional genomic analysis.

184. Other data interpretation issues discussed in the literature include consideration of factors complicating interpretation of DEGs in the liver, for example zonation of hepatic gene expression, nutritional status of animal subject and mixed cell population (Irwin et al, 2004). Morgan et al (2004) notes the value of having an understanding of mathematics to interpreting, comprehending and providing further analytical insight, particularly with regards to interpreting the temporal nature of gene changes and subsequent links with the mathematical discipline of dynamics.

185. An IPCS Workshop on TGX and Risk Assessment for the Protection of Human Health, organised by an international steering group[37] was convened in Germany in 2003 to consider the different methods of evaluating TGX studies in risk assessment and identify and discuss gaps in knowledge, issues and challenges (IPCS, 2003). The overriding conclusion was the need to strengthen dialogue across disciplines i.e. between those generating TGX information, traditional toxicologists and risk assessors to foster development and application of TGX. The workshop also considered it particularly important that molecular epidemiologists and bioinformatics experts are included. Various initiatives have since developed and are discussed in the following section.

C. DATABASE MANAGEMENT

186. This short section focuses on the storage and processing of high throughput TGX data to enable further data analysis and comparison by other researchers.

(i) Databases

187. Various database systems and repositories have been developed that provide molecular expression datasets from omic technologies to facilitate information-sharing and pattern recognition, which also aids the predictive power of TGX (Morgan et al 2004; Hayes & Bradfield, 2005). These databases allow comparison of array files e.g. transcriptional profiles and conventional toxicological approaches with toxicological pathway and gene regulatory network information relevant to environmental toxicology and human disease (Ju et al., 2007; Zhou et al 2009). Databases commonly used/developed in Europe, Asia (Japan) and USA can be categorised as either local or public (Ahmed, 2006b; Ju et al 2007). Local databases can be locally installed and hold species-, genus-, topic-specific data.

<u>1. General genomic databases</u>

188. Public databases are available for either public query or submission of data. There have been a number of initiatives to address specific database issues e.g. the Microarray Gene Expression Data (MGED) Society was developed to enable efficient cross database communication via the formulation of a conceptual framework. This led to development of several databases based on international data communication standards (e.g. MIAME guidelines) developed by MGED Society (Morgan et al 2004). Databases also containing non-TGX data include (Hayes & Bradfield, 2005):

(i) Gene Expression Omnibus (GEO) – the largest microarray repository worldwide, storing MIAME standard quality data and allows users to assess quality of data for themselves (Hayes & Bradfield 2005). Considered a powerful tool for data mining and

[37] The steering group included the US EPA, the US National Institute of Environmental Health Sciences (NIEHS), the Japanese National Institute of Health Sciences (NIHS) and the OECD.

hypothesis generation although contains few data relevant to toxicologists.

(ii) Array Express – The European Bioinformatics Institute (EBI) provides a suite of databases and applications which includes Array Express, considered the second largest international open source repository for microarray data (one of 3 repositories recommended by the MGED Society for storing data). Array Express provides public and password protected access to well-annotated raw and normalised submitted data. NB. MIAMExpress is a web-based MIAME supportive data-submission tool. Data can be submitted either directly from local databases or online and accession numbers are used to retrieve data queried on the basis of species, author or array platform.

(iii) BioGPS - a successor to Symatlas[38] that currently focuses on annotation for human, mouse, and rat genes (Wu et al 2009). Based on a simple, unstructured plugin interface that allows for simple community extensibility. Also implements a powerful user interface that enables precise customisability .

(iv) Centre for Information Biology gene Express database (CIBEX) – a Japanese developed repository for a wide range of high throughput gene expression data (e.g. microarray, SAGE) and MS proteomic data. CIBEX complies with MGED standards for microarray data and uses the open source database software MySQL. Although CIBEX is mainly used by East Asian scientists there are plans to collaborate with ArrayExpress and other Western systems.

2. TGX-specific databases

189. Morgan et al (2004) notes that a key database limitation is the fact that most are populated with rat toxicant-rich data, although other non-rat databases are in development. In 2004, the HESI Committee on the Application of Genomics considered the lack of publicly available toxicogenomic databases as one of several key hurdles (Petit, 2004). The development of public databases housing toxicologically relevant microarray data was considered necessary in order to accommodate the significant amount of TGX data, to help address the complexity of comparing different gene annotations and splice variants across platforms, to provide a resource for complex informatics analyses of traditional toxicology/ pathology and microarray data thus providing the scientific community with easy access to integrated data in a structured standard format (Corvi et al 2006). A list of publicly available TGX-databases is provided below:

(i) Environment, Drugs and Gene Expression (EDGE) database – a publicly accessible microarray database devoted to TGX research. Developed by the Bradfield Laboratory, McArdle Lab for Cancer Research, Wisconsin USA. EDGE is considered unique in that it avoids the problems that afflict other databases (wrt data comparison) by operating under standardised microarray platforms.

[38] Symatlas database was managed by Novartis and provided a catalogue of gene expression in varying tissues and cell cultures enabling researchers to find tissues/cell culture models expressing their gene of interest.

Researchers can add their data to the database and compare their results with the large volume of data under similar conditions. The database also offers clustering tools for data mining. However EDGE is limited by the fact it contains only murine, hepatic-rich TGX data although efforts are underway to include other organs (Hayes et al (2005).

(ii) ArrayTrack– an integrated software system for managing, mining and visualising microarray gene expression data was developed by the US National Center for Toxicological Research (NCTR) (Fang et al 2009). ArrayTrack stores a full range of information related to DNA microarrays and clinical and non-clinical studies as well as summarised data from proteomics and metabonomics experiments and has been used in routine reviews of genomic data submitted to the US FDA.

(iii) Chemical Effects in Biological System (CEBS) Knowledge database – developed by the Toxciogenomics Research Consortium (TRC) for genomic, proteomic and metabonomic studies on chemicals (displays microarray data in the context of study design and assay measures such as clinical chemistry and histopathology) of the National Center for Toxicogenomics (NCT) ;

(iv) Comparative Toxicology Database (CTD) – aims to develop a comparative database that links sequence information for genes that are relevant to toxicology ;

(v) dbZach – an emerging database that houses microarray data relating to endocrine distruption and testicular toxicity;

(vi) Toxicogenomics for Efficient Safety Test (TEST) database management system – an intelligent database system, capable of handling heterogeneous and complex data from many different experimental and information sources (Lee et al 2004).The intelligent query feature enables users to obtain relevant, useful information from complex data sets and conduct multiple comparisons. Information can be retrieved for compounds, animal experimental data, gene expression data and annotation. At the time of publication the system housed information for 16 compounds, 45 microarrays, 190 animal experiments, and had a customised 4.8K rat clone set. Data can be accessed online via http://istech.info/TEST/ and users requiring gene level data can enter their query into the annotation database with the gene's name and ID no of the relevant database, and functional key words. Expression profile information is obtained via links to a microarray database.

(vii) Profiles of Chemical Effects on Cells (pCEC) - a Japanese TGX (gene expression) database with a system of classifying chemicals that have effects on human health (Sone et al 2010). pCEC classifies chemicals according to specific tissues and cells they affect, the gene expression changes they induce and their toxicity and biological functions. The database also analyses the relationships between chemicals and the genes they affect in specific cells and tissues. The developers hope this database will

help support decision making within the context of environmental regulation.

190. Despite progress in many areas, an EPA published article considers that the current state of the majority of public microarray databases is inadequate for supporting predictive toxicology and meta-analysis (Williams-Devane et al 2009). This is particularly with regard to chemical indexing, considered to be a first important step toward integrating chemical, toxicological and genomics data into predictive toxicology.

(ii) Data comparison

191. A good TGX database would enable easy comparison of the expression profiles of one compound with another to help contextualise the significance of the data. Elashoff (2008) notes the importance of the having comparable study and database samples and outlines an approach to comparing expression profiles of a study compound with that from database compound(s). Firstly, an assessment of the similarity of the type of vehicle controls, sex/strain of animals, sample and chip processing methods used is conducted, as these are known to alter the baseline expression level of genes but not expression regulation induced by compound. Next, within-study data normalisation is conducted as experience shows that comparing unnormalised data between groups is limited, whereas within-study normalisation removes much of the cross study differences while preserving the underlying biological responses. The comparability of the study and database samples is then assessed using the following methods: QC metrics, PCA (to determine whether study samples group with database samples on the first several principal components), and clustering (to determine whether those study samples grouping with database samples cluster with an acceptable level of correlation). Data exploration is performed to search for patterns i.e. grouping of compounds which may arise due to similar mechanisms of toxicity, induction of high level toxic effects (e.g. necrosis), similar non-toxic effects followed by gene level analysis (includes gene similarity analysis and pathway fold change).

192. A recent attempt to enable researchers to compare in-house data with data contained in the Japanese TGX Project (TGP) database led to the development of the similar compounds searching system (SCSS) by Toyoshiba et al (2009). In practice, however, it is not easy to compare in-house microarray data with those in a database. This is largely due to differences in the experimental conditions i.e. too much inter-laboratory variation, and too many strategies developed to annotate targets (although it is thought that annotation problems should resolve in time as more genomes undergo careful sequencing and curation (Mattes, 2004; Hayes & Bradfield, 2005; Toyoshiba et al 2009). Furthermore, the ease and convenience of data accessibility and comparability is influenced by the type of system used for microarray data analysis. Web-based systems are highly regarded due to their ease of access from any computer on the internet, which thereby facilitates sharing of data (Ahmed, 2006b).

193. One approach to address inter-laboratory variability is to standardise the experimental protocols e.g. use common platforms, however, the range of platforms and protocols presents a significant challenge (Baker et al 2004; Mah et al 2004), although two methods for combining information across different versions of Affymetrix oligonucleotide arrays have been devised (Morris et al., 2006).

(iii) Standardisation

194. The standardisation of TGX protocols (and data) represents a significant contribution to enabling data comparability and subsequently ensuring reproducibility, which greatly impacts on the wider application and general acceptance of TGX as a risk assessment tool. Various standardisation initiatives exist, the most established arising from the MGED Society, an international organisation of scientists specialising in biology, computer science and data analysis that facilitate the sharing of large microarray datasets generated by high throughput functional genomics and proteomics experiments (http://mged.org).

195. The MGED Society have established several standards for data quality, data management, data annotation, data exchange/communication, the most notable being the Minimum Information About a Microarray Experiment (MIAME) Guidelines from which modified offshoots arose. MIAME is discussed further in the next section, however those relating to database management include the MIAME/Tox described as an array based TGX standard developed by various organisations (ILSI-HESI, NIEHS, NCR, FDA NCTR and EBI) (Sansone et al 2005). MIAME/Tox principally aims to guide the development of TGX databases and data management software although it also provides a set of guidelines describing minimal information required to correctly interpret and replicate the experiments or retrieve and analyse microarray data of toxicological significance. The MGED facilitates the creation of these and other tools including the MGED Ontology which aims to develop standard terms (ontologies) for annotating (describing) samples used in microarray experiments. The Society also participates in community development of ontologies supporting the Open Biological and Biomedical Ontologies (previously known as Global Open Biological Ontologies (Gobo)) and works with other standards organisations e.g. EBI and HESI to develop the Tox-MIAMExpress – an annotation and submission tool to ArrayExpress database.

196. Sansone et al (2004) identifies a potential challenge to standardisation initiatives, in which different sector needs may complicate the development of a unified approach. For example, the main objective for the regulatory sector is regulatory submission of data and therefore standardisation initiatives should ideally accelerate the review process, facilitate proprietary data submission, and optimise data visualisation. In contrast to this, the research community is mainly concerned with discovering genes and identifying mechanisms and the need for databases and tools. As such, standardisation would ideally ease the deposition of data into public databases and facilitate data mining via the use of common annotation standards and ontologies. The

only way forward, it seems, is to focus on areas of overlap and commonality and the Reporting Structure for Biological Investigation (RSBI) is proposed as a the solution (Sansone et al 2006).

197. The RSBI represents a working group under the MGED Society umbrella that brings together several communities including the TGX community. It aims to tackle the challenges associated with integrating data and representing complex biological investigations employing multiple omics technologies i.e. duplication and incompatibility (Sansone et al 2006). Such challenges are thought to arise because each community is developing databases and establishing their own data communication standards. In this era of functional genomics and systems biology such efforts cannot be developed in isolation, as failure to deliver will increase burden and cost of data management tasks. Therefore, the RSBI aims to provide a single point of focus for the various omic communities to synergise their insular approachs into one common solution. RSBI hopes to produce technology centred data communication standards that not only stand alone but also are able to function together.

198. Mattes (2008) accurately sums up the value of shared efforts to addressing the problems affecting the analysis of TGX data in the following abstract:

> *"Public consortia provide a forum for addressing questions requiring more resources than one organization alone could bring to bear and engaging many sectors of the scientific community. They are particular well suited for tackling some of the questions encountered in the field of toxicogenomics, where the number of studies and microarray analyses would be prohibitively expensive for a single organization to carry out. Five consortia that stand out in the field of toxicogenomics are the Institutional Life Sciences Institute (ILSI) Health and Environmental Sciences Institute (HESI) Committee on the Application of Genomics to Mechanism Based Risk Assessment, the Toxicogenomics Research Consortium, the MicroArray Quality Control (MAQC) Consortium, the InnoMed PredTox effort, and the Predictive Safety Testing Consortium. Collectively, these consortia efforts have addressed issues such as reproducibility of microarray results, standard practice for assays and analysis, relevance of microarray results to conventional end points, and robustness of statistical models on diverse data sets. Their results demonstrate the impact that the pooling of resources, experience, expertise, and insight found in consortia can have."*

SECTION 4. QUALITY CONTROL OF TRANSCRIPTOMIC BASED STUDIES

INTRODUCTION

199. Quality control (QC) should comprise an integral part of any TGX study. The deployment of QC measures is vital to guaranteeing data quality and consistency and ensuring data accuracy and reproducibility (Hartmann, 2005). QC is principally concerned with data quality (i.e. approaches that help ensure the generation of optimal and accurate data at each technical procedural endpoint) and can, therefore, be deployed at each data generation and acquisition stage of a TGX study. A closely related term, validation, describes the process of ensuring that a test reliably measures and reports the determined endpoints. It can be applied at several levels e.g. the technological/ platform level, software/data analysis level, biological and generalisability level and regulatory level (NAS, 2007ab). This section highlights the various QC measures and related approaches used to ensure the production and interpretation of accurate, reproducible and consistent data.

A. QUALITY CONTROL MEASURES

200. QC measures can be defined according to the stage at which they are applied e.g. pre-hyrbidisation QC measures assessing the quality/efficiency of RNA extraction and labelled probe sample preparation, and post-hybridisation QC measures which either assess quality/efficiency of (a) individual microarray spots or genes via filtering approaches (b) individual hybridisations or chips (c) whole batches of hybridisations (Hartmann, 2005).

(i) Pre-hybridisation quality control measures:

1. RNA Quality

201. The quality of RNA derived from tissues and cell samples strongly influences the type of data produced in a TGX study. High quality RNA is necessary for reproducible and reliable data while low level RNA quality reduces the statistical power of a study (Thompson & Hackett 2008).

202. Three different approaches are typically used to measure RNA Quality. The RNA Quality Index (RQI) considers both RNA purity and integrity. RNA can be contaminated by protein, genomic DNA and chemicals and an optical density ratio[39] of 2 (260 cf 280 nm) is used to indicate RNA of sufficient purity. Because RNA is extremely sensitive to degradation by RNases, it is particularly important to confirm whether RNA degradation has occurred in samples or tissues for gene expression analysis, such as after long term storage. Although a moderate degree of RNA degradation does not preclude meaningful results for microarray analysis or RT-PCR, more extensive

[39] Quantified using spectrophotometers

degradation necessitates the exclusion of affected samples from further study (Sumida et al 2007).

203. RNA integrity can be assessed via use of microfluidics platforms for nucleic acid analysis. The procedure involves RNA separation (via electrophoresis), followed by quantification (via fluorescence) and calculation of 28S/18S RNA ratio where intact RNA would have a value greater than 2 although the value of this approach has been questioned i.e. the ratio is not predictive of sample quality and it not considered to be a useful indictor of sample integrity when total RNA is only partially purified (Thompson & Hackett 2008). RNA degradation is typically indicated by incomplete full length cDNA. For certain protocols intact (undegraded) RNA is required, indicated by a greater than 3 value for the 3' to 5' probe ratio of universally expressed genes (e.g. GADPH).

204. Other approaches used to measure RNA quality include electropherograms (graphic outputs of electrophoresis devices) which are considered to provide a more accurate assessment of RNA quality, and RNA yield, described as the expected yield calculated from a given weight of tissue, which essentially measures the effectiveness of the RNA isolation protocol.

205. Copois et al (2007) compared the ability of four different RNA quality assessment methods to detect reliable RNA samples and found that the 28S/18S ratio leads to a misleading categorisation, while two computer methods (RIN and degradometer) and an in-house RNA quality scale had similar capacities to detect reliable RNA samples. Furthermore, the authors developed a new approach based on clustering analysis of full chip expression that controls RNA quality after hybridisation experiments and found that monitoring RNA quality after hybridisation experiments in addition to before ensures reliable and reproducible microarray data.

2) Target preparation

206. Efficiency assessment of cDNA synthesis can be done by monitoring the yield and size of cRNA product (Thompson & Hackett 2008). Efficiency of cRNA fragmentation is similarly evaluated by monitoring the shift in size of product.

207. External controls aim to monitor the performance of technical procedures and produce data that can be used to assess the overall quality of the starting RNA (sampling), labelling, hybridisation and grid alignment (microarray) and are also referred to as calibration standards. They are often commercially available non-mammalian RNA sequences that are not from the species being analysed (i.e. have no similarity to the genome of the species under study) and measure system performance independent of the quality of the RNA sample being analysed. These controls are either spiked into samples (hence also referred to as spiked-in targets which hybridise onto their corresponding sequences on arrays) or microarrays (aka exogenous spike-in controls that correspond to probes on the microarray surface). For example, in the labelling step, replicate arrays are run using reverse dye-incorporation

orientations for samples. Success of labelling can then be determined by spiking polyA RNAs into RNA samples (before the reverse transcription step).

(ii) Post-hybridisation QC measures

1. QC of individual microarray spots/genes

208. Hartmann (2005) notes two possible options for quality control at the individual gene or spot level: quality measures based on general spot properties (e.g. spot size, shape, pixel distribution, intensity) or on spot replicates (within the chip or between technical replicates). Spot quality can vary especially for spotted cDNA microarrays. Image analysis software often offers spot quality assessment which requires intensive learning steps, in particular a set of spots with known quality status which can make it lab-dependent and labour intensive. However, not all image analysis software offers meaningful quality measures and most spot properties are usually difficult to evaluate for genes expressed at low intensities. Consequently, use of quality control measures based on spot features should be carefully considered. Quality control measures for probes on Affymetrix GeneChips are not considered as critical, as there is little intra probe cell variation. Few probe set variability estimates have been used to eliminate individual, poorly measured probe sets since such measures are sequence-dependent and hence not comparable between probe sets.

209. Spot intensities can also be used as a quality control measure for microarray platforms (Hartmann, 2005). Low intensity range spots result in poorer signal-to-noise ratios as it becomes difficult to distinguish the signal from background. Hence high intensity range spots are thought to be more reliable. However, investigators are advised against automatically eliminating low intensity spots by QC as this can result in loss of valuable information regarding the identity of non-regulated genes in particular samples, for example. Such genes can also be of value as controls for subsequent data analysis

210. Quality control measures based on spot replicates depend on microarray platform used (Hartmann, 2005). For Affymetrix GeneChips, each gene is represented by a probeset (11 – 20 perfect matches and an equal number of mismatches). Robust methods are available that summarise the probe set into one signal intensity thereby precluding additional filtering. For cDNA spotted microarrays spot replicates are either non-existent or very limited (up to 2-3 per gene) on one chip. If a reasonable number of replicates are available, robust methods can be applied to deal with outliers.

2. QC of individual hybridisations/chips

211. These measures can be grouped into two categories, depending on whether they are based on a selection of quality control spots (e.g. housekeeping genes, hybridisation controls, empty spots) or on a measure derived from all spots on a chip (Hartmann, 2005).

a) Measures based on QC of a selection of spots (within chips)

212. Corvi et al (2006) considers use of housekeeping genes and external RNA controls as the routine element of the approaches used to validate TGX platforms. They define 'best practice' for the experimenter (or array manufacturer see below) to employ on a regular ongoing basis. It is thought that the use of these 'biological standards' is required to address inherent technological and biological noise in these systems. When these control spots with their known concentrations in two channels are used (for calibrating an experiment) they improve normalisation and provide valuable information about experimental variation (Wang et al., 2007)

~ External controls

213. The success of the hybridisation step can be determined by adding an external control after the cDNA synthesis step. These in-vitro polyadenylated RNA transcripts each composed of random unique and non-mammalian sequences are spiked into RNA samples of interest (in either one or two channel RNA formats). They are considered to facilitate comparisons among laboratories and platforms and provide a way to assess the quality of experiments over time. However, there is concern that control spots usually lack the sensitivity for thorough post-hybridisation quality control, probably due to the limited number of available spots (a few hundred at most) (Hartmann, 2005). The value of external controls is being assessed by a several groups including the External RNA Control Consortium (ERCC) who are developing better control RNAs.

214. The ERCC[40] have been developing external RNA controls to assess technical performance in gene expression assays. Their overall aim is to produce a standard set of 96 well-characterised, tested external RNA controls with demonstrated acceptable performance on major microarray platforms and with commonly used QRT-PCR methods (Baker et al, 2005; Thompson & Hackett, 2008). It is hoped these controls can be added into a test sample and tested in neutral background with probes for these sequences included in new commercial arrays. The scope and goals of the ERCC are discussed in a commentary by Warrington et al (2005) and readers should refer to Baker et al (2005) for a further description of the proposed experiments and informatics processes that will be followed to test and qualify individual controls.

~ Housekeeping genes

215. The use of housekeeping genes (or common reference RNA standards) as internal controls for real time RT-PCR and microarrays[41], Northern analysis and RNAse protection assays is discussed elsewhere in this paper. A common misconception/assumption is that their expression is constant regardless of experimental conditions. However, the fact that their expression

[40] The ERCC is an adhoc group of approximately 70 members from private, public and academic organisations

[41] In microarrays RNA standards are competitively hybridised with the sample of interest in two channel array formats.

can vary implies the possibility of an erroneous interpretation of the expression profile of a target gene. Arukwe (2006) suggestion of validating potential reference genes prior to their use provides a credible approach to addressing the possibility of variability in expression. Arukwe (2006) evaluated the suitability of the most commonly used housekeeping genes in toxicology to provide researchers with a summary of the key information needed to re-evaluate housekeeping genes used. The expression pattern of beta-actin, beta-tubulin, 18S ribosomal RNA (18S rNA) and elongation factor-lalpha (EF-lalpha) were found to be modulated on the basis of random exposure condition and time, in both in-vivo and in-vitro test systems of Atlantic salmon (Salmo salar). Although use of aquatic models potentially limits the value of the study in human risk assessment, the authors concluded that the choice of internal control gene should be determined empirically on the basis of the individual exposure condition and by the individual researcher.

216. Commonly used housekeeping genes may vary in stability depending on the cell type or disease being studied. Therefore, Lee et al (2007) sought to identify additional housekeeping genes that show sample-independent stability. Statistical methods were used to search a large human microarray database for genes that were stably expressed in various tissues, disease states and cell lines. Those selected were expressed at different levels because the authors considered that reference and target genes should be present in similar copy numbers to achieve reliable quantitative results. Lee et al (2007) identified three new reference genes CGI-119, CTBP1 and GOLGAI alongside three well-known housekeeping genes that were more stably expressed in individual samples with similar ranges and concluded that statistical analysis of microarray data can be used to identify new candidate housekeeping genes showing consistent expression across tissues and diseases. The authors proposed that CGI-119, CTBP1 and GOLGAI represented novel candidate housekeeping genes that could prove useful for normalisation across a variety of RNA-based techniques.

b) Whole (between) chip measures

217. The inspection of the whole chip image is recommended as a post-hybridisation quality control as it can instantly reveal a lot of information about background, foreground or spatial effects (Hartmann, 2005). Other useful quality control measures include total background, the ratio of total signal over total background to measure hybridisation efficiency, and signal or ratio distribution for one and two channel data respectively.

218. A key conclusion reached at the 2003 IPCS Workshop on TGX and Risk Assessment for the Protection of Human Health was the need to develop data quality standards in order to ensure confidence in data generated from different sources and platforms. Percent present calls (PPC) is used to assess data quality (discussed elsewhere), while the multi-array approach is used to identify poor quality arrays that should be removed from further consideration (Hartmann, 2005). The multi-array approach uses median normalised unscaled standard errors (NUSE) that provides a measure of relative chip quality i.e. for Affymetrix Genechips they measure the

heterogeneity within a probeset. Plotting the NUSE yields a summary of the quality of chips in a single figure whereby arrays with large NUSEs may be suspect and examined further by single array exploratory analysis.

219. Another useful way to assess quality is to compare intensities and expression ratios across chips. Pairwise scatter plots or unsupervised clustering methods can help identify failing hybridisations because they produce intensities/ratios that deviate significantly from those of any other hybridisation and therefore show up as outliers. However, these methods do not necessarily distinguish between a quality outlier and a biological outlier and, therefore, the elimination of individual chips should not be based on such an assessment alone.

220. Fluorescence standards are used to assess the limits of performance or range of scanning software (via fluorescence calibration slides a typical output range 0 – 65.5K relative fluorescence units per pixel), and distinguish hybridisation failure from scanner defect (via the use of software programmes). With regards to older models the photomultiplier tube is considered to be a source of variability and it is recommended that scanners undergo regular software-run inspections to identify artefacts (NAS, 2007b).

3. QC of whole hybridisation batches

a) Statistical Process control

221. This robust PCA based approach proposed by Model et al (2002) describes a statistical example of a quality control that monitors the hybridisation process which typically arises in high throughput labs conducting experiments with several hundreds of chips (Hartmann, 2005). These hybridisation processes can take several weeks or months and are therefore particularly susceptible to systematic changes such as changes in scanner calibration, room temperature, ozone concentration or buffer solutions which can impact on measured spot intensity and chip quality. Systematic changes in the process are reflected by increases in the distance to chips from the initial stable process (historical data set). Plotting the distances wrt the process parameter under investigation enables easier visual detection thereby serving as early warnings for changes in the production process.

222. In summary, although some of the methods described above serve as absolute measures of quality, most serve as controls that pinpoint spots/chips most likely to bias data, and deciding the correct cut-off for quality control measures is not always easy (Hartmann, 2005). Clearly, the value of quality control measures lies in its ability to reduce the random error term, however the application of too stringent criteria can reduce the power due to reduced sample size. Whether a quality control measure is applied at a particular level depends on the number of hybridisations performed and the availability of measurement replicates. For example, QC for individual hybridisations becomes necessary for experiments with small sample sizes (n<10-20) but for large experiments (with tens to hundreds of chips) process control becomes the priority. It is recommended that the filtering of chips should always be

carefully done with strong and explicit indications that the observed peculiarity is due to quality and not biology (Hartmann, 2005).

B. VALIDATION OF TGX PLATFORMS

223. In 2004, the HESI Committee on the Application of Genomics considered the lack of validation of available technologies one of several key hurdles (Pettit, 2004). Validation is necessary to identify and reduce technological artefacts and procedures used to control for microarray quality and instrumentation are the responsibility of array manufacturer/provider and have been defined as one-off validation (Corvi et al 2006). Routine validation allows for data comparability and encompasses QC aspects of critical experimental components. These include the random sequence verification of gene targets (to ensure no errors are introduced between batches), the use of biological standards (as discussed above), and quality assurance and good laboratory practice (GLP) (intended to promote proper documentation, quality and authenticity of data as is required for data acceptance by regulatory authorities). Corvi et al (2006) notes that most large scale TGX efforts were not (at the time) arising from GLP-complaint studies and suggests identifying procedural aspects of GLP compliance not currently captured in MIAME/Tox and incorporating them over time as a way forward (establishing best practices for TGX until formal procedures are adopted).

224. Qin et al (2004) highlight the lack of realistic empirical validations of TGX data analysis methods (i.e. the fact it is impossible to know whether a given methodology is better at revealing the right answer). It is thought this is due to data analysis methods being introduced either by examining their performance in real microarray experiments (where the truth is unknown) or in simulated data (that rely on distributional assumptions /idealised models for the error structure). Therefore, a study was conducted to evaluate the relative effectiveness of two data transformations (i.e. intensity-based normalisation[42] and local background subtraction) and to assess the performance of six different ranking statistics for detecting DEGs, method (mean, median, t-statistic, S-statistic (related to SAM software), B-statistic, and BL-statistic) and different image analysis programmes (GenePix ®, SPOT, ArraySuite, QuantArray). Qin et al (2004) tested the analytical methods using ten spike in dye swap experiments (in which the truth was known – analogous to the Latin Square dataset) conducted by six different laboratories within the TRC. The authors found that the most favourable conditions for identifying DEGs were to apply intensity normalisation without background adjustment (which they suggest may possibly be detrimental for effective detection of DEGs), and use robust alternatives to the t-statistic such as S-, B- or BL- statistic or the median. However, they note that the outcome from using robust statistics may be influenced by the fact only four technical replicates were used. The authors also concluded that the choice of image analysis software substantially influences experimental conclusions with SPOT offering some improvement over GenePix® in detecting DEGs.

[42] i.e. Intensity dependent selection whereby the threshold for selecting DEGs varies with spot intensities.

SECTION 5. SOURCES OF VARIATION IN TRANSCRIPTOMIC-BASED ANALYSES

INTRODUCTION

225. The multi-step nature of a TGX study protocol makes variation an unavoidable and expected phenomenon (Chen et al 2004). As a source of bias, sources of variation must be identified and characterised and its magnitude estimated to ensure cost-efficient microarray experiments are designed. Better characterisation of sources of variation would also enhance the use of gene expression profiling in clinical and laboratory settings. This section discusses the different types of variation and approaches used to identify, characterise and control them.

A. BIOLOGICAL VARIATION

226. Biological variation relates to individual factors that produce variation arising from the use of different animals, cell lines and tissues. Biological variation is intrinsic to all organisms and is influenced by genetic and/or environmental factors and by whether the samples were pooled or processed individually (Chen et al 2004). Examples of biological variation previously discussed include circadian rhythm regulation. Novak et al (2002) refers to biological variation as physiological variation which comprises one of three types of background variation (i.e. variation not directly related to the pathology or stimulus). Pooling is often used to minimise the effects of biological variability as previously discussed in section 1. However, unlike other types of variation, biological variation may be of interest in its own right.

B. TECHNICAL VARIATION

227. Technical or experimental (non-biological) variability arises from use of the microarray system and is considered the most significant challenge of microarray data analysis. Microarray experiments are subject to additional sources of technical variability (in addition to those inherent in toxicology experiments) which can be categorised at each step of a microarray experiment (Ju et al, 2007). These fluctuations arise independently of the RNA source and are beyond the experimenter's control (Novak et al., 2002). The various sources of technical variation are listed in Table 3. NB. In addition to technical and physiological (biological) variation, Novak et al (2002) considers sampling variation as a third type of variation (background). Sample variations are defined as differences in sample characteristics arising from sampling adjacent or contaminating tissues, tissues with heterogeneous cell population, and gene expression differences resulting from animals having minor infections, or suffering environmental stress activity or feeding behaviour differences.

Table 3. A list of sources of technical variation encountered in TGX microarray studies

Experimental stage	Variable	Notes
Sample preparation	Amounts of mRNA used	Relates to tissue/mRNA extractions
	RNA purity	
	RNA amplification method used	Including RNA priming efficiency contributing to nonlinear amplification of expressed genes during probe synthesis
	cDNA preparation	
	Labelling method Dye label incorporation	
Microarray construction	Gene target type	
	Amount target applied to slides	
	Spot shape	
	Pin geometry	
	Gene target printing/deposition methods	Impacts on fixation of spotted DNA onto slides
	Matrix quality	
	Gene annotation across chips	
	technical platform sensitivity	
Hybridisation/washing	Manual vs automated protocol	
	Reaction wash/buffer component	NB. These can be controlled for via use of stock solutions/ master mixes
	Amount applied to slides	
	Temperature	
Detection	Reading	
	Scanning parameter differences	Influences signal-noise ratio, data resolution and reproducibility
	Scanner power	
	Scanner artefacts	Either visual or automated
Data analysis	Cross hybridisation (within gene families)	
	Outshining from neighbouring spots	
	Range of methods used	
	Different technical settings on analytical equipment	
	Between array variation	Stochastic variation across replicate slides
Statistical analysis	Range of methods used	
Other	Laboratory environmental conditions	E.g. room temperature, ozone (Aka time/block effects).

Source: HESI-MINS of invitational meeting (2003); Lee et al (2005); Ju et al (2007); Yauk & Berndt (2007); Thompson & Hackett (2008)

C. IDENTIFICATION/ESTIMATION

228. Analysis of Variance (ANOVA) is typically used to characterise variation arising in TGX datasets. This statistical method models sources of variation by firstly considering all sources of variation (or variance components) that can arise in an experiment and summarising them into an equation (Chen et al, 2004). These can then be corrected for nuisance effects automatically. Use of ANOVA enables estimation of the magnitude of each variance component. Several studies have evaluated the impact of variation on subsequent gene expression and most report the need to better characterise variation arising from different sources.

(i) Studies investigating sources of variability

229. One of the early studies to address the impact of variation on gene expression described an approach to estimating sources of variation and their relative contributions to the overall variation (Chen et al., 2004). The authors used ANOVA to identify and estimate variability in two data sets: 1) a TGX study that generated cisplatin-induced gene expression changes in rat kidney; and 2) a circadian study that evaluated circadian associated gene expression changes in the rat liver. A mixed-effects (ANOVA) variance component model was used to estimate technical variances in the TGX study and technical and residual[43] variances in circadian study (with replicates used to investigate biological variance). The authors found that the greatest source of variation in the TGX study arose between arrays (due to batch to batch variation in array quality and manufacture and array to array hybridisation variance) while week-to-week variance accounted for the greatest variance in the circadian study. The authors concluded that overall data variability was due to the performance of weekly procedures and more precise estimates of gene expression changes are generated with reduced week-to-week variance.

230. Novak et al (2002) sought to assess the relative importance of various sources of variation (described as background variation i.e. technical, physiological (biological) and sampling variation) via use of a novel method that estimates sample dispersion. To test for technical variability the authors compared expression profiles from replicates tests using the same RNA sample. Physiological (biological) variability was tested by comparing expression profiles generated on HuGeneFL Affymetrix GeneChips with RNA samples from replicate cultures of the same cell line i.e. SK-BR-3 breast carcinoma (or IMR90 diploid fibroblasts). To test for sampling variability the authors compared expression profiles generated on Mu11kSubA/B Affymetrix Gene chips (containing perfect match and mismatch oligonucleotides) with RNA samples obtained from tissue samples of different mice. A linear characteristic function was used to provide a measure of dispersion i.e. data points which deviate from the mean. This novel method incorporates SD of differences in gene expression, mean signal intensity and sample mean gene expression. It reportedly redresses the fact that most expression studies obtain inadequate measures of SDs for each gene detected (due to lack of appropriate number of replicates). The authors observed similar dispersion patterns between RNA samples used to test for technical and physiological variability, suggesting that under carefully controlled conditions the size of the basal physiological variability is similar to that solely attributable to technical aspects of microarray studies. However, the authors reported higher levels of dispersion in genes playing a role in generalised stress response when testing for sample variability reflecting possible undetected infection, undernourishment or physical trauma before tissue sampling. This led the authors to conclude that when samples from different subjects were used variation induced by the stimulus may be masked by non-stimuli-related differences in the subjects biological state (since the fact that seemingly

[43] Residual variance is defined as a third type of variance relating to experimental unaccountable factors (Chen et al., 2004).

identical tissues from distinct animals may have difference gene expression profiles stresses the need for replica experiments in any comparative study). The authors further suggested that sample pooling when used as a means to reduce biological variation (as opposed to a means of obtaining sufficient RNA material) is of limited value.

231. The lack of data on baseline fluctuations in gene expression presents a particular challenge and major efforts are underway to determine the level of background variation (biological) in control animals. The challenge presented by the lack of control animal microarray expression datasets is further compounded by the fact that these datasets are not in a form best served for data mining. Boedigheimer et al (2008) reported on the HESI Committee on the Application of Genomics to Mechanism-based Risk Assessment attempts to assemble datasets for control rat liver and kidney generated from more than 500 Affymetrix microarrays. The HESI Committee assessed biological and technical factors and identified gender, organ section, strain and fasting state as particular key sources of variability. It was concluded that these and other factors should be included in MIAME study guidelines and that better characterisation of sources of variation in control animals would enhance TGX study design and data interpretation.

232. Members should note that the HESI Committee's Baseline Animal Database currently holds microarray data from 536 Affymetrix arrays from rat liver and kidney samples of control groups used in TGX studies produced by 16 different institutions (of which 48 where in-life studies). Their findings, published in BMC Genomics (2008) and in the book chapter Sources of Variance in Rat Liver and Kidney Baseline Gene Expression in a Large Multi-Site Dataset. In: Batch Effects and Experimental Shift in Microarray Analysis: Sources and Solutions (2009), suggested that bias correction has minimal effect on results of analyses of major sources of variance and noted that the identification of genes associated with certain study factors was affected if significant smooth bias was present.

233. The assessment of biological variability is also considered a worthwhile approach to validating TGX methods as described by Corvi et al (2006). Measuring the range of biological variability of gene responses for a given test system under baseline and toxicant-challenged conditions enables regulators to better discriminate biologically relevant responses from baseline homeostatic functions. This is considered an important TGX issue as studies conducted on cell culture populations reportedly demonstrate a wide range of biological variability in gene expression measurements for individual cells under both baseline and challenged conditions. To enable assessment of cross species differences that often hamper risk assessments only one species, tissue, and endpoint should be used at a time.

D. IMPLICATIONS FOR REPRODUCIBILITY

234. As part of a series of ongoing research projects the Toxicogenomics Research Consortium (TRC), NIEHS are performing standardisation experiments to identify and address sources of technical variation in gene expression experiments across multiple technology platforms and research centres (TRC website). Through this Co-operative Research Program (CRP) will evaluate variation arising in different aspects of a microarray experiment, in particular RNA labelling and hybridisation, data analysis (bioinformatics), RNA extraction and animal husbandry. The proposed overall outcome is to develop research standards for scientist within the TRC and scientific community as a whole in the hope this will lead to high quality data that are reproducible and comparable and also lays the foundations for their Star Projects (collaborative toxicology research using gene expression profiling). Details of the first standardisation experiment were published by Bammler et al (2005). The study sought to identify sources of error and data variability between 7 laboratories and across 12 DNA microarray platforms (that were either spotted or commercial), and also explore methods to accommodate the above variabilities identified. An ANOVA random effects model was used to assess to relative contribution of different sources of technical variability in gene expression measurements. The authors found that more than half the variability observed in the data was attributable to the microarray platform, with commercial microarrays yielding results that were more comparable between laboratories (differences between different laboratories contributed less). The study also observed increased interlaboratory reproducibility after implementing standardised protocols for RNA labelling, hybridisiation, microarray processing, data acquisition and data normalisation. The authors concluded that comparability is highest when technical variables are standardised and microarray results can be compared across multiple labs when a common platform and set of procedures are used.

235. Members are reminded that the COT secretariat plans to discuss issues relating to the reproducibility of TGX data as a separate discussion paper at the next COT meeting in September 2010. These will include consideration of factors affecting reproducibility (i.e. specific aspects of TGX design and analysis that either enhance or reduce reproducibility); comparative studies (e.g. cross platform correlation studies); the MAQC Project (evaluation of inter and intra-platform reproducibility); and the findings from inter-laboratory studies.

ANNEX I

Literature Search Strategy

1. An initial attempt to conduct a systematic review revealed the extensive amount of work published since 2004. For example, a search using basic general terms such as toxicogenomics, genomics and toxicity/toxicology, transcriptomics and toxicity/toxicology, for studies published between 2004 and 2008 yielded over 2.5K references (minus duplicates).

2. To reduce the amount of papers to a more manageable level, it was decided that [as an initial step] literature searches should be based on review papers that would provide the basis to identify key individual studies, which themselves would be summarised to update the review (NB. A similar approach was used to update the COM on the use of TGX in toxicology).

3. Details of the literature search strategy used are as follows:-
   - Date of search: 25/11/08
   - Database: Pubmed (via Endnote)
   - Limits: Reviews; published between 2004-2008; English
   - Search terms : Basic and specific (see below)

4. Basic search terms:
   - Toxicogenomics (TGX)
   - Genomics (GX) AND toxicity/toxicology
   - Transcriptomics (TRSX) AND toxicity/toxicology
   - Proteomics (PTX) AND toxicity/toxicology
   - Metabolomics (MTBLX) AND toxicity/toxicology AND risk assessment
   - Metabonomics (MTBNX) AND toxicity/toxicology AND risk assessment

5. Specific search terms (based on abstracting key words from COT conclusions as documented in the 2004 Joint Statement on TGX). Those relating to categories discussed in this paper (a, d and e) are shown below:
   - Design AND TGX/GX/TRSX/PTX/MTBLX/MTBNX
   - Reproducib* AND TGX/GX/TRSX/PTX/MTBLX/MTBNX
   - Statistic* AND TGX/GX/TRSX/PTX/MTBLX/MTBNX
   - Gene expression AND TGX/GX/TRSX/PTX/MTBLX/MTBNX
   - Microarray AND TGX/GX/TRSX/PTX/MTBLX/MTBNX

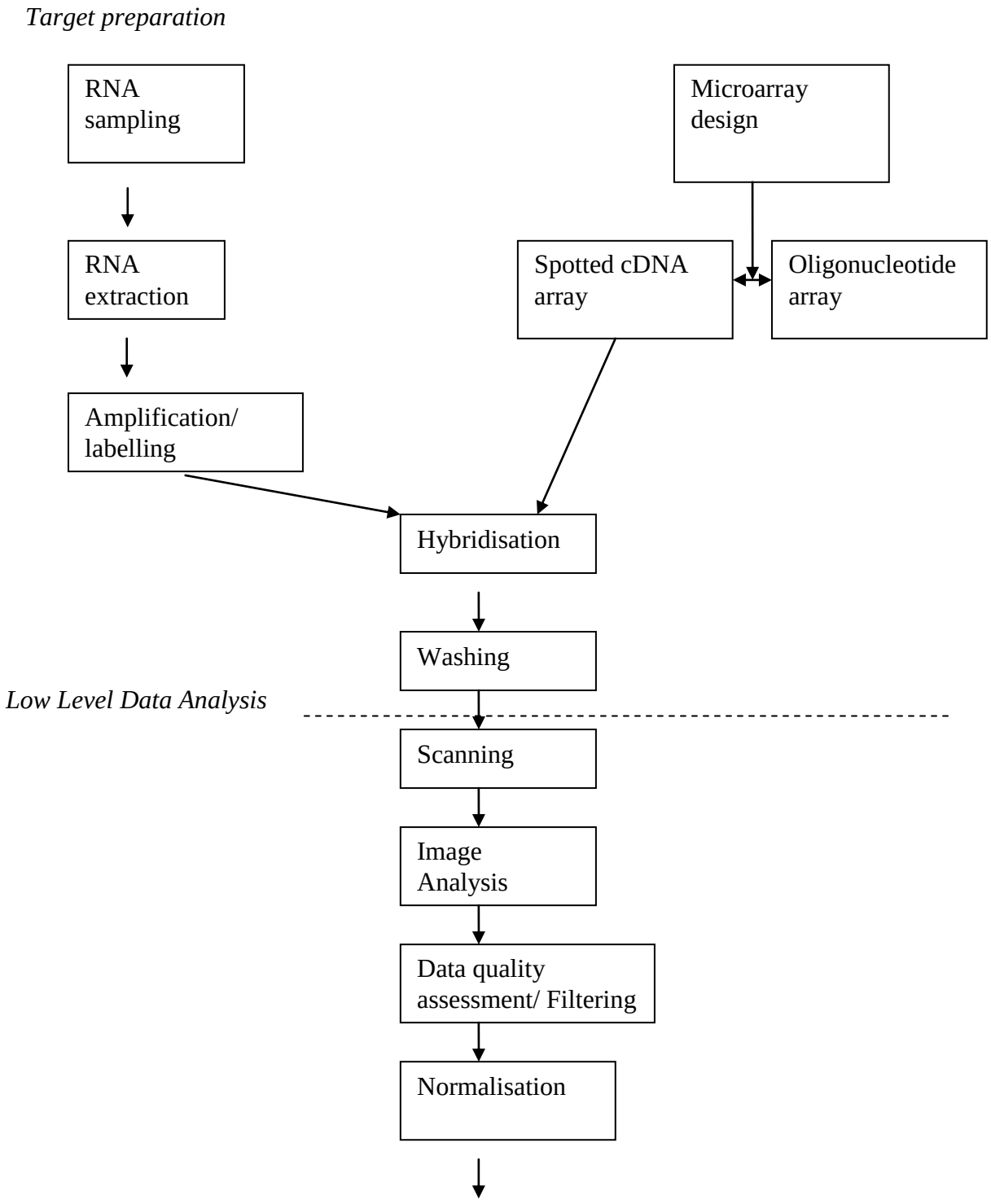
6. A total of 847 references were obtained with 144 references relevant to issues discussed in this paper. Review papers were selected on the basis that they address/update issues relevant to design, analysis and statistics Thirteen review articles were identified for further summary (see Annex III). Individual study papers cited in these reviews were identified and further expanded on the basis that they provide additional relevant information that could be used to build a balanced discussion paper for the Committee.

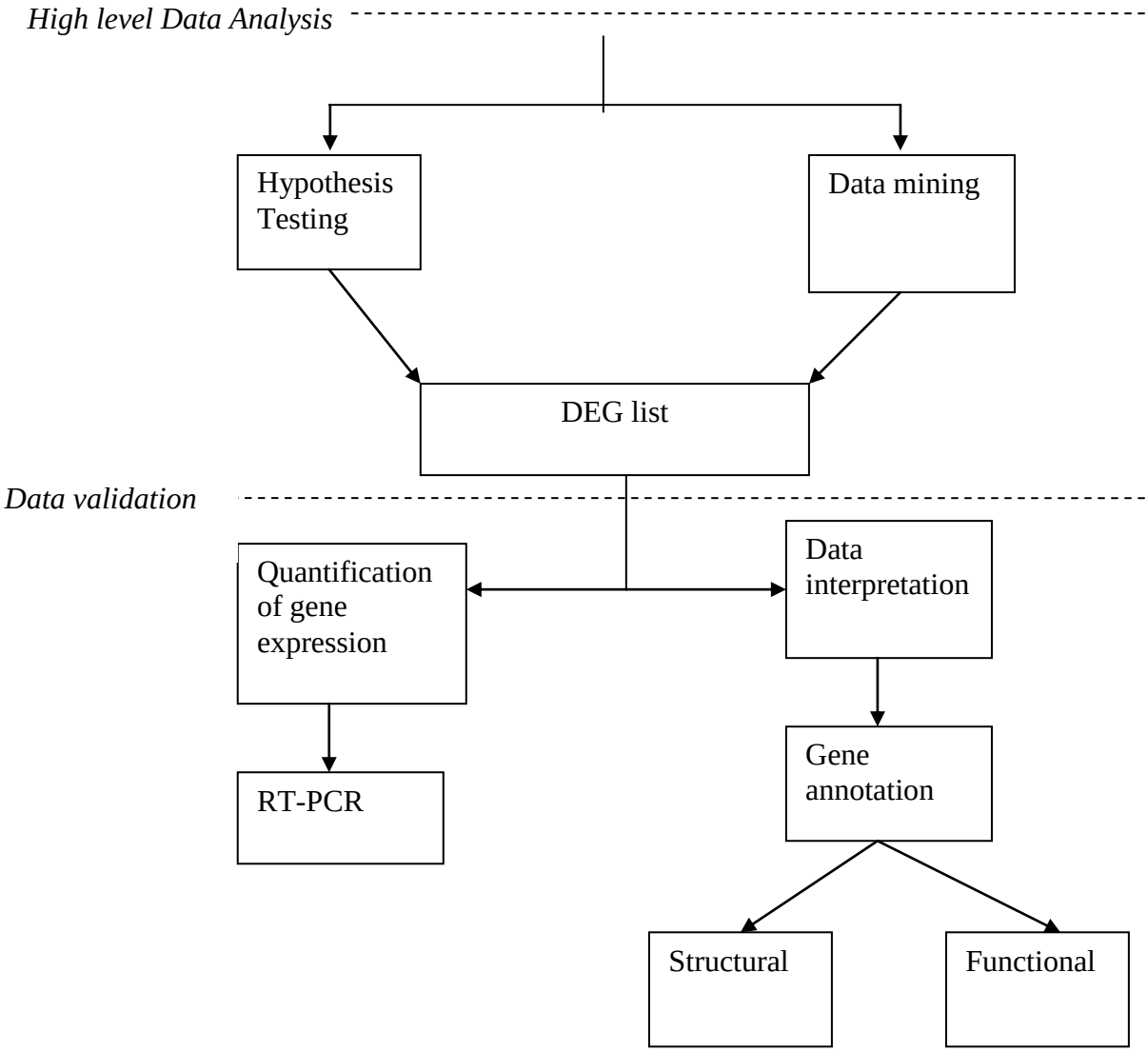
7. Updated literature searches were also conducted using the above basic and specific search terms to identify relevant individual studies.

8. Note. Issues relating to risk assessment i.e. target organ toxicity, regulatory submission, systems biology, etc will be considered in a separate discussion paper – the overriding consideration being whether the application of TGX demonstrates added value to risk assessment and also whether it provides any mechanistic insights.

ANNEX II

A schematic of a typical two-colour channel microarray experiment.




High level Data Analysis
Hypothesis Testing
Data mining
DEG list
Data validation
Quantification of gene expression
Data interpretation
RT-PCR
Gene annotation
Structural
Functional

ANNEX III

Narrative Summaries of Selected Review Papers

Morgan et al 2004. Complementary roles for toxicologic pathology and mathematics in toxicogenomics with special reference to data interpretation and oscillatory dynamics. Toxicological Pathology, 32(Suppl 1):13-25
Topic covered: This review considers the role of mathematics in toxicogenomics (TGX) and bioinformatical approaches. It takes a pathologists view of interpreting TGX data and highlights several study design issues for consideration. The role of statistics in TGX data analysis is briefly covered.
Design: The authors briefly mention the process used to measure the state of the transcriptome. The liver (as an example) is homogenised and RNA extracted and hybridised onto a large scale gene expression array platform. Signal intensities derived from the array are then processed generating a table of normalised signal intensity data. This is achieved via use of statistical software such as NLR (Normalisation by Local Regression), **Kepler et al 2002**) and requires considerable mathematical manipulation. Treatment related changes can then be determined. Whole organ experiments are limited by the fact that any gene expression changes merely represent an average of all cell types and locations within the organ. However, progress is being made to generate transcriptome data from single cells (**Tietjen et al 2003**). The authors note that in-vitro based studies have improved our mechanistic understanding of how the transcriptome responds to toxicants (**Morgan et al 2002**). They report a study by **Boess et al (2003)** which demonstrated marked differences in gene expression patterns in hepatocyte culture, liver slices and intact liver (temporal effects associated with the length of time in culture were also noted). The authors consider that this study informs on the design of gene expression based toxicology screens. Spatial and temporal issues of the liver are also discussed and should always be considered when designing/interpreting TGX studies. The authors recommend that liver studies document the lobe sampled as gene expression differences in hepatic lobes have been reported (**Irwin et al 2003**). Furthermore, zonal differences should also be considered in study designs, as decreased expression of a particular transcript/gene could be due to the absence of cells within that particular zone expressing the gene (e.g. the differential expression of glutamine synthetase and its restriction to a particular zone of hepatocytes). Temporal issues refer to the dynamicity of the liver in which the transcriptome activity of a living liver (as opposed to a static one) changes due to its dynamic nature and function. The authors also note how interactions between the liver and other organs can affect gene expression e.g. via action of chemical mediators, metabolites and hormones. The review discusses how biological oscillations such as circadian rhythms can potentially affect the design of TGX experiments. The authors report a study by **Kita et al (2002**) which examined the influence of circadian rhythms on gene expression in rat liver and kidney. Gene expression was affected by both time of day and feeding state. The authors refer to a branch of mathematics known as Fourier Analysis of Time Series, which considers dynamic responses. Through dense time series experiments, it is believed one acquires a better understanding of the structure and behaviour of a system following exposure to compounds (i.e. an understanding of background variation or toxic effects), which leads to better designed studies and ultimately more reliable data. The authors consider that such approaches give insight into the nature of the underlying control circuitry and note the value of framing biological systems as 'complex integrated circuits'.
Analysis: The authors recommend that TGX studies consider the spatial and temporal contexts of any gene expression changes observed. The authors link the temporal aspects of how the activity of the transcriptome evolves to the mathematical discipline, Dynamics, which studies events as they unfold. The authors further emphasise the dependence of TGX data set interpretation to the application of mathematics in the following areas: eliminating noise via statistical procedures; detecting patterns of behaviour in the data in relation to treatment or their relevance to other endpoints; and [in particular] discovering regulatory/signalling networks and cascades controlling these events. A lack of comprehension of the underlying mathematics in bioinformatical procedures (i.e. approaches to data normalisation, pattern

recognition, singular value decomposition, principal component analysis, and clustering algorithms) is considered likely to generate a suboptimal interpretation. The authors suggest the necessary mathematical skills required to ensure optimal interpretation of TGX (bioinformatical) data include geometry, algebra and statistics. It is thought such a skill set enables researchers to follow the narrative and gain further insight into the analysis.
The authors describe the approach pathologists use to interpret/analyse TGX data i.e. via the application of statistics, assessment and sorting of data according to quality (aka data triage via use of bioinformatics tools), and use of an initial computer generated gene expression list to analyse genes one gene at a time (especially on well-characterised mRNA transcripts). Gene expression lists contain tens to thousands of transcripts that may be significantly up/down-regulated compared to reference population. These genes undergo further analysis to identify and understand their function. Identifying a gene's function is considered time consuming and tedious but software is available to help annotate them i.e. online bioinformatics databases and data-triage tools. Literature searches and textbooks are used to further understand the gene's function(s), and an automated method called Expression Analysis Systematic Explorer (EASE) is also available which searches bioinformatics databases simultaneously to produce triaged statistically relevant information that is pooled into a single spreadsheet with hyperlinks to selected databases. Further understanding of any associated physiological events can be achieved by integrating the information with other endpoints. Genes are categorised into functional classes (e.g. those involved in fatty acid metabolism, immune regulation, cell proliferation and apoptosis) to aid interpretation of changes (ranging from molecular to clinical). Identifying unaltered gene transcript expression within each functional class is considered worthwhile as it expands the overall understanding of the response. Diagrammatically linking transcripts in a pathway is thought to help illustrate the effects of gene transcript expression. However, the authors suggest the use of caution since these pathways are only speculative and would require further confirmatory investigations. The ability of TGX to predict the potential for toxic responses to chemicals is considered a valuable feature of this technology. This can be achieved through use of commercially available (and expensive) databases containing transcriptome expression patterns of well characterised treatments. It is noted that most of the data in these databases is generated using rat liver toxicants. However, other databases with data for other species and tissues are being developed.

Statistics: The authors explain that statistics is applied to TGX data analysis to determine the probabilities that the intensities of gene transcript expression changes (observed in the treatment group) are truly different to those of the control group. With regard to gene expression changes, the authors consider that the statistical level of change is more relevant than fold change and suggest that use of fold-change cut-off approaches should be avoided.

Comments: The authors consider that the effective application of TGX depends on the deployment of a range of skills derived from an understanding of molecular biology, biochemistry, toxicology, bioinformatics, statistics, mathematics and pathology – the latter two disciplines being particularly important. However, interdisciplinary communication issues associated with the use of different languages is thought to limit the rate of progress. The authors suggest that researchers become versed in the most essential areas and note that the Transnational College of LEX is attempting to address this issue.
The authors also infer that a multidisciplinary contribution to TGX analysis can help identify important toxicologically relevant patterns of transcriptome expression that would otherwise be missed. This is because biochemists can identify important metabolic pathways, while toxicologic pathologists can provide important morphological links to regional and cell specific protein expression.

Refs:

1. Boess et al 2003. Gene expression in two hepatic cell lines, cultured primary hepatocytes, and liver slices compared to the *in-vivo* liver gene expression in rats: possible implications for toxicogenomics use of *in-vitro* systems. Toxicol Sci 73: 386-402.

2. Irwin et al 2004[44]. Application of toxicogenomics to toxicology: basic concepts in the analysis of microarray data. Toxicol Pathol. 32(Suppl 1). 72-83.

3. Kepler et al 2002. Normalization and analysis of DNA microarray data by self consistency and local regression. Genome Biol. 3. 1-12.

   Expanded summary: Normalisation by local regression (NLR) is a statistical software which works on the assumption that: (a) the expression levels for a majority of genes will not change appreciably from one treatment to the next, such that a stable background pattern of activity (or transcriptional core) exists. Consequently, the constituent genes of this transcriptional core can be identified statistically for each experiment (i.e. from the data itself and not in advance); (b) any differences in expression level vs. signal intensity are small but significant. The authors illustrate the use of NLR in a study comparing the expression profiles of rat mesothelioma cells exposed to a potent inducer of oxidative stress (potassium bromate) against control cells. Validation of expression changes were confirmed by quantitative PCR on a selected set of genes. The authors also conducted simulation studies (under various error models) to test the normalisation method and demonstrate the technique's satisfactory performance.

4. Kita et al 2002. Implications of circadian gene expression in kidney, liver and the effects of fasting on pharmacogenomic studies. Pharmacogen. 12:55-65

5. Morgan et al 2002. Application of cDNA microarray technology to in-vitro toxicology and the selection of genes for real time RT-PCR-based screen for oxidative stress in Hep-G2 Cells. Toxicol Pathol. 30. 435-51.

   Expanded summary: This study had two objectives. The first was to better understand how the transcriptome responds to toxicity, and secondly, to use the information obtained to develop a high throughput RT-PCR based assay to detect one or more selected mechanisms of toxicity (in which the genes act as a marker of single mechanism of action). The authors used cDNA microarrays to examine chemically induced alterations of gene expression in HepG2 cells exposed to a diverse group of toxicants. Equitoxic concentrations of the following agents where used: oubain, lauryl sulphate, dimethylsulfoxide, cyclohexamide, tolbutamide, sodium fluoride, diethyl maleate, buthionine sulfoximine, potassium bromate, sodium selenite, alloxan, adriamycin, hydrogen peroxide and heat stress. The authors found that gene expression patterns correlated with morphological and biochemical indicators of toxicity (i.e. the responses corresponded with cell cycle arrest, DNA damage, diminished protein synthesis and oxidative stress). Also, each treatment yielded characteristically different gene expression responses (although certain genes failed to respond in an expected consistent or meaningful manner following treatment). It was decided that oxidative stress be used in the second part of the study as it yielded the most promising data, and is also considered a particularly significant mechanism of toxicity. The authors incorporated primers and probes for seven genes modified by oxidative stress into the design of a 7-gene plate for RT-PCR (5 genes were upregulated and 2 downregulated). Linear regression and ranking (Pearson product) procedures were used to correlate a simple oxidative stress score (0-1) (which was based on the responses by the 7 genes on the RT-PCR plate) with the GSH:GSSG ratio (which provides a measure of oxidative stress). The authors observed a good correlation between biochemical measures of oxidative stress (i.e. GSH:GSSG ratio) and transcriptional measures (i.e. oxidative stress score) – statistical analysis yielded correlation coefficients of 0.74 and 0.87 respectively). The authors conclude by highlighting the importance of measuring the mechanism of interest directly in the test system being used for any studies assessing the use of gene expression as a tool for toxicology, (as their findings show that selecting

[44] One of the reviews selected for summarising – see relevant narrative summary.

genes based on published literature is insufficient for marker gene selection, although it can provide an essential guide)..

6. Tietjen et al 2003. Single cell transcriptional analysis of neuronal progenitors. Neuron 38. 161-75.

   Expanded summary: This study sought to understand the mechanisms involved in neuronal differentiation and diversification, which is considered particularly challenging given the extraordinary cellular heterogeneity of the mammalian nervous system and the paucity of molecular data on the single-cellular level. The authors also note that the complexity of various tissues makes it difficult to detect highly specific precursor populations by simple homogenisation of whole tissue/organ areas to isolate RNA. The authors monitored expression profiles of individual neurons and progenitor cells of the highly heterogenous mammalian olfactory system, (i.e. specifically mature olfactory sensory neurons and olfactory progenitor cells – mitral cells of the olfactory bulb). The authors collected single cells from either dissociated tissue or from intact slices using laser capture mediated cell isolation (microdissection) techniques. Transcriptome data was generated by picking individual cells at random and seeding them into individual PCR tubes, DNA lysis and synthesis of first cDNA strand, followed by PCR amplification. cDNA samples were then hybridised to Affymetrix genechip probe arrays. NB. The authors also determined the identity and developmental stage of cell by PCR Southern blot analysis of the single cell cDNAs. The authors identified hundreds of transcriptional differences between olfactory progenitors and mature sensory neurons within the olfactory system, which further enabled them to define the large variety of signal pathways expressed by individual progenitors at a precise developmental stage. The authors conclude that their technique provides a sensitive and reproducible representation of the single cell transcriptome. Their findings suggest that a genome wide transcriptional analysis can be performed successfully at the single cell level. Furthermore, regional differences in gene expression can be predicted from transcriptional analysis of single neuronal precursors isolated by laser capture from defined areas of the developing brain.

Gant (2007). Novel and future applications of microarrays in toxicological research. Expert Opin. Drug Metab. Toxicol. 3(4):599-608
Topic covered: This review explores alternative applications of microarray technology and its role in toxicity assessment/drug development. The paper summarises various design issues relating to the application of nucleic acid-based arrays and provides a personal commentary from the author in the final section. There is limited discussion of data analysis issues and statistical topics are not addressed.
Design: The author notes that data generation microarray technologies have matured and become more robust, as demonstrated by studies conducted by the US FDA (**Guo et al 2006**). These studies show that different microarray platforms are producing quantitatively similar data. The adaptability of the microarray technology format is considered to be a useful feature in all experimental applications involving hybridisation. While a major application of microarray technology is to determine mRNA levels for many genes simultaneously other alternative applications do exist. In genomics these include those applications providing information on events upstream of mRNA synthesis (e.g. Array Comparative Genome Hybridisation (ArrayCGH) – a technique which detects variations in genomic copy number, epigenetic analysis, Chromatin Immunoprecipitation (ChIP) analysis and transcription rate analysis) and applications providing information on events downstream of mRNA synthesis (e.g. mRNA translation assays). The first alternative application of microarray technology was determining gene changes (amplification and deletion) in the genome. Using microarray technology to determine chromosomal changes can be considered analogous to measuring mRNA transcript levels except that the probe is genomic DNA (gDNA). For a two-colour system the probes are hybridised onto the same microarray producing a red/green spot image after scanning. The ratio of the fluorescent dye indicates either an amplification or deletion in the genome. After plotting the data against the chromosomal location of the probe a map of the chromosomes is produced. The author provides an example of a single gene deletion in rats where gDNA in the test rat (bearing a mutation that leads to Wilson's disease) is hybridised against a control Fisher rat to reveal a deletion of cadherin 11 gene. The author considers that the area of genome assayed (and the resolution) is dependent on the targets present on the microarray (and the number of probes used). The significance of epigenetic modifications in inducing transmissible genomic changes is summarised below (see comments section). Cytosine methylation is described as a type of epigenetic modification, which can be assessed using microarrays and immunoprecipitation methods (**Van Steensel, 2005**). The methods used in ChIP analysis are considered to be similar to those used in epigenetic analysis. ChIP analysis involves the use of antibodies raised against a transcription factor of interest; therefore, any microarrays used must have target sequences from gene promoter regions. The author notes several chemical agents that regulate gene expression by influencing binding of transcription factors to promoter regions of genes. These include TCDD, phenobarbital and retinoic acid. ChIP analysis is considered a useful tool to understand mechanisms, which can ultimately inform risk assessment. ChIP analysis also shows binding of transcription factors to gene promoter regions under different conditions. For example, **Rubins et al (2005)** and **Grass et al (2006)** examined transcription factor binding sites (for HNF6) in liver samples and GAT complexes respectively. The author stresses that there is still a lack of microarrays (targets) with suitable promoter fragments, although companies are developing more appropriate microarrays to address this. The author considers transcription rate analysis as a compliment to the ChIP assay, which aids an increased mechanistic understanding of toxicants. Transcription rate analysis adopts a nuclear run-on assay and microarray to determine increased transcriptional rates of genes in certain circumstances (e.g. after chemical exposure). **Gant et al (1991)** observed increased transcriptional rate of ABCB1 gene in rat liver following chemical exposure. The author describes the key steps involved in microarray transcription rate analysis as follows: isolating nuclei from test and control samples; incorporating labelled nucleotide into RNA (during transcription); isolating the RNA and hybridising it onto microarray (containing gene coding regions); detecting differential gene transcription by increased hybridisation to relevant target sequence on microarray. The author explains that criticisms levied against use of mRNA levels as a measure of gene expression are based on two key issues. Firstly, proteins are thought to be the most relevant

biomolecules with respect to mechanistic toxicological assessment of a compound. Secondly, there is no evidence that mRNA is translated into proteins (since increases in mRNA transcription do not necessarily follow with increases in protein levels). The author notes that although several attempts have been made to provide such evidence attempts are hindered by the technical limitations associated with 2D-gel resolution, quantification and other issues. Ultimately this makes its difficult or even impossible to quantitatively compare genomic and proteomics (PTX) data. The author states that the mRNA microarray translation assay aims to determine if mRNA is translated and if translation occurs differentially. The assay works by using density to separate the different types of mRNA i.e. those with ribosomes attached (heavier polysomal fraction) from whole RNA. mRNA with no ribosomes attached together with ribosomal RNA is known as the monosomal fraction. It is assumed that polysomal mRNA undergoes active translation in which the number of bound ribosomes is proportional to the amount of protein formed. By inhibiting mRNA species from translation and separating it onto a polysomal gradient a UV tracer can be used to illustrate the two separate monosomal and polysomal layers. The mRNA content can then be assessed via use of a microarray (containing consensus sequences for gene coding regions). Comparing the proportion of RNA in both layers can inform on whether the mRNA is translated. Measuring the ratio of RNA in the two layers informs whether the mRNA is differentially translated **(Mazan-Mamczarc et al 2005)**. Since fractionated RNA is used it is suggested that the above steps are performed with care. Therefore, during the hybridisation step, care should be taken with the amount of mRNA used. In the normalisation step, the authors recommend hybridising monosome fractions and polysome fractions separately onto different microarrays (i.e. test + control samples (monosomes) on one, and test + control samples (polysomes) on another). The author comments that although density RNA fractionation with microarrays is used in reproductive and cancer research, there is limited use of this method in toxicology. Some studies have examined the translational response to redox stress in yeast (**Shenton et al 2006)** and mammalian cells exposed to UV light (**Mazan-Mamczarz et al 2005**).

MicroRNA (miRNA) are single stranded RNA molecules that control gene (mRNA) translation. They are transcribed from polycistronic regions of the genome via RNA polymerase II and III to produce immature transcripts. These transcripts are processed in the nucleus/cytoplasm to produce mature miRNA (21-23 nucleotides long). MiRNA regulates translation by interacting with a multiprotein complex called RNA-inducing silencing complex (RISC) which essentially represses translation. MiRNA store mRNA in P-bodies within the cytoplasm which are later retrieved for translation. Such actions can increase protein levels without new transcription. The author discusses the potential regulation of miRNA expression by chemicals. In these scenarios any alteration in miRNA expression could alter the cell protein complement and a cells subsequent response to chemical exposure. The author suggests that the pattern of miRNA expression (miRNA profiling) could be used to identify specific toxicities, although to date there has been no application of this technology in toxicology.

MiRNA profiling is fraught with technical challenges associated with the short nature of mature miRNA species. The author suggests using the RNA tailing method for labelling and modified targets on microarrays (aka locked nucleic acid nucleotides) for hybridisation (**Castoldi et al 2006**).

The author stresses that these alternative techniques are not a replacement but addendum for established expression profiling applications. These techniques provide additional information to further understand mechanisms of chemical toxicity at the gene level. This can lead to identifying gene expression biomarkers that are specific for certain xenobiotic types. To demonstrate any genomic changes arising from xenobiotic exposure the author suggests using ArrayCGH in the same biological samples used for transcriptomics (TRSX). MiRNA analysis could also be conducted on the same samples from short term exposure studies to detect any xenobiotic related translational effects and whether this resulted from differential miRNA expression.

Variability in biological background is thought to affect the assessment of a chemical's toxicity and result in low dimension data sets. Such variability typically arises in *in-vitro* settings whereby cells become genetically unstable or are easily contaminated or responsive to environmental changes. Adopting *in-vivo* approaches e.g. using in-bred strains (to control for genetic background variability) could help reduce variability.

Analysis: The author considers that interpreting microarray data is challenging. There does not appear to be much discussion of issues relating to data analysis. However, the author does note the value of comparing arrayCGH data with mRNA microarray data, which helps

determine whether genes were amplified whole and if the increased copy number is reflected in mRNA levels. An example is provided showing certain genes that were either amplified but not overexpressed or amplified and overexpressed (and vice versa).
Statistics: This review does not discuss issues related to the statistical analysis of TGX data.
Comments: The author suggests that genotoxicity assays need to improve their predictive power to make them more relevant to actual carcinogenicity and non-genotoxic carcinogenicity. This can be achieved by deployment of arrayCGH techniques which could inform on a chemical's genotoxic potential and identify where genotoxic effects occurs on the genome. The author notes the application of ArrayCGH in drug efficacy and safety and for characterising cells and animal strains for testing purposes. Quantitative assessment of cells affected by genotoxic agents is considered problematic. Epigenetic modifications are considered important mechanisms of toxicity due to the inherited nature of their effects. Recent sequencing of the human epigenome has aided understanding of the effects drugs/chemicals have on DNA methylation patterns and subsequent gene expression changes leading to toxicity (including transgenerational toxicity). Such changes are thought to account for differences in susceptibility and resistance to drugs and chemical agents. The author considers that although epigenetic modification has a clear role in cancer development this is not the case with chemically-mediated toxicity as there is little work conducted in this area. A better understanding of the relationship between epigenetic change and phenotype is suggested as a way forward. Transgenerational toxicology is defined as a genome alteration where the phenotype is present in progeny. It is distinct from reproductive toxicity which occurs when the fetus is directly exposed to a toxicant *in-utero*. Transgenerational toxicology involves germline transmission of mutations in exposed parents. These minisatellite mutations in germ cells are thought to arise from parental exposure to chemicals. The author uses diethylstilbestrol (DES) as an example of an agent that causes reproductive toxicity and subsequent transgenerational toxicity. DES is thought to act by DNA methylation.

Refs

1. Castoldi et al (2006). A sensitive array for microRNA expression profiling (miChip) based on locked nucleic acids (LNA). RNA. 12(5):913-20.

2. Gant et al (1991). Regulation of 2-acetylaminofluorene and 3-methylcholanthrene-mediated induction of multidrug resistance and cytochrome P4501A gene family expression in primary hepatocyte cultures and rat liver. Mol Carcinogen. 4(6):499-509.

3. Grass et al (2006). Distinct functions of dispersed GATA factor complexes at an endogenous gene locus. Mol Cell Biol. 26(10):7056-67

4. Guo et al (2006)[45]. Rat toxicogenomic study reveals analytical consistency across microarray platforms. Nat. Biotechnol. 24(9):1162-9.

5. Mazan-Mamczarz et al (2005). En masse analysis of nascent translation using microarrays. Biotechniques. 39(1):61-7.

    Expanded summary: This paper describes the development of an approach for measuring en-masse changes in translation via cDNA microarrays. Human carcinoma cells were exposed to short wavelength UV light and the relative distribution of mRNAs were monitored along polysome gradients. Each gradient fraction was analysed via cDNA array analysis and regression analysis was used to quantify the mRNA translational status. The findings showed that steady state mRNA levels increased or remained unchanged while the translational status decreased (and vice versa). The authors report that the robust and predictive nature of their strategy enabled them to identify and verify a subset of 17 translationally induced mRNAs and 69 translationally repressed mRNAs. The authors concluded that the

[45] Also cited by Yauk & Berndt (2007).

assessment of total mRNA levels provides an incomplete account of gene expression changes. Instead critical information regarding which genes are ultimately expressed into protein is obtained by determining the degree of translational engagement.

6. Rubins et al (2005). Transcriptional networks in the liver: hepatocyte nuclear factor 6 function is largely independent of foxa2. Mol Cell Biol. 25(16):7069-77

7. Shenton et al (2006). Global translational responses to oxidative stress impact upon multiple levels of protein synthesis. J. Biol. Chem. 281(39):29011-29021.

   Expanded summary: This paper describes the analysis of protein synthesis regulation in response to oxidative stress. Yeast *Saccharomyces cerevisiae* were exposed to H2O2 followed by analysis of protein synthesis via incorporation of radiolabelled amino acids. Translational activity was analysed via measurement of the distribution of polysomes and ribosomal transit times. Polysome- and monosome-associated mRNA pools were then analysed using microarrays to identify mRNAs that are translationally regulated in response to oxidative stress conditions. The authors found that H2O2 inhibits translation initiation dependent on the protein kinase Gcn2 (which phosphorylates and thereby inhibits the initiation factor eIF2-alpha). A Gcn2 independent inhibitory mechanism was also observed (arising via inhibition of ribosomal transit). Other changes induced by H2O2 include the slower rate of ribosomal run-off (consistent with an inhibitory effect on the elongation or termination stages of translation) and H202 concentration-dependent effects on protein production, with low [H202] increasing protein production while high [H2O2] promoting polyribosome association without an automatic increase in protein production. The authors suggest the latter response may represent an mRNA store that can become rapidly activated following relief of the stress condition. The authors also found that oxidative stress increased the average mRNA transit time confirming post-initiation inhibition of translation. Global gene expression profiling revealed that certain mRNAs were translationally maintained following oxidative stress (i.e. increased in level in association with ribosomes) thereby indicating that translational control is a key component of the cellular response to oxidative stress. The authors concluded that oxidative stress elicits complex translational reprogramming that is fundamental for adaption to the stress.

8. Van Steensel (2005). Mapping of genetic and epigenetic regulatory networks using microarrays. Nat. Genet. 37:518-24. Review.

Magglioli, J. et al (2006) Toxicogenomic Analysis Methods For Predictive Toxicology. Journal of Pharmacological and Toxicological Methods, Vol 53:31-7
Topic covered: This paper focuses on the analysis of gene expression data i.e. the computational methods of class prediction, and the different steps that inform it (i.e. data preparation, class comparison, class discovery and evaluation). Statistical techniques used in each of these different steps are briefly discussed. Design issues are briefly addressed in relation to data preparation step.
Design: See analysis section below.
Analysis: The authors consider data preparation a necessary step to correct data sets for sources of variability arising from random and systematic error. To reduce random error the authors suggest generating many replicates and performing data analysis on the combined replicates. To reduce systematic error the application of background subtraction or normalisation is suggested. The authors describe class comparison as the method used to define a set of genes indicative of a particular class of toxicant (aka discriminatory gene set). These gene sets are defined by analysing prepared data from a training set. The authors describe the approach/procedure used to define a discriminatory gene set and cite **Tsai et al (2005)** as an example via the application of statistical methods such as ANOVA F-Test and One Versus All (OVA) test. As there are many variables the data set is often highly dimensional. Therefore, dimension reducing techniques such as Principal Component Analysis (PCA), Multidimensional Scaling (MDS) and wavelet transformation are applied (**Yang et al 2004**), although these techniques are themselves limited by the fact that producing a smaller number of weighted variables obscures information about which genes are extensively modified. A combination approach is proposed as a possible way forward i.e. use of ANOVA and wavelet transformation (Yang et al 2004). The authors define class discovery as the application of various techniques (e.g. clustering techniques) to visualise the similarity of gene expression present in a training set between individual treatments (of a chemical) or multiple treatments of chemicals from different toxicological classes. This approach is considered to be subjective as the results are influenced by selection of clustering algorithms and similarity metrics (**Simon et al 2003**). The authors highlight the two most common clustering algorithms used i.e. hierarchical algorithms (which results in a dendogram tree) and partitioning algorithms such as K-Means (that produces data bins based on *a priori* specified no. of clusters). Tsai et al (2005) examined the ability of hierarchical clustering algorithms to cluster datasets generated from rats treated with toxic metals. Clustering techniques where also used to examine how gene response varies with time (**Hamadeh et al 2002**). The use of K-Means is thought to be limited due to the bins preventing inferences being drawn on the relationships between each data points within a cluster. The authors define class prediction as the use of a toxin's gene expression signature to predict the toxicological class of an unknown toxicant (aka predictive toxicogenomics/toxicology). This is achieved by applying a classifier (or supervised learning method) to gene signatures of a training set, generating a mathematical model that can be used to predict the toxicological class of the unknown chemical. Class prediction is limited by the fact it can only indicate possible relationships between gene responses and phenotypes. The authors highlight various challenges for predictive toxicology for e.g. the cost of creating databases containing relevant gene expression data from studies of known toxicants (**Luhe et al 2005; Van Delft et al 2005**); dividing known toxins into toxicant classes distinguishable by their expression data, and comparing gene expression data collected using different technologies (**Hayes et al 2005**). The authors note that class prediction is preceded by class comparison and class discovery steps. The use of these steps to classify toxicants into particular groups has been documented in the published literature. Hamadeh et al (2002) used these steps to classify known hepatotoxins as either peroxisome proliferators or enzyme inducers; **Thomas et al (2001)** classified known toxins to one of five characterised toxicological classes; and Tsai et al (2005) classified toxic metals into seven or nine distinct groups. The authors describe the different types of classifiers/classification methods available, which include Linear Discriminant Analysis (LDA), Fisher's LDA (FLDA), Nearest Neighbour (NN), K-Nearest Neighbour (kNN) and Support Vector Machines (SVM). Tsai et al (2005) used FLDA and kNN to predict the class of gene signatures from the liver tissue of rats exposed to various toxic metals. These classification methods are limited by their tendency to

overfit the data in the training set, which limits their ability to predict profiles outside the training set. Filtering out gene expression profiles before applying classifiers is considered a useful way to deal with invariant gene expression profiles (Van Delft et al 2005). Different filtering methods are available and the authors note that their ability to influence the performance of classifiers has been investigated by Van Delft et al (2005). The final evaluation step involves evaluating the model produced from class prediction and estimating its ability to predict the toxicological class of unknown chemicals. The authors consider this acts as a validation step to characterise the ability of a classifier to predict the toxicological class of unknowns. This is done using individual and blinded samples from the training set.
Statistics: The authors consider the statistical challenges for predictive toxicology, which include calculating significant differences for datasets with many variables, and the developing statistical techniques that can accommodate the complexity of a toxin's effect on gene expression. Use of statistical techniques in each of the above five steps presents a particular challenge since each step has its own set of statistical methods that do overlap and also studies differ in the method they use. The authors identify a need for developing statistical methods (in class prediction) that can address the limitations of classifiers/supervised learning methods.
Comments: This review focuses on class prediction in relation to drug development. No other areas in TGX are discussed to any great length. The authors note their value as providers of commercial gene expression data analysis and management software for all relevant stakeholders in TGX. The authors conclude on the future of predictive TGX and suggest the computational methods used in steps that inform class prediction need to be standardised. This they propose can be achieved by narrowing the choice of classifier used or by conducting research comparing the merits and performance of different classifiers (Van Delft et al 2005).

Refs

1. Hamadeh et al (2002)[46]. Prediction of compound signature using high density gene expression profiling. Toxicological Sciences, 67:232-40

2. Hayes et al (2005). EDGE: A centralised resource for the comparison, analysis, and distribution of toxicogenomic information. Molecular Pharmacology, 67:1360-68

    Expanded summary: This paper relates to the development of the Environment, Drugs and Gene Expression (EDGE) public database, which the authors created to address the challenges of comparing gene expression data collected via different technologies. The idea to develop a low cost centralised resource (where researchers can share TGX data generated on a common platform) arose from the observation that different types of platforms, protocols and informatics produce different data, which hinder meaningful comparisons of transcriptional profiling data across laboratories. A key objective of the EDGE database is to map transcriptional changes from chemical exposure for future use as a diagnostic "fingerprint" to predict toxicity and provide valuable insights into the basic molecular changes responsible. The authors describe the approach used to develop the database for the analysis of liver gene expression in the mouse. This involved creating a standardised set of microarray reagents and reproducible protocols i.e. a cDNA-based microarray enriched for responsive targets of hepatotoxicants (e.g. TCDD, cobalt chloride and phenobarbital) in the mouse model. This would enable researchers to compare transcriptional profiles arising from chemicals and other stimuli. The authors developed a pipeline to generate transcriptional profiles for a set of prototype chemicals and pathological states e.g. inflammatory cytokines, aryl hydrocarbon receptor agonists and peroxisome proliferators. At the time of writing 117 treatments, doses and timepoints were publically available. The authors anticipate an additional 400 treatments within a further 4 months. Platforms containing unique genes from mouse skin, lung, kidney, palate and tendon have been developed, and at the time of writing similar

[46] See Hayes et al (2005) narrative summary

platforms for thymus, heart and ureter were in development. Finally, the authors propose that EDGE would serve as a prototype resource for the sharing of TGX information. They implemented an online database containing tools enabling researchers to query and interpret large numbers of transcriptional profiles. Two modes of researcher interaction are described: (1) Investigators can query a database of toxicant induced transcriptional profiles generated from a common microarray protocol/platform (various informatic tools available to help via web interface available at http://edge.oncology.wis.edu/edge; (2) Investigators can submit RNA samples to develop novel toxicant-induced transcriptional profiles for inclusion in the database.

3. Luhe et al (2005). Toxicogenomics in the pharmaceutical industry: Hollow promises or real benefit? Mutation Research, 575(1-2):102-115. Review

The following selected references cited in the above paper are further summarised

Butte (2002). The use and analysis of microarray data. Nat Rev. Drug. Discov. 1: 951-60.

[Append]

Chen et al (2004). Analysis of variance components in gene expression data. Bioinformatics. 20:1436-46

This paper presents an approach to estimating sources of variation and their relative contributions (magnitude) to the overall variation in microarray studies. The authors note that variability in microarray data is unavoidable and expected and that many potential sources of variation exist due to the multistep nature of the system. Identifying and estimating different sources of variation is essential for designing cost-efficient microarray experiments. Three types of variation are considered: biological, technical and residual variation. Biological variation is defined as variation arising from the use of different animals, cell lines or tissues i.e. variation from different RNA sources that reflect differences in host characteristics. Technical variation arises from use of the microarray system and the authors list the different sources that arise at each step of a microarray experiment i.e: sample preparation procedures (e.g. RNA extraction and purification, cDNA synthesis, and incorporation extent of dyes/specific batch of dyes used); microarray construction procedures (e.g. amount of probe applied to the slide, spot shape, pin geometry and fixation of the spotted DNA to the slides); hybridisation and washing procedures (e.g. amount of labelled cDNA applied to the slides, and hybridisation temperatures); detection method (e.g. scanner setting parameters); cross hybridisation within gene families; outshining from neighbouring spot; laboratory environmental conditions (aka time/block effect e.g. room temperature); and other sources such as concentrations of reaction/wash buffer components (can be effectively controlled using stock solutions and master mixes). Residual variance refers to sampling or other experimental unaccountable factors. The authors note that all three types of variation are mutually independent and their summation is the variation in a measured fluorescent intensity.

The paper states that variation is identified and estimated using the statistical technique analysis of variance (ANOVA) which models sources of variation and provides an automatic correction for the nuisance effects in estimating the relative expression of genes across experimental samples. ANOVA considers all sources of variation in an experiment and summarises them into one equation. Although initially developed to analyse differences between means, ANOVA was later adapted into a model capable of explicitly estimating the magnitude of the sources of variation and variance

components. Analysis of variance components of a data set involves attributing variability to various factors (e.g. treatment, dye, animal or array). Each factor has a different level of effect that impacts on the measurement of interest. These effects are considered as being either fixed or random. Fixed effects for dye or treatment factors relate to the fact the dye is either red or green and that treatment groups are either exposed or unexposed. Random effects are attributed to animal and array factors as they are randomly chosen from an infinite population. The authors note that analysis of variance components involves estimating the variance of random effects that requires separating the variance of random components (animal, array) from treatment (fixed) effects. This can be achieved by use of a variance component model of a repeated dye-flip experiment (see below).

The authors note that replication also contributes to sources of variation assessments and forms one of three basic concepts of experimental design (the other two being randomisation and blocking – see below). Since the number of replications integrated into an experiment determines the quality of the statistical method used to analyse variation, it is important to carefully design an experiment with appropriate replications to provide a sound basis for statistical analysis, This will help towards generating more accurate and reliable data, and thus enabling better understanding and interpretation of the significance of the observed changes for thousands of genes. Replication is incorporated at different levels of an experiment e.g. wrt sample (using a replicate number of tissues or cell types), array (using a replicate no of arrays) or spot (using a replicate number of spots of the same gene). The number of replicates necessary depends on the noise level of the system. The greater the number of replicates used the greater the ability to detect very small differences and distinguish differentially expressed genes from noise. Two types of replications are described. Technical replication involves the use of replicates where the mRNA is from the same pool (same extraction). Technical replicates are used to reduce experimental variabilities such as those arising from measurements. Biological replication refers to hybridisations that involve mRNA from different extractions i.e. different biological samples that reflect variability in the population of interest. Biological replicates are used to obtain averages of independent data (and to validate generalisations of conclusions). Experiments that pool biological samples minimise biological variation without affecting variation of the technical or residual component.

The authors note that the type of replicate used is dependant on the aim of the experiment which will thereby determine the statistical test used to assess variability in the data. For example, consider an experiment that aims to determine the effects of treatment on different biological populations, to detect if the variability in the data obtained is statistically significant, since biological replicates (different biological samples) were used the statistical test would need to be based on biological replicate samples. If an experiment aims to detect the variation within a particular experimental group, since the same sample is used any variations would be due to technical procedures and so tests used to determine the statistical significance of these variations would need to be based on technical replicate samples. A widely used technical replicate (used in two colour spotted array analysis) is the dye reversal or dye flip design. This process aims to compensate for dye bias (i.e. all biases occurring during labelling or hybridisation or due to the physical properties of the dyes themselves e.g. heat and light sensitivity or half life). Dye flip uses two arrays with treated samples labelled with a Cy5 red dye and control samples labelled with a Cy3 green dye and both hybridised onto one array. On the second array the reverse labelling of samples occurs. A schematic is provided which illustrates the design of a replicated dye-flip experiment.

The authors investigated variation in two data sets. The first data set was retrieved from a toxicogenomic (TGX) study that measured gene expression changes of kidney samples from rats dosed with 5 mg/kg cisplatin (a known kidney toxin). The array used consisted of a 700 gene cDNA rat chip from Phase-1 Molecular Toxicology. Plant and bacterial genes were also spotted on the array as replicates to monitor non-specific background binding of labelled cDNA. Each gene had four replicate values on each of the six arrays used (labelled A1-A6). To estimate only technical and residual variances samples were pooled for either treated or control RNA thereby minimising the effect of biological variation. Treated samples were derived from kidneys of 5 rats, 7 days after treatment. Both labelling and hybridisation were performed on one date with each replicate sample labelled independently. A dye flip design was used to minimise dye bias. A1-A3 arrays had control samples assigned to the green (Cy3) dye and treated samples to the red (Cy5) dye. A4-A6 arrays had the reverse labelling assignment. Fluorescence intensity was assessed by subtracting the local background intensity from each raw fluorescent value using the GenePix software package (Axon instruments Inc, 1999).

The authors estimated the following technical variances: between-array and within-array (i.e. between-section and within-section variances). This was done using a mixed effects (ANOVA) variance-component model. Between-array variance relates to variation from one array batch to another (e.g. variance in quality and homogeneity in manufacturing an array including gene sequence variance) or variance in hybridisation from one array to another (e.g. due to sample preparation being performed on different dates). Between array variance can observed can be due to either biological or technical factors. To assess between-array variance due to technical factors replicates based on mRNA samples from the same extraction were used in more than one array. (NB. Between-array variance due to biological factors arises due to use of mRNA samples from different biological samples hybridised onto different arrays, however, this was not investigated in this paper). Within-array variance is defined as variation originating from either array-specific spot effects (e.g. scratches or dust on surface of an array or due to printing, washing or image extraction) or systematic effects (e.g. differences in labelling efficiency, intensity or spatial dependency biases). Within –array variance was assessed using replicates derived from mRNA samples from the same extraction at different locations on each array. To investigate the distribution of variance components across genes the authors consider a mixed effects model for gene by gene analysis.

The second data set was obtained from a study investigating circadian changes in gene expression in liver samples from rats. The study design was as follows: rats were fed an ad libitum NIH-31 diet with a 12 hr light/dark cycle; 4 rats were sacrificed at 52 weeks of age at 4 different time points; total RNA was extracted from the livers generating 16 samples; reference RNA was produced by mixing equal amounts of 16 samples; RNA samples were divided into 4 experimental blocks, each block containing a sample from each of the 4 sacrifice times and each sample within each block labelled and hybridised on a single day (dye flip reversal labelling/hybridisation was also performed on the next consecutive day); each block was run on 4 different weeks. The authors note that the key effect of interest was detecting differences in gene expression among the 4 sacrifice times. Biological, technical and residual variance were investigated. Animal-to-animal (between-rat variation) was estimated via use of biological replicates in the test samples for five housekeeping genes (i.e. genes that are not expected to change with time among rats). Since the reference samples came from the same pool there was no biological effect to detect. A mixed effects model is used to estimate the following variance components: block (week to week), between-day, between-array, within-array and residual variance.

Findings from the toxicogenomics data set showed that between array variance was larger than the between-section (within array) variance which itself was larger than within-section (within array) variance. For the circadian data set in the reference sample ANOVA analysis and the variance components estimates showed that the week-week variance was larger than the between array variance which was larger than the within section variance. For the test sample ANOVA analysis and variance component estimates revealed that the week by week (block) variance was larger than the animal-animal (biological) variation which itself was greater than the two technical variations (between-array and within array). Gene by gene analysis of the five housekeeping genes revealed that the technical variation (between array) was larger than the biological variation (animal to animal) for 4/5 genes.

The authors concluded that the week-to-week effect (week by week variations) were the largest i.e. the overall variability in the data was largely due to the performance of procedures from one week run (block) to the next. The authors suggest that reducing this variability would increase the precision of the estimates of gene expression changes. Also animal-to-animal variation was considered a key source of variability that can be effectively reduced by replicating biological samples.

Additional noteworthy points

The authors consider randomisation, blocking and replication as basic principles of experimental design. The purpose of randomisation is to reduce the likelihood of systematic biases caused by selection or assignment. This would involve randomising biological samples to a treatment to equally represent underlying characteristics of subjects or randomising dye assignments in technical replicates. The authors define blocking or a block as a subset of experimental units that are more homogenous than the entire experimental itself. Blocking is often incorporated into a study to increase the precision of estimates made.

The authors consider a microarray experiment as a comparative experiment that compares relative expression levels among samples rather than determining absolute intensity measures of each sample.

Chu et al 2004. Cross-site comparison of gene expression data reveals high similarity. EHP.112:449-55

This study was conducted to evaluate data quality and statistical models that facilitate comparison of high-density gene expression data sets at the probe level. The authors consider data quality and choice of statistical models as one of several factors that influence the consistency and coherence of gene expression data across multiple test sites. The authors note that the Hepatotoxicity Working Group of the ILSI HESI consortium on the application of genomics to mechanism-based risk assessment is investigating these and other factors: [The Consortium's investigation involved comparing high-density gene expression data sets generated on two sets of RNA obtained from two independent in-vivo rat experiments conducted at seven different laboratory sites (pharmaceutical companies). The hepatotoxicant methapyrilene (MP) was administered by gavage to male Sprague Dawley rats for 1, 3 or 7 days at doses of 0, 10, 100 mg/kg/day. A single platform (Affymetrix Rat Genome U34A GeneChip) was used for transcript profiling]. The authors present methods for exploring and quantitatively assessing differences in the above data to generate lists of site insensitive genes (i.e. genes that are invariant across sites) that are responsive to low and high doses of MP. To increase the power of statistical inferences, the authors pooled data sets across all test sites, however, this is known to compromise

the comparability of the data, which the authors state can be rectified by adopting a robust normalisation method and they subsequently describe three possible approaches. The authors highlight the advantages of using interquartile range normalisation, which not only makes data comparable across sites but also preserves a certain level of site effects when combining the data. Other approaches summarised include using a universal reference sample or an invariant portion of data across arrays. The authors found that using both numerical and graphical techniques reveals important patterns and partitions of variability in the data (including the magnitude of site effects i.e. effects/differences in datasets due to study being conducted at a particular site e.g. platform differences, environmental conditions, data quality). The authors report that these site effects were primarily additive, and can be adjusted in the statistical calculations in a way that does not bias conclusions regarding treatment differences. The authors used a mixed model approach, which they concluded provides a flexible method to adjust site effects and use different array variations between sites. The authors also note that each site tends to generate similar lists of significantly differentially expressed genes.

Additional noteworthy points

The authors note that a common scenario in microarray design is having large “p” (number of genes) and small “n” (number of arrays aka. sample size), which is associated with low statistical inference power. To overcome this and minimise both false positive and false negative rates investigators can increase sample size, although this can be costly. To help alleviate the cost investigators incorporate data sets generated at disparate sites and times but such an approach raises questions over the consistency of data generated across multiple sites and whether the same or similar conclusions be drawn.

Novak et al (2002). Characterisation of variability in large-scale gene expression data: implications for study design. Genomics. 79:104-113

Rihl et al (2004). Technical validation of cDNA based microarray as screening technique to identify candidate genes in synovial tissue biopsy specimens from patients with spondyloarthropathy. Ann. Rheum. Dis. 63. 498-507.

This study sought to validate the use of cDNA-based microarrays on synovial biopsies by analysing the technical sources of experimental variability. Previous work in the field has typically used peripheral blood and synovial fluid mononuclear cells from patients with active spondyloarthropathy (SpA), a prototype of chronic inflammatory arthritis. However, as the synovium is the primary site of inflammation of the arthritic joint the authors considered that this tissue would be the relevant target structure. Such an approach raises a number of technical questions in relation to the reproducibility of methods, sample heterogeneity, use of statistical analysis with thresholds of an arbitrary nature, and the quality and quantity of RNA/cDNA (which may require amplification, and thereby potentially introduce further bias and distortion of gene expression profiles). The authors hoped that the study would enable them to better characterise the mRNA transcripts mediating active SpA (i.e. identify candidate genes in synovial tissue) to increase their understanding of the mechanisms of this particular disease.

The authors analysed the reproducibility of this screening technique in three ways. Firstly, they compared the effect of two different amplification approaches (exponential and linear RNA amplification) since the type of amplification approach used can introduce variability due to associated technical limitations. Exponential amplification was performed using the Switch Mechanism At the 5’ end of the RNA Template (SMART PCR), which

allows reverse transcription of small amounts of total RNA and subsequent amplification of the entire cDNA. Linear amplification (an in-vitro transcription (IVT) based approach), is typically based on T7 RNA polymerase IVT. Secondly, the authors studied the variability between two different cDNA based nylon membrane microarray systems (Atlas [1126 genes] and GeneFilters (GF211) [4K genes]) on peripheral blood mononuclear cells (PBMC) (with duplicate experiments conducted). Their final analysis of reproducibility involved studying the run-to-run variability on synovial tissue biopsies (i.e. variability arising from different runs of the microarray procedures).

The effect of sample heterogeneity on the microarray results was assessed by analysing the gene expression profiles of the SpA synovium (with the total RNA sampled from 3 SpA patients) and comparing this with two control groups specimens: the oestoarthritis (OA) synovium – which is phenotypically quite heterogenous (total RNA sampled from 3 OA patients), and PBMCs – which is phenotypically quite homogeneous (total RNA sampled from 4 healthy controls). The authors identified the genes expressed and compared the microarray results for the two sets of arthritic patients with their histological findings (to evaluate correlation between histology and gene expression).

Finally, the authors examined whether use of the classic statistical methods (Analysis of Variance and Students t test with Bonferroni adjustment) represented a valid approach to analyse the data i.e. whether these methods could reliably identify statistically significant differences between samples. To analyse the appropriateness of the chosen thresholds, permutation tests of the SpA and OA synovium tissue data were conducted. The authors hoped this would show that their analysis produces a low number of false positives thereby confirming that the results obtained reflect gene expression and were not a random result due to multiple comparisons. It was also hoped that this would emphasize the need to use stringent thresholds to avoid increasing the number of false positive genes detected.

The authors found that 86 per cent of the cytokine/chemokine genes identified were expressed in both microarrays and both RNA amplification systems. Furthermore, in one microarray system the expressed genes were 78-95% concordant in duplicate experiments. Cluster analysis revealed a higher degree of similarity between gene expression profiles of SPA synovium than between PBMCs (more homogeneous) or OA synovium despite the fact that the synovial samples had clear histopathological differences. The authors suggested that tissue heterogeneity did not bias the results since comparisons made between the SpA synovium and OA synovium and with PBMCs yielded 11 and 18 expressed transcripts respectively, (of which six were shared in both comparisons) and permutations of SpA and OA samples yielded only one expressed gene in 19 comparisons.

The authors concluded that microarrays can be used for the analysis of synovial tissue biopsies with high reproducibility and low variability of the generated gene expression profiles.

Lee et al (2004). The intelligent data management system for toxicogenomics. J Vet. Med. Sci. 66:1335-38.

This is essentially an information paper that presents the TEST (Toxicogenomics for Efficient Safety Test) database management system (DBM). The authors consider that this represents an intelligent database system, capable of handling heterogeneous and complex data from many different experimental and information sources. The intelligent query feature

enables users to obtain relevant useful information from complex data sets and conduct multiple comparisons. Information can be retrieved for the following: (i) compounds, which are classed into either anti-cancer, antibiotic, hypertension and gastric ulcer groups; (ii) animal experimental data, such as food consumption, histopathology, statistical analysis results and pathologist's comments; (iii) gene expression data , e.g. annotation and expression information for each clone, statistical data, data quality of each slide and differentially expressed gene lists (with 6 array data sets per compound); and (iv) annotation. The authors note that at the time of publication the system housed information for 16 compounds, 45 microarrays, 190 animal experiments, and had a customised 4.8K rat clone set. Data can be accessed online via http://istech.info/TEST/ and users requiring gene level data can enter their query into the annotation database with the gene's name and ID nos of relevant database, and functional key words. Expression profile information is obtained via links to a microarray database. The authors also describe the design of a toxicogenomic (TGX) array (to demonstrate the application/value of the DBM system) which involved screening candidate toxin related genes from several public databases (via searches of genes that are functionally related to known toxins genes (based on Gene Ontology)), annotating the selected genes and dividing them into their respective grade sets (A-D) based on their confidence intervals; grades A and B representing the major clone sets for toxicity testing.

In conclusion, the authors consider that the TEST database represents a useful information source for studying toxicology mechanisms on a genome-wide level, which can also be applied to the design of microarrays for toxicity testing.

Zhang & Gant (2004). A statistical framework for the design of microarray experiments and effective detection of differential gene expression. Bioinformatics. 20: 2821-28.

This paper describes the development of an approach to measure the success rate of differential gene expression (DGE). The authors consider that the unsatisfactory detection rate of DGE together with the large number of false positives represent significant challenges for microarray studies, and that these challenges arise due to the large data sets and the intrinsic variability of the system. Two sets of variation associated with gene expression experiments are described. Biological variation is defined as inter-individual differences between members of a population while technical variation refers to errors arising from the experimental procedure. The authors also describe approaches used to account for these variations, for example, the effect of biological variation can be reduced by using sufficient number of biological individuals. However, reducing the effects of technical variation is more complicated as it depends on whether the variability arises from random error (whose effects can be reduced by making multiple measurements) or systematic bias (which requires the deployment of correct experimental designs to reduce their effects). A significant source of systematic bias considered throughout this paper is imbalances in the measured fluorescence intensities (for microarray experiments using dual label hybridisations). The authors note that normalisation procedures are often applied to remove systematic biases before statistical analysis of microarray data and describe two types used to adjust the measured fluorescence levels: globalisation normalisation and an alternative intensity-dependent normalisation method.

The authors propose a mathematical model to help investigators alter experimental design (i.e. to help select an appropriate number of microarrays for an experiment which can then provide the desired detection power of DGE) and in doing so thereby account for fluorescent label bias. (NB. The

model also takes into account other major variables associated with microarray data). The model is applied to measured data in microarray experiments and works under the following assumptions i.e. is designed for experiments where: there are two sample groups (treated vs. control); each gene is spotted only once on each microarray; the following factors contribute to the log intensity fluorescence for a feature (single spot): the amount of corresponding mRNA in biological sample, the effect of quality of the feature spot, the effect from the labelling fluorchrome and random measurement error). These parameters are incorporated into an equation enabling calculation of magnitude of differential expression. The authors note that this model can be used for developing a t-based statistical procedure to determine DGE (per gene printed on the microarray). The t-test compares expression in the treated vs. control group with the null hypothesis being that the gene has same expression level in two groups.

The authors derive a formula to determine the success rate of DGE detection (i.e. the rate at which DGE is correctly identified (either up or downregulated)). The formula assists in the design of microarray experiments and takes into account the number of microarrays and genes, the magnitude of DGE and the variance from biological and technical sources. The authors note that a look-up table based on the above formula is used to help investigators determine the percentage of true DGEs which can be detected by the experimental design used. A software link for investigators wishing to calculate the success rate of DGE detection is provided http://www.le.ac.uk/mrctox/microarray_lab/Microarray_Softwares/Microarray_Softwares.htm [doesnt work].

Finally the authors propose an adhoc method for estimating the fraction of non-differentially expressed (null) genes within a set of genes being tested (N_o/N). This method increases the power of DGE detection and is based on an equation which can be used independently from the specific form of statistical tests being used. However, since it makes a number of assumptions the authors note that the method should serve as an approximation for estimating the fraction of null statistics.

Overall the authors propose that this measure could be routinely used in the design of microarray experiments (or post-experiment assessment).

Additional noteworthy points

In this paper ‘a desired number of experiments’ refers to the number of forward and reverse labelled microarrays required to achieve a desired power of DGE detection with control on the number of false DGE calls.

The authors note that global normalisation seeks to adjust the effect of global factors that could generally affect the fluorescence. Such factors include the difference between overall concentrations of two mRNAs and the difference of photoamplifier voltages used between two fluorescent channels when the microarray image is scanned. Global normalisation globally (uniformly) adjusts fluorescence levels of all the features. However, this can introduce possible local feature specific bias, which is accounted for by reverse labelling and statistical testing (see below). Despite this, global normalisation is limited by its inability to account for different magnitudes of imbalances from feature to feature. In contrast, an alternative (intensity-dependent) normalisation method adjusts fluorescence level according to local properties of the feature spot and fits measured data to a smooth non-linear curve. Although this approach is unlikely to correct for all features, this can be rectified by removing systematic bias via experimental means i.e. using reverse labelling dyes when hybridising some microarrays (dye swapping) and using ANOVA methods.

Steiner et al (2004). Discriminating different classes of toxicants by transcript profiling. EHP. 112. 1236-48.

This study sought to determine whether biological samples from rats treated with various compounds could be classified into different classes of hepatotoxicant based on gene expression profiles. The study also aimed to see whether the mode of toxicity could be predicted i.e. cholestasis, steatosis, direct-acting and peroxisome proliferator-activated receptor-α. Wistar rats were exposed to high doses of 28 hepatotoxicants (that were either already established or pre-clinically tested) and 3 non-hepatoxicants with their corresponding time-matched controls. NB. Sprague-Dawley rats were also used (see below). High doses were used to ensure conventional endpoints could assess the toxicity produced. Hepatic gene expression profiles were analysed via Support Vector Machines (SVM), a supervised learning method that generates classification rules. To enhance its classification performance SVM is used in combination with another method (recursive feature elimination) creating sets of informative genes. The authors note that SVM is particularly well suited for the analysis of microarray expression as it can recognise informative gene patterns in input data and make generalisations on previously unseen samples. However, a training set of examples must be provided i.e. a database of model compounds that produce a particular toxicological endpoint. The authors tested two different SVM algorithms to produce predictive models, which underwent compound based external cross validation. To assess toxicity the authors performed a complete serum chemistry profile on each animal, and liver and kidney histopathology. As different strains are widely used in toxicology and known to vary in their susceptibilities to toxicants, the authors investigated the effect of strain differences of Wistar and Sprague-Dawley rats for classification based on the transcript profiles. The authors generated a SVM algorithm using Wistar rat data and assessed whether the model would correctly classify individual animals from another strain by using the vehicle control and WY14643 (a peroxisome proliferator)-treated Sprague-Dawley livers. Transcript profiles for SVMs were then assessed. Finally, the authors tested the hypothesis that unknown blinded compounds can be accurately classified based solely on gene expression profiles using compounds with mechanisms of toxicity not represented in their training set (e.g. liposaccharide, phenobarbital and indomethacin). The authors found that combined use of recursive feature elimination enabled them to identify a compact subset of probe sets with potential use as biomarkers. Furthermore, the SVM models were able to predict toxicity as well as the mode of toxicity, enabling discrimination of hepatotoxic and nonhepatoxic compounds and correct prediction of the toxicants class. Finally, the authors found that the predictive model (built on transcripts from one rat strain) could successfully classify profiles from another rat strain. In conclusion, their findings confirmed the hypothesis that compound classification based on gene expression data is feasible and that toxicogenomics is a powerful tool for classifying compounds according to their toxicity mechanism (assuming a well-designed database is combined with appropriate bioinformatic tools).

4. Simon et al (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute, 95:14-18. Commentary.

5. Thomas et al (2001)[1]. Identification of toxicologically predictive gene sets using cDNA microarrays. Molecular Pharmacology, 60:1189-94

Expanded summary: The authors developed an approach to test the hypothesis that toxicants can be classified according to how they affect mRNA transcript profiles. This centres on the ongoing requirement for alternative approaches for

the safety testing of chemicals as current methods cannot be applied to all chemicals of concern. The authors exposed male C57BL/6J mice to 24 known (model) treatments (toxins) that fall into the following 5 toxicological classes: non-coplanar PCBs (e.g. PCB-153, Arocolor-1260, Phenobarbital); Peroxisome proliferators (e.g. Cipro, Wy-16,463); Inflammatory agents (TNF-a, LPS, IL-6); Hypoxia-inducing agents (e.g Cobalt, Phenylhydrazine); and Aryl Hydrocarbon Receptor agonists (e.g. TCDD, BNF). The authors also examined changes in expression of genes obtained from liver samples. The authors were able to classify toxicants with up to 70% accuracy (after analysing 1200 transcripts). They also identified a diagnostic set of 12 transcripts that allow for an estimated 100% predictive accuracy for the toxicological classes chosen in the study. The authors concluded that their findings support the accurate classification of toxic chemicals based on their transcript expression profiles (i.e where transcript expression is an endpoint), which as an alternative testing strategy would provide huge savings in terms of cost, animals, time. Furthermore, the authors considered that once a diagnostic gene set of indicator transcripts are identified, large arrays with thousands of transcripts are unnecessary to make these classifications.

6. Tsai et al (2005). Multi-class clustering and prediction in the analysis of microarray data. Mathematical Biosciences, 193:79-100.

7. Van Delft et al (2005). Comparison of supervised clustering methods to discriminate genotoxic from non-genotoxic carcinogens by gene expression profiling. Mutation Research, 575:1-3

8. Yang et al (2004). Toxicogenomics in drug discovery: From preclinical studies to clinical trials. Chemico-Biological Interactions, 150:71-85. Review

Hayes, K.R. & Bradfield, C.A. (2005). Advances In Toxicogenomics. Chemical Research In Toxicology. Vol 18, No. 3:403-14
Topic covered: This paper provides a general overview of the application and interpretation of transcriptional profiling. Design issues are considered in relation to microarray fluorescence and the experimental design of microarray studies. The analysis of raw microarray data are briefly described in relation to clustering techniques. Statistical approaches are not discussed to any great extent, except in relation to hierarchical clustering and as approaches to classifying chemicals based on their transcriptional profiles.
Design: The authors note two microarray approaches used in transcriptional profiling: single and double fluor protocols. Single fluor protocols define approaches where control and experimental samples are hybridised against separate microarrays. Expression ratios are calculated from each microarray to relate the data. Double fluor protocols define approaches where both control and experimental samples are hybridised against the same microarray. Fluorescent tags with different excitation and emission spectra are used and the expression ratio is calculated for the two samples at the same location. The data generated is then presented as a heat map. The authors acknowledge the significance of experimental design in terms of its ability to affect the information that can be obtained from a microarray (particularly for two colour hybridisations). Three common types of design for microarray experiments are described. Direct design involves hybridisation of the control sample directly against the experimental sample (thus requiring the use of different fluors to label each sample). Reference design involves hybridising both control and experimental samples directly against a common reference sample. Loop design involves hybridising biological replicates (from multiple groups) against each other to connect all samples. The authors consider that the aim of an experiment will ultimately determine which design is appropriate (**Churchill (2002); Yang & Speed (2002); Kerr (2003)**). The direct design is most sensitive and thus appropriate for studies seeking to identify a list of differentially expressed targets at a particular time (i.e. identity of target is important). This is done by comparing each time point to a time-matched control. Loop design is considered most appropriate for studies seeking to understand how targets change over time (i.e. where the temporal nature of the target is important). Loop design is thought to provide better information on the influence of time on a particular target. However, comparing each time point against the preceding and subsequent time points compromises the sensitivity of detecting changes in transcriptional profiling. Reference design is deemed most appropriate for studies seeking to determine both the identity of targets and temporal nature of their changes. Reference design provides information on both the magnitude of the change and temporal relationship, although at the expense of reducing statistical power for detecting both.
Analysis: The authors define clustering as an unsupervised approach to presenting microarray data that is often used to organise global expression profiles. Three types of clustering techniques are briefly described. Hierarchical clustering (HC) groups targets based on their similarity of expression and can be applied to treatments also. HC generates a dendogram (via statistical approaches – see below). To fully visualise the data the dendogram is positioned together with the heat map. K-Means clustering groups targets based on a preset number of cluster groups (i.e. number of clusters are defined a priori). Principal Component Analysis (PCA) is used to analyse data with multiple values. Data from matrices of gene expression data are collapsed into eigenvectors and plotted to give relative locations of profiles. The distance between profiles correlates with similarity.
Statistics: The authors note that to create a hierarchical tree / dendogram in hierarchical clustering the Euclidian distance (correlation coefficient) must be calculated to rank similarities of gene expression profiles. The following statistical approaches identified from published literature illustrate the ability of transcriptional profiling to classify chemicals i.e. Bayesian probability (**Thomas et al 2001**) and Linear Discriminant Analysis (LDA), genetic algorithm (GA)/ K-Nearest Neighbours (KNN) (**Hamadeh et al 2002**). The authors suggest that a robust method is required that is capable of incorporating and predicting toxicity using much larger data sets.
Comments: Other areas addressed include an explanation of how chemicals can invoke similar transcriptional profiles e.g. inducing cognate adaptive metabolic pathways, and the benefit microarray technology brings to the process of classifying chemicals. The authors comment on the value of toxicogenomic (TGX) repositories of online

transcriptional profiles and highlight the significance of information sharing and the impact Minimum Information About Microarray Experiment (MIAME) guidelines have made to TGX studies. Comparing gene expression data is considered problematic due to extensive interlaboratory variation and the existence of multiple strategies to annotate targets (**Mattes 2004**). The authors suggest that possible solutions include standardisation of protocols (**Thompson et al 2004**) – although the large number of platforms and protocols is a limitation (**Baker et al 2004; Mah et al 2004**) – and careful sequencing and curation of genomes.

The review summarises the various TGX transcriptional profiling resources (i.e. National Center for Toxicogenomics (NCT), Environment, Drugs and Gene Expression (EDGE), Pharmacogenomics Knowledge Base (PharmGKB), dbZach and Comparative Toxicology Database (CTD)) and basic (non-TGX specific/general) transcriptional profiling resources (i.e. GEO, ArrayExpress and Symatlas). Comparative and computational genomics and systems biology/pathway mapping are also summarised. A brief comment is given regarding the value of TGX in drug development and the regulation of environmental/industrial chemicals. The authors state that the use of TGX data in predicting the toxicity of environmental/industrial chemicals is a grey area, although TGX will be particular useful in verifying the toxicities of compounds regulated under one rubric e.g. Toxic Equivalency Factor (TEF) for dioxins, dibenzofurans and PCBs. In conclusion, the following areas are identified as problematic: platform and data compatibility; completeness of information; assimilation into usable databases and statistical power.

Refs

1. Baker et al (2004). Clofibrate-induced gene expression changes in the rat liver: A cross laboratory analysis using membrane cDNA arrays. Env Health Perspect, 112:428-38;

   [Abstract]: Microarrays have the potential to significantly impact our ability to identify toxic hazards by the identification of mechanistically relevant markers of toxicity. To be useful for risk assessment, however, microarray data must be challenged to determine reliability and interlaboratory reproducibility. As part of a series of studies conducted by the International Life Sciences Institute Health and Environmental Science Institute Technical Committee on the Application of Genomics to Mechanism-Based Risk Assessment, the biological response in rats to the hepatotoxin clofibrate was investigated. Animals were treated with high (250 mg/kg/day) or low (25 mg/kg/day) doses for 1, 3, or 7 days in two laboratories. Clinical chemistry parameters were measured, livers removed for histopathological assessment, and gene expression analysis was conducted using cDNA arrays. Expression changes in genes involved in fatty acid metabolism (e.g. acyl-CoA oxidase), cell proliferation (e.g. topoisomerase II-a), fatty acid oxidation (e.g. cytochrome P450 4A1), consistent with the mechanism of clofibrate hepatoxicity, were detected. Observed differences in gene expression levels correlated with the level of biological response induced in the two in vivo studies. Generally, there was a high level of concordance between the gene expression profiles generated from pooled and individual RNA samples. Quantitative real-time polymerase chain reaction was used to confirm modulations for a number of peroxisome proliferator marker genes. Though the results indicate some variability in the quantitative nature of the microarray data, this appears due largely to differences in experimental and data analysis procedures used within each laboratory. In summary, this study demonstrates the potential for gene expression profiling to identify toxic hazards by the identification of mechanistically relevant markers of toxicity..

2. Churchill (2002). Fundamentals for experimental design for cDNA microarrays. Nat. Genet. 32:490-5. Review

3. Hamadeh et al (2002). Prediction of compound signature using high density gene expression profiling. Toxicological Sciences, 67:232-40

4. Kerr (2003). Design considerations for efficient and effective microarray studies. Biometrics. 59:822-8. Review

5. Mah et al (2004). A comparison of oligonucleotide and cDNA-based microarray systems. Physiol. Genomics 16:361-70.

   [Concluding paragraph]: The results of this study indicate that oligonucleotide-based arrays, such as those produced by Affymetrix, and full-length clone-based arrays may be too different in experimental design to be expected to give global expression results that can be directly correlated. This suggests that microarray technologies should not be used as an absolute quantitation method and that pooling of global expression profiles from different microarray platforms for the purposes of large-scale data mining should be undertaken with caution. The observation that there is only moderate overlap and no correlation in the expression data warrants the simultaneous use of complementary approaches to obtain a complete expression profile in complex tissue

6. Mattes (2004). Annotation and cross-indexing of array elements on multiple platforms. Env Health Perspect, 112:506-10

7. Thomas et al (2001). Identification of toxicologically predictive gene sets using cDNA microarrays. Molecular Pharmacology, 60:1189-94

   [Abstract]: The authors developed an approach to classify toxicants based upon their influence on profiles of mRNA transcripts. Changes in liver gene expression were examined after exposure of mice to 24 model treatments that fall into five well-studied toxicological categories: peroxisome proliferators, aryl hydrocarbon receptor agonists, non-coplanar polychlorinated biphenyls, inflammatory agents, and hypoxia-inducing agents. Analysis of 1200 transcripts using both a correlation-based approach and a probabilistic approach resulted in a classification accuracy of between 50 and 70%. However, with the use of a forward parameter selection scheme, a diagnostic set of 12 transcripts was identified that provided an estimated 100% predictive accuracy based on leave-one-out cross validation. Expansion of this approach to additional chemicals of regulatory concern could serve as an important screening step in a new era of toxicological testing.

8. Thompson et al (2004). Identification of platform-independent gene expression markers of cisplatin nephrotoxicity. Env Health Perspect, 112:488-94

9. Yang & Speed (2002). Design issues for cDNA microarray experiments. Nat. Rev. Genet. 3. 579-88. Review

Irwin, R.D., et al. (2004). Application of Toxicogenomics to Toxicology: Basic Concepts in the Analysis of Microarray Data. Toxicologic Pathway, Vol 32(Suppl. 1):72-83
Topic covered: This review addresses issues relating to microarray design, analysis and to a lesser extent statistics. The paper provides general guidelines about the design and analysis of microarray data using the liver as an example.
Design: The authors explain the main steps involved in a typical 2 colour microarray experiment. These are fluorescent labelling (which involves labelling treatment and control group mRNA samples with red (cy5) or green (cy3) dyes respectively), hybridisation (the authors note the competitive nature of the process and the variable nature of the type of probe/element (gene) used i.e. cDNA or an oligonucleotide), measuring gene expression (in terms of the significance of increases or decreases in red or green transcripts; scanning the fluorescence signal and calculating the expression ratio) and the evaluation of raw data (via image analysis software to correct for background fluorescence/eliminate poor measurements, and standard transformations such as normalisation and logarithm transformation – see below). Other platforms (aside from DNA microarrays) used in transcriptional profiling are summarised e.g. nylon cDNA arrays (described as an early form of microarray that uses radioactive rather than fluorescent labels although limited by the high rate of false positives) and high density synthetic oligonucleotide arrays (e.g. Affymetrix arrays whereby each gene (element/probe) is represented by a set of oligonucleotide probe pairs to ensure specificity; labelling is achieved via use of an antisense copy of RNA with biotinylated nucleotides; after hybridisation gene expression is measured by treating the chip with streptavidin-labelled with phycoerthyrin dye and scanned; hybridisation intensities are then calculated). These alternative microarray platforms are considered challenging particularly in relation to gene annotation, the location of gene on a chip and programmes generating gene lists. However, it is believed these problems could be resolved by considering the biological plausibility of the results and using quantitative PCR to confirm a subset of genes identified. The authors discuss the significance of biological and technical replicates. Biological replicates refer to the number of animals used per treatment/control group. Technical replicates refer to the number of measurements made with each sample from one animal. Biological replicates allow investigators to determine the extent to which individual animal responses vary between treated and control groups. Technical replicates enable assessment of experimental variation associated with sample handling. This is done via use of the 'fluor-flip' or 'dye reversal' method (swapping dyes used to label treated and control groups). Technical replicates ultimately help reveal bias in the labelling reaction or in the fluorescence yield. The National Center for Toxicology National Institute of Environmental Health Sciences (NCT NIEHS) recommends that 3 animals should be used per group with a dye reversal run for each animal. The authors demonstrate the application of transcriptional profiling via an *in-vivo* rat study that examined gene expression in the liver of animals treated with acetaminophen. The authors consider the liver a useful target organ for gene expression profiling experiments due to the ease of sample removal and preparing high quality RNA from liver tissue. Their first study objective was to exam gene expression patterns in rat liver associated with varying acute acetaminophen exposures and correlate specific toxic phenotypes/ histological changes with signature patterns of gene expression. Male F344 rats were exposed to 1500 and 2000 mg/kg acetaminophen via oral (gavage) (no information was provided re: number of rats used). Histological sections of liver were obtained at 6, 24 and 48 hours after dosing. The study also aimed to extract information about the mechanism of toxicity and categorise the genes involved (see below). To reduce variability and improve interpretation of microarray experiments the authors suggest that toxicogenomics (TGX) study designs should consider zonation of the liver (or kidney) sample selected and whether the toxicant is a zone specific hepatotoxicant (see below). Other recommendations include monitoring the test animal's food and water consumption rate as this can be affected by treatment; any resulting reduced body weight can influence the pattern of gene expression in the liver and confound findings.
Analysis: The authors acknowledge the inability of microarrays to provide strict measurements of the magnitude of change for genes identified as being differentially expressed. This is because microarrays provide only semi-quantitative information about changes in gene expression. The authors suggest that quantitative PCR can help verify the expression levels of a representative sample of identified genes.

The authors describe two necessary standard transformations conducted on raw data to prepare it for statistical and biological analysis. Normalisation adjusts data for possible differences arising from technical aspects of the experiment (i.e. those associated with sensitivity and efficiency). Log transformation is performed to numerically equilibrate similar magnitudes of increases and decreases. This involves converting raw data to logarithm base 2 (log(2)) of the ratio of treated to control for each array element. Image analysis software is also used on raw microarray data to correct for background fluorescence and eliminate poor measurement quality data.

The authors define cluster analysis as a general term used to describe a group of statistical methods for ordering/organising/visualising data into groups or clusters. It is considered particularly useful for grouping genes sharing similar patterns of expression where there are no *a priori* hypotheses about how the data should be grouped (provides an initial unsupervised ordering). Several types of clustering algorithms exist (**Dougherty et al 2002**) and the authors note the use of hierarchical clustering especially in tumour biology.

Two-dimensional hierarchical clustering analysis of gene expression data (generated from the *in-vivo* acetaminophen-exposed rat liver study described earlier) revealed dose-related differences in gene expression pattern (i.e. 2000 vs. 1500mg/kg). Within each dose, gene expression patterns differed with time after exposure (e.g. at the 1500mg/kg dose both up- and downregulated genes occurred 6hrs after dosing compared to the 2000mg/kg dose resulting in a large group of upregulated genes (red) in the middle of the cluster, which were highest at 24 and 48 hrs after dosing). Histological sections taken at different time points suggest these changes may reflect hepatocellular injury (a spectrum of adverse changes occurred 24 hrs after dosing at both dose groups; there was little change 6hrs after dosing). This indicated a possible correlation between the molecular events occurring in the liver and histological observations at corresponding times and doses. The authors noted that there was no systematic way of extracting mechanistic toxicology information aside from directly examining single nodes in the clustergram and inspecting individual genes that have clustered together. Examination of nodes in the 2000 mg/kg clustergram showed downregulation of genes coding for two enzymes of lipid biosynthesis. The authors recognised that commercially available software products are available to assign/categorise genes into metabolic or signal transduction pathways but considered most limited by their inability to provide comprehensive information. The global nature of gene expression is considered a particular challenge for categorising genes. Although gene expression reflects the state of most of the genome all pathways must still be fitted into a biologically meaningful result. From their experience the authors suggest that work should be conducted at the level of the individual differentially expressed genes.

Key issues relating to the analysis of liver-specific gene expression data are also discussed following the unexpected variation in severity of necrosis observed in the *in-vivo* rat liver study. Zonation of hepatic gene expression, the nutritional status of and the mixed cell population in liver are all factors believed to complicate interpretation of differential hepatic gene expression. The authors suggest that studies acknowledge the contribution different liver cell types may have to subsequent analyses of samples. Different cell types may present different targets for toxicants and play different roles in certain pathologies.

Statistics: The authors extol the superiority of statistical over threshold-based approaches for identifying genes whose expression is altered by treatment with the agent under study. Approaches based on relative changes occurring above a threshold (as determined by spot intensities) are considered limited given the arbitrary nature of thresholds and their inability to provide any level of confidence about statistical significance. Statistical approaches provide the most reliable and unbiased way of selecting differentially expressed genes. They enable precise measurement of genes exhibiting even a small fold increase or decrease in expression (to which many important genes fall into this category). Two types of statistical methods are defined: a simple calculation of mean and standard deviation of distribution of log(2) intensity ratios, and selecting differentially expressed genes that fall outside the 95% confidence interval; Analysis of variance (ANOVA) (used to determine the statistical significance of increases/decreases in gene expression) is thought to provide a solid statistical basis for identifying differentially expressed genes (based on *p*-values).

Comments: Genomics and transcriptional profiling are the main focus of the review. Concepts are discussed from a non-specialist perspective. Other areas discussed (not highlighted above) include a brief account of the origins of transcriptional profiling and a comparison of the two main transcriptional profiling methods used i.e. Reverse Transcriptase Polymerase

Chain Reaction and microarray. The review concludes with the statement that transcriptional profiling using microarrays is only a tool and so will only provide useful information when properly applied.

Refs

1. Dougherty et al 2002. Inference from clustering with application to gene expression microarrays. Nat Genet 21, 10-14.

KM Lee et al. (2005). Design issues in toxicogenomics using DNA microarray experiment. *Toxicology and Applied Pharmacology*: 207 S200-08
Topic covered: This review focuses on design of microarray studies. Data analysis and use of statistics are only briefly mentioned. Background and definitions for key topic areas are provided when possible.
Design: The authors comment on the importance of addressing design issues to ensure toxicogenomic (TGX) data is correctly interpreted, which should aid greater use of TGX and assert its value in toxicology. The authors identify the following five areas as significant to design: (a) Experimental objectives: Poorly designed studies are particularly challenging and researchers are advised to state their objectives in advance in terms of what they hope to get from their microarray study; (b) Selection of genes for microarray: The authors describe the selection criteria used which is based on whether or not the toxicants mechanism of action is known. The authors also explain how selected genes are categorised on the basis of their biochemical/pathological roles i.e. xenobiotic metabolism, DNA repair, etc. The resources used to categorise these genes i.e.GeneCards/Weizmann Institute; Kyoto encyclopedia of genes and genomes (KEGG) are also noted. (c) Selection of microarray platform: The authors describe the procedures used to conduct a cDNA microarray. This involves obtaining the relevant tissues, isolating RNA/mRNA, producing labelled cDNA probes, hybridising probes to arrays, analysing the data by measuring signal intensity and determining the ratio of signals between samples. The different types of DNA microarray platforms available are also considered i.e. cDNA, spotted oligonucleotide and Affymetrix arrays (**TABPWG, 2004**) – including the advantages and disadvantages of using each; (d) Design of DNA microarray: The authors list possible sources of variation in microarray experiments (i.e. animals/subjects, tissue/mRNA extraction, cDNA preparation and labelling, hybridisation, washing, reading, DNA spot, between array variation, nuisance variables and matrix quality). The Minimum Information About A Microarray Experiment (MIAME) guidelines and its significance are also introduced in relation to information on experimental design and protocol. This section is further subdivided into the following five areas: (i) Experimental design – which identifies the following as particular design challenges: cost of labelling and hybridising mRNA; the amount of RNA available and the number of slides to use; deciding whether to use a direct, reference or loop design (**Kerr and Churchill, 2001**) and incorporating dose response and time course parameters into the experiment. (ii) Species and sample types – which advises Investigators to select the most appropriate species and samples to maximise the likelihood of true positives and minimise false negatives (**Ezendam et al 2004**). (iii) Replicates – which describes the significance of using replicates and the numbers and different types used (**Lee et al 2000**). States that the most common approach for using replicates involves putting replicates of the same spot (cDNA probe) on each slide. Replicates put between slides are categorised as either technical or biological. Defines technical replicates as target mRNA taken from the same extraction or pool, which produces less variation in measurements. Defines biological replicates as target mRNA from different extractions (e.g. different samples of cells from a particular cell line or tissue), which is often referred to as the sample size. Dye sway (dye-flip) replications, (that involves reversing the dye assignment in one of two hybridisations using two mRNA samples from the same extraction) is considered to reduce systematic bias. (iv) Sample sizes – which comments on the confusion caused by the various definitions of 'sample size' in relation to a microarray experiment; also notes the complexity of the methods used to calculate sample size (**Wei et al 2004**), to which at least four components exist/must be considered: 1) the variance of individual measurements, 2) the magnitude of the effect to be detected, 3) the acceptable false positive rate, and 4) the desired power (i.e. probability of detecting an effect of the specified (or greater) magnitude. The authors note that large false positive rates are a possible consequence of multiple tests and suggests adjustment (according to the study's objectives) via use of Bonferroni correction. Practically up to 10 inbred mice are required per treatment group and for human samples a large number of individuals are needed per exposure group (**Lampe et al 2004**), which can be very expensive. The authors recommend pooling tissue samples from individuals in the same treatment group. (v) Data analysis and interpretation – the authors define normalisation as the process of removing systematic

variation in microarray data. They describe the approaches used, how the resulting gene expression matrices are analysed, and the methods used to determine gene regulation and function (i.e. via bioinformatic approaches such as clustering, classification and pattern discovery); (e) Design issues in epidemiological studies: This is further subdivided into two areas: (i) Bias and confounding – which describes the causes of selection and information bias and confounding (and how it is controlled). (ii) Sample size – which comments on the appropriate sample sizes required for population studies investigating the effect of genetic variation in specific diseases.
Analysis: See design section above, part (d), (v)
Statistics: Issues relating to statistics are not specified (exception: wrt use of Bonferroni correction in relation to adjusting large false positive rates – see design section)
Comments: There is a strong descriptive element to the review. The paper focuses on explaining the basic issues than suggesting possible remedies/ways forward. The authors conclude that there is a lack of data on baseline gene expression in human samples, and consider the detection of environmental chemical-induced TGX expression changes comprises a significant challenge. This they feel is due to the wide variability of baseline gene expression among individuals. Subsequently, this may make it impossible to detect changes due to environmental chemical exposure.

Refs

1. Ezendam, J et al (2004). Toxicogenomics of subchronic hexachlorabenzene exposure in Brown Norway rats. Environ. Health Perspect. 112(7):782-91

2. Kerr, MK & Churchill, GA. (2001). Statistical design and the analysis of gene expression microarrays. Genet. Res. 77, 123-8. Review

3. Lampe, JW et al (2004). Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. Cancer Epidemiol., Biomarkers Prev. (13(3):445-3

   [Abstract]: Functional biological markers of environmental exposures are important in epidemiological studies of diseases risk. Such markers not only provide a measure of the exposure, they also reflect the degree of physiological and biochemical response to the exposure. In an observational study, using DNA microarrays, the authors report that it is possible to distinguish between 85 individuals exposed and unexposed to tobacco smoke on the basis of mRNA expression in peripheral leukocytes. Furthermore, the authors report that active exposure to tobacco smoke is associated with a biologically relevant mRNA expression signature. The authors conclude that these findings suggest that expression patterns can be used to identify a complex environmental exposure in humans.

4. Lee, MT et al (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc. Natl. Acad. Sci. USA 97: 9834-9

   [Abstract]: The authors present statistical methods for analysing replicated cDNA microarray expression data and report the results of a controlled experiment. The study was conducted to investigate inherent variability in gene expression data and the extent to which replication in an experiment produces more consistent and reliable findings. The authors introduced a statistical model to describe the probability that mRNA is contained in the target sample tissue, converted to probe, and ultimately detected on the slide. The authors also introduce a method to analyse the combined data from all replicates. Of the 288 genes considered in this controlled experiment, 32 would be expected to produce strong hybridisation signals because of the known presence of repetitive sequences within them. The authors report that the results based on individual replicates, however, show that there are 55, 36, and 58 highly expressed genes in replicates 1, 2, and 3

respectively. On the other hand, an analysis by using the combined data from all 3 replicates reveal that only 2 of the 288 genes are incorrectly classified as expressed. The authors consider that the experiment shows that any single microarray output is subject to substantial variability and pooling data from replicates makes it possible to provide a more reliable analysis of gene expression data. Therefore, the authors conclude that designing experiments with replications will greatly reduce misclassification rates. The authors recommend that at least three replicates be used in designing experiments by using cDNA microarrays, particularly when gene expression data from single specimens are being analysed.

5. (TABPWG, 2004). The Tumor Analysis Best Practices Working Group, 2004. Expression profiling – best practices for data generation and interpretation in clinical trials. Nature Reviews Genetics. 5:229-237.

6. Wei, et al (2004). Sample size for detecting differentially expressed genes in microarray experiments. BMC Genomics 5 (1). 87

   [Abstract]: The authors conclude that factors affecting power and sample size calculations include variability of the population, the desired detectable differences, the power to detect differences, and an acceptable error rate. In addition, experimental design, technical variability and data pre-processing play a role in the power of the statistical tests in microarrays. The authors show that the number of samples required for detecting a 2-fold change with 90% probability and a p-value of 0.01 in humans is much larger than the number of samples commonly used in present day studies, and that far fewer individuals are needed for the same statistical power when using inbred animals rather than unrelated human subjects.

7. Yang YH & Speed T (2002). Design issues for cDNA microarray experiments. Nat. Rev. Genet. 3: 579-88

Yauk & Berndt (2007). Review of the Literature Examining the Correlation Among DNA Microarray Technologies. Environmental & Molecular Mutagenesis 48:380-94
Topic covered: This review summarises cross platform studies and discusses possible reasons for the discrepancies reported in earlier comparative studies and the subsequent methodological changes that led to improved correlations. Issues relating to design, reproducibility and correlation of toxicogenomic (TGX) microarray studies are addressed. Data analyses topics are only briefly covered. This paper does not review issues relating to the statistical analyses of TGX data.
Design: The authors consider that Affymetrix chips restrict users to Affymetrix-based technologies (i.e. from the choice of scanner to image analysis). The construction of Affymetrix chips (via oligonucleotide synthesis and photolithography) is described whereby specific oligonucleotide probes are positioned onto the array in a predetermined spatial orientation. Single genes are represented by a series of probes that span the coding region. Each probe is paired with a mismatch probe in which the central base in the sequence is changed. This renders adherence to manufacturer's recommendations necessary. Spotted microarrays using glass microscopic slides are also considered. The authors suggest that the range of choices available for these arrays contribute to the variation in data acquired and they briefly discuss sources of technical variation for each stage of a microarray experiment. Variation linked to probes arises due to the different types of probes available, and different methods for probe printing and deposition onto glass slides. Variation in target preparation arises due to the variable amounts of starting RNA that can be used and the different amplification and labelling methods that exist. The different designs available in two colour labelling procedures and the fact hybridisation can either be manual or automated also contributes to variation. The scanner power used also varies, with different settings that can be adjusted to maximise the linear dynamic range. The authors cite a review by **Ahmed (2006a)** which provides comprehensive information relating to design issues. The authors separate the studies reviewed into two time periods: those conducted pre-2004 (between 2000-2003) and post-2004 (between 2004-2007). Comparative studies conducted pre-2004 were considered to be of little value as they compared only 2-3 technologies, used small sample sizes and focussed on cDNA microarrays. The authors also report that issues related to the platform, protocol and the type of experimental design used contributed to the discrepancies in these early cross platform comparison studies. For example, many cDNA platforms were contaminated or probes were incorrectly annotated. Probes on Affymetrix platforms were also subject to annotation errors (e.g. lack of correspondence to appropriate mRNA reference sequence (which impacts on signal intensity)). There were also inaccurate probe set definitions, probes hybridising to multiple splice variants or showing cross hybridisation to other genes in the same family and probes hybridising to non-specific probes. Furthermore, early studies (which matched genes based on the manufacturers annotation) largely examined incorrectly matched gene sets. Probe annotation remains a concern primarily because it continues to be an issue. The authors suggest the way forward would be to improve annotation via curation, validation, and annotation of more sequence information, and less reliance on manufacturer's gene identification. Probe sequence information is now widely available and Minimum Information about Microarray Experiment (MIAME) guidelines require submission of probe sequences for each spot on a microarray. Users can also cross check probe sequence annotation to validate expression changes. Other platform issues noted include suboptimal printing, labelling, hybridising and washing methods; the lack of technical expertise in one of the platforms being compared (leading to poor quality data); environmental influences such as ozone which affects fluorescent chemicals (although this can be controlled for); and improvements in printing quality of cDNA and oligonucleotide microarrays. Design issues thought to contribute to the discrepancies of early studies include the fact that data were generated in different labs, at different times using different samples. Small sample sizes were often used and studies failed to use both biological and technical replicates. The authors recommend that the same RNA sample should be used for all experiments. The authors consider that addressing the above issues will decrease technical variability and increase performance and thus improve the subsequent correlation among technologies. Studies conducted post-2004 improved with the use of larger sample sizes and more microarray platforms were included. Studies also examined the relationships among laboratories and employed sophisticated bioinformatics approaches. Annotation-driven and

sequence driven matching are highlighted as the two different approaches used in post 2004 comparative studies to analyse data from different platforms. However, annotation driven approaches are limited by the effect that any errors can have on subsequent analyses. In comparison, sequence-driven approaches are reportedly able to eliminate errors introduced by mis-annotation and also ensure probe pairs examine similar gene regions i.e. within the same axon. Re-examination using sequence-driven probe matching is therefore considered a worthwhile approach and is reported to improve correlation among technologies (**Kuo et al 2002**).

Analysis: The authors note that the final critical steps of a microarray experiment are filtering, data quality assessment, normalisation and data analysis. Filtering is defined as an important pre-processing step designed to remove unreliable data from experiments prior to analysis. It eliminates noise and cleans signals within the background range resulting in stronger signals. Indeed comparative studies conducted post-2004 reported greater correlation for probes with strong expression signals (Kuo et al 2006). Commercial image acquisition programmes are used to rid image data of poor quality, saturated and low signal spots. The authors state that failure to conduct appropriate/stringent filtering methods would result in an inaccurate representation of gene expression, a flaw of most early cross platform studies. NB. Using relative ratios of gene expression rather than signal intensity constituted another flaw of these early studies.

The authors consider that the different filtering methods available are a possible source of technical variation. Sources of technical variation arising in data analysis steps are briefly discussed and the authors note that variation can arise in data acquisition from images due to the different algorithms available from different commercial packages. The authors cite a review by **Ahmed (2006b)** which provides comprehensive information relating to data analytical issues.

The authors consider that studies should apply appropriate 'statistical' tools such as clustering and normalisation (applied both within and between technologies). Furthermore, to identify differentially expressed genes, studies should use correct tools such as fold change ranking as opposed to other methods which may not result in reproducible and thus comparable gene lists. However, further research is needed to develop more accurate and reproducible methods for deriving lists of differentially expressed genes from different technologies. The authors suggest that studies should not compare the absolute magnitude of gene expression changes across platforms as microarrays are not precise or accurate wrt quantifying gene expression changes. Rather, approaches should focus on the direction of change i.e. whether expression has increased or decreased. Tissue heterogeneity and biological variation are also considered as sources of variation between datasets.

Statistics: This review does not consider issues relating to statistical analyses of TGX data.

Comments: The authors note that standards for microarray experiments developed as a consequence of the range of microarray technologies and methods of data analysis, which raised concern over the impact different approaches would have for data comparability (**Kawasaki et al 2006**). Various projects embarked on developing standards which include the MIAME guidelines, External RNA Controls Consortium (ERCC) and Microarray Quality Control (MAQC). The authors note how these standards have helped improve the evaluation of microarray data quality and reproducibility of results obtained by different labs and/or platforms. Furthermore, adherence to established standards are a requirement of microarray databases/repositories, as well as proven reproducibility and correlation between (and within) datasets produced by different microarray technologies.

The authors consider that research exploring correlation and reproducibility among microarrays helps validate micoarrays as robust, sensitive and accurate detectors of gene expression. They report that 40 studies have investigated the subject to determine the extent to which data produced by different microarray technologies correlate. Reviews examining factors influencing accuracy and reproducibility across time, laboratories and platforms are also highlighted (**Reimers 2005; Brietling, 2006**), which essentially flagged up the importance of procedures such as normalisation and the detection of differential gene expression.

The authors noted the different rationale used to conduct comparative studies i.e. to determine the best platform to use (which is dependant on the type of experiment being conducted); which platforms generate comparable/reproducible data; how commercially made and in-house microarrays differ in terms of their accuracy (proximity to true value), sensitivity (ability to detect changes at low concentrations) and specificity (ability to hybridise to the

correct gene) (**Draghici et al 2006**).
The authors question the validity of some early studies conducted between 2000-2003 which produced results supporting the reproducibility and concordance of data across microarray technologies. However, these studies are thought to have helped identify potential sources of discrepancies between microarray datasets. Given the notable discrepancies in the published literature, the authors decided to perform their own cross platform study to evaluate gene expression in the following way: using replicates of three different RNA sources (mouse whole lung, lung cell line, reference RNA –Strategene Universal); using technologies encompassing different reporter systems/probes (short/long oligonucleotides, cDNA) with different labelling techniques and hybridisation protocols; applying rigorous filtering and normalisations; and using an adequate sample size. Their findings showed that top performing platforms had an increased ability to detect differential expression due to low levels of technical variability. Biological rather than technological differences are thought to account for the most of the variation in the data.
The authors noted that 32 studies examined correlation among microarray technologies post 2004. Three of these studies concluded poor correlation between microarray platforms (**Mah et al 2004; Severgnini et al 2006; Gwinn et al 2005**). Potential study author errors are thought to account for these conclusions as well as the use of expression intensities as opposed to examining log ratios (**Park et al 2004**). The remaining 29 studies generated results showing moderate to high degree of correlation among microarray technologies. The approaches used varied with the most comprehensive employing many platforms, one or two colours, different probes spotted both inhouse and commercially, and using data from the same samples analysed via different labs. These studies helped identify methods that produce high correlation among labs and platforms.
Generating reproducing data requires optimisation and standardisation of protocols which is achieved by performing intra-platform reproducibility tests prior to inter-platform reproducibility tests. The best performing laboratories were noted for their use of optimised protocols and technical expertise (which would occur in labs that routinely use a technology), and increased standardisation (i.e. using/developing commercially available microarrays rather than in house microarrays). The authors consider that more relatable data will come with improved methods, and developments in quality control standards and references, and implementation of standards for data analysis.
Various examples of comprehensive studies evaluating microarray performance are provided (**Irizarry et al 2005**; Kuo et al 2006; **Wang et al 2006**). The Toxicogenomic Research Consortium (TRC) examined data produced by seven laboratories and 12 microarray platforms. Each laboratory was supplied with two samples of RNA (taken from the liver and other tissues). Although poor correlation across platforms and laboratories (and between raw intensity values) were reported, reproducibility increased after implementing standardised protocols for RNA labelling, hybridisation, filtering, processing, data acquisition and normalisation. The highest levels of reproducibility were achieved in laboratories using commercial arrays and applying standard protocols (correlation coefficients ranged from 0.87-0.92). The study concluded that microarray platform contributes significantly to variability in data and standardisation is necessary for achieving reproducible data across laboratories. Furthermore, high reproducibility among platforms were achieved when analyses were conducted on biological categories identified by gene ontology analysis. Another comprehensive comparative study - Microarray Quality Control (MAQC) - led by the US FDA and involving 137 participants from 51 organisations, evaluated inter- and intra-platform reproducibility via a series of scientific papers (e.g. **Shi et al 2006** and **Guo et al 2006**). The findings supported inter-platform consistency and reproducibility and the use of microarray platforms for quantitative characterisation of gene expression.

Refs:

1. Ahmed 2006a. Microarray RNA transcriptional profiling. I. Platforms, experimental design and standardization. Expert Rev Mol Diagn. 6:535-50 (Review)

   Read/consider. Relates to its provision of comprehensive design issue info

2. Ahmed 2006b. Microarray RNA transcriptional profiling. I. Analytical considerations and annotation. Expert Rev Mol Diagn. 6:703-15. (Review)

Read/consider. Relates to its provision of comprehensive data analysis issue info

3. Brietling 2006. Biological microarray interpretation: The rules of engagement. Biochem Biophys Acta 1759:319-27. (Review)

4. Draghici et al 2006. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet. 22:101-109. (Review)

5. Guo et al 2006. Rat toxicogenomic study reveals analytical consistency across microarray platforms. Nat Biotechnol. 24:1162-1169.

   [Abstract]: To validate and extend the findings of the MicroArray Quality Control (MAQC) project, the authors generated a biologically relevant toxicogenomics data set using 36 RNA samples from rats treated with three chemicals (aristolochic acid, riddelliine and comfrey) and each sample was hybridised to four microarray platforms. The MAQC project assessed concordance in intersite and cross-platform comparisons and the impact of gene selection methods on the reproducibility of profiling data in terms of differentially expressed genes using distinct reference RNA samples. The authors consider that the real-world toxicogenomic data set reported here showed high concordance in intersite and cross-platform comparisons. Furthermore, the authors consider that the gene lists generated by fold-change ranking were more reproducible than those obtained by t-test P value or Significance Analysis of Microarrays. Finally, the authors report that the gene lists generated by fold-change ranking with a nonstringent P-value cutoff showed increased consistency in Gene Ontology terms and pathways, and thereby conclude that the biological impact of chemical exposure could be reliably deduced from all platforms analysed.

6. Gwinn et al 2005. Transcriptional signatures of normal human mammary epithelial cells in response to benzo[a]pyrene exposure: A comparison of three microarray platforms. Omics 9: 334-50.

   [Abstract]: Microarrays are used to study gene expression in a variety of biological systems. A number of different platforms have been developed, but few studies exist that have directly compared the performance of one platform with another. The goal of this study was to determine array variation by analysing the same RNA samples with three different array platforms. Using gene expression responses to benzo[a]pyrene exposure in normal human mammary epithelial cells (NHMECs), the authors compared the results of gene expression profiling using three microarray platforms: pholithographic oligonucleotide arrays (Affymetrix), spotted oligonucleotide arrays (Amersham) and spotted cDNA arrays (NCI). While most previous reports comparing micorarrays have analysed pre-existing data from different platforms, this comparison study used the same sample assayed on all three platforms, allowing for analysis of variation from each array platform. The authors report that in general, poor correlation was found with corresponding measurements from each platform. Each platform yielded different gene expression profiles, which lead the authors to suggest that while microarray analysis is a useful discovery tool, further validation is needed to extrapolate results for broad use of the data. The authors also consider that microarray variability needs to be taken into consideration, not only in the data analysis but also in specific probe selection for each array type.

7. Irizarry et al 2005. Multiple laboratory comparison of microarray platform. Nat Methods 2:345-50.

   [Abstract]: Microarray technology is a powerful tool for measuring RNA expression for thousands of genes at once. Various studies have been published comparing competing platforms with mixed results: some find agreement, others

do not. As the number of researchers starting to use microarrays and the number of cross-platform meta-analysis studies rapidly increases, appropriate platform assessments become more important. The authors present results from a comparison study that they consider offers important improvements over those previously described in the literature. In particular, the authors report that none of the previously published papers consider differences between labs. For this study, a consortium of ten laboratories from the Washington, DC-Baltimore, USA area was formed to compare data obtained from three widely used platforms using identical RNA samples. The authors report use of appropriate statistical analysis to demonstrate that there are relatively large differences in data obtained in labs using the same platform, but that the results from the best-performing labs agree rather well.

8. Kawasaki et al 2006. The end of the microarray tower of babel: Will universal standards lead the way? J Biomol Tech 17:200-06. Review.

9. Kuo et al 2002. Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 18: 405-12.

10. Park et al 2004. Current issues for DNA microarrays: Platform comparison, double linear amplification, and universal RNA reference. J Biotechnol 112:225-45.

    [Abstract]: DNA microarray technology has been widely used to simultaneously determine the expression levels of thousands of genes. A variety of approaches have been used, both in the implementation of this technology and in the analysis of the large amount of expression data. However, several practical issues still have not been resolved in a satisfactory manner, and among the most critical is the lack of agreement in the results obtained in different array platforms. In this study, the authors present a comparison of several microarray platforms [Affymetrix oligonucleotide arrays, custom complementary DNA (cDNA) arrays, and custom oligo arrays printed with oligonucleotides from three different sources] as well as analysis of various methods used for microarray target preparation and the reference design. The authors report that the results indicate that the pairwise correlations of expression levels between platforms are relatively low overall but that the log ratios of the highly expressed genes are strongly correlated, especially between Affymetrix and cDNA arrays. The microarray measurements were compared with quantitative real-time-polymerase chain reaction (QRT-PCR) results for 23 genes, and the varying degrees of agreement for each platform were characterised. The authors also developed and tested a double amplification method which reportedly allows the use of smaller amounts of starting material. The authors note that the added round of amplification produced reproducible results as compared to the arrays hybridised with single round amplified targets. Finally, the authors tested the reliability of using a universal RNA reference for two-channel microarrays and report that the results suggest that comparisons of multiple experimental conditions using the same control can be accurate.

11. Reimers 2005. Statistical analysis of microarray data. Addict Biol 10: 23-35. (Review)

12. Severgnini et al 2006. Strategies for comparing gene expression profiles from different microarray platforms: Application to a case control experiment. Anal Biochem. 353: 43-56.

    [Abstract]. Meta-analysis of microarray data is increasingly important, considering both the availability of multiple platforms using disparate technologies and the accumulation in public repositories of data sets from different laboratories. The authors addressed the issue of comparing gene expression profiles from two microarray platforms by devising a standardised investigative strategy. The authors tested the procedure by studying MDA-MB-231 cells, which undergo apoptosis on treatment with resveratrol. Gene expression profiles were obtained

using high density, short oligonucleotide, single-colour microarray platforms: GeneChip (Affymetrix) and CodeLink (Amersham). Interplatform analyses were carried out on 8414 common transcripts represented on both platforms, as identified by LocusLink ID, representing 70.8% and 88.6% of annotated GeneChip and CodeLink features, respectively. The authors identified 105 differentially expressed genes (DEGs) on CodeLink and 42 DEGs on GeneChip. Among them, only 9 DEGs were commonly identified by both platforms. Multiple analyses (BLAST alignment of probes with target sequences, gene ontology, literature mining and quantitative real-time PCR) permitted the authors to investigate the factors contributing to the generation of platform-dependent results in single colour microarray experiments. The authors conclude that an effective approach to cross-platform comparison involves microarrays of similar technologies, samples prepared by identical methods, and a standardised battery of bioinformatic and statistical analyses.

13. Shi et al 2006. The MicroArray Quality Control (MAQC) project shows inter and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 24:1151-1161.

14. Wang et al 2006. Large scale real-time PCR validation on gene expression measurements from two commercial long-oligonucleotide microarrays. BMC Genomics. 7:59.

Ju, z., et al. (2007). DNA microarray technology in toxicogenomics of aquatic models: Methods and applications. Comparative biology and physiology, Part c 145: 5-14
Topic covered: This review describes the use of gene expression profiling in relation to its application to aquatic model research. Although the paper leans toward the application of microarray technology to aquatic toxicogenomics (ecotoxicogenomics) and environmental safety assessment, the authors provide a general discussion of the fundamental principles of microarray research in relation to study design, data and statistical analysis.
Design: The authors describe the design aspects of microarray construction and the two different types of microarray platforms typically used: Affymetrix GeneChip and customised (spotted) cDNA microarrays. Affymetrix chips are commercially produced and use short oligonucleotide probes (approx 25-mers) that are directly synthesised onto a silicon chip. Many different types of species ranging from prokaryotes to humans have been tested on Affymetrix chips including the zebrafish (*Danio rerio*). Although Affymetrix chips are ready-made and cover a wide number of genes, their use is limited by their high cost. Furthermore, probe synthesis for zebrafish genechip is considered uneconomical as the species is not commonly used. In comparison, spotted cDNA microarrays are considered to be more cost-effective and have less background as they use longer probes which give stronger signals that enable more stringent washing conditions. Spotted arrays use glass slides or nylon membranes and two different types of probes/subplatforms i.e. cDNA fragments and synthetic oligonucleotides. cDNA fragments are amplified via PCR, therefore, their arrays are also referred to as amplicon arrays. Synthetic oligonucleotides are much longer (50-70 mer) than cDNA fragments and are referred to as long oligonucleotide microarray (LOM). Their design is based on expressed sequence tags (EST). The density of spots in spotted arrays range from low (i.e. 100s of transcripts) to mid-high (1000s of transcripts). The two subplatforms are noted for their differences although LOMs are considered more advantageous as they produce expression data that correlates better with quantitative real time PCR (QRT-PCR) (cf. full length amplicon arrays). Furthermore, LOM data is thought to be more concordant with that produced by Affymetrix Genechips. Spotted arrays are limited by the significant set up time required, for example, amplicons must be produced, followed by design and purchase of oligonucleotides from vendors, and quality control of slides, membrane printing, etc. Also, if the cDNA sequences are not readily available then the cDNA must be constructed, and the EST sequenced and annotated. The authors note that several aquatic species have been tested using spotted arrays including zebrafish (**van der Ven et al 2005**). The construction of LOM is described. EST sequences are randomly selected from cDNA libraries and collected. Unigene sequences are then determined allowing probes to be designed in a batch manner via open source or commercial software (Array Designer). The authors note that synthesis of LOM requires consideration of the following quality assurance parameters, which impact on both reproducibility and reliability of fabricated arrays: (a) oligonucleotide length – which affects probe uniqueness and thus the possibility of cross hybridisation, (where the longer the probe the greater the specificity); (b) location of probes within each mRNA sequence – which affects downstream signal intensity (probes derived from 3' ends of mRNA are considered better than those derived from 5' ends (**Brentani et al 2005**)); (c) simple repeats – which affects signal intensity (a maximum of 6 bases of repeats should be designed); (d) cross homology determination to evaluate the potential of cross hybridisation – this can be done by aligning probe sequences with EST databases and conducting BLAST search of each probe sequence against DNA sequences of target aquatic and other species. Design of microarray slides requires high quality intact RNA. The authors note that this can be obtained by using validated sample handling and RNA extraction procedures. Two types of target preparation protocols are described. The standard method uses more RNA (10ug total RNA), while the amplification method starts with smaller amounts of RNA (i.e. 0.01-2ug total RNA) and is typically used when RNA isolation is insufficient for standard array protocol (i.e. when isolating RNA from small samples such as fish organs). The amplification method adopts two types of in-vitro transcription methods: single rounds (IVT) – where RNA is used to produce double stranded cDNA and eventually cRNA; and double rounds (dIVT) – where the cRNA is used to initiate another round of cRNA synthesis. cRNA is then labelled to hybridise with the array. The authors note that RNA can be amplified by about 250 times via the IVT protocol. However, the amplification method is limited by its reduced sensitivity and introduction of minor biases (**Schindler et al 2005**). Two types of dye labelling are

highlighted: direct labelling, which arises when the dye (Cy) is anchored directly onto the nucleotide; and indirect labelling that involves incorporating an aminoallyl labelled nucleotide into the target during reverse transcription (which then is coupled with the dye). Indirect labelling is considered to yield more reproducible data (**Yu et al 2002**).

Analysis: The authors define image analysis as a necessary procedure that considers various factors such as, irregularities of spot position and shape, quantitative quality control, signal variability, grid segmentation and image reconstruction. Open source software e.g. Automated Microarray Image Analysis (AMIA) is available for researchers.
The authors consider the following non-biological sources of variation to be the most significant challenges of microarray data analysis: selective incorporation of Cy dyes; variable amounts of mRNA used; differences in scanning parameters; stochastic variation occurring across replicate slides; hybridisation conditions and human error. Normalisation is always performed before statistical hypothesis tests are conducted. It is considered an important component of data analysis as it removes or minimises the influence of non-biological effects, which thus makes it easier to detect biological differences. However, it is limited by the fact that the final results are influenced by the type of algorithm used. The authors suggest different analytical methods should be used with the same data sets to determine which best suit the experimental design. Normalisation methods include trimmed mean and global mean, local mean, Bayesian analysis and locally weighted scatter plot smoothing (LOWESS). LOWESS is noted for its ability to normalise intensity dependent dye bias arising in experiments that use two-colour microarray platforms. Intensity dependent dye bias occurs when fluorescent dye chemicals (Cy3 and Cy5) emit unequal light resulting in low correlation of signals between Cy dyes.
The authors consider cluster analysis as a statistical tool i.e. a more sophisticated statistical analytical method. Cluster analysis provides a way of grouping objects that are similar and is therefore an ideal data exploration method to look for patterns or structure in data of interest. Cluster analysis is conducted after data normalisation and hypothesis testing (see below). It is used to extract gene expression patterns and define relationships between gene expression profiles across different experiments and data points. The authors also note its ability to depict co-regulated clusters of genes. Two types of clustering mechanisms are described: those for clustering genes or samples. Gene clustering approaches identify gene expression patterns across multiple timepoints or tissues. This allows similar gene expression patterns to be established, and suggestions made on genes with similar responses or gene sharing regulatory circuits. Sample clustering approaches are conducted to obtain gene expression profiles which when clustered themselves can indicate samples that have a biological relationship. The authors briefly highlight the three methods used in clustering analysis i.e. hierarchical clustering, self organising maps and principal component analysis (PCA). Three essential steps for cluster analysis are described: (a) Eulicidean distance provides a measure of the similarity between genes or samples; (b) average linkage distance measures dissimilarity between clusters; (c) selection of clustering method type e.g. hierarchical clustering tree or self organising maps. Correlation tests are also performed to compare gene expression patterns in the same or other experiments/conditions.
Intrinsic sources of variability in gene expression levels such as physiological stages, sex, age, natural genetic polymorphisms in populations are considered as key challenges in microarray research. The authors suggest that in order to generate reliable biological conclusions experimental individuals should be carefully selected and straightforward normalisation algorithms used. Other challenges include the cost and time consumption associated with developing the microarray and analysing the data.
The authors provide a schematic representation of microarray data analysis as follows: (1). Scan Genechip or array slides; (2). Save the image and text data in a local or public database; (3). Discard poor quality images and clean data to filter out extremes (by flowing files through image evaluation software); (4). Conduct further analyses of qualified data e.g: normalisation/statistical analyses (to compare classes and identify differentially expressed genes – these are validated either statistically or via biological procedure); bioinformatical data mining (to identify potential biomarkers, gene signature and biological pathways - via annotation, gene ontology grouping and pathway analysis (using public databases and software tools)). Further validation/investigation will help establish significance in clinical practice and safety assessment.

Statistics: The authors state that statistical analysis/hypothesis testing aims to identify significant differences in the gene expression under different conditions. However, it is noted

that there is no single statistical tool or method capable of adequately meeting all the needs of microarray researchers. As with normalisation methods, the outcome of these statistical tests can be dependent on the algorithm used. Therefore, different analytical methods should be used with the same data sets to determine which best suit the experimental design. Three standard methods are described: Analysis of Variance (ANOVA) is used to determine significance effects of both biological and non-biological variation and can distinguish signal from noise; T-test detects the significance of biological effects and since it is a parametric test is used to determine p-values when the assumption of normality holds true (if it does not hold true then a permutation t-test is used); significant analysis of microarray (SAM) is a non-parametric test that detects significance of biological effects. The authors consider p-values, fold changes and gene expression patterns as necessary statistics that help make sense of data in terms of their biological relevance.
Comments: The authors note that fundamental questions still exist in microarray research, particularly in relation to environmental gene regulation, cell-specific gene expression level differences, and gene function. Only by high throughput assessment of genes and proteins can these questions be addressed. Traditional approaches which are necessary to validate interesting gene regulatory circuits are limited by their inability to sufficiently reveal functional genomics of intact organisms under various experimental conditions. Use of high throughput technologies such as DNA microarrays is considered a way forward. An essential requirement for data management is the database/repository that is used to store array files and track information related to genes and experiments. Two different types exist: local (stores specific data based either on species, genus, topic, etc); and public (examples include open source Gene Expression Omnibus (GEO) and ArrayExpress). The authors highlight the benefits of bioinformatics in terms of its ability to predict possible gene function, hallmark potential gene interactions, identify biomarkers and targets and elucidate molecular networks and pathways. Bioinformatical data mining aims to reveal further biological meaning of microarray data and pinpoints the best gene candidate to focus on thereby providing time and cost savings.

Refs:

1. Brentani et al 2005. Gene expression arrays in cancer research: methods and applications. Crit. Rev. Oncol. Hematol. 54:95-105. (Review).

2. Schindler et al 2005. cRNA target preparation for microarrays: comparison of gene expression profiles generated with different amplification procedures. Anal. Biochem. 344: 92-101.

3. van der Ven et al 2005. Development and application of a brain-specific cDNA microarray for effect evaluation of neuro-active pharmaceuticals in zebrafish (*Danio rerio*). Comp. Biochem. Physiol. B Biochem. Mol. Biol. 141: 408-17.

4. Yu et al 2002. Evaluation and optimization of procedures for target labelling and hybridization of cDNA microarrays. Mol. Vis. 8:130-7

Thompson & Hackett (2008). Quality Control of Microarray Assays for Toxicogenomic and In Vitro Diagnostic Applications. From: Methods in Molecular Biology, vol. 460: Essential Concepts in Toxicogenomics. Edited by: D.L. Mendrick and W.B. Mattes © Humana Press, Totowa, NJ.
Topic covered: This review focuses on quality control (QC) measures for samples generated from rats. The authors provide a detailed discussion of design-related QC issues and briefly consider QC issues related to the analysis of raw toxicogenomic (TGX) data. There is no discussion in relation to the statistical analysis of TGX data.
Design: The role of QC in TGX is described. QC ensures results are reproducible and accurate. It is essential that QC operates via standards, metrics and methods to produce high quality reliable data to aid open access to TGX knowledge base. The authors schematically illustrate the application of QC at each of the following steps involved in generating TGX samples i.e. study design, animal tissue handling, RNA isolation, microarray sample processing, sample hybridisation, microarray scanning and data analysis. It is noted that since 2003 microarray design has evolved and led to an increased and improved annotation of the rat genome and optimisation in methods. The authors suggest that experiments are designed to avoid introducing bias in data due to non-randomised processing of treatment groups. However, most studies vary in their study protocol (e.g. diet, vehicle type, vehicle route of administration, dosing frequency, method of sacrifice, use of anaesthetics during study), and the authors note that variation in study design can affect gene expression. More research is therefore needed to determine the possible effects of these variables. The Health & Environmental Sciences Institute's (HESI) Technical Committee on Genomics are reportedly producing a resource for identifying/analysing baseline fluctuations in gene expression due to biological/technical replicates (**HESI, 2004**). Bias can be introduced either via biological variance (caused by for e.g. circadian cycle regulation (**Boorman et al 2005**), fasting (**Morgan et al 2005**), vehicle/anaesthesia (control animals) (**Takashima et al 2006**, **Sakamoto et al 2005**) and individual animal variability) or through differences in tissue handling. Individual animal variability is considered to be low for rodents but for other species such as dogs and monkeys, it is recommended that the baseline level is predetermined to ensure appropriate group sizes and to also establish a clear understanding of the limits of statistical power. A study by **Whitney et al (2003)** reported on the confounding effects of variation in humans and observed differential gene expression associated with gender, age, time of day at which sample (blood) was taken and other factors. Tissue handling introduces bias in data when different regions of a tissue are sampled between or among control and test animals. Bias also arises when only a particular region known to be sensitive to injury is sampled e.g. liver. The authors cite studies observing differential gene expression in different lobes of rodent livers treated with toxicants i.e. acetaminophen (**Irwin et al 2005**) and furan (**Hamadeh et al 2004**). This emphasises the importance of conducting TGX and histopathological analyses on the same lobe. Preserving tissues for RNA isolation (via use of liquid nitrogen or immersion in appropriate solutions) is considered problematic due to RNA's degradation liability. The authors report that although archival tissue is useful in retrospective analyses of gene expression, tissue fixative and processing methods compromise RNA integrity. The authors cite a review by **Lewis et al (2001)** which suggests methods to extract and use RNA from formalin fixed paraffin-embedded tissue. Laser capture microscopy (LCM) is used to isolate RNA from specific regions of tissue for microarray analysis. However, several limitations have been observed with this method, and a study by **Michel et al (2003)** examined whether the LCM procedure affected detection of gene expression changes induced by clofibrate. The study found that LCM muted the lower-fold changes. It is suggested that loss of sensitivity should be a matter of judgment weighed against increased sensitivity of analysing only the tissue region associated with toxicity. The review also discusses issues relating to the use of peripheral blood as the tissue source of RNA. There are concerns over the adequacy of sample preservation for RNA extraction because prolonged storage is thought to reduce sample comparability. Use of total blood could potentially decrease sensitivity due to the predominance of globin mRNA. The removal of blood components is known to interfere with microarray results and use of fractionated blood for peripheral blood mononuclear cells (PMBCs) is thought to constitute a potential source of bias. Various techniques are used to prepare blood for gene expression analysis. Examples include PAXgene, QIAamp and Ficoll-

Hypaque method. The authors note that the overnight storage of blood significantly affects gene expression compared to blood processed immediately. **Debey et al (2004)** compared the effects of different blood isolation techniques on the quality of results produced on Affymetrix GeneChip arrays. Differences in expression profiles were observed and it was thought this was due to differential isolation of blood cell populations. The study authors also reported improvements in the quality of Affymetrix assay results in protocols that reduce levels of globin mRNA in whole blood samples. The importance of collecting blood, storing and isolating RNA under standardised conditions was noted as a way to help integrate data across laboratories.

RNA quality is considered a critical factor in achieving useful data. High quality RNA helps achieve reproducibility and interpretable results on microarrays, while low level RNA quality reduces statistical power of a study (due to increased measurement error). The RNA Quality Index (RQI) considers the purity and integrity of RNA, which is often contaminated by protein, genomic DNA or chemicals. Pure RNA should have an optical density ratio of 2 at 260 and 280 nm (quantified using spectrophotometers). RNA integrity is assessed via microfluidics-based platforms for nucleic acid analysis. This involves separating RNA (via electrophoresis) and quantifying it (via fluorescence). Calculation of the 28s/182 RNA ratio provides a measure of RNA integrity with intact RNA usually of a value > 2. There are concerns over the usefulness of this calculation and the authors note that electropherograms (a graphical output of electrophoresis devices) provide a more complete picture of RNA quality. An additional RNA quality check involves calculating the expected RNA yield from a given weight of tissue (guidance is available to improve quality/yield for different tissue types). This approach essentially measures the effectiveness of an RNA isolation protocol. RNA degradation is indicated when cDNA's are not of full length. For certain protocols, RNA is considered to be undegraded if the 3' to 5' ratio's of probes derived from universally expressed genes such as GADPH are > 3.

The authors suggest that when comparing microarray data, only those protocols using optimised protocols and reagents should be used. Technical proficiency can be assessed via use of internal/external controls or sample metrics.

Several processing steps are involved in hybridising RNA to arrays and these can serve as checkpoints for monitoring the entire process. The efficiency of the cDNA synthesis/amplification step can be determined by monitoring the yield and size of cRNA product (good quality products range from 500-3000bp). The efficiency of the cRNA fragmentation step can be determined by monitoring the shift in size of products e.g. from 50-200 nucleotides.

The authors consider that the two-colour labelling step is liable to introduce bias for several reasons. This can be due to the different rates at which dyes are incorporated into a sample, or differences in quantum efficiency between two dyes, or the differential sensitivity of Cy5 and Cy3 dyes to quenching, photobleaching and degradation. A potential control would be to run replicate arrays where the orientation of dye incorporation is switched between treated and control samples.

External controls for process monitoring are used to assess the quality of different aspects of the technical performance of sampling, labelling, hybridisation, grid alignment, etc. These controls tend to be non-mammalian sequences (selected from prokaryotic/plant gene sequences) that are spiked into samples (hence aka spike-in targets) which hybridise onto corresponding probe sequences on their arrays. External controls are commercially available for use on in-house spotted arrays. The authors note that external controls have been evaluated in 4 different commercially available arrays (**Tong et al 2006**), and the External RNA Control Consortium (ERCC) are developing better control RNAs. Labelling controls are also available and involve spiking polyA RNAs into RNA samples before the reverse transcription step. External controls are often added after synthesis to assess the success of the hybridisation and staining steps.

Microarray scanners must also undergo quality control assessment and fluorescence standards are used to assess their limits of performance. Various software programmes are used to discriminate between hybridisation failures and scanner defects. The range of a scanner is evaluated using fluorescence calibration slides. For a typical array scanner the output range is between 0 and 65,535 relative fluorescence units per pixel. However, the extent of fold-change differences that can be observed between two samples is limited. It is noted that the HESI Genomics Committee Study identified the photomultiplier tube (PMT) setting as a source of variability (for Affymetrix assays). However, this is not considered

problematic as newer models are available, although this does not apply to investigators using data generated from older models as it would be difficult to compare archival data with recent data. The authors suggest that scanners undergo regular inspections (e.g. after scanning an array) to identify artefacts which can be automated or visual. A software program is available to do this.

Analysis: Microarray data also undergoes quality assessment and various procedures are described. Comparing the intensities of signal data from a sample array with that from technical or biological replicates is one approach. Another approach involves applying principal component analysis (PCA) to signal data, which essentially provides a measure of the quality of data precision (it visualises the similarity of samples within and between groups). Percent present calls (PPC) is a quality index that compares the no. of probes that fall above a threshold on a hybridised (Affymetrix) array with results typically obtained from similar RNA sources. PPC is based on a statistical algorithm that uses perfect match and mismatch pairs. As a guide, the authors note that variation should be less than 10% between samples in the same project. Negative probes are used to estimate global or local background on a microarray. These probes flag up signals that are not significantly above background using feature extraction software. The number of non-significant signals can be used as a strategy to exclude poor quality data. Outliers in groups of microarrays can also be identified using dChip Bioconductor packages.

Statistics: This review does not address issues related to the statistical analysis of TGX data.

Comments: Standards and metrics must be applied to TGX data generation (and analysis) to ensure microarray data is of high quality and to also assess the performance of microarray assays. However, establishing and translating standards and metrics to omic technologies is considered quite a challenge owing to the large no of measurable endpoints in a single omics assay, and the platforms, instruments, reagents and protocols used to generate TGX data. Various papers published between 2003-4 describe best practice for conducting microarray assays, and the National Institute of Standards and Technology (NIST) is working with the FDA to address this issue. Since 2003 data comparability and reproducibility have been enhanced largely due to the increased availability/use of reagent kits and automated systems to process samples/ arrays. High overall levels of reproducibility can be achieved by carefully mapping probes to curated cDNA sequence databases and using standardised protocols to generate data in high performing labs. Two studies compared the reproducibility of data derived from commercial vs. in-house spotted arrays (**Bammler et al 2005, Shi et al 2006**). Their findings showed that data from commercial arrays were more reproducible and it was thought this was possibly due to a more consistent use of manufacturing practices and the application of advanced levels of quality assessments. Various consortia are designing RNA reference materials to assess performance on microarrays across platforms to establish limits of accuracy, precision and linear range. The US FDA Center for Drug Evaluation and Research are also collaborating with government agencies and industry to design and test a reagent for use in several performance assessments on rat whole genome expression microarrays. It is vital that regulatory as well as scientific requirements are met for the commercialisation of in-vitro diagnostic devices (IVDs) arising from the marketing of gene sets associated with a toxic outcome. Regulation is required in particular with regards to assessing the risk from use of the medical devices to patients being tested. The FDA provide guidance documents on how to prepare medical device submissions. The authors conclude that the establishment of universal standards are an important goal to help improve lab performance, protocol optimisation and methods standardisation. They suggest further research should focus on the effect of variations in animal study protocols on gene expression level variance.

Refs:

1. Bammler et al 2005. Standardizing global gene expression analysis between laboratories and across platforms. Nat. Methods.2. 351-6.

   [Abstract]: To facilitate collaborative research efforts between multi-investigator teams using DNA microarrays, the authors identified sources of error and data variability between laboratories and across microarray platforms, and methods to accommodate this variability. RNA expression data were generated in seven laboratories, which compared two standard RNA samples using 12 microarray platforms. At least two standard microarray types (one spotted, one commercial)

were used by all laboratories. The authors report that reproducibility for most platforms within any laboratory was typically good, but reproducibility between platforms and across laboratories was generally poor. Reproducibility between laboratories reportedly increased markedly when standardised protocols were implemented for RNA labelling, hybridisation, microarray processing, data acquisition and data normalisation. Reproducibility was noted to be highest when analysis was based on biological themes defined by enriched Gene Ontology (GO) categories. The authors conclude that these findings indicate that microarray results can be comparable across laboratories, especially when a common platform and set of procedures are used.

2. Boorman et al 2005. Hepatic gene expression changes throughout the day in the Fischer Rat: Implications for toxicogenomics experiments. Toxicol. Sci. 86:185-93.

   [Abstract]: There is increasing use of transcriptional profiling in hepatotoxicity studies in the rat. Understanding hepatic gene expression changes over time is critical, since tissue collection may occur throughout the day. Furthermore, when comparing results from different data sets, times of dosing and tissue collection may vary. Circadian effects on the mouse hepatic transcriptome have been well documented. However, limited reports exist for the rat. In one study approximately 7% of the hepatic genes showed a diurnal expression pattern in a comparison of rat liver samples collected during the day versus livers collected at night. The results of a second study comparing liver samples collected at multiple time points over a circadian day suggest only minimal variation of the hepatic transcriptome. The authors of this paper studied temporal hepatic gene expression in 48 untreated F344/N rats using both approaches employed in the above previous studies. Statistical analysis of microarray (SAM) identified differential expression in day/night comparisons, but was less sensitive for liver samples collected at multiple times of day. However, a Fourier analysis identified numerous periodically expressed genes in these samples including period genes, clock genes, clock-controlled genes, and genes involved in metabolic pathways. Furthermore, rhythms in gene expression were identified for several circadian genes not previously reported in the rat liver. Transcript levels for twenty genes involved in circadian and metabolic pathways were confirmed using quantitative RT-PCR. The authors conclude that the results of this study demonstrate a prominent circadian rhythm in gene expression in the rat that is a critical factor in planning toxicogenomic experiments.

3. Debey et al 2004. Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and role of different cell types. Pharmacogenomics J. 4:193-207.

   [Abstract]: Owing to its clinical accessibility, peripheral blood is probably the best source for the assessment of differences or changes in gene expression associated with disease or drug response and therapy. Gene expression patterns in peripheral blood cells greatly depend on temporal and interindividual variations. However, technical aspects of blood sampling, isolation of cellular components, RNA isolation techniques and clinical aspects such as time to analysis and temperature during processing have been suggested to affect gene expression patterns. The authors therefore assessed gene expression patterns in peripheral blood from 29 healthy individuals by using Affymetrix microarrays. When RNA isolation was delayed for 20-24 h – a typical situation in clinical studies – gene signatures related to hypoxia were observed, and down regulation of genes associated with metabolism, cell cycle or apoptosis became dominant preventing the assessment of gene signatures of interindividual variation. Similarly, gene expression patterns were strongly dependent on choice of cell and RNA isolation and preparation techniques. The authors conclude that for large clinical studies, it is crucial to reduce maximally the time to RNA isolation. Furthermore, prior to study initiation, the cell type of interest should already be defined. The authors consider that their data will therefore help to optimise clinical studies applying

gene expression analysis of peripheral blood to exploit drug responses and to better understand changes associated with disease.

4. Hamadeh et al 2004. Integration of clinical and gene expression endpoints to explore furan-mediated hepatotoxicity. Mutat Res. 549:169-183.

5. HESI (2004)[47]. Committee on the Application of Genomics in Mechanism-Based Risk Assessment, Baseline Animal Data Working Group. http://www.hesiglobal.org/Committees/TechnicalCommittees/Genomics/default.htm Related paper: Boedigheimer et al (2008). Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. BMC Genomics. (9):285-300.

   [Abstract]: Background. The use of gene expression profiling in both clinical and laboratory settings would be enhanced by better characterisation of variance due to individual, environmental and technical factors. Meta-analysis of microarray data from untreated or vehicle-treated animals within the control arm of toxicogenomics studies could yield useful information on baseline fluctuations in gene expression, although control animal data has not been available on a scale and in a form best served for data-mining. Results. A dataset of control animal microarray expression data was assembled by a working group of the Health and Environmental Sciences Institute's Technical Committee on the Application of Genomics in baseline gene expression. Data from over 500 Affymetrix microarrays from control rat liver and kidney were collected from 16 different institutions. Thirty-five biological and technical factors were obtained for each animal, describing a wide range of study characteristics, and a subset were evaluated in detail for their contribution to total variability using multi-variate statistical and graphical techniques. Conclusion. The authors note that the study factors that emerged as key sources of variability included gender, organ section, strain, and fasting state. These and other study factors were identified as key descriptors which the authors consider should be included in the minimal information about a toxicogenomics study needed for interpretation of results by an independent source. Genes that are most and least variable, gender selective, or altered by fasting were also identified and functionally categorised. The authors conclude that better characterisation of gene expression variability in control animals will aid in the design of toxicogenomics studies and in the interpretation of their results.

6. Irwin et al 2005. Transcriptional profiling of the left and median liver lobes of male f344/n rats following exposure to acetaminophen. Toxicol Pathol. 33. 111-117.

7. Lewis et al 2001. Unlocking the archive-gene expression in paraffin-embedded tissue. J Pathol. 195:66-71.

8. Michel et al 2003. Liver gene expression profiles of rats treated with clofibric acid: comparison of whole liver and laser capture microdissected liver. Am. J. Pathol. 163:2191-9.

9. Morgan et al 2005. The hepatic transcriptome as a window on whole body physiology and pathophysiology. Toxicol. Pathol. 33, 136-45.

10. Sakamoto et al 2005. Influence of inhalation anaesthesia assessed by comprehensive gene expression profiling. Gene. 356. 39-48.

[47] See Boedigheimer et al 2008. BMC Genomics. 9.

11. Shi et al 2006[48]. The MicroArray Quality Control (MAQC) project shows inter- and intra-platform reproducibility of gene expression measurements. Nat Biotechnol. 24:1151-1161.

12. Takashima et al 2006. Effect of differences in vehicles on gene expression in the rat liver – analysis of the control data in the Toxicogenomics Project Database. Life Sci. 78:2787-96.

    [Abstract]: The Toxicogenomics Project is a 5-year collaborative project by the Japanese government and pharmaceutical companies in 2002. Its aim is to construct a large-scale toxicology database of 150 compounds orally administered to rats. The test consists of a single administration test (3, 6, 9 and 24 h) and a repeated administration test (3, 7, 14 and 28 days), and the conventional toxicology data together with the gene expression data in liver as analysed by using Affymetrix GeneChip are being accumulated. In the project, either methylcellulose or corn oil is employed as vehicle. The authors examined whether the vehicle itself affects the analysis of gene expression and found that corn oil alone affected the food consumption and biochemical parameters mainly related to lipid metabolism, and this accompanied typical changes in the gene expression. Most of the genes modulated by corn oil were related to cholesterol or fatty acid metabolism (e.g. CYP7A1, CYP8B1, 3-hydroxy-3-methylglutaryl-Coenzyme A reductase, squalene epoxidase, angiopoietin-like protein 4, fatty acid synthase, fatty acid binding proteins), suggesting that the response was physiologic to the oil intake. The authors note that many of the lipid-related genes showed circadian rhythm within a day, but the expression pattern of general clock genes (e.g. period 2, arylhydrocarbon nuclear receptor translocator-like, D site albumin promoter binding protein) were unaffected by corn oil, and suggest that the effects are specific for lipid metabolism. The authors consider that these results would be useful for usage of the database especially when drugs with different vehicle control are compared.

13. Tong et al 2006. Evaluation of external RNA controls for the assessment of microarray performance. Nat. Biotechnol. 24: 1132-9.

14. Whitney et al 2003. Individuality and variation in gene expression patterns in human blood. Proc. Natl. Acad. Sci. U.S.A. 100, 1896-1901.

    [Abstract]: The nature and extent of inter-individual and temporal variation in gene expression patterns in specific cells and tissues is an important and relatively unexplored issue in human biology. The authors surveyed variation in gene expression patterns in peripheral blood from 75 healthy volunteers by using cDNA microarrays. Characterisation of the variation in gene expression in healthy tissue is an essential foundation for the recognition and interpretation of the changes in these patterns associated with infections and other diseases, and peripheral blood was selected because it is a uniquely accessible tissue in which to examine this variation in patients or healthy volunteers in a clinical setting. The authors report that specific features of inter-individual variation in gene expression patterns in peripheral blood could be traced to variation in the relative proportions of specific blood cell subsets; other features were correlated with gender, age, and the time of day at which the sample was taken. An analysis of multiple sequential samples from the same individuals allowed the authors to discern donor specific patterns of gene expression. The authors conclude that these data help to define human individuality and provide a database with which disease-associated gene expression patterns can be compared.

[48] Cited by other reviews (Yauk & Berndt 2007)

Merrick A. (2008). The plasma proteome, adductome and idiosyncratic toxicity in toxicoproteomics research. Briefings in Functional Genomics and Proteomics. Vol 7(1):35-49
Topic covered: This review discusses toxicoproteomics (TPX), a new discipline defined as proteomics applied to toxicology. The paper considers the different proteomic platforms available to TPX research and various approaches designed to elucidate how specific chemical exposures alter protein expression, behaviour and host response leading to injury and disease. The paper considers the different strategic approaches used to generate TPX data but does not address evaluation of raw TPX data (wrt data analyses or statistics).
Design: The authors note that two tiers of TPX research exist: Tier 1 involves identifying and quantifying proteins and their cellular location; Tier 2 involves the detailed investigation of a proteins function, its interaction with other proteins/macromolecules, its 3D structure and any specific post-translational modifications. The different types of PTX platforms used in TPX studies are also considered (e.g. gel affinity and chromatography). Chromatography is further subdivided into adsorptive, liquid and SELDI (surface enhanced laser desorption ionisation). The authors note that the type of platform used can depend on whether the PTX analyses is part of a larger 'omic' investigation, as this would influence the amount of sample ultimately available. Many TPX studies are considered to have served as proof of principle experiments that examine a well characterised toxicant and associate proteomics data output with known toxicological endpoints (i.e. serum and urine chemistries and histopathology). However, these studies report the following as particular challenges: the lack of dose response relationships and time course (in early experiments), the lack of confirmation analysis of differential protein expressions (i.e. via ELISA, western blot, immunohistochemistry, etc), a lack of validation studies of proposed biomarkers, lack of organising, integrating and communicating data within organisations and across laboratories. The following study design modifications are recommended as worthwhile: use of multiple doses, several time points, positive and negative control compounds, use of non-toxic chemical isomers, single and multiple dosing, confirmation of results, and validation in blind studies. Factors preventing the inclusion of some or all of these elements include limited resources, realities of incremental research objectives and the nature of TPX (wrt long data analysis times for interpreting mass spectra and the large data volumes generated per experiment). Biomarker development is considered a specific TPX research objective. The authors note that although the proteomic analysis of blood is a common approach, a more ideal approach would be analysing target tissue. Several advantages for analysing the blood proteome are highlighted however a significant challenge arises with the masking of proteins of interest by more abundant proteins in the blood (often by 10 orders of magnitude). Other potential biomarkers include microparticles and the adductome. Microparticles are intact vesicles derived from cell membranes formed from various events such as apoptosis or membrane activation processes. Microparticles have physiological roles in coagulation, angiogenesis and inflammation and as such change in response to chemical exposure and injury. The adductome refers to proteins/specific amino acid residues covalently bound to reactive chemicals/intermediates. The adductome is considered a useful measure of protein adduction and thus bioactivation of xenobiotics. Current approaches to measuring protein adduction include the use of radiolabelled compounds to track protein adduct formation in liver microsomes. New approaches include the application of high resolution mass spectrometry (MS) instruments using biotin-tagged model electrophiles (**Shin et al 2007**). This approach allows for the identification of adducted proteins with the exact structure of adducted chemical groups/amino-acids identified.
Analysis: The different types of PTX analyses that exist reflect the complexity of the different properties and structures of proteins. Global protein analysis enumerates all proteins identifiable within a sample while TPX analysis enumerates only those proteins that change in accordance with exposure to a particular toxic agent. Protein change is determined by measuring fold or absolute change in protein expression. The authors consider whole proteome analyses a particular challenge as often only portions of a proteome in a sample are analysed. Other key areas of TPX research identified include the analysis of the blood proteome, the interference of abundant proteins in plasma/serum analysis, the presence of soluble microparticles and combining PTX and TRSX analysis of blood. The authors report that many immunosubtraction matrices/devices are commercially available to remove abundant proteins and increased numbers of detectable/identifiable proteins were observed in

a study by **Pieper et al 2003**.
Statistics: The authors do not discuss statistical analysis of TPX data.
Comments: The findings of several TPX reviews published since 2004 are also considered. Many of the reviews covered issues relating to development of serum protein pattern diagnostics, biomarkers and toxicological signatures, and the achievements and limitations associated with biomarker development. The ability of PTX analyses to delineate the potential role of protein adduction in the toxicity of various chemical and drug exposures is reported via illustrated examples i.e. for bromobenzene, acetaminophen, monocrotaline, acrylamide and small molecule electrophiles such as acrolein and nitrative oxidants.

Refs:

1. Shin et al 2007. Protein targets of reactive electrophiles in human liver microsomes. Chem Res Toxicol; 20:859-67.

2. Pieper et al 2003. The human serum proteome: display of nearly 3700 chromatographically separated protein spots on two dimensional electrophoresis gels and identification of 325 distinct proteins. Proteomics;3:311-26.

Elashoff., M (2008). Role of Statistics in Toxicogenomics. From: Methods in Molecular Biology, vol. 460: Essential Concepts in Toxicogenomics. Edited by: D.L. Mendrick and W.B. Mattes © Humana Press, Totowa, NJ.
Topic covered: Although this review aims to address the application of statistics in toxicogenomics (TGX), the term statistics also applies to the various methods used to identify and evaluate toxicologically relevant gene pattern changes, and not just approaches that evaluate the statistical significance of changes/differences in gene expression data. The review considers issues related to the assessment of data quality, data exploration and gene analysis (and predictive modelling) under two main sections: individual TGX studies and TGX databases. Design issues are discussed as a separate section.
Design: The authors note that a typical TGX study will include several time points and doses (3-6 animals per control or dose group) with a single RNA sample run for each animal. Studies employ 3-4 time points as there is no clear answer on which single timepoint is best to use. From a statistical point of view the more doses used the better as this enables dose response to be analysed in relation to the overall toxicity, as well as on an individual gene and pathway level. Sample size calculations are conducted to work out the best way of maximising scientific information at minimal cost. This involves making assumptions in relation to the degree of gene regulation and biological variability between samples. Authors provide an illustrative example. Biological rather than technical replicates are preferred because an average response to treatment is more informative for drawing conclusions than individual responses (also biological variation in gene expression tends to be 2-4-fold higher than technical variation). Technical replicates are preferred when the gene expression profile of an individual is the target (as occurs in diagnostic situations). Intermediate approaches are possibly required when multiple measurements are taken for each animal if biological variability increases. Thus the design is dependent on the ratio of biological vs. technical variability. A key aim of the study requires information on the average response, and how this response varies across a set of individuals. Pooled samples provide information on the average response but such samples do not inform on the variation in that response which precludes a meaningful analysis of the study data. The authors note that studies using only technical replicates do not provide information on the average response.
Analysis: [Individual TGX studies]: Many gene expression analysis tools exist but their value is limited by a lack of standardisation and the extensive number of analytical approaches available. The first step toward analysing gene expression data is to consider the quality of data, treatment effects, identity of regulated genes/pathways and the toxicological context of gene expression changes. Assessing data quality helps distinguish poor quality data from useful data. Two main types of approaches are used: quality metrics and correlation. Quality metrics can detect variation in data quality and involves 3 basic approaches: application of a threshold benchmark (where metrics that fall short of a predetermined value fail); consistency (failing metric values that lie outside the norm within a study); and balance (comparing distribution of metric values between study groups). The most informative measure of gene expression data quality is the percent present (PP). PP measures the percentage of genes present (expressed) in a sample, as a fraction of genes deemed present / total no genes present on chip. PP is limited by the fact its value depends on the type of chip and sample used. Other quality metrics include 5'3' ratio for specific control genes (measures RNA degradation); scale factor (involves scaling unnormalised gene expression mean values); Affymetrix specific MM > PM (provides an informative measure of Affymetrix chip quality in which perfect match (PM) probe pairs should be greater than mismatch (MM) probe pairs); and signal distribution (compares the distribution of expression signal strength that reveals differences in data quality). Pearson correlation measures the similarity of expression log values between a pair of samples and uses the entire set of genes. A correlation matrix is produced by deriving correlation values (or average correlation values) for each sample relative to another within the same study. Principle component analysis (PCA) is a multivariate technique used in the data exploration phase of TGX. PCA visualises multidimensional data sets for gene sets and generates principal components plotted against each other. The gene sets can be based on all genes, or changing genes or genes for particular pathways. Interpreting all genes PCA involves highlighting predominant patterns in gene expression data (where outliers correspond to poor quality data). Genes driving the different prominent patterns can be further investigated. The

authors emphasis that prominent patterns may also be caused by differential sample processing. Changing genes PCA reveal dose and time effects and provide information on samples that do not fit the pattern and they themselves may be further investigated. Percent variance corresponds to a principal component and enables a rough assessment of how the PCA plots of the first several components reflect the entire gene set.

[TGX databases]: The authors consider that comparing the expression profiles of both similar and dissimilar compounds helps put things in context. The study sample must be comparable with samples in database. Data comparability is determined by firstly considering the similarity of various study design features e.g. vehicle controls, sex/strain of animals, sample and chip processing methods. Such features are thought to alter the baseline expression level of genes. Normalising the data (within a study) is the next step followed by an assessment of data comparability between study and database samples. Normalising data (within a study) is considered better for comparing data because it removes much of the cross-study difference while preserving the underlying biological responses. The authors report that experience suggests unnormalised data is of limited value for comparing data between different groups. Assessing data comparability between study samples and the database samples is recommended as a next step. Possible reasons for the clustering/grouping of compounds observed during data exploration stage are suggested i.e. due to compounds sharing similar mechanisms of toxicity, or having high level toxic effects resulting in grouping (necrosis may be ultimately induced but arises via different mechanisms), or having similar non toxic effects, or sharing study effects (i.e. shared control group). Assessing the similarity of genes during the gene analysis stage helps inform on the toxicological significance of genes. Gene similarity analysis uses gene expression profiles to identify genes that act similarly to known toxicity markers. Similar genes are identified using a statistical algorithm (Match X). Predictive modelling aims to link gene expression profiles of multiple compounds to an expected behaviour of the compound when used in humans. Model validation is considered the final phase of this model building process but it is considered difficult in practice. A multistep validation procedure (for a predictive model) is described in relation to the compounds of the test and training set, the building and evaluation of the model, and reporting accuracy rates. The authors note the significance of comparing training and test sets for developing or assessing a model, and provide illustrated examples based on a hypothetical data set. The authors describe how to set parameters for building models and describe the different methods used to categorise the selection of genes, which themselves can be ranked in terms of how useful they are. Several different types of classification methods are used in TGX modelling (e.g. clustering, classification tree, logistic regression, K-Means, partial least squares, support vector machine (SVM), neural networks and discriminant analysis). Each have their own pro's and cons in relation to their fitting capability, their tendency to overfit and ability to isolate gene contributions.

Statistics: [Individual TGX studies]: Fold change (aka two sample t-test) is considered the most basic analytical method used in gene level analysis. It tests (statistically) for a difference in mean logged expression levels between two groups. It generates a P value and fold change. Two groups of samples are required (e.g. high dose vs. control samples). Other methods used in gene level analysis include cut-offs, filtering, ANOVA and pathway analysis. Cut-offs describe a statistical approach to determining which genes are regulated by a compound and involve calculating the false discovery rate (FDR). The FDR is a ratio of the number of false positive genes (genes that appear to be significant but are not regulated by the compound) to the total of true positive and false positive genes. The level of FDR is decided by investigators e.g. a 5% FDR threshold means no more than 5% genes in the list of significant genes to be false positive. An example is provided. The authors introduce a related calculation 'power', defined as the ratio of the number of true positives to the total number of truly regulated genes. As an example a list of significant genes is said to have 90% power when 90% of its genes are truly regulated by a compound. The best cut-off is one that achieves maximal power while minimising the FDR at some preset level. Filtering removes genes that are not called present by the compound via use of image processing algorithms. However, filtering out of low expressing genes that are also regulated by the compound is a key concern. ANOVA analyses whether a dependent variable (gene expression level) changes/varies in response to a particular parameter or independent variable (e.g. time, dose). ANOVA allows for the simultaneous analysis of several independent variables and describes the overall effect of a particular independent variable (i.e. whether gene expression

level varies with time or dose). Pathway analysis involves the detection of pathways through the identification of sets of genes with common characteristics. An equation is used to measure whether a compound can affect a pathway. However, since it relies on counts of genes, there is the assumption that the genes are independent which is a limitation of this technique. Use of a measure that accounts for the correlation between genes within a pathway would be a viable alternative approach. The overarching view is that statistics helps identify genes/pathways regulated or associated with compound under study.

[TGX databases]: WRT gene level analysis, fold-change (t-test) analysis helps establish which pathways are differentially regulated. It does this by comparing the mean fold change of a single gene for two groups of samples. It takes into account how genes within a pathway correlate.

Comments: The authors consider that statistics could solve the problem of there being too many gene expression analytical approaches. Readers are referred to the Elashoff Consulting website www.elashoffconsulting.com where appropriate statistical software is available. Elashoff Consulting is a biostatistics company specialising in genomics/genetics analysis and clinical trials. The company purports to have experience analysing TGX data wrt predictive models, cross platform prediction, phenotypic anchoring and regulatory aspects.

Refs.

There were no refs cited within the text.

Rho et al (2008). From proteomics toward systems biology: integration of different types of proteomics data into network models. BMB reports; 41(3):184-93.
Topic covered: This review aims to demonstrate how current proteomic (PTX) technologies can improve our understanding of how complex biological networks operate at a systems level. The authors address some aspects of PTX study design and discuss a few of the tools used to analyse data generated from mass spectrometry (MS). However, there is a strong emphasis on application of PTX data to systems biology research. The review does not discuss issues related to the statistical analysis of PTX data.
Design: The authors categorise PTX technologies as either antibody or mass-spectrometry (MS) based. MS-based techniques are considered more useful as they provide extensive (global) information on proteins in terms of their function, abundance, modifications and interactions on different levels. The authors define MS as an analytical tool that measures the mass-to-charge (m/z) ratios of ionised analytes (proteins or peptides). In a two scan MS/MS procedure, the first scan produces peptide fragments. In a second scan these fragments undergo further isolation and fragmentation (via collision-induced dissociation) for identification purposes. The authors note that the intensity of the measured peaks is proportional to the abundance of certain peptides. A key disadvantage associated with these techniques is undersampling i.e. the inability to manage the several thousand proteins in complex samples and thus detect less abundant proteins. Approaches to address this issue and improve the performance of these technologies include adopting sample preparation methods, liquid chromatography (LC) or applying additional fractionation methods such as 2-D electrophoresis and multidimensional protein identification technology (MudPIT) (**Prakash et al 2006**). These approaches enable thousands of proteins to be detected in complex samples. Three stages of MS-based PTX analysis are noted: sample preparation, LC-MS/MS analysis and computational analysis of MS data (to quantify and identify the proteins); different combinations of the above can be used. Sample preparation methods include isotope labelling and subproteome capture methods. Isotope labelling methods (such as ICAT, SILAC and iTRAQ) measure the abundance of proteins in complex samples. Label-free methods also exist and both methods require the deployment of various computational tools to estimate protein abundances. Subproteome capture methods (e.g. of phosphoproteome or ubiquinated proteome) measure the extent of post-translational modifications (PTM) of proteins and these can be quantified if used in combination with isotope labelling. Other sample preparation methods measure protein interaction e.g. immuno-precipitation-based methods, tandem affinity purification (TAP)-tagging and chemically conjugated bead-based methods. LC separates peptides by molecular weight and is often used in combination with MS/MS (LS-MS/MS) to effectively analyse complex samples.
Analysis: Various data analysis systems are available to deal with the huge amounts of data generated from LC-MS/MS. These systems convert, visualise, store and exchange PTX data, and also conduct basic analytical functions associated with protein quantification and identification. The authors note the development of their Integrative Proteomics Data Analysis Pipeline (IPDAP), a PTX data analysis system used in systems biology research. IPDAP is built on two platforms: a computational PTX laboratory database (CPAS) and a systems biology experiment analysis management system (SBEAMS). NB. Includes other tools such as Trans Proteomic Pipeline (TPP). IPDAP operates in the following way: raw LC-MS/MS data is stored onto CPAS and converted into a standard data format (mzXML). Next, the data is analysed via a database search to identify proteins (using X! Tandem or SEQUEST), and a best match is identified via use of Peptideprophet (in TPP) which performs the probability calculations. Computational tools such as XPRESS, ASAPRatio (in conjunction with isotopic labelling methods) are used to quantify the identified proteins. This results in the generation of a list of proteins and other associated data i.e. protein abundance and PTM. NB. The identified protein data are also stored on CPAS. Proteins are then mapped onto operational biological networks to determine how these proteins interact with each other and how they enrich pathways and functional groups, including information on any temporal effects on key pathways. Various systems biology software tools analysing particular aspects of the data are used e.g. for interaction (BIND, HPRD), statistical analyses (clustering, PCA), network modelling and analysis (STRING, Cytoscape), pathway analysis (KEGG).
Statistics: The review does not address issues related to the statistical analysis of raw TGX

Comments: The authors define a biological network as a composite of nodes (e.g. DNA, mRNAs, proteins and metabolites of cellular systems) and edges (e.g. the interactions between these nodes, which can be between proteins i.e. protein-protein (PPI), protein-DNA (PDI), chemical-protein (CPI) and chemical-DNA (CDI)). These networks enable systems (defined as organs, tissues, cells and subcellular compartments) to function and they receive signals from these systems. The authors describe network modules as a particular portion of a biological network. They are activated to execute certain functions to offset perturbations caused by environmental or genetic events. Disease arises when these network modules malfunction and are unable to offset perturbations.
The authors state that systems biology approaches aim to understand how complex networks operate via a three-step process. The first step generates global data following perturbation of a system. The data is then integrated into network models that provide information on key events arising from these perturbations. The final step involves generating a testable hypothesis to determine associated mechanisms. Prior to these steps, the authors note that key network modules must first be identified and any transitions arising from these perturbations also examined.
The authors comment that PTX studies have enhanced our understanding of how biological networks function. Use of PTX data in systems biology approaches involves 3 steps: step 1 defines problems for biological/medical systems. Step 2 involves perturbing relevant biological systems while the final step involves generating comprehensive PTX data. This is followed by a series of computational steps to determine biological networks and modules. Finally, the authors consider network modelling as a key step for processing PTX data in systems biology. IPDAP provides a general solution for network modelling and visualising pipelines although various technical challenges exist.

Refs:

1. Prakash et al 2006. Signal maps for mass spectrometry-based comparative proteomics. Mol.Cell. Proteomics. 5:423-32.

Annex IV

REFERENCES

1. Ahmed (2006a). Microarray RNA transcriptional profiling. I. Platforms, experimental design and standardization. Expert Rev Mol Diagn. 6:535-50.

2. Ahmed (2006b). Microarray RNA transcriptional profiling. I. Analytical considerations and annotation. Expert Rev Mol Diagn. 6:703-15.

3. Andersen et al (2008). Genomic signatures and dose-dependent transitions in nasal epithelial responses to inhaled formaldehyde in the rat. Toxicol Sci;105(2):368-83..

4. Arukwe (2006). Toxicological housekeeping genes: do they really keep the house? Environ Sci Technol;40(24):7944-9.

5. Baker et al (2004). Clofibrate-induced gene expression changes in the rat liver: A cross laboratory analysis using membrane cDNA arrays. Env Health Perspect, 112:428-38.

6. Baker et al (2005). External RNA Controls Consortium. The External RNA Controls Consortium: aprogress report. Nat Methods. 2(10):731-4.

7. Bammler et al (2005). Members of the Toxicogenomics Research Consortium. Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods;2(5):351-6.

8. Barlow et al (2003). Report of a symposium on the use of genomics and proteomics in toxicology. Mutagenesis;18(3):311-7.

9. Battershill (2005). Toxicogenomics: regulatory perspective on current position. Hum Exp Toxicol;24(1):35-40.

10. Boedigheimer et al (2008). Sources of variation in baseline gene expression levels from toxicogenomics study control animals across multiple laboratories. BMC Genomics. (9):285-300.

11. Boess et al (2003). Gene expression in two hepatic cell lines, cultured primary hepatocytes, and liver slices compared to the *in-vivo* liver gene expression in rats: possible implications for toxicogenomics use of *in-vitro* systems. Toxicol Sci 73: 386-402.

12. Boes & Neuhauser (2005). Normalization for Affymetrix GeneChips.
13. Methods Inf Med;44(3):414-7.

14. Boorman et al (2005). Hepatic gene expression changes throughout the day in the Fischer Rat: Implications for toxicogenomics experiments. Toxicol. Sci. 86:185-93

15. Brors (2005) Microarray annotation and biological information on function. Methods Inf Med;44(3):468-72.

16. Burgoon & Zacharewski (2008). Automated quantitative dose-response modeling and point of departure determination for large toxicogenomic and high-throughput screening data sets. Toxicol Sci. Aug;104(2):412-8.

17. Butte et al (2002). The use and analysis of microarray data. Nat Rev. Drug. Discov. 1: 951-60.

18. Castoldi et al (2006). A sensitive array for microRNA expression profiling (miChip) based on locked nucleic acids (LNA). RNA. 12(5):913-20.

19. Chen et al (2004). Analysis of variance components in gene expression data. Bioinformatics. 20:1436-46.

20. Chou et al (2007). Extracting gene expression patterns and identifying co-expressed genes from microarray data reveals biologically responsive processes. BMC Bioinformatics;8:427

21. Cipriany et al (2010). Single molecule epigenetic analysis in a nanofluidic channel. Anal Chem;82(6):2480-7.

22. Collas (2009). The state-of-the-art of chromatin immunoprecipitation. Methods Mol Biol;567:1-25.

23. Copois et al (2007) Impact of RNA degradation on gene expression profiles: assessment of different methods to reliably determine RNA quality. Biotechnol;127(4):549-59.

24. Corvi et al (2006). Meeting report: Validation of toxicogenomics-based test systems: ECVAM-ICCVAM/NICEATM considerations for regulatory use. Environ Health Perspect;114(3):420-9.

25. Daston (2008). Gene expression, dose-response, and phenotypic anchoring: applications for toxicogenomics in risk assessment. Toxicol Sci;105(2):233-4.

26. Debey et al (2004). Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and role of different cell types. Pharmacogenomics J. 4:193-207.

27. De Wit et al (2008). Molecular targets of TBBPA in zebrafish analysed through integration of genomic and proteomic approaches. Chemosphere;74(1):96-105.

28. Du et al (2008). lumi: a pipeline for processing Illumina microarray. Bioinformatics;24(13):1547-8.

29. Durinck (2008). Pre-processing of microarray data and analysis of differential expression. Methods Mol Biol;452:89-110.

30. EPA (2004). Potential Implications of Genomics for Regulatory and Risk Assessment Applications at EPA. Prepared for the US EPA by members of the Genomics Task Force Workgroup, a group of EPA's Science Policy Council. EPA 100/B-04/002.

31. Elashoff et al (2008). Role of Statistics in Toxicogenomics. From: Methods in Molecular Biology, vol. 460: Essential Concepts in Toxicogenomics. Edited by: D.L. Mendrick and W.B. Mattes © Humana Press, Totowa, NJ.

32. Elferink et al (2008). Microarray analysis in rat liver slices correctly predicts in vivo hepatotoxicity. Toxicol Appl Pharmacol;229(3):300-9.

33. Fang et al (2009). ArrayTrack: an FDA and public genomic tool. Methods Mol Biol;563:379-98.

34. Fujita et al (2009). Quality control and reproducibility in DNA microarray experiments. Genome Inform. 23(1);21-31.

35. Gant (2007). Novel and future applications of microarrays in toxicological research. Expert Opin. Drug Metab. Toxicol. 3(4):599-608

36. Hartmann (2005) Quality control for microarray experiments. Methods Inf Med;44(3):408-13.

37. Hayes & Bradfield (2005). Advances In Toxicogenomics. Chemical Research In Toxicology. Vol 18, No. 3:403-14

38. Hayes et al (2005). EDGE: A centralised resource for the comparison, analysis, and distribution of toxicogenomic information. Molecular Pharmacology, 67:1360-68

39. Hurd & Nelson (2009). Advantages of next-generation sequencing versus the microarray in epigenetic research. Brief Funct Genomic Proteomic;8(3):174-83.

40. IPCS (2003). Workshop Report. Toxicogenomics and the risk assessment of chemicals for the protection of human health. WHO International Labour Organiszation, UN Environment Programme. IPCS/Toxicogenomics/03/1.

41. Irwin et al (2004). Application of Toxicogenomics to Toxicology: Basic Concepts in the Analysis of Microarray Data. Toxicologic Pathway, Vol 32(Suppl. 1):72-83

42. Ittrich (2005). Normalization for two-channel microarray data. Methods Inf Med;44(3):418-22.

43. Jafari & Azuaje (2006). An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. BMC Med Inform Decis Mak;6:27.

44. Ju et al (2007). DNA microarray technology in toxicogenomics of aquatic models: Methods and applications. Comparative biology and physiology, Part c 145: 5-14

45. Kavlock et al (2008). Computational toxicology--a state of the science mini review. Toxicol Sci;103(1):14-27.

46. Kerr & Churchill (2001). Statistical design and the analysis of gene expression microarrays. Genet. Res. 77, 123-8.

47. Kepler et al 2002. Normalization and analysis of DNA microarray data by self consistency and local regression. Genome Biol. 3. 1-12.

48. Kuhn et al (2004). A novel, high performance random array platform… Genome Res;14(11):2347-56.

49. Kulkarni et al (2008). Assessing chronic liver toxicity based on relative gene expression data. J Theor Biol;254(2):308-18.

50. Kwon et al (2008). Time- and dose-based gene expression profiles produced by a bile-duct-damaging chemical, 4,4'-methylene dianiline, in mouse liver in an acute phase. Toxicol Pathol;36(5):660-73.

51. Lam et al (2008). Zebrafish whole-adult-organism chemogenomics for large-scale predictive and discovery chemical biology. PLoS Genet;4(7):e1000121.

52. Lampe et al (2004). Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. Cancer Epidemiol., Biomarkers Prev. (13(3):445-3

53. Lee et al (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc. Natl. Acad. Sci. USA 97: 9834-9

54. Lee et al (2004). The intelligent data management system for toxicogenomics. J Vet. Med. Sci. 66:1335-38

55. Lee et al (2005). Design issues in toxicogenomics using DNA microarray experiment. *Toxicology and Applied Pharmacology*: 207 S200-08

56. Lee et al (2007). Identification of novel universal housekeeping genes by statistical analysis of microarray data. J Biochem Mol Biol;40(2):226-31.

57. Magglioli et al (2006). Toxicogenomic Analysis Methods For Predictive Toxicology. Journal of Pharmacological and Toxicological Methods, Vol 53:31-7

58. Ma et al (2006). Application of Real-time Polymerase Chain Reaction (RT-PCR). The Journal of American Science, 2(3): 1-15

59. Mah et al (2004). A comparison of oligonucleotide and cDNA-based microarray systems. Physiol. Genomics 16:361-70.

60. Maier et al (2009) Correlation of mRNA and protein in complex biological samples. FEBS Lett;583(24):3966-73.

61. Mattes (2004). Annotation and cross-indexing of array elements on multiple platforms. Env Health Perspect, 112:506-10

62. Mattes (2008). Public consortium efforts in toxicogenomics. Methods Mol Biol;460:221-38.

63. Mayo et al (2006). Some statistical issues in microarray gene expression data. Radiat Res;165(6):745-8.

64. Mazan-Mamczarz et al (2005). En masse analysis of nascent translation using microarrays. Biotechniques. 39(1):61-7.

65. Mei et al (2009). Application of microarray-based analysis of gene expression in the field of toxicogenomics. Methods Mol Biol;597:227-41.

66. Melamed et al (2009). Exploring translation regulation by global analysis of ribosomal association. Methods;48(3):301-5.

67. Model et al (2002). Statistical process control for large scale microarray experiments. Bioinformatics;18 Suppl 1:S155-63.

68. Morgan et al (2002). Application of cDNA microarray technology to in-vitro toxicology and the selection of genes for real time RT-PCR-based screen for oxidative stress in Hep-G2 Cells. Toxicol Pathol. 30. 435-51.

69. Morgan et al (2004). Complementary roles for toxicologic pathology and mathematics in toxicogenomics with special reference to data interpretation and oscillatory dynamics. Toxicological Pathology, 32(Suppl 1):13-25

70. Morris et al (2006). Alternative Probeset Definitions for Combining Microarray Data Across Studies Using Different Versions of Affymetrix Oligonucleotide Arrays.In: Meta-Analysis in Genetics. Ed. Rudy Guerra and David Allison. New York: Chapman-Hall.

71. Muellner et al (2010). Human cell toxicogenomic analysis of bromoacetic acid: a regulated drinking water disinfection by-product. Environ Mol Mutagen;51(3):205-14.

72. Naciff et al (2007). Uterine temporal response to acute exposure to 17alpha-ethinyl estradiol in the immature rat. Toxicol Sci; 97(2):467-90.

73. NAS (2007a). Validation of Toxicogenomic Technologies: A Workshop Summary. Committee on Validation of Toxicogenomic Technologies: A focus on Chemical Classification Strategies, Committee on Emerging Issues and Data on Environmental Contaminants, National Research Council. ISBN: 0-309-66825-5. National Academy of Sciences. (http://www.nap.edu/catalog/11804.html)

74. NAS (2007b). Applications of Toxicogenomic Technologies to Predictive Toxicology and Risk Assessment. Committee on Applications of Toxicogenomic Technologies to Predictive Toxicology and Risk Assessment, National Research Council. ISBN: 0-309-11299-0. National Academy of Sciences. (http://www.nap.edu/catalog/12037.html)

75. NRC, (2007). Toxicity Testing in the 21st Century: A Vision and a Strategy. Committee on Toxicity Testing and Assessment of Environmental Agents, Board on Environmental Studies and Toxicology, Institute for Laboratory Animal Research, Division on Earth and Life Studies; National Research Council of the National Academies; The National Academies Press. Washington, D.C. www.nap.edu. ISBN-10: 0-309-15173-2

76. Novak et al (2002). Characterisation of variability in large-scale gene expression data: implications for study design. Genomics. 79:104-113

77. Page et al (2006). The PowerAtlas: a power and sample size atlas for microarray experimental design and research. BMC Bioinformatics;7:84.

78. Patterson et al (2006). Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nat Biotechnol;24(9):1140-50.

79. Pettit (2004). Toxicogenomics in risk assessment: communicating the challenges. Environ Health Perspect;112(12):A662

80. Pettit et al (2010). Current and Future Applications of Toxicogenomics: Results Summary of a Survey from the HESI Genomics State of Science Subcommittee. Environ Health Perspect. 2010 Jan 25.

81. Plummer et al (2007). Time-dependent and compartment-specific effects of in utero exposure to Di(n-butyl) phthalate on gene/protein expression in the fetal rat testis as revealed by transcription profiling and laser capture microdissection. Toxicol Sci;97(2):520-32.

82. Qin et al (2004). Empirical evaluation of data transformations and ranking statistics for microarray analysis. Nucleic Acids Res. 2004 Oct 12;32(18):5471-9.

83. Rahnenfuhrer (2005a). Image analysis for cDNA microarrays. Methods Inf Med;44(3):405-7.

84. Rahnenfuhrer (2005b). Clustering algorithms and other exploratory methods for microarray data analysis. Methods Inf Med;44(3):444-8.

85. Repsilber et al (2005). Tutorial on microarray gene expression experiments. An introduction. Methods Inf Med;44(3):392-9.

86. Repsilber & Ziegler (2005). Two-color microarray experiments. Technology and sources of variance. Methods Inf Med;44(3):400-4.

87. Ringner (2008). What is principal component analysis? Nat Biotechnol. 2008 Mar;26(3):303-4.

88. Rho et al (2008). From proteomics toward systems biology: integration of different types of proteomics data into network models. BMB reports; 41(3):184-93.

89. Roach et al (2009). High throughput single cell bioinformatics. Biotechnol Prog;25(6):1772-9.

90. Rosenzweig et al (2004). Dye bias correction in dual-labeled cDNA microarray gene expression measurements. Environ Health Perspect;112(4):480-7.

91. Ruetze et al (2010). Analysis of tissue-specific gene expression using laser capture microdissection. Methods Mol Biol;585:183-92.

92. Sansone et al (2004). Standardization Initiatives in the eco)toxicogenomics Domain: A Review. Comp Funct Genomics;5(8):633-41.

93. Sansone et al (2006). RSBI Members. A strategy capitalizing on synergies: the Reporting Structure for Biological Investigation (RSBI) working group. OMICS;10(2):164-71

94. Scholz et al (2008). The zebrafish embryo model in environmental risk assessment--applications beyond acute toxicity testing. Environ Sci Pollut Res Int;15(5):394-404.

95. Shenton et al (2006). Global translational responses to oxidative stress impact upon multiple levels of protein synthesis. J. Biol. Chem. 281(39):29011-29021.

96. Shi et al (2006). The MicroArray Quality Control (MAQC) project shows inter and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 24:1151-1161.

97. Shinawi & Cheung (2008). The array CGH and its clinical applications. Drug Discov Today;13(17-18):760-70

98. Skibbe et al (2006). Scanning microarrays at multiple intensities enhances discovery of differentially expressed genes. Bioinformatics;22(15):1863-70.

99. Slonim & Yanai (2009). Getting started in gene expression microarray analysis. PLoS Comput Biol;5(10):e1000543.

100. Sone et al (2010). Profiles of Chemical Effects on Cells (pCEC): a toxicogenomics database with a toxicoinformatics system for risk evaluation and toxicity prediction of environmental chemicals. J Toxicol Sci;35(1):115-23.

101. Steiner et al (2004). Discriminating different classes of toxicants by transcript profiling. EHP. 112. 1236-48.

102. Sumida et al (2007). Optimization of an animal test protocol for toxicogenomics studies (I); requirement study of a protocol. J Toxicol Sci;32(1):19-32.

103. Suarez et al (2009). Microarray data analysis for differential expression: a tutorial. P R Health Sci J;28(2):89-104.

104. Suzuki et al (2008). In vitro gene expression analysis of hepatotoxic drugs in rat primary hepatocytes. J Appl Toxicol; 28(2):227-36.

105. Szyf (2007). The dynamic epigenome and its implications in toxicology. Toxicol Sci;100(1):7-23.

106. Takishima et al (2006). Effect of differences in vehicles on gene expression in the rat liver – analysis of the control data in the Toxicogenomics Project Database. Life Sci. 78:2787-96.

107. Thomas et al (2001). Identification of toxicologically predictive gene sets using cDNA microarrays. Molecular Pharmacology, 60:1189-94

108. Thompson et al (2004). Identification of platform-independent gene expression markers of cisplatin nephrotoxicity. Env Health Perspect, 112:488-94

109. Thompson & Hackett (2008). Quality Control of Microarray Assays for Toxicogenomic and In Vitro Diagnostic Applications. From: Methods in Molecular Biology, vol. 460: Essential Concepts in Toxicogenomics. Edited by: D.L. Mendrick and W.B. Mattes © Humana Press, Totowa, NJ.

110. Tietjen et al (2003). Single cell transcriptional analysis of neuronal progenitors. Neuron 38. 161-75.

111. Toyoshiba et al (2009) Similar compounds searching system by using the gene expression microarray database. Toxicol Lett;186(1):52-7.

112. Tsai et al (2005). Multi-class clustering and prediction in the analysis of microarray data. Mathematical Biosciences, 193:79-100.

113. Uehara et al (2008) Species-specific differences in coumarin-induced hepatotoxicity as an example toxicogenomics-based approach to assessing risk of toxicity to humans. Hum Exp Toxicol;27(1):23-35.

114. Usenko et al (2008). Fullerene C60 exposure elicits an oxidative stress response in embryonic zebrafish. Toxicol Appl Pharmacol;229(1):44-55.

115. Van den Berg (2010). Re-annotation is an essential step in systems biology modeling of functional genomics data. PLoS One;5(5):e10642.

116. VanGuilder et al (2008). Twenty-five years of quantitative PCR for gene expression analysis. Biotechniques; 44(5):619-26.

117. Vlaanderen et al (2010). Application of OMICS technologies in occupational and environmental health research; current status and projections. Occup Environ Med;67(2):136-43.

118. Wahl et al (2008). A technical mixture of 2,2',4,4'-tetrabromo diphenyl ether (BDE47) and brominated furans triggers aryl hydrocarbon receptor (AhR) mediated gene expression and toxicity. Chemosphere;73(2):209-15.

119. Wang et al (2007). Systematic approaches for incorporating control spots and data quality information to improve normalization of cDNA microarray data. J Biopharm Stat.17(3):415-31.

120. Warrington et al (2005). The External RNA Controls Consortium: a progress report. Nat Methods;2(10):731-4.

121. Wei et al (2004). Sample size for detecting differentially expressed genes in microarray experiments. BMC Genomics 5 (1). 87

122. Williams & Thomson (2010). Effects of scanning sensitivity and multiple scan algorithms on microarray data quality. BMC Bioinformatics;11:127.

123. Williams-Devane et al (2009). Toward a public toxicogenomics capability for supporting predictive toxicology: survey of current resources and chemical indexing of experiments in GEO and ArrayExpress. Toxicol Sci;109(2):358-71.

124. Whitney et al (2003). Individuality and variation in gene expression patterns in human blood. Proc. Natl. Acad. Sci. U.S.A. 100, 1896-1901.

125. Wu et al (2009). BioGPS: an extensible and customizable portal for querying and organizing gene annotation resouces. Genome Biology, 10:R130-R130.8

126. Yang & Speed (2002). Design issues for cDNA microarray experiments. Nat Rev Genet; 3(8):579-88.

127. Yang et al (2007). Transcriptional profiling reveals barcode-like toxicogenomic responses in the zebrafish embryo. Genome Biol;8(10):R227.

128. Yauk & Berndt (2007). Review of the Literature Examining the Correlation Among DNA Microarray Technologies. Environmental & Molecular Mutagenesis 48:380-94

129. Yu et al (2007). PARE: a tool for comparing protein abundance and mRNA expression data. BMC Bioinformatics;8:309.

130. Zhang & Gant (2005). A statistical framework for the design of microarray experiments and effective detection of differential gene expression. Bioinformatics. 20: 2821-28.

131. Zhou et al (2009). Toxicogenomics: transcription profiling for toxicology assessment. EXS;99:325-66.

Groups/Organisations/Initiatives

Name		URL: http://
BTS	British Toxicology Society	thebts.org
EBI	European Biomarkers Institute	ebi.ac.uk/
ERCC	External RNA Control Consortium	cstl.nist.gov
ECVAM	European Centre for the Validation of Alternative Methods	ecvam.jrc.ec.europa.eu/
EMBL	European Molecular Biology Laboratories	embl.org
EPA	US Environmental Protection Agency	epa.gov
FDA	US Food & Drug Administration	fda.gov
FP6/7		
GO Consortium	Gene Ontology Consortium	geneontology.org/
HESI	Health and Environmental Sciences Institute	hesiglobal.org
ICCVAM	Interagency Coordinating Committee on the Validation of Alternative Methods	iccvam.niehs.nih.gov/
IMI	Innovative Medicines Initiatives	imi.europa.eu/index_en.html
ILSI	International Life Sciences Institute	ilsi.org
IPCS	WHO International Programme on Chemical Safety	who.int/ipcs/en/
MAQC	MicroArray Quality Control	fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/default.htm
MGED Society	Microarray Gene Expression Data Society	mged.org/
NCT	National Center for Toxicogenomics	niehs.nih.gov/research/atniehs/nct.cfm
NCTR	National Center for Toxicological Research	fda.gov
NICEATM	National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods	iccvam.niehs.nih.gov/
NIEHS	National Institute of Environmental Health Sciences	niehs.nih.gov/research
NIHS	Japanese National Institute of Health Sciences;	nihs.go.jp/english/index.html
NRC	US National Research Council	sites.nationalacademies.org/NRC/
OECD	Organisation for Economic Co-operation and Development	oecd.org
REACH	EU Registration, Evaluation, Authorisation and restriction of CHemicals	echa.europa.eu/home_en.asp
SOT	Society of Toxicology	toxicology.org
TRC	Toxicogenomic Research Consortium	niehs.nih.gov/research/supported.centers/trc

Software/Databases

Name		URL http://www.
AMIA	Automated Microarray Image Analysis	?
Array Express		ebi.ac.uk/microarray-as/ae/
ArrayTrack		fda.gov/nctr/sceince/centers/toxicoinformatics/ArrayTrack/index.htm
BASE	BioArray Software Environment	base.thep.lu.se/
Baseline Animal Database		hesiglobal.org/
Bioconducter Project		bioconductor.org/
BioGPS	A successor to Symatlas	biogps.gnf.org/
BRB-Array Tools	Biometric Research Branch Array Tools	linus.nci.nih.gov/BRB-ArrayTools.html
CEBS	Chemical Effects in Biological Systems	cebs.niehs.nih.gov
CIBEX	Centre for Information Biology gene Express	
CTD	Comparative Toxicology Database	ctd.mdibl.org
dbZach	Zacharewski Lab database	dbzach.fst.msu.edu/
dChip	DNA-Chip Analyzer	biosun1.harvard.edu/complab/dchip/
DDBJ	DNA Data Bank of Japan	ddbj.nig.ac.jp/
EDGE	Environment, Drugs and Gene Expression	edge.oncology.wisc.edu/edge.php
EMBL Bank	EMBL Nucleotide Sequence Database	www.ebi.ac.uk/embl/
GenBank		ncbi.nlm.nih.gov/genbank/
Genecards		genecards.org/
GEO	Gene Expression Onimbus	ncbi.nlm.nih.gov/geo/
GOBO	Global Open Biological Ontologies now called Open Biological and Biomedical Ontologies	obofoundry.org/
GO database	Gene Ontology Database	geneontology.org/
GSEA	Gene Set Enrichment Analysis	broadinstitute.org/gsea/
KEGG	Kyoto Encyclopedia of Genes and Genomes	genome.jp/kegg/
IPA	Ingenuity Pathway Analysis	

Locus Link		ncbi.nlm.nih.gov/LocusLink
MeSHer	Integrated into MeV (multi Experiment Viewer)	tm4.org/mev/
MGED Ontology		mged.sourceforge.net/ontologies/MGEDontology.php
MIAME Express	Tox MIAMExpress no longer supported	ebi.ac.uk/miamexpress/
Onto-Tools	(Onto-Express, Onto-Compare, Onto-Design, Onto-Translate and Onto-Miner)	vortex.cs.wayne.edu/projects.htm
pCEC	Profiles of Chemical Effects on Cells	project.nies.go.jp/eCA/cgi-bin/index.cgi
PIPE	Protein-Protein Interaction Prediction Engine	cgmlab.carleton.ca/PIPE2/
Prosite		expasy.org/prosite/
Pubmed		ncbi.nlm.nih.gov/pubmed/
RACE	Remote Analysis Computation for gene Expression data	race.unil.ch.
RefSeq	Reference Sequence	ncbi.nlm.nih.gov/RefSeq/
SCOP	Structural Classification of Proteins	scop.mrc-lmb.cam.ac.uk/scop/
SOFG	Standards and Ontologies for Functional Genomics	sofg.org/
SwissProt		expasy.org/sprot/
TEST	Toxicogenomics for Efficient Safety Test	istech.info/TEST/
TGP	Japanese Toxicogenomics Project database	??
TM4 software		tm4.org/
Unigene		ncbi.nlm.nih.gov/unigene

Commercial products

Name	**URL: http://**
Affymetrix	affymetrix.com/
Agilent	agilent.com
Molecular Beacons	molecular-beacons.org/
Nimblegen	nimblegen.com/
Scorpions ® Primers	premierbiosoft.com/tech_notes/Scorpion.html
SYBR® Green	appliedbiosystems.com
TaqMan Probes ®	