nature biotechnology

Rat toxicogenomic study reveals analytical consistency across microarray platforms

Lei Guo[1], Edward K Lobenhofer[2], Charles Wang[3], Richard Shippy[4], Stephen C Harris[1], Lu Zhang[5], Nan Mei[1], Tao Chen[1], Damir Herman[6], Federico M Goodsaid[7], Patrick Hurban[2], Kenneth L Phillips[2], Jun Xu[3], Xutao Deng[3], Yongming Andrew Sun[8], Weida Tong[1], Yvonne P Dragan[1] & Leming Shi[1]

To validate and extend the findings of the MicroArray Quality Control (MAQC) project, a biologically relevant toxicogenomics data set was generated using 36 RNA samples from rats treated with three chemicals (aristolochic acid, riddelliine and comfrey) and each sample was hybridized 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 real-world toxicogenomic data set reported here showed high concordance in intersite and cross-platform comparisons. Further, gene lists generated by fold-change ranking were more reproducible than those obtained by *t*-test *P* value or Significance Analysis of Microarrays. Finally, gene lists generated by fold-change ranking with a nonstringent *P*-value cutoff showed increased consistency in Gene Ontology terms and pathways, and hence the biological impact of chemical exposure could be reliably deduced from all platforms analyzed.

To validate and extend the findings of the MAQC project[1], described elsewhere in this issue, we generated a toxicogenomics data set using a rat chemical exposure study. One of the objectives of the MAQC project was to assess the reproducibility of gene expression profiling data across laboratories and platforms. Analysis of the MAQC data set shows the high reproducibility of microarray data under well-controlled conditions and further indicates that the criteria used to define differentially expressed genes can have a dramatic impact on the overlap of the resulting gene lists. In particular, lists of differentially expressed genes generated using fold change, rather than *t*-test *P* value for gene selection have been previously proposed to be more reproducible[1,2]. The two RNA samples used in the MAQC project were reference samples with no explicit biological connection: the Stratagene Universal Human Reference RNA (comprised of RNA from ten different cell lines) and Ambion Human Brain Reference RNA[1]. The availability of these data provides an invaluable resource for benchmarking laboratory performance and for testing and validating new procedures, equipment and reagents, for example. Although data from these reference samples address technical performance and reproducibility of results from microarray technology, they cannot address whether microarray data from different laboratories or platforms would result in the same biological interpretation of real-world samples. We therefore sought to apply the findings of the MAQC study to a set of experimental toxicogenomic data to validate the approach.

Several recent publications have investigated the genotoxicity of three botanical carcinogens: aristolochic acid, riddelliine and comfrey[3–6]. In the present study, 36 RNAs were isolated from the kidney and/or liver of rats exposed to one of these compounds or a control group. To corroborate the findings of the MAQC project and to determine whether the same biological interpretations would result from cross-platform comparisons, we hybridized these samples to four commercially available platforms (Affymetrix, Agilent, Applied Biosystems and GE Healthcare). To address intersite performance, we used the Affymetrix platform at two different test sites.

The results from this study are consistent with those of the MAQC project in that good concordance is found between data generated at different sites, as well as from different platforms. Furthermore, when fold-change ranking is used as the primary criterion for selecting differentially expressed genes, the overlap between gene lists from different laboratories using either the same or different platforms is high. In contrast, when a *t*-statistic (*P*-value) ranking is used as the primary criterion the cross-site or cross-platform overlap is substantially lower[1,2]. The selection criteria for differential expression can thus affect both the apparent reproducibility of microarray data, as well as

[1]National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Road, Jefferson, Arkansas 72079, USA; [2]Cogenics, A Division of Clinical Data, 100 Perimeter Park Drive, Suite C, Morrisville, North Carolina 27560, USA; [3]UCLA David Geffen School of Medicine, Transcriptional Genomics Core, Cedars-Sinai Medical Center, 8700 Beverly Boulevard, Los Angeles, California 90048, USA; [4]GE Healthcare, 7700 S. River Parkway, Suite #2603, Tempe, Arizona 85284, USA; [5]Solexa, 25861 Industrial Boulevard, Hayward, California 94545, USA; [6]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894, USA; [7]Center for Drug Evaluation and Research, US Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, Maryland 20993, USA; [8]Applied Biosystems, 850 Lincoln Centre Drive, Foster City, California 94404, USA. Correspondence should be addressed to L.G. (lei.guo@fda.hhs.gov) or L.S. (leming.shi@fda.hhs.gov).

Received 5 June; accepted 18 July; published online 8 September 2006; doi:10.1038/nbt1238

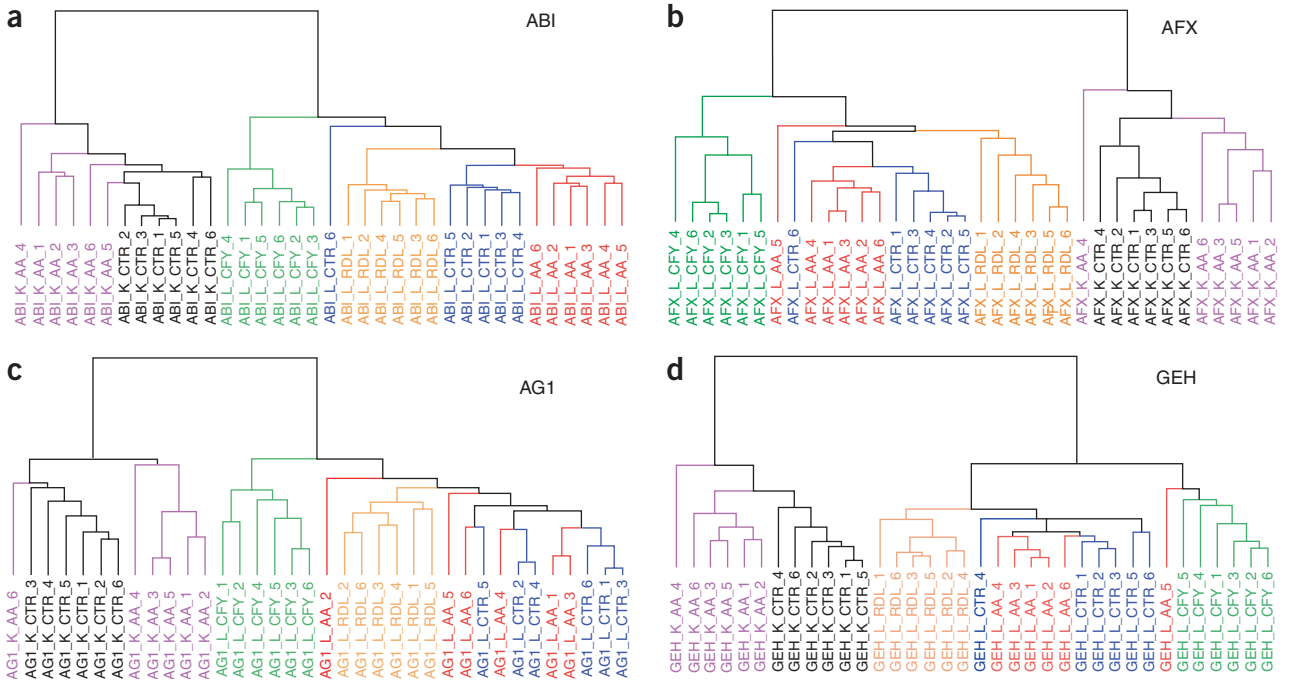


Figure 1 Hierarchical clustering of platform-specific microarray data separates samples by tissue and treatment. For each platform, the $\log_2$ intensity data from all 36 microarrays after filtering for genes flagged as below the detection level were hierarchically clustered using an average linkage algorithm and Euclidean distance as the distance metric. (**a**) Data from the Applied Biosystems platform (ABI). (**b**) Affymetrix site 1(AFX). (**c**) Agilent (AG1). (**d**) GE Healthcare (GEH). The sample labels are colored based on treatment/tissue group. Black, control kidney; purple, aristolochic acid–treated kidney; blue, control liver; red, aristolochic acid–treated liver; orange, riddelliine-treated liver; green, comfrey-treated liver.

the biological interpretation of the data. By using fold-change ranking plus a nonstringent *P*-value cutoff, the overlap of differentially expressed gene lists is increased, leading to improved agreement of the biological interpretation of the data in terms of enriched Gene Ontology (GO) nodes and pathways. Furthermore, data generated by this approach led to novel biological findings concerning chemical exposure. These findings are reproducible across laboratories and platforms when the preferred gene selection criteria are used. Together, these results further support the findings of the MAQC project, highlight the importance of appropriate data analysis procedures and demonstrate that microarray data generated from different platforms not only result in similar biological interpretation, but also reveal novel findings.

RESULTS

RNA was isolated from the target organs of rats exposed to aristolochic acid, riddelliine or comfrey, from studies that have been detailed previously[3–6]. In total there were six treatment/tissue groups: kidney from aristolochic acid–treated rats, kidney from vehicle control, liver from aristolochic acid–treated rats, liver from riddelliine-treated rats, liver from comfrey-treated rats and liver from vehicle control. Within each treatment/tissue group there were six biological replicates. Aliquots of these samples were prepared and distributed to each of the test sites for gene expression profiling using microarrays from four different platforms. Laboratory procedures were identical to those in the MAQC project[1]. Unless otherwise stated, the platform manufacturer's recommendations were used for data processing.

Hierarchical clustering analysis

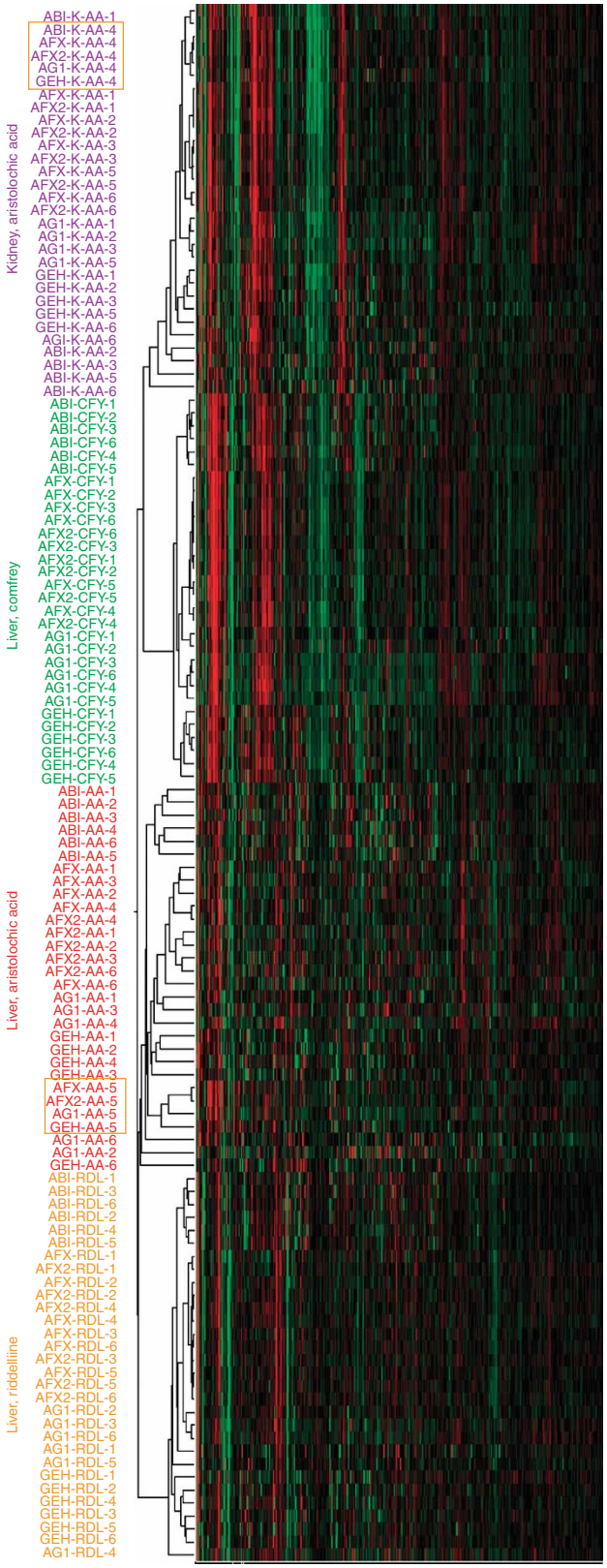
To assess the overall reproducibility of microarray data from the four platforms, we performed hierarchical clustering analyses for each platform. Within each platform, samples were largely clustered first by tissue type and then by treatment (**Fig. 1**). Within each platform there are individual samples that did not cluster with the other members of their respective treatment/tissue group; however, the only sample that was consistently different across all platforms was sample no. 4 from the aristolochic acid–treated kidney

Table 1 Average Pearson correlation coefficients of $\log_2$-normalized intensity data for each treatment/tissue group

Test site	No. of Probe(set)s	Aristolochic acid kidney[a]	Control kidney	Aristolochic acid liver	Comfrey liver	Control liver	Riddelliine liver
Applied Biosystems (ABI)	26,857	0.9586 (0.9623)	0.9742	0.9636	0.9737	0.9634	0.9705
Affymetrix no. 1 (AFX)	31,099	0.9748 (0.9828)	0.9881	0.9871	0.9861	0.9876	0.9867
Affymetrix no. 2 (AFX2)	31,099	0.9736 (0.9818)	0.9879	0.9860	0.9827	0.9862	0.9836
Agilent (AG1)	41,071	0.9610 (0.9711)	0.9701	0.9642	0.9659	0.9740	0.9675
GE Healthcare (GEH)	35,129	0.9697 (0.9739)	0.9761	0.9690	0.9690	0.9687	0.9734

[a]Numbers in parentheses represent data after excluding sample no. 4.

group. Similar results were obtained using principal components analysis (**Supplementary Fig. 1** online). DNA adduct data indicate that this sample had 50% fewer DNA adducts compared to the other five animals in the same treatment group (Mei, N. *et al.*, unpublished data), suggesting that the consistent failure of this sample to cluster with its treatment/tissue group may be biologically based. It was also determined that aristolochic acid–treated liver samples showed a relatively small difference in expression profiles when compared to their tissue-matched control group. This result was reproduced across all platforms and is consistent with previous observations that kidney, not liver, is the target organ of aristolochic acid–mediated carcinogenesis[7].



The reproducibility of the microarray data was further explored by calculating the Pearson correlation coefficients of the $\log_2$ intensity data for all pair-wise sample comparisons within a treatment/tissue group for each platform. **Table 1** shows the average correlation of biological replicates within each treatment/tissue group for each platform and further demonstrates the high degree of similarity of these data. Because of the presence of an animal that had a diminished treatment response, the aristolochic acid–treated kidney group had a significantly lower correlation, as expected, compared to other groups (e.g., $P = 0.0024$, two-sided, paired *t*-test compared to the control kidney group). Removal of sample no. 4 from the aristolochic acid–treated kidney group resulted in a less significant difference ($P = 0.085$, two-sided, paired *t*-test compared to the control kidney group). These data coupled with the DNA adduct data consistently indicate that sample no. 4 from the aristolochic acid–treated kidney group has a different response relative to the other group members. Therefore, for the assessment of cross-platform data consistency, the data from this sample have been excluded.

Overlap of differentially expressed gene lists across sites

One of the fundamental goals of a gene expression profiling experiment is to identify those genes that are differentially expressed within the system being studied. There are a large number of methods for selecting such genes, and ultimately, the genes that are identified have a fundamental impact on the biological interpretation of the data. Therefore, this toxicogenomics study was used to validate the findings in regard to gene selection methods by employing different selection criteria and determining the percentage of overlap between different laboratories or platforms[1,2]. The overlap across the two sites that generated data using the Affymetrix platform is high (85–90%) when the genes (from a few up to ~2,000) are selected by rank ordering the genes based on fold change (**Supplementary Fig. 2** online). As more genes are considered differentially expressed (that is, moving to the right on the *x*-axis) the percentage of overlap begins to decline because of the inclusion of more genes demonstrating smaller fold changes, which are less likely to be reproducible across sites. There is a small decrease in the overlap when a *P*-value cutoff of 0.01 or 0.05 is applied to the fold change–based,

Figure 2 Hierarchical clustering of all individual sample data from all microarray platforms separated by tissue and treatment. Within each platform/site, a fold change was calculated and $\log_2$ transformed for all 5,112 common genes that did not have any missing values ($n = 4,609$) for each of the 24-treated individual samples compared to a tissue-match control. These values were then hierarchically clustered using Euclidean distance metric and average linkage. Each row represents the results from an individual treated animal assayed on a particular platform. Each row is labeled with a platform designation first, followed by the organ assayed for kidney samples, and then the treatment and unique animal identifier (1–6). ABI, Applied Biosystems platform; AFX, Affymetrix site 1; AFX2, Affymetrix site 2; AG1, Agilent; and GEH, GE Healthcare. K, kidney; AA, aristolochic acid; RDL, riddelliine; and CFY, comfrey. The yellow boxes highlight areas in which replicates of the same sample across all multiple platform and/or sites have clustered together.

gene-selection methods. This results from the *P*-value threshold altering the composition of the total list of genes such that each test site has a different list of genes to begin with in the gene selection process, thereby increasing the intersite inconsistency.

Supplementary Figure 2 also illustrates the overlap when genes are selected based on *P*-value rank ordering alone or with a fold-change criterion of 2.0 or 1.4. For *P* value–based gene-selection methods, the overlap gradually increases as the number of differentially expressed genes increases. An increase in the overlap is also observed when a fold-change cutoff of 1.4 or 2.0 is applied in conjunction with the *P*-value criterion. This is understandable since the larger fold changes are more easily reproduced than smaller ones.

The impact of different normalization methods on the overlap of gene lists was also assessed by comparing the overlap of gene lists derived from two normalization methods using the same gene selection method on the same sample pair comparison from data generated at the same test site (**Supplementary Fig. 3** online). When *P* value is used as the criterion for gene selection, the overlap from different normalization methods is relatively low. However, when genes are ranked and selected based on fold change with or without a *P*-value cutoff, the overlap between different normalization methods is very high (>90%). Furthermore, global scaling methods do not alter the rank order of genes based on fold change (hence the gene lists); therefore, the overlap between raw, mean-, or median-scaled data is 100% when using the fold change for ranking and selecting genes. However, these scaling factors can affect the magnitude of the fold changes and the *P* values and thus will only affect the gene list when a *P*-value criterion is involved in gene selection. Our results are consistent with those reported elsewhere[8].

In addition to the standard *t*-test, numerous different statistical tests have been used for the identification of differentially expressed genes[9]. One commonly used method is Significance Analysis of Microarrays (SAM)[10]. **Supplementary Figure 4** online illustrates the intersite concordance results of differentially expressed genes selected based on fold-change ranking, SAM, *t*-test and random selection when the data from the comfrey-treated liver samples are compared to their corresponding controls. The site-site concordance based on SAM was clearly improved over that based on a simple *t*-test, but did not achieve the same level of concordance as that reached based on fold-change ranking. Similar results were obtained when other sample pairs or cross-platform data were analyzed in the same manner (data not shown). Cumulatively, these results illustrate that fold change–based

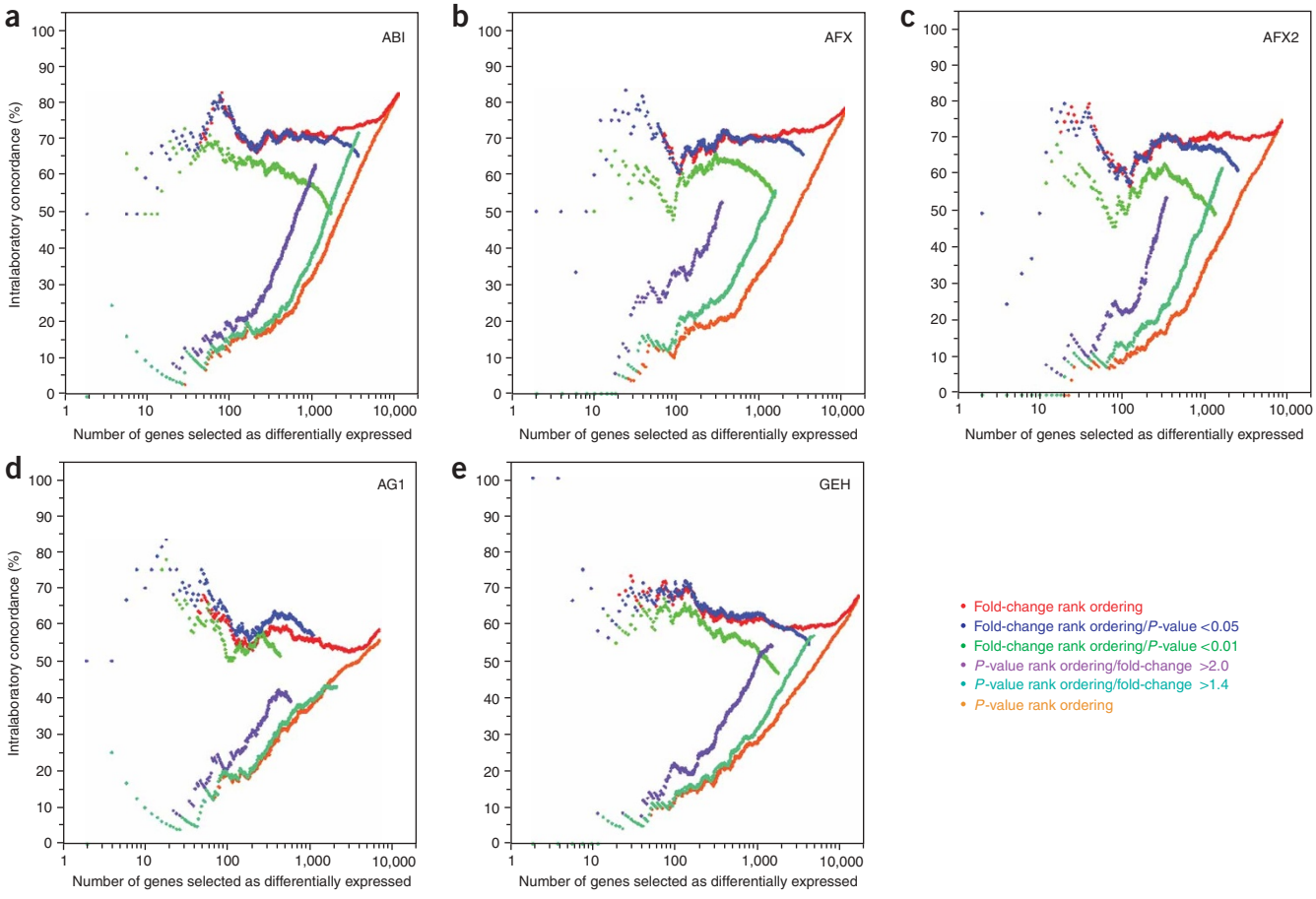


Figure 3 Intralaboratory overlap of differentially expressed gene lists generated using different selection criteria. For each platform, the liver control and comfrey treatment groups were equally and randomly divided into two experiments and the differentially expressed genes were identified independently from the two experiments using different gene selection criteria. Differentially expressed genes were selected from a subset of genes that are detectable by both experiments. The *x*-axis represents the number of genes selected as differentially expressed, and the *y*-axis represents the overlap (%) of two gene lists for a given number of differentially expressed genes. Each line on the graph represents the intralaboratory overlap of differentially expressed gene lists based on one of six different gene ranking/selection methods. Red, fold-change rank ordering only; orange, *P*-value rank ordering only; light green, fold-change rank ordering and $P < 0.01$; blue, fold-change rank ordering and $P < 0.05$; teal, *P*-value rank ordering and fold change >1.4; and purple, *P*-value rank ordering and fold change >2.0. (**a**) Applied Biosystems (ABI). (**b**) Affymetrix site 1 (AFX). (**c**) Affymetrix site 2 (AFX2). (**d**) Agilent (AG1). (**e**) GE Healthcare (GEH).

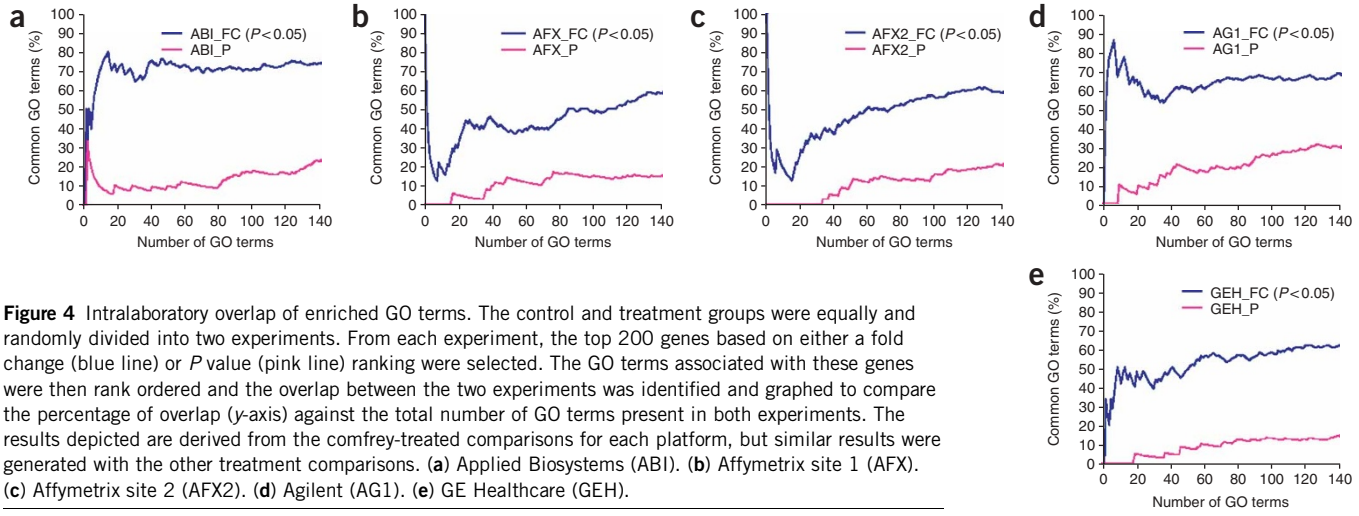


Figure 4 Intralaboratory overlap of enriched GO terms. The control and treatment groups were equally and randomly divided into two experiments. From each experiment, the top 200 genes based on either a fold change (blue line) or *P* value (pink line) ranking were selected. The GO terms associated with these genes were then rank ordered and the overlap between the two experiments was identified and graphed to compare the percentage of overlap (*y*-axis) against the total number of GO terms present in both experiments. The results depicted are derived from the comfrey-treated comparisons for each platform, but similar results were generated with the other treatment comparisons. (**a**) Applied Biosystems (ABI). (**b**) Affymetrix site 1 (AFX). (**c**) Affymetrix site 2 (AFX2). (**d**) Agilent (AG1). (**e**) GE Healthcare (GEH).

selection methods usually offer a higher level of consistency of lists of differentially expressed genes.

Overlap of differentially expressed gene lists across platforms

To assess the reproducibility of data across multiple microarray platforms, we identified the list of genes that was measured by all four of the microarray platforms using the March 2006 version of the RefSeq database and the methods described by the MAQC project[1]. This resulted in the identification of 5,112 common genes, which were used in all subsequent cross-platform comparisons. Consistent with results from intersite comparisons (**Supplementary Fig. 2** online), the cross-platform data comparisons reveal the same trends. Specifically, the percentage of overlap for differentially expressed gene lists is highest when fold change–based gene selection methods are used (**Supplementary Fig. 5a** online). Not surprisingly, the cross-platform overlap is higher (~80%) in all instances when genes that are not reproducibly detected on the microarrays are omitted (e.g., those probes that are flagged as 'not present') (**Supplementary Fig. 5b** online). These results combined with intersite results further corroborate the findings of the MAQC project that fold change–based selection criteria for differentially expressed genes generate more reproducible results[1,2]. No measure of sensitivity or specificity of the approach was included in the analysis.

Within each platform/site, the fold change was calculated for all 5,112 common genes that did not have any missing values ($n = 4{,}609$) for each of the 24-treated individual samples compared to a tissue-match control and these values were then hierarchically clustered (**Fig. 2**). The resulting dendrogram illustrates that the samples are separated by tissue and then by treatment. Each of the four major branches of the dendrogram contain all of the biological replicate data for a given treatment/tissue group regardless of the site or platform that was used to generate the data. Within each of these branches, the platform as opposed to the biological replicate is the next major division. There are a few notable exceptions to this observation. When the same platform is performed at different test sites, the replicates of the same sample assayed at different sites cluster more closely together. In a few instances, the results from multiple different platforms for the same biological sample cluster together (e.g., aristolochic acid–treated liver sample no. 5). Because no gene selection criteria were used to generate this visualization, these results further indicate that interlaboratory and cross-platform data are highly reproducible.

Agreement of biological interpretation with GO and pathways

Typically, a microarray-based experiment is performed in a single laboratory using a single platform. Furthermore, it is relatively common to use three biological replicates in a toxicogenomic study when multiple groups of samples are involved. To explore whether or not a similar biological response was obtained when comparing results within a given laboratory, we generated data from six biological replicates. The control and treatment groups were then equally and randomly divided into two artificial experiments. Consistent with the interlaboratory and cross-platform results, the overlap of differentially expressed genes using different gene selection criteria from the intra-laboratory results revealed the same trend, namely that fold change–based selection criteria generate more reproducible results (**Fig. 3**). For each of the ABI, AFX and AFX2 intralaboratory comparisons, the overlap of gene lists was almost identical with or without a *P* cutoff (<0.05) for up to ~1,000 genes selected as differentially expressed; for AG1 and GEH, the use of a *P* cutoff (<0.05) slightly increased the overlap of gene lists. However, the use of a more stringent *P* cutoff (<0.01) decreased the overlap of gene lists. These intralaboratory comparison results are consistent with those of interlaboratory comparisons (**Supplementary Fig. 2** online). Therefore, a modest *P* cutoff (<0.05) appeared to be reasonable for data sets of this small sample size (3). Furthermore, the use of a fold-change threshold increased the overlap of gene lists derived from *P*-value ranking; a more stringent fold-change threshold leads to higher overlap of gene lists (**Fig. 3** and **Supplementary Fig. 2** online).

The differences in overlap of gene lists based on selection criteria were further investigated by assessing the impact on the associated GO terms. From each artificial experiment, the top 200 genes based on either a fold-change (with $P < 0.05$ cutoff) or *P*-value ranking were selected. The *P* value from the Fisher's exact test was calculated for each GO term associated with these genes. For each artificial experiment, the GO terms were then rank-ordered based on the *P* value. The overlap between the two artificial experiments was determined by dividing the number of GO terms commonly meeting a *P*-value ranking criterion in both of the artificial experiments by the total number of GO terms meeting the *P*-value criterion for either experiment. **Figure 4** illustrates the percentage of overlapping GO terms plotted against a defined number of the highest ranking GO terms from both experiments. Clearly, the overlap of GO terms was much higher when genes are selected by fold change compared

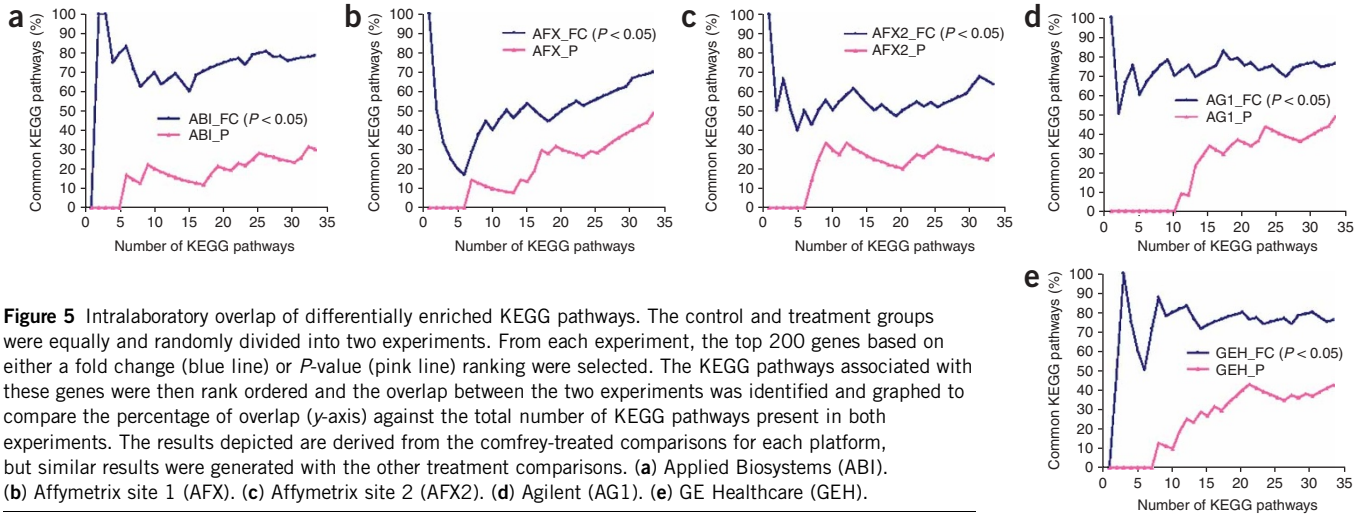


Figure 5 Intralaboratory overlap of differentially enriched KEGG pathways. The control and treatment groups were equally and randomly divided into two experiments. From each experiment, the top 200 genes based on either a fold change (blue line) or *P*-value (pink line) ranking were selected. The KEGG pathways associated with these genes were then rank ordered and the overlap between the two experiments was identified and graphed to compare the percentage of overlap (*y*-axis) against the total number of KEGG pathways present in both experiments. The results depicted are derived from the comfrey-treated comparisons for each platform, but similar results were generated with the other treatment comparisons. (**a**) Applied Biosystems (ABI). (**b**) Affymetrix site 1 (AFX). (**c**) Affymetrix site 2 (AFX2). (**d**) Agilent (AG1). (**e**) GE Healthcare (GEH).

to those selected by *P* value. Similar results were obtained when the gene lists are mapped to KEGG pathways (**Fig. 5**) or other pathway databases (e.g., Ingenuity) (data not shown). These results clearly show that common biological responses are evident when genes are selected by criteria that lead to reproducible gene lists. Nonoverlapping lists of differentially expressed genes generally lead to inconsistent biological interpretation of microarray results in terms of GO terms and pathways.

Agreement of biological response

To further explore the agreement of biological response across the microarray platforms, we combined data from the cross-platform common gene list (5,112 genes) and from the six comfrey-treated liver samples and compared them to the data from the six control liver samples for each platform. A *t*-test was performed and genes with $P < 0.05$ were identified. This filtered gene set was then rank ordered by fold change and for each platform the top 250 up- and down-regulated genes were selected, generating a list of the top 500 differentially expressed genes for each of the five platform/site combinations (the overlap in genes between the gene lists for any two platforms is >70%).

A GO enrichment analysis was performed for each platform by comparing the content of the top 500 differentially expressed genes to the content of the 5,112 common gene list using a Fisher's Exact Test in GoMiner[11,12], resulting in an enrichment *P* for each GO term. A comparison of *P* values across platforms identified 101 nodes that were significantly over- or underenriched ($P < 0.05$) in at least four of five platforms, with nearly 60% of these terms being significant in all five platforms. Inspection of these enriched categories confirmed that the different microarray platforms were reporting the same biological responses in these samples, and also provided novel insight into the effects of comfrey exposure.

Comfrey is a perennial plant that has been widely used for over 2,000 years as an herbal medicine for a wide variety of ailments. However, comfrey has been shown to be both genotoxic and hepatotoxic[13]. The exact molecular mechanism underlying these toxicities is not fully understood, but is known to be associated with the pyrrolizidine alkaloids present in comfrey, which can be metabolically activated and bind to DNA[6,14]. Considering that there are >350 different pyrrolizidine alkaloids found in over 6,000 different species[15], it has been suggested that pyrrolizidine alkaloids are "probably the most common poisonous plant constituents that poison livestock, wildlife, and humans, worldwide[14]." Examination of the 101 significant GO terms revealed at least two that were noteworthy: copper ion homeostasis (GO:0006878) and vitamin A metabolism (GO:0006776). Dietary or medicinal exposure to several pyrrolizidine alkaloid–containing plants has been shown to result in decreased levels of vitamin A in the liver and increased liver levels of copper[16–18], but there is no indication that these effects have been observed in response to comfrey exposure. These results suggest that comfrey influences copper and vitamin A levels similar to other pyrrolizidine alkaloid–containing plants. Furthermore, these data are the first indication that changes in liver vitamin A and copper levels in response to pyrrolizidine alkaloid–exposure are transcriptionally regulated. Interestingly, only four genes associated with copper ion homeostasis are present in the common gene list and in all instances each platform identified two of these genes as significantly upregulated: amyloid beta (A4) precursor protein (*APP*) and prion protein (*PRNP*). Previously, both of these genes were shown to bind copper and were shown to be upregulated in response to chronic copper exposure[19–21]. Cumulatively, these findings indicate that comfrey, like several other pyrrolizidine alkaloid–containing plants, may affect liver levels of vitamin A and copper. Importantly, these data demonstrate that different microarray platforms can consistently report novel biological findings at the level of biological processes and of individual genes.

DISCUSSION

In this study, a data set was created that could validate and extend the findings of the MAQC project by focusing on a biologically relevant set of samples. Specifically, a large toxicogenomics data set was generated using 36 RNA samples from rats treated with three chemicals and four commercial microarray platforms to investigate the agreement of intersite and cross-platform gene lists.

When a few or up to 2,000 genes are selected as differentially expressed from different sites using the same microarray platform, the percentage of overlap is ~85% based on a fold-change criterion for gene ranking and selection (**Supplementary Fig. 2** online). A lower percentage of overlap is observed using *P* as the criterion for gene ranking and selection, in particular when fewer genes are selected as differentially expressed. This same trend was also observed when gene selection methods were compared across platforms using the subset of 5,112 common genes (**Supplementary Fig. 5** online). In addition,

concordance offered by the widely used SAM approach did not achieve the same high level of concordance generated by fold-change ranking (**Supplementary Fig. 4** online). These results are also consistent with those based on MAQC human samples and highlight the problems with commonly used gene selection methods that are solely based on *t*-test *P* values[1,2]. As expected, the degree of overlap of gene lists directly affects the ability to consistently identify the same biological response in regard to GO terms (**Fig. 4**) and KEGG pathways (**Fig. 5**). Therefore, to ensure reproducible biological interpretation of microarray results, it is important that criteria for generating lists of differentially expressed genes are selected properly.

The lack of overlap of lists of differentially expressed genes selected using a *P*-value criterion may be explained by the fact that fold change is calculated by comparing signal intensity for a given gene as directly measured using a microarray, whereas the *P*-value calculation incorporates the signal-to-noise ratio. Therefore, if the signal intensity for the gene is more reproducible across laboratories or platforms than the associated noise level, this would result in the finding that fold change–based, gene-selection methods are more reproducible. However, the impact of the proposed analysis method on two other parameters, sensitivity and specificity, will also have to be assessed before any final conclusions can be drawn regarding the generalizability of this approach.

Sample size is another important factor that impacts concordance of lists of differentially expressed genes. It is interesting to compare the results of **Figure 3** (AFX and AFX2) with those of **Supplementary Figure 2b** online in which for the same microarray platform, one can observe an overall increased level in the overlap of differentially expressed genes when six replicates from different laboratories are compared as opposed to the three replicates from within the same laboratory. This increase is observed despite the potential for interlaboratory variation, which would affect the six-replicate comparisons but not comparisons of three replicates within a laboratory. This demonstrates the relationship between increases in statistical power and the resulting gain in reliable detection of differential expression that occurs with increased sample sizes. It is worth noting that differences between individual biological replicates also contribute to the relatively lower overlap observed in **Figure 3**.

To illustrate the importance of using gene selection criteria that maximize overlap of gene lists, we first filtered the data (comfrey compared to control) using a relatively nonstringent *P* cutoff (<0.05) and then the remaining genes were rank ordered using fold change. By selecting the top 250 up- and downregulated genes from each platform and performing a GO enrichment analysis, not only was the cross-platform reproducibility of GO terms demonstrated, but a novel biological finding was also revealed on all platforms and at all sites. Specifically, comfrey, like several other pyrrolizidine alkaloid–containing plants, affects liver levels of vitamin A and copper; furthermore, these changes are, at least in part, transcriptionally regulated.

Microarray technology has had a profound impact on biological research partially from its ability to identify differentially expressed genes that may be used to develop potential biomarkers, elucidate molecular mechanisms and group similar samples based on gene signatures. Therefore, the reproducibility and reliability of the data from a study and the choice of methods that lead to the identification of concordant lists of differentially expressed genes are critical for biological interpretation. Concerns have been raised regarding the reliability of microarray results due to the apparent lack of overlap of the lists of differentially expressed genes[22–28]. The results from this study suggest that the disappointingly low concordance reported in some earlier publications can be attributed in large part to the practice of deriving differentially expressed gene lists based on the ranking of genes solely by a statistical significance measure. Furthermore, these results demonstrate that microarray data generated from different platforms can not only result in a similar biological interpretation, but also reveal novel findings.

METHODS

Microarray processing. Details on the description of the *in vivo* portion of this study has been described[3–6]. Briefly, groups of six 6-week-old Big Blue rats were gavaged with riddelliine (1 mg/kg body weight) or aristolochic acid (10 mg/kg body weight) five times a week for 12 weeks or Big Blue rats were fed a diet of 8% comfrey roots for 12 weeks. The animals were sacrificed after 12 weeks of treatment, and the tissues were isolated, frozen quickly in liquid nitrogen and stored at –80 °C. RNA was isolated from tissues of rats that had been exposed to aristolochic acid (liver and kidney), riddelliine (liver), comfrey (liver) or a control group (liver and kidney). There were six biological replicates for each treatment/tissue group for a total of 36 samples. The samples were randomly labeled and each test site was provided an aliquot of each sample. To avoid potential confounding factors in experimental implementation, the identity of the RNA samples was kept unknown to the test sites before data were submitted to FDA/NCTR. The sample ID, RNA Integrity Number, OD ratio, microarray ID and data file names are provided in the **Supplementary Table 1** online.

Each of the RNA samples was labeled and hybridized to a microarray from one of four commercial platforms: Affymetrix (Rat Genome 230 2.0), Agilent (Whole Rat Genome Oligo Microarray, G4131A), Applied Biosystems (Rat Genome Survey Microarray) and GE Healthcare (Rat Whole Genome Bioarray, 300031). Except for Affymetrix, which was performed at two independent test sites, each platform was used at one single test site with 36 microarrays using biological replicate RNA samples. The labeling and hybridizations were performed according to the manufacturer's recommendation using methods detailed in the MAQC project[1].

Data analysis. Unless otherwise stated, the manufacturer's recommended normalization methods were used: quantile normalization for Applied Biosystems, PLIER with an offset value of 16 for Affymetrix and median-scaling for both Agilent and GE Healthcare[1]. To assess the impact of normalization methods on microarray results, we compared a limited number of commonly used normalization methods: raw, mean, median and quantile (**Supplementary Fig. 3** online). The toxicogenomics data set generated in this study has also been used for the evaluation of microarray assay performance based on external RNA controls[29].

Six different gene selection methods were used: (i) fold-change rank ordering only, (ii) fold-change rank ordering and $P < 0.01$, (iii) fold-change rank ordering and *P*-value cutoff < 0.05, (iv) *t*-test *P* value (assuming equal variance) rank ordering only, (v) *P*-value rank ordering and fold change >1.4, (vi) *P*-value rank ordering and fold change >2.0. The percentage of overlapping genes from these differentially expressed gene lists was then calculated in the same way as was described elsewhere[1].

ArrayTrack[30] was used for GO and KEGG pathway mapping, whereas GO enrichment analyses were performed using High Throughput GoMiner[11,12].

Cross-platform sequence mapping to RefSeq. Probe sequences from each microarray platform were mapped onto the NCBI-curated rat RefSeq database from March 2006. The same mapping criteria as reported for the main MAQC study was used[1]. The primary mapping criterion is a perfect match between a probe sequence and the target transcript sequence: a probe perfectly matches a transcript provided that a completely homologous sequence of length equal to the probe length is found anywhere on the transcript. The only exception to this rule is from the Affymetrix platform in which a ProbeSet is considered a perfect match to a transcript as long as 80% of probes within the ProbeSet (usually nine out of 11) perfectly match the same transcript. To simplify the cross-platform data analysis, a mapping table was generated with one probe per gene. Consistent with the MAQC main study[1], if more than one probe from a platform perfectly matches the same gene, the probe closest to the 3′ UTR was considered, resulting in 5,204 common non-model RefSeq mRNAs (NMs) mapped across 5,112 common genes (**Supplementary Table 2** online).

Accession numbers. All data are available through GEO (series accession number: GSE5350), ArrayExpress (accession number: E-TABM-132), ArrayTrack (http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/), and the MAQC web site (http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

E.K.L., K.L.P. and P.H. acknowledge Agilent Technologies, Inc. and Affymetrix, Inc. for their material contributions to this work, thank John Pufky, Stephen Burgin and Jennifer Troehler for their outstanding technical assistance, and gratefully acknowledge the Advanced Technology Program of the National Institute of Standards and Technology, whose generous support provided partial funding of this research (70NANB2H3009). C.W. acknowledges Affymetrix, Inc. for material contributions to this work. R.S. acknowledges technical support of Alan Brunner for generating GE Healthcare microarray data. L.G. and L.S. thank X. Megan Cao, Stacey Dial, Carrie Moland and Feng Qian for their superb technical assistance.

DISCLAIMER

This work includes contributions from, and was reviewed by, the FDA and the NIH. Certain commercial materials and equipment are identified in order to adequately specify experimental procedures. In no case does such identification imply recommendation or endorsement by the FDA or the NIH, nor does it imply that the items identified are necessarily the best available for the purpose.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Biotechnology* website for details).

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24**, 1151–1161 (2006).
2. Shi, L. *et al.* Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics* **6** (Suppl 2), S12 (2005).
3. Chen, L., Mei, N., Yao, L. & Chen, T. Mutations induced by carcinogenic doses of aristolochic acid in kidney of Big Blue transgenic rats. *Toxicol. Lett.* **165**, 250–256 (2006).
4. Mei, N., Chou, M.W., Fu, P.P., Heflich, R.H. & Chen, T. Differential mutagenicity of riddelliine in liver endothelial and parenchymal cells of transgenic Big Blue rats. *Cancer Lett.* **215**, 151–158 (2004).
5. Mei, N., Heflich, R.H., Chou, M.W. & Chen, T. Mutations induced by the carcinogenic pyrrolizidine alkaloid riddelliine in the liver *cII* gene of transgenic Big Blue rats. *Chem. Res. Toxicol.* **17**, 814–818 (2004).
6. Mei, N., Guo, L., Fu, P.P., Heflich, R.H. & Chen, T. Mutagenicity of comfrey (Symphytum Officinale) in rat liver. *Br. J. Cancer* **92**, 873–875 (2005).
7. Arlt, V.M., Stiborova, M. & Schmeiser, H.H. Aristolochic acid as a probable human cancer hazard in herbal remedies: a review. *Mutagenesis* **17**, 265–277 (2002).
8. Patterson, T.A. *et al.* Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. *Nat. Biotechnol.* **24**, 1140–1150 (2006).
9. Allison, D.B., Cui, X., Page, G.P. & Sabripour, M. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* **7**, 55–65 (2006).
10. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
11. Zeeberg, B.R. *et al.* GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* **4**, R28 (2003).
12. Zeeberg, B.R. *et al.* High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics* **6**, 168 (2005).
13. Stickel, F. & Seitz, H.K. The efficacy and safety of comfrey. *Public Health Nutr.* **3**, 501–508 (2000).
14. Fu, P.P., Xia, Q., Lin, G. & Chou, M.W. Pyrrolizidine alkaloids–genotoxicity, metabolism enzymes, metabolic activation, and mechanisms. *Drug Metab. Rev.* **36**, 1–55 (2004).
15. Betz, J.M., Eppley, R.M., Taylor, W.C. & Andrzejewski, D. Determination of pyrrolizidine alkaloids in commercial comfrey products (*Symphytum* sp.). *J. Pharm. Sci.* **83**, 649–653 (1994).
16. Cheeke, P.R. Toxicity and metabolism of pyrrolizidine alkaloids. *J. Anim. Sci.* **66**, 2343–2350 (1988).
17. Huan, J. *et al.* Dietary pyrrolizidine (Senecio) alkaloids and tissue distribution of copper and vitamin A in broiler chickens. *Toxicol. Lett.* **62**, 139–153 (1992).
18. Moghaddam, M.F. & Cheeke, P.R. Effects of dietary pyrrolizidine (Senecio) alkaloids on vitamin A metabolism in rats. *Toxicol. Lett.* **45**, 149–156 (1989).
19. Armendariz, A.D., Gonzalez, M., Loguinov, A.V. & Vulpe, C.D. Gene expression profiling in chronic copper overload reveals upregulation of Prnp and App. *Physiol. Genomics* **20**, 45–54 (2004).
20. Hesse, L., Beher, D., Masters, C.L. & Multhaup, G. The beta A4 amyloid precursor protein binding to copper. *FEBS Lett.* **349**, 109–116 (1994).
21. Varela-Nallar, L., Toledo, E.M., Chacon, M.A. & Inestrosa, N.C. The functional links between prion protein and copper. *Biol. Res.* **39**, 39–44 (2006).
22. Tan, P.K. *et al.* Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* **31**, 5676–5684 (2003).
23. Ramalho-Santos, M., Yoon, S., Matsuzaki, Y., Mulligan, R.C. & Melton, D.A. "Stemness": transcriptional profiling of embryonic and adult stem cells. *Science* **298**, 597–600 (2002).
24. Ivanova, N.B. *et al.* A stem cell molecular signature. *Science* **298**, 601–604 (2002).
25. Fortunel, N.O. *et al.* Comment on " 'Stemness': transcriptional profiling of embryonic and adult stem cells" and "a stem cell molecular signature". *Science* **302**, 393 author reply 393 (2003).
26. Marshall, E. Getting the noise out of gene arrays. *Science* **306**, 630–631 (2004).
27. Miller, R.M. *et al.* Dysregulation of gene expression in the 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine-lesioned mouse substantia nigra. *J. Neurosci.* **24**, 7445–7454 (2004).
28. Frantz, S. An array of problems. *Nat. Rev. Drug Discov.* **4**, 362–363 (2005).
29. Tong, W. *et al.* Evaluation of external RNA controls for the assessment of microarray performance. *Nat. Biotechnol.* **24**, 1132–1139 (2006).
30. Tong, W. *et al.* ArrayTrack–supporting toxicogenomic research at the US Food and Drug Administration National Center for Toxicological Research. *Environ. Health Perspect.* **111**, 1819–1826 (2003).