nature
biotechnology

The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance

Charles Wang[1,27], Binsheng Gong[2,27], Pierre R Bushel[3,4,27], Jean Thierry-Mieg[5], Danielle Thierry-Mieg[5], Joshua Xu[2], Hong Fang[6], Huixiao Hong[2], Jie Shen[2], Zhenqiang Su[2], Joe Meehan[2], Xiaojin Li[7], Lu Yang[7], Haiqing Li[7], Paweł P Łabaj[8], David P Kreil[8,9], Dalila Megherbi[10], Stan Gaj[11], Florian Caiment[11], Joost van Delft[11], Jos Kleinjans[11], Andreas Scherer[12], Viswanath Devanarayan[13], Jian Wang[14], Yong Yang[14], Hui-Rong Qian[14], Lee J Lancashire[15], Marina Bessarabova[15], Yuri Nikolsky[16], Cesare Furlanello[17], Marco Chierici[17], Davide Albanese[17,18], Giuseppe Jurman[17], Samantha Riccadonna[17,18], Michele Filosi[17], Roberto Visintainer[17], Ke K Zhang[19], Jianying Li[3,20], Jui-Hua Hsieh[21], Daniel L Svoboda[22], James C Fuscoe[23], Youping Deng[24], Leming Shi[2,25], Richard S Paules[26], Scott S Auerbach[21] & Weida Tong[2]

The concordance of RNA-sequencing (RNA-seq) with microarrays for genome-wide analysis of differential gene expression has not been rigorously assessed using a range of chemical treatment conditions. Here we use a comprehensive study design to generate Illumina RNA-seq and Affymetrix microarray data from the same liver samples of rats exposed in triplicate to varying degrees of perturbation by 27 chemicals representing multiple modes of action (MOAs). The cross-platform concordance in terms of differentially expressed genes (DEGs) or enriched pathways is linearly correlated with treatment effect size ($R^2 \approx 0.8$). Furthermore, the concordance is also affected by transcript abundance and biological complexity of the MOA. RNA-seq outperforms microarray (93% versus 75%) in DEG verification as assessed by quantitative PCR, with the gain mainly due to its improved accuracy for low-abundance transcripts. Nonetheless, classifiers to predict MOAs perform similarly when developed using data from either platform. Therefore, the endpoint studied and its biological complexity, transcript abundance and the genomic application are important factors in transcriptomic research and for clinical and regulatory decision making.

Emerging technologies facilitate basic science research, but their value in clinical and regulatory settings requires rigorous assessment and consensus within the research community. The US Food and Drug Administration (FDA)'s initiative on advancing regulatory science embraces collaborations among various stakeholders to expedite translation of advancement in basic science to regulatory application[1]. In the past decade, microarray has been one of the principal technologies for analyzing transcriptomes to support drug development and safety evaluation[2]. The FDA launched the community-wide MicroArray Quality Control (MAQC) Consortium to investigate the reliability and utility of microarrays in identifying DEGs and predicting patient or toxicity outcomes based on gene-expression data in the first (MAQC-I)[3,4] and second (MAQC-II)[5,6] phases of the project, respectively. MAQC-I and MAQC-II demonstrated the critical roles of a comprehensive

[1]Center for Genomics and Division of Microbiology & Molecular Genetics, School of Medicine, Loma Linda University, Loma Linda, California, USA. [2]Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas, USA. [3]Microarray and Genome Informatics Group, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA. [4]Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA. [5]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA. [6]The Office of Scientific Coordination, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas, USA. [7]Functional Genomics Core, Department of Molecular Medicine, Beckman Research Institute, City of Hope, Duarte, California, USA. [8]Chair of Bioinformatics Research Group, Boku University Vienna, Vienna, Austria. [9]University of Warwick, Coventry, UK. [10]CMINDS Research Center, Department of Electrical and Computer Engineering, Francis College of Engineering, University of Massachusetts, Lowell, Massachusetts, USA. [11]Department of Toxicogenomics, Maastricht University, Maastricht, the Netherlands. [12]Australian Genome Research Facility Ltd., The Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia. [13]AbbVie, Inc., North Chicago, Illinois, USA. [14]Research Informatics and Statistics, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, Indiana, USA. [15]Thomson Reuters, IP & Science, Carlsbad, California, USA. [16]Vavilov Institute of General Genetics, Russian Academy of Science, Moscow, Russia. [17]Fondazione Bruno Kessler, Trento, Italy. [18]Computational Biology Department, Research and Innovation Centre, Fondazione Edmund Mach (FEM), San Michele all'Adige, Italy. [19]Bioinformatics core, Department of Pathology, University of North Dakota, Grand Forks, North Dakota, USA. [20]Kelly Government Solutions, Inc., Durham, North Carolina, USA. [21]Biomolecular Screening Branch, Division of the National Toxicology Program, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA. [22]SRA International, Durham, North Carolina, USA. [23]Division of Systems Biology, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, Arkansas, USA. [24]Department of Internal Medicine and Biochemistry, Rush University Medical Center, Chicago, Illinois, USA. [25]State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, Schools of Life Sciences and Pharmacy, Fudan University, Shanghai, China (L.S.'s primary affiliation). [26]Laboratory of Toxicology and Pharmacology, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA. [27]These authors contributed equally to this work. Correspondence should be addressed to W.T. (weida.tong@fda.hhs.gov) or S.S.A. (AuerbachS@niehs.nih.gov) or P.R.B. (bushel@niehs.nih.gov).

Received 2 December 2013; accepted 29 July 2014; published online 24 August 2014; doi:10.1038/nbt.3001

study design and crowdsourcing model to reach community-wide consensus on the fit-for-purpose use of emerging technologies.

High-throughput sequencing technologies provide new methods for whole-transcriptome analyses of gene expression[7]. Recently published studies have compared data obtained from microarrays and RNA-seq in terms of technical reproducibility, variance structure, absolute expression, and detection of DEGs or gene isoforms[8–20] (**Supplementary Table 1**). Some of these studies suggested that RNA-seq is less precise for weakly expressed genes owing to the nature of sampling[21,22], whereas others found higher sensitivity of RNA-seq for gene detection[23,24]. The varied conclusions can be attributed to the fact that the researchers used few treatment conditions, and hence they do not cover a wide range of biological complexity. Furthermore, the question has not been adequately addressed about whether predicting toxicity outcomes on the basis of gene-expression data could be enhanced using RNA-seq instead of microarray data.

Under the umbrella of the third phase of the MAQC Consortium[3–6], also known as the Sequencing Quality Control (SEQC) Consortium[25], we conducted a comprehensive study to evaluate RNA-seq in its differences and similarities to microarrays in terms of identifying DEGs and developing predictive models. In contrast to data generated as part of the SEQC project using reference RNA samples[25], our data and study design allowed us to compare transcriptional responses detected by each platform in terms of extensive chemical treatments, biologic replication and breadth of shared MOA of the chemicals, which complements the performance metrics monitored in other SEQC studies.

Specifically, we report the results of a comparative analysis of gene expression responses profiled by Affymetrix microarray and Illumina RNA-seq in liver tissue from rats exposed to diverse chemicals. We used either microarray or RNA-seq data to generate DEGs and predictive models of the MOA of each chemical. This allowed us to assess the influence of the chemical (referred to hereafter as the 'treatment effect') on the concordance between RNA-seq and microarrays and on the performance of predictive models generated using each technology. Treatment effect is characterized by the number of DEGs and the overexpressed pathways underlying the MOA of the chemical.

We found that (i) the concordance between microarray and sequencing platforms for detecting the number of DEGs was positively correlated with the degree of the perturbation elicited by the treatment, (ii) RNA-seq was better than microarrays at detecting weakly expressed genes, and (iii) gene expression–based predictive models generated from RNA-seq and microarray data were similar. The experimental design also allowed us to identify positive correlations in differentially expressed RNA elements (mRNA, splice variants, noncoding RNA and exon-exon junction) with the degree of the perturbation elicited by the treatment, and to examine treatment-induced alternative splicing and shortening of 3′ untranslated regions (UTRs).

RESULTS

Study design

We exposed male Sprague-Dawley rats to one of 27 chemicals (three rats per chemical with matched controls), isolated RNA from the livers and analyzed these samples using Affymetrix microarrays and Illumina RNA-seq (**Fig. 1a** and **Supplementary Table 2**). To examine the performance of RNA-seq in predicting toxicity with independent validation, we divided the 27 chemicals into a training set (15 chemicals were used to develop the predictive models) and a test set (12 chemicals were used to validate the models). The 15 chemicals in the training set elicited varying strengths of transcriptional responses in the rat liver, which allowed us to examine the concordance between microarray and RNA-seq in DEGs and pathways in response to such variations in treatment strength. Furthermore, sets of three chemicals share one of five MOAs. Three MOAs are associated with well-defined receptor-mediated processes—peroxisome proliferator-activated receptor alpha (PPARA), orphan nuclear hormone receptors (CAR/PXR) and aryl hydrocarbon receptor (AhR). The other two are nonreceptor-mediated DNA damage (DNA damage) or cytotoxicity (cytotoxic). The MOAs serve as endpoints for investigation of the predictive models and cross-platform concordance.

RNA-seq data were processed with six bioinformatics pipelines (**Supplementary Note 1**) that used four gene models (**Fig. 1b**). The mapping details, such as the total number of reads and the mapping rate for each sample, are summarized in **Supplementary Table 3**. The effect of sequencing depth varied between the six pipelines (**Supplementary Fig. 1**).

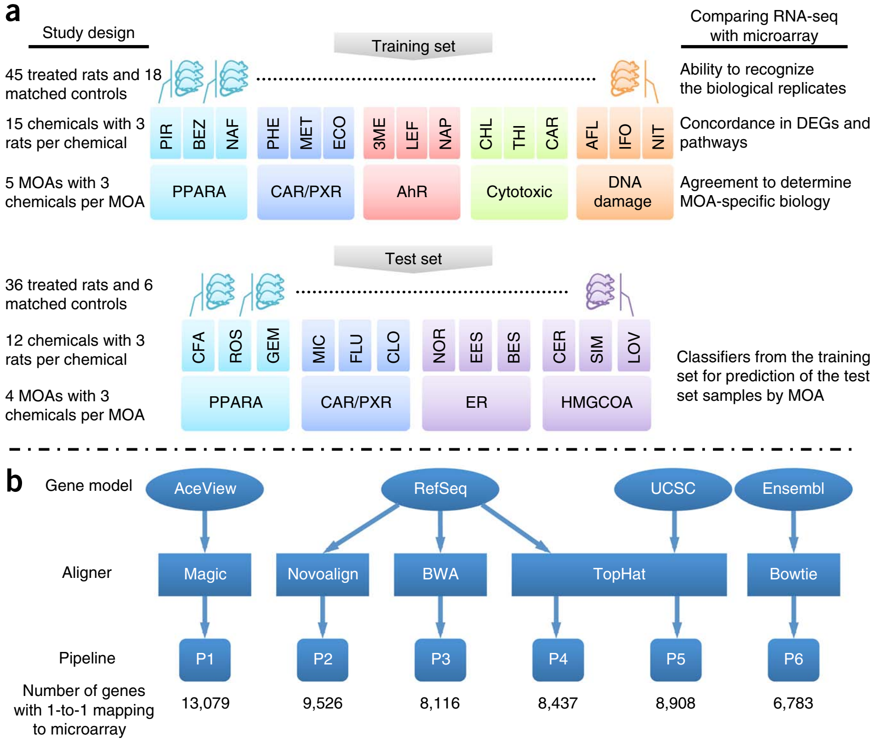
a
Study design
Training set
Comparing RNA-seq with microarray
45 treated rats and 18 matched controls
Ability to recognize the biological replicates
15 chemicals with 3 rats per chemical
PIR
BEZ
NAF
PHE
MET
ECO
3ME
LEF
NAP
CHL
THI
CAR
AFL
IFO
NIT
Concordance in DEGs and pathways
5 MOAs with 3 chemicals per MOA
PPARA
CAR/PXR
AhR
Cytotoxic
DNA damage
Agreement to determine MOA-specific biology
Test set
36 treated rats and 6 matched controls
12 chemicals with 3 rats per chemical
CFA
ROS
GEM
MIC
FLU
CLO
NOR
EES
BES
CER
SIM
LOV
Classifiers from the training set for prediction of the test set samples by MOA
4 MOAs with 3 chemicals per MOA
PPARA
CAR/PXR
ER
HMGCOA
b
Gene model
AceView
RefSeq
UCSC
Ensembl
Aligner
Magic
Novoalign
BWA
TopHat
Bowtie
Pipeline
P1
P2
P3
P4
P5
P6
Number of genes with 1-to-1 mapping to microarray
13,079
9,526
8,116
8,437
8,908
6,783

Figure 1 Overview of study design. (**a**) The study comprised a training set and a test set with the text on the left detailing the experimental design and the text on the right listing the key analyses conducted. Both microarray and RNA-seq were used to profile transcriptional responses induced by treatment of rats by each chemical; each is associated with a specific MOA. For each MOA there were three representative chemicals and three biological replicates per chemical. We evaluated cross-platform concordance at multiple levels: DEGs, mechanistic pathways and MOAs. To compare the predictive potential of RNA-seq and microarray as gene-expression biomarkers, we analyzed four MOAs by both platforms as a test set, two of the MOAs (PPARA and CAR/PXR) appear in the training set whereas the other two do not. (**b**) Overview of RNA-seq analysis pipelines.

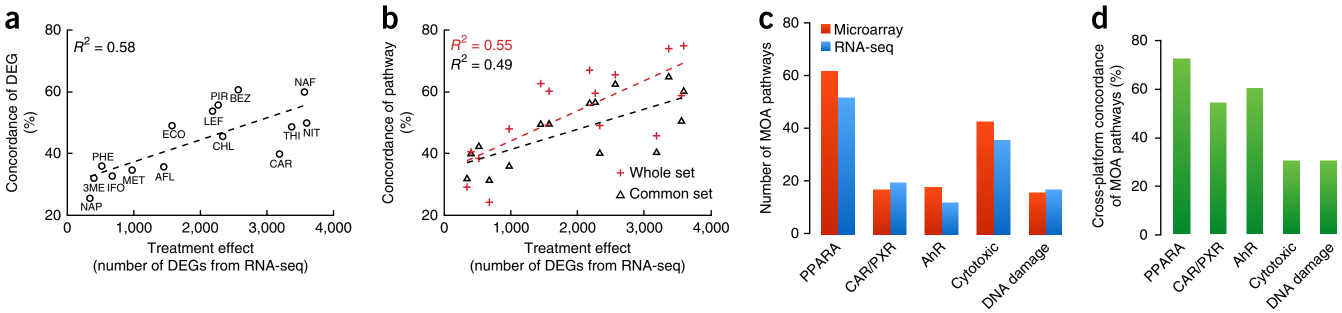

Figure 2 Concordance between RNA-seq and microarray. (**a**) Between-platform concordance of DEGs against the number of DEGs identified by RNA-seq. Concordance is the overlap of the DEGs from the two platforms with agreement in the direction of fold change. The concordance was adjusted to remove the contribution of random chance. (**b**) Between-platform concordance of pathways against the number of DEGs identified by RNA-seq. The concordance was adjusted to remove the contribution of random chance. Two sets of pathways were analyzed, one obtained from the DEGs derived from the analysis based on the common set of genes and the other based on the platform-independent analysis. (**c**) Common pathways shared by three chemicals with the same MOA (*x* axis) for both platforms (left bar for microarray and right for RNA-seq). (**d**) Concordance between pathways identified by both platforms for each MOA. Concordance is the percentage of common pathways (described in **c**) shared by the two platforms.

The study design included 15–60 million 100-bp paired-end reads for each sample. Increased sequencing depth did not markedly increase gene detection for most pipelines except pipeline P1 (AceView gene model[26] with the MAGIC aligner) (**Supplementary Fig. 1**).

Microarray data from the Affymetrix GeneChip Rat Genome 230 2.0 array were normalized with both robust multi-array average[27] (RMA) and Microarray Analysis Suite 5.0 (ref. 28; MAS5). We conducted a pairwise analysis of the number of DEGs for the 15 chemicals in the training set among all the data processing methods for both platforms (six pipelines for RNA-seq and two normalization methods for microarray data). The pairwise correlation (R^2) was 0.78–1.00 (**Supplementary Fig. 2**). Thus, the bulk of the presentations in the main text are based on normalized microarray data from RMA.

We evaluated three methods (limma[29], edgeR[30] and DESeq[31]) for identifying DEGs from RNA-seq data. All three were highly consistent (linearly correlated with $R^2 = 0.87–1.00$) in determining the number of DEGs across the 15 chemicals (**Supplementary Fig. 3**). Thus, the bulk of the presentations in the main text are based on limma, which was originally developed for microarray and has recently been found to be equally applicable to RNA-seq data analysis.

Treatment effect dictates concordance

We first investigated the concordance between RNA-seq and microarrays across 15 chemicals in the training set, which elicited a wide range of treatment effects in the rat livers. We sought to assess cross-platform concordance in terms of both DEGs and MOA. Concordance is the percentage of DEGs shared by the two platforms with agreement in the direction of fold change (Online Methods).

To eliminate the potential bias that could be introduced because the two platforms assayed different numbers of genes, we focused the analysis of DEGs on genes assayed by both platforms. This common set comprises 13,079 genes based on the AceView gene model (Online Methods). We observed a wide range of treatment effects for 15 chemicals; the number of DEGs exhibited at least a tenfold difference between the least and most perturbing chemical treatment (**Fig. 2a**). The agreement in DEGs between the two platforms was linearly correlated to the treatment effect (**Fig. 2a**; the concordance was adjusted against random chance). Agreement is higher for chemicals with a marked treatment effect size (nafenopin: NAF = 60%) than for those where the response is weak (beta-naphthoflavone: NAP = 25%).

We next mapped the DEGs found by each platform to pathways (Online Methods). This was done for the DEGs derived from either the common set or the whole gene set of each platform. We observed the linear relationship between treatment effect (i.e., number of DEGs) and cross-platform concordance in enriched pathways (Online Methods) (**Fig. 2b**). These results suggest that a large discrepancy is expected between the two platforms when the treatment effect is small (that is, comparing two similar biological conditions), but their agreement will be improved significantly if the treatment effect is large (that is, comparing two distinct biological conditions such as cancer versus normal).

To assess common transcriptional responses elicited by chemicals with the same MOA, we analyzed the overlap in pathways differentially expressed in response to each set of three chemicals that share one of the five MOAs (**Fig. 1a**). Statistically significant pathways were derived from the DEGs for each chemical using the whole gene set for both platforms. The number of enriched pathways shared by each of the MOA chemical groups was similar between two platforms across all the MOAs (**Fig. 2c**). Some of the key pathways associated with xenobiotic activities in liver were observed for both platforms (**Supplementary Table 4**). For three well-defined receptor-mediated MOAs (PPARA, AhR and CAR/PXR), the overlap between the two platforms in the MOA-based pathways was over 50% (**Fig. 2d**). For the two MOAs involving general, nonspecific toxicity (cytotoxic and DNA damage), however, the overlap between the two platforms is much lower (30%). These results suggest that the concordance between the two platforms is higher for an endpoint involving specific mechanisms rather than complex ones.

Low abundance contributes to discrepancies

Another feature of this study design is the sets of three animals treated by one chemical, which allows examination of the ability of both platforms to recognize the replicates exposed to the same chemical using the gene expression data. Both platforms performed equally well when genes with above-median expression were used for this assessment; a low variation measured by root mean squared distance (RMSD) among the replicates was observed (**Fig. 3a**). However, a large variation for the genes with below-median expression was observed, and the RMSD values also varied significantly according to the choice of data processing methods (that is, the normalization methods for microarray and the pipelines for RNA-seq). This observation was consistent in the log ratio versus mean average plots for both platforms (**Supplementary Fig. 4**).

For gene groups with expression both above and below median, we assessed the agreement between the two platforms in terms of

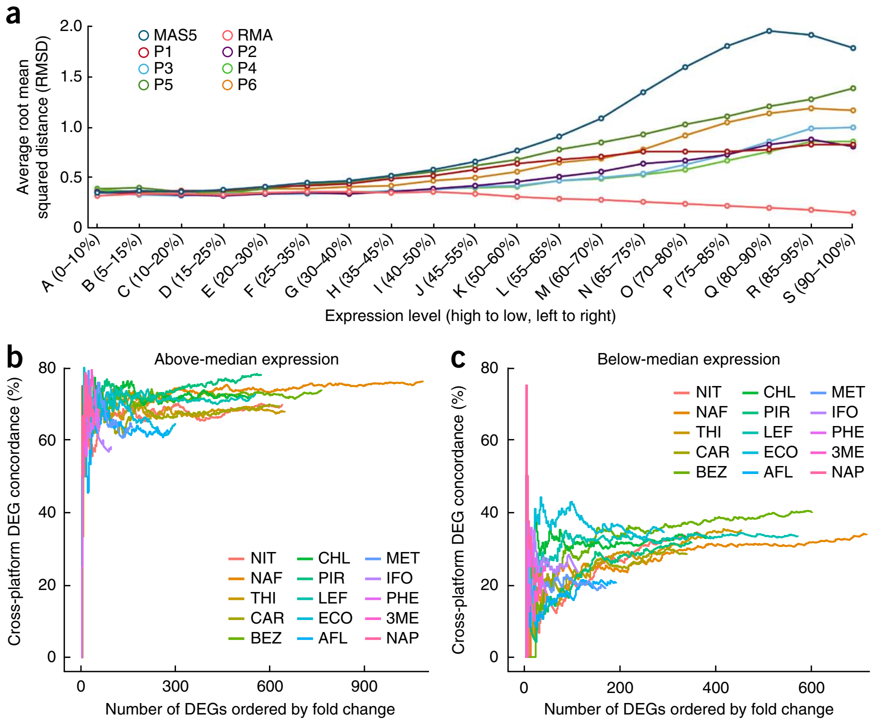

Figure 3 Transcript abundance–dependent concordance between RNA-seq and microarray. (**a**) Root mean squared distance (RMSD in *y* axis) between pairs of rats for each chemical and averaged over all the chemicals by bins of genes. Expression levels ranged from high (0%) to low (100%) and each bin, A to S, contained 10% of the expressed genes. The analysis was performed on RNA-seq with six pipelines and the microarray with two normalization methods (RMA and MAS5). (**b**,**c**) For each chemical, the *x* axis represents the number of DEGs top ranked by the fold change with $P < 0.05$ for both platforms with equal numbers of up- and downregulation. The *y* axis represents the overlap (%) between platforms for a given number of ranked DEGs. Each line on the graph represents the overlap of DEG lists between two platforms for one chemical for above-median expressed genes (**b**) and below-median expressed genes (**c**).

the order of differential expression using the percentage of overlapping genes plot[4]. Specifically, genes detected as differentially expressed (*P*-value < 0.05) by a platform were first ranked by their fold changes to generate an ordered DEG list, one for RNA-seq and another for microarrays. We then calculated the overlap percentage between two ordered DEG lists for different numbers of DEGs taken from the top of both lists (half are upregulated genes and another half are downregulated genes). The plot of the percentage of overlapping genes provides a measure of the order preservation (concordance) between the two platforms in a cutoff-free fashion. We observed a much higher concordance (~75% for most chemicals) for the above-median expressed genes (**Fig. 3b**) than for the below-median expressed genes (20–40%) (**Fig. 3c**). As a point of reference, in our previous investigation comparing two sets of DEGs from treated rats and generated from identical experimental conditions with microarrays, we observed that the maximum within-laboratory overlap in DEGs was ~70% when no abundance-based filtering was applied[4]. This suggested that a high degree of concordance between two platforms is expected only if the above-median expressed genes are considered.

These results also suggested that the difference between the two platforms is largely determined by how accurately genes expressed at low levels are quantified. Therefore, we selected 18 cancer-related genes (**Supplementary Table 5**) and used quantitative polymerase chain reaction (qPCR) to verify their responses to a subset of eight chemicals (a total of 65 chemical-gene pairs were assayed). These chemicals were selected to ensure that the validation results reflect the objectives of the overall study design (Online Methods). Fold change and *P*-value of differential expression for each chemical-gene pair are listed in **Supplementary Table 6**.

Both platforms exhibited a high concordance to qPCR in DEGs (94% for RNA-seq and 88% for microarray) for the above-median expressed genes (**Fig. 4a**). However, for the below-median expressed genes, although the high concordance remained for RNA-seq (89%; 8 out of 9 DEGs verified), it was drastically lower for microarrays (17%; 1 out of 6 DEGs verified). We also compared the fold change of qPCR versus RNA-seq and qPCR versus microarray for the DEGs detected by qPCR. The RNA-seq data were correlated better with qPCR data ($R^2 = 0.97$) than were the microarray data ($R^2 = 0.85$) (**Fig. 4b**). Furthermore, the regression slope for RNA-seq is close to 1, whereas the slope for microarray deviates from 1 and is tilted toward the *x*-axis, showing the classic ratio compression phenomenon[32]. The analysis for each chemical separately yielded similar results (**Supplementary Fig. 5**).

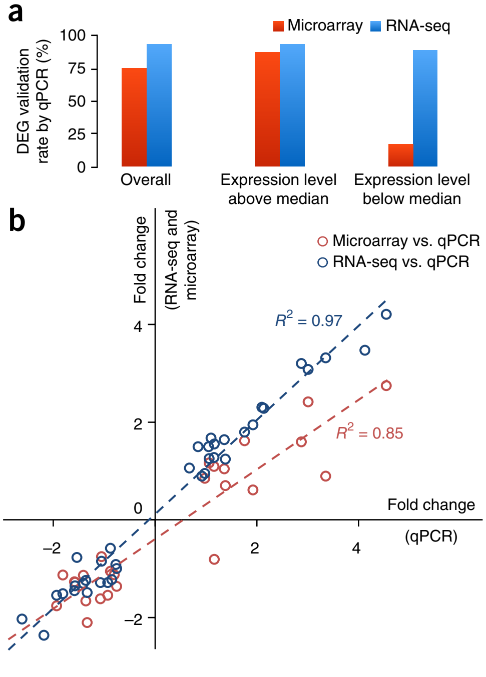

Figure 4 Concordance of RNA-seq and microarray data with qPCR data. (**a**) The qualitative agreement (differentially expressed or not) of RNA-seq and microarray against qPCR for the DEGs with above-median expression or below-median expression is shown along with the overall average results. Differential expression was determined by absolute fold change > 1.5 and *P*-value < 0.05. (**b**) Correlation of RNA-seq and microarray (*y* axis) with qPCR data (*x* axis) using the $\log_2$ fold change measure of the genes differentially expressed across the two gene-expression platforms under correlation analysis.

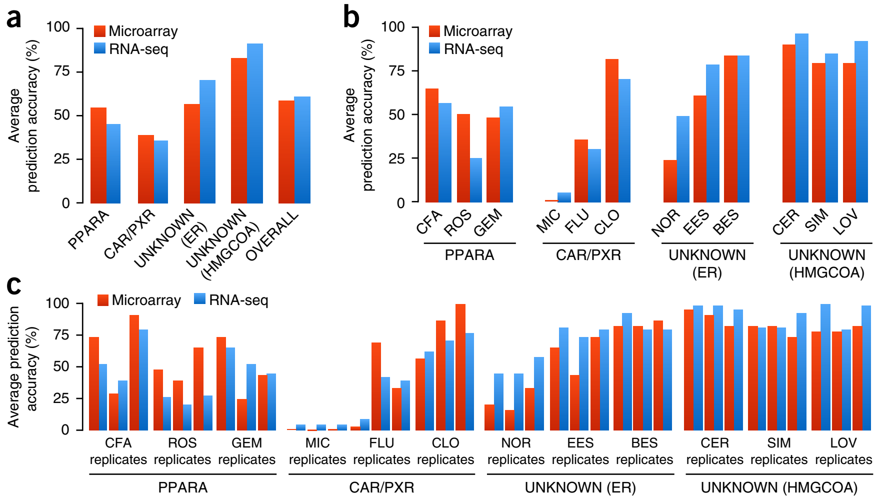

Figure 5 Cross-platform comparisons of prediction results between two platforms. Sixty-one classifiers were generated using the training set and subsequently predicted the test set chemicals in a blind fashion. (**a**) The bar chart displays the average (percent) prediction accuracy (*y* axis) achieved by a platform for a particular chemical group (*x* axis: chemicals grouped by PPARA, CAR/PXR, unknown to the training set, and Overall). The left bar (red) shows the average prediction accuracy using microarray data; the right bar (blue) is for RNA-seq data. (**b**) Comparison of prediction accuracy at individual chemical level grouped by MOAs. (**c**) Comparison of prediction accuracy at the individual animal level. The *x* axis lists all animals grouped by treatment chemical and then MOA.

Prediction of samples by MOA is similar

To compare the use of RNA-seq and microarray data for developing gene-expression-based classifiers, we identified sets of genes from the training set that were predictive of the MOAs of the chemical treatments in the test set. The test set chemicals represented four MOAs, two of which are shared with the training set (CAR/PXR and PPARA) and two of which are not (estrogen receptor (ER) and HMG-CoA reductase (HMGCOA)) (**Supplementary Table 2**). For CAR/PXR or PPARA, prediction accuracy for a MOA was based on the correct classification of individual samples that were from the chemicals grouped by the MOA. Otherwise, the prediction was designated “UNKNOWN” (because ER or HMGCOA were not included in the training set.) Thus, we assessed the ability to correctly select chemicals that exert effects similar to what was used in the training set and exclude those that work through other mechanisms. Sixty-one classifiers were generated with most of the approaches equally applied to the data from both platforms (**Supplementary Table 7** and **Supplementary Note 2**). The overall average accuracy was 61% for classifiers profiled by RNA-seq and 58% for those profiled by microarray (**Fig. 5a** and **Supplementary Table 8**). For PPARA, both platforms had modest predictability with microarray performing slightly better (random chance = 33%). In contrast, both platforms performed poorly when using gene expression signatures to predict CAR/PXR. By comparing the prediction results at the level of individual chemicals (**Fig. 5b**) or per replicate sample (**Fig. 5c**), both platforms exhibited similar prediction accuracies with excellent ability to identify samples from the two MOAs not seen previously.

We also performed rank-based scoring of the 15 training set chemical signatures derived from RNA-seq against 657 microarray liver chemical signatures deposited in the NextBio database[33]. This analysis takes into account rank-based weighting of genes derived from the amplitude of differential expression. In 9 out of 15 queries with the RNA-seq signatures, the corresponding microarray signature was the top-scoring signature (**Supplementary Table 9**). The remaining six RNA-seq signatures matched the corresponding microarray signatures within the top ten out of 657 profiles.

RNA-seq–specific features show the dependency of treatment effect

We investigated how different RNA elements, including noncoding RNAs, splice variants and exon-exon junctions, were affected by different degrees of treatment effect. Changes in differential expression of these RNA elements were proportional to changes in mRNA (**Fig. 6**), suggesting that the overall level of perturbation of different RNA species is similar and is dictated by the nature of the perturbing chemicals.

To investigate features of transcripts that could be analyzed using RNA-seq but not microarray data, we selected two chemicals (phenobarbital (PHE) and pirinixic acid (PIR)) with low and high treatment effect sizes, respectively. We first used a Poisson hidden Markov model[34] to detect 345 and 381 transcripts with shortened 3′ UTRs (Fisher’s exact test *q*-value ≤ 0.01) when compared to control (**Supplementary Table 10**). There are 278 transcripts in common, which enrich for spliceosome- and ubiquitin-mediated proteolysis KEGG pathways. For 11 of the 278 transcripts, distinct alternatively

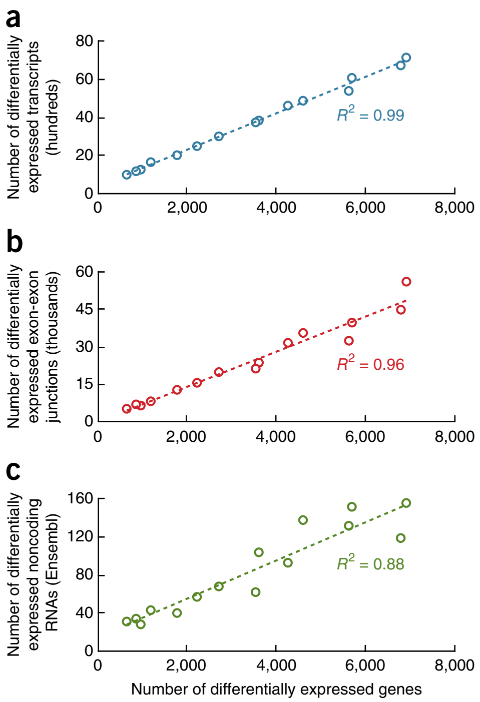

Figure 6 Systemic trends of differentially expressed RNA elements. All results were based on Ensembl annotation and pipeline P6. The differential expression test was done using edgeR with the cutoffs fold change >1.5 and uncorrected $P < 0.05$. (**a**–**c**) Numbers of differentially expressed (**a**) transcripts, (**b**) exon-exon junctions, and (**c**) noncoding RNAs across all 15 chemicals in the training set are positively correlated with the number of DEGs.

polyadenylated isoforms were found in the control, PHE or PIR samples. We also used a mixture of isoforms probabilistic model[35] to identify 408 and 449 transcripts with isoforms that are significantly (Fisher's exact test q-value < 0.05) differentially spliced between control and PHE- or PIR-treated samples, respectively (**Supplementary Fig. 6** and **Supplementary Table 11**).

DISCUSSION

Using a comprehensive study design involving chemical treatments grouped within several MOAs, training and test sets, and biological replication, we pursued a collaborative effort, under the umbrella of the SEQC project, to evaluate Illumina RNA-seq and compare its performance with Affymetrix microarrays for determining DEGs and developing predictive models. Various bioinformatics approaches were applied to microarray processing, RNA-seq data analysis and DEG identification, and they had minimum impact on the results.

One of the advantages of our study design is that the rat liver was perturbed to varying degrees by 15 chemicals. We observed a linear relationship between the cross-platform concordance and the degree of perturbation (**Fig. 2a**). Similar results were obtained from comparisons of the two platforms using pathways (**Fig. 2b**) and MOAs (**Fig. 2c**,**d**). These findings provide evidence that the treatment effect (which includes both the strength of the transcriptional response and the complexity of the underlying biology) is an important factor that determines the consistency between RNA-seq and microarrays. To date, most, if not all, studies have compared the two platforms using two distinct biological conditions (e.g., one organ versus another, or cancer versus normal tissue). Such a design reveals only one aspect of comparative characteristics of the two platforms. Our results demonstrated that when investigating two similar biological conditions, a lower concordance is expected between the two platforms, and the discrepancy is derived from the measurement of the genes expressed at low levels, for which RNA-seq performs better. However, if two distinct biological conditions are compared, the two platforms are in better agreement. These results are consistent with the MAQC-II finding that the prediction performance of classification models is dependent on the inherent nature of the endpoints, with some endpoints more predictive than others[5].

We found that the two platforms perform equally well for genes with expression levels above the median of all the assayed genes. The agreement between the two platforms was around 75% when the above-median expressed genes were used for determining concordance (**Fig. 3b**). We expect that the unknown degree of false discovery and other variations specific to each platform precludes the possibility of 100% cross-platform concordance. For instance, the MAQC-I study compared two sets of DEGs from treated animals profiled in identical experimental conditions in the same laboratory and observed that the maximum within-laboratory overlap in DEGs was ~70%[4] and can serve a point of reference to evaluate our findings of the cross-platform comparison. Thus, the cross-platform concordance reported here in our study is high, particularly when the highly expressed genes are used. The findings suggest that the above-median expressed genes have a good transferability between the two platforms. Consequently, in future studies to identify biomarkers transferrable between two gene-expression measurement platforms, an emphasis should be placed on the above-median expressed genes.

Clearly, the discrepancy between the two platforms rests largely on the genes with below-median expression. Using these genes, the concordance between two platforms was below 40% (**Fig. 3c**). The qPCR results suggest that RNA-seq is better at detecting differential gene expression at low expression levels than is microarrays (**Fig. 4a**). However, owing to the small number of DEGs expressed at low levels from both platforms (six for microarrays and nine for RNA-seq) that were tested by qPCR, this finding needs to be further validated with a large sample size.

Identifying genomic biomarkers to predict disease states or biological conditions is of high priority in research, regulatory and clinical settings. We compared the ability of microarray and RNA-seq to profile gene expression signatures as predictors of a biological response elicited from treatment with a chemical that falls into a known MOA or chemicals in MOAs unknown to those in the training data set. This was conducted in a unbiased fashion. (i) The sequencing data for the test set were generated independently from the training set. (ii) To minimize the impact of the choice of modeling methods, each modeler involved in the analysis had the freedom to choose any methodology, and this resulted in a broad range of classifier approaches, feature selection methods and RNA-seq pipelines applied (**Supplementary Note 2**). (iii) Finally, the labels of the chemicals in the test set were not revealed to the modelers. The average prediction accuracy for all treated samples was ~60% but slightly higher for RNA-seq–derived classifiers than for those identified by microarray (**Fig. 5a**). By closely examining the predictions for individual chemicals (**Fig. 5b**) and individual samples (**Fig. 5c**), both platforms exhibited similar patterns. Of note, one classifier derived from RNA-seq data recognized all sample replicates and predicted all the test samples with 100% accuracy but its application to microarrays was not successful. Overall, both platforms exhibited similar prediction performance when all classifiers were considered.

We observed that the two platforms profile very similar gene expression changes when compared at the mechanistic level. For example, the correlation in the number of DEGs detected by the two platforms was relatively high ($R^2 > 0.9$), indicating that both platforms have similar capacity to rank different treated samples based on transcriptional response. In particular, evaluating RNA-seq and microarray in terms of signal transduction pathway gene correlations revealed that the two platforms performed similarly when profiling gene sets that share co-regulation by transcription regulators or regulatory networks. For example, both platforms had 12 significantly affected transcription regulators in common (**Supplementary Table 12**). However, RNA-seq detected signaling from the full-length version of NF-1 (T13803), whereas microarray detected the signaling from splice variants NF-1A and NF-1C. Conversely, RNA-seq uniquely detected several significant isoforms of HNF1 and HNF4. Moreover, on the basis of reconstructed gene regulator networks from each platform separately, we found several genes and interactions in common between the two platforms (**Supplementary Fig. 7**). These included *Cyp4a10*, *Cyp4a22*, *Acot1*, *Dci* (also known as *Eci1*) and *Ech1*, each of which has been shown to play some role in MOAs related to the exposure to one of the chemicals.

In addition, we also included ERCC[36] spike-ins in this study for the purpose of quality controls. Although ERCC spike-ins have been suggested as a means for data normalization across samples and conditions[37], we observed a considerable variation in the fraction of reads aligning to ERCC spike-ins for a given sample between libraries (**Supplementary Fig. 8**). This observation is also confirmed in the SEQC main study[25]. This variation in ERCC controls across library batches led us to adopt data normalization approaches that did not involve external spike-in controls.

We also demonstrated that the increase of sequencing depth did not remarkably increase gene detection for most pipelines except the AceView-based pipeline P1 (**Supplementary Fig. 1**). As pipeline P1 was extensively studied in the comparative analysis of DEGs between the two platforms based on the common set of genes (**Fig. 1b**), we evaluated whether the increase of sequencing depth had any significant effect on DEGs in the comparative analysis. An increase of almost four times more reads per sample (>90 million more reads per sample) resulted in an increase of only ~8% more DEGs detected by RNA-seq (**Supplementary Table 13**), which is very small compared to the observed range of about sixfold for the number of DEGs across 15 chemicals (**Fig. 2a**) where the comparative analysis was conducted. Notably, the number of DEGs in concordance between two platforms was virtually unchanged, indicating that ~25 million RNA-seq reads per sample is adequate for the comparative analysis of DEGs and a higher sequence depth would not likely alter our conclusions.

In summary, treatment effects and abundance of the genes affect many observations in RNA-seq and its comparison with microarrays. Many RNA-seq–specific features followed a systematic trend correlated with the strength of perturbation of the samples, which highlights the importance of having a comprehensive study design to systematically evaluate these features. We also observed that the major difference between RNA-seq and microarrays is their respective sensitivity for genes with low expression, where RNA-seq offers advantages over microarrays. Consequently, RNA-seq outperforms microarrays substantially when profiling two similar biological conditions (e.g., weakly treated samples versus control) compared to profiling two distinct biological conditions such as cancer versus the normal state. If one's goal is to develop a genomic biomarker with statistical performance as the sole consideration, presumably either platform is suitable.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. SRA: SRP039021; GEO: GSE55347, GSE47875.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank M. Arana and D. Mendrick for their critical review of the manuscript. This research was supported, in part, by the Intramural Research Program of the National Institutes of Health (NIH), National Institute of Environmental Health Sciences (NIEHS) (ES102345-04 and ES023026) and National Library of Medicine. P.P.Ł. and D.P.K. acknowledge support by the Vienna Scientific Cluster (VSC), the Vienna Science and Technology Fund (WWTF), Baxter AG, Austrian Research Centres (ARC) Seibersdorf and the Austrian Centre of Biopharmaceutical Technology (ACBT).

DISCLAIMER

The views presented in this article do not necessarily reflect current or future opinion or policy of the US Food and Drug Administration. Any mention of commercial products is for clarification and not intended as an endorsement. This article may be the work product of an employee or group of employees of the National Institute of Environmental Health Sciences (NIEHS), National Institutes of Health (NIH), however, the statements, opinions or conclusions contained therein do not necessarily represent the statements, opinions or conclusions of NIEHS, NIH or the United States government.

AUTHOR CONTRIBUTIONS

W.T. coordinated the consortium study and manuscript preparation. W.T., S.S.A. and C.W. designed the study. C.W. conducted sequencing and qPCR experiments. S.S.A. provided rat tissue samples, gene expression data and contributed to the data analysis. P.R.B. was involved heavily in manuscript preparation and data analysis. B.G. and J.X. conducted the majority of data analysis and prepared various figures and supplementary materials. J.T.M. and D.T.M. constructed the mapping table between microarray and RNA-seq along with other data analysis and interpretation. All the co-authors contributed to various components of the study, including data analysis and preparation of text, figures, tables and supplementary materials.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Hamburg, M.A. Advancing regulatory science. *Science* **331**, 987 (2011).
2. Chen, M., Zhang, M., Borlak, J. & Tong, W. A decade of toxicogenomic research and its contribution to toxicological science. *Toxicol. Sci.* **130**, 217–228 (2012).
3. Shi, L. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24**, 1151–1161 (2006).
4. Guo, L. *et al.* Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat. Biotechnol.* **24**, 1162–1169 (2006).
5. Shi, L. *et al.* The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* **28**, 827–838 (2010).
6. Fan, X. *et al.* Consistency of predictive signature genes and classifiers generated using different microarray platforms. *Pharmacogenomics J.* **10**, 247–257 (2010).
7. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
8. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
9. Bottomly, D. *et al.* Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS ONE* **6**, e17820 (2011).
10. Bradford, J.R. *et al.* A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genomics* **11**, 282 (2010).
11. Giorgi, F.M., Del Fabbro, C. & Licausi, F. Comparative study of RNA-seq- and microarray-derived coexpression networks in *Arabidopsis* thaliana. *Bioinformatics* **29**, 717–724 (2013).
12. Malone, J.H. & Oliver, B. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.* **9**, 34 (2011).
13. Merrick, B.A. *et al.* RNA-seq profiling reveals novel hepatic gene expression pattern in Aflatoxin B1 treated rats. *PLoS ONE* **8**, e61768 (2013).
14. Nookaew, I. *et al.* A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **40**, 10084–10097 (2012).
15. Raghavachari, N. *et al.* A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. *BMC Med. Genomics* **5**, 28 (2012).
16. Sirbu, A., Kerr, G., Crane, M. & Ruskin, H.J. RNA-Seq vs dual- and single-channel microarray data: sensitivity analysis for differential expression and clustering. *PLoS ONE* **7**, e50986 (2012).
17. Su, Z. *et al.* Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. *Chem. Res. Toxicol.* **24**, 1486–1493 (2011).
18. Subramaniam, S. & Hsiao, G. Gene-expression measurement: variance-modeling considerations for robust data analysis. *Nat. Immunol.* **13**, 199–203 (2012).
19. Xiong, Y. *et al.* RNA sequencing shows no dosage compensation of the active X-chromosome. *Nat. Genet.* **42**, 1043–1047 (2010).
20. Xu, W. *et al.* Human transcriptome array for high-throughput clinical studies. *Proc. Natl. Acad. Sci. USA* **108**, 3707–3712 (2011).
21. Łabaj, P.P. *et al.* Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics* **27**, i383–i391 (2011).
22. McIntyre, L.M. *et al.* RNA-seq: technical variability and sampling. *BMC Genomics* **12**, 293 (2011).
23. Mooney, M. *et al.* Comparative RNA-Seq and microarray analysis of gene expression changes in B-cell lymphomas of Canis familiaris. *PLoS ONE* **8**, e61088 (2013).
24. Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956–960 (2008).
25. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* doi:10.1038/nbt.2957 (24 August 2014).
26. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.* **7** (suppl. 1), S12.1–14 (2006).
27. Irizarry, R.A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
28. Li, C. & Wong, W.H. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98**, 31–36 (2001).
29. Smith, G.K. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* (eds. Gentleman, R., Carey, V., Huber, W., Irizarry, R. & Dudoit, S.) 397–420 (Springer, 2005).
30. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
31. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
32. Shi, L. *et al.* Microarray scanner calibration curves: characteristics and implications. *BMC Bioinformatics* **6** (suppl. 2), S11 (2005).
33. Kupershmidt, I. *et al.* Ontology-based meta-analysis of global collections of high-throughput public data. *PLoS ONE* **5**, e13066 (2010).
34. Lu, J. & Bushel, P.R. Dynamic expression of 3′ UTRs revealed by Poisson hidden Markov modeling of RNA-Seq: implications in gene expression profiling. *Gene* **527**, 616–623 (2013).
35. Katz, Y., Wang, E.T., Airoldi, E.M. & Burge, C.B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
36. Baker, S.C. *et al.* The External RNA Controls Consortium. a progress report. *Nat. Methods* **2**, 731–734 (2005).
37. Lovén, J. *et al.* Revisiting global gene expression analysis. *Cell* **151**, 476–482 (2012).

ONLINE METHODS

Samples. All biological samples employed in the studies described here were derived from the DrugMatrix tissue/RNA bank that is now owned by the National Toxicology Program (NTP, https://ntp.niehs.nih.gov/drugmatrix/index.html). Details on the design and in-life portion of these studies can be found elsewhere[38]. In short, male Sprague-Dawley rats (aged 6–8 weeks and weighing 200–260 g) were purchased from Charles River Laboratories (Portage, MI). Test chemicals were administered orally (10 ml/kg body weight in corn oil or water) or via intraperitoneal, intravenous or subcutaneous injection (5 ml/kg body weight in saline). In order to ensure a maximal transcriptional response, 5-d maximum tolerated doses (MTD) of test chemicals were administered to the study animals. The MTD was determined in a 5-d range-finding study in which an MTD was determined as a 5 to 10% reduction in body weight relative to control. Animals were dosed once daily for 3, 5 or 7 d, depending on the chemical, and livers were harvested 24 h after the last dose. Animals were randomly assigned to test chemicals using a computerized body-weight stratification procedure. All toxicity studies described here were done in a contract laboratory and therefore investigators from the sponsor were blinded to group allocation during the study. Animals were handled in accordance with United States Department of Agriculture and Code of Federal Regulations Animal Welfare Act (9 CFR Parts 1, 2, and 3) and housing of animals is detailed in **Supplementary Note 3**.

Sample selection. RNA samples for RNA-seq were derived from the NTP DrugMatrix Frozen Tissue Library. The criteria to select samples are detailed in **Supplementary Note 4**. The samples were split into two sets, training and test, which were sequenced in two separate batches and run on different days, to allow for the evaluation of classifiers derived from the data. There were 63 samples in the training set and 42 in the test set. Of the 63 samples in the training set 45 were derived from rats treated with test chemicals and 18 were control samples (3 sets of 6). Thirty-six of the test set samples were derived from chemical treated animals and six were from vehicle- and route-matched controls. For each test chemical there were three rats treated, whereas for each MOA there were three representative test chemicals. Five MOAs were represented in the training set and four MOAs were in the test set. Two of the MOAs were duplicated from the training set with different test chemicals and two were without representation in the training set. (**Fig. 1a** and **Supplementary Table 2**).

Paired-end RNA-seq and analysis. RNA-seq of 63 training and 42 test set samples on Illumina HiScanSQ or HiSeq2000 systems was performed according to the manufacture's protocol using the Illumina TruSeq RNA Sample Preparation Kit and SBS Kit v3 (San Diego, CA). Samples were handled in a blinded fashion during the library preparation and sequencing process. Details of the method are in the **Supplementary Note 5**. Depths of ~23–25 million paired-end 100-bp reads were generated for each sample. Data were deposited in the Sequence Read Archive (NCBI) under accession number SRP039021. One sample supposedly treated by carbon tetrachloride (CAR) was excluded from analysis due to a mislabeling discovered through common QC procedures for RNA-seq. Six bioinformatics pipelines were used to analyze the RNA-seq data. The details of the analysis pipelines are in the **Supplementary Note 1**. The processed data from pipeline P1 were deposited in Gene Expression Omnibus (NCBI) under accession number GSE55347.

Microarray analysis. Fragmented cRNA prepared from liver RNA was hybridized to the Affymetrix whole genome GeneChip Rat Genome 230 2.0 Array according to protocols outlined in the Affymetrix GeneChip Expression Analysis Technical Manual (subsection: Eukaryotic Target Hybridization) using the GeneChip Hybridization, Wash, and Stain Kit (P/N 900720). Details of the hybridization are in **Supplementary Note 6**. Scanning of arrays was performed on the GeneChip Scanner 3000 7G. CEL files were generated using the GCOS software. All arrays met the minimal recommended quality parameters as described in the Affymetrix Data Analysis Fundamentals Guide (P/N 701190). For purposes of data normalization, all 2,218 arrays run on the liver samples, collected as part of the data generation process for the DrugMatrix Database, were normalized with GeneSpring GX 11.5.1 (Agilent Technologies, Santa Clara, CA) by the MAS 5.0 (ref. 39), PLIER16 (ref. 40), RMA[27], GCRMA[41] and Li-Wong[28] algorithms. All normalized data sets are available on the DrugMatrix ftp site (ftp://anonftp.niehs.nih.gov/drugmatrix/Affymetrix_data/Normalized_data_by_organ/). Unless otherwise stated, the bulk of the analyses were done on the RMA normalized data in $\log_2$ scale. Raw data and MAS5 normalized data were also deposited to Gene Expression Omnibus (NCBI) under the accession number GSE47875. All test data were initially gathered in a blinded fashion to allow for unbiased evaluation of MOA classifiers generated using the training data set.

qPCR analysis. We selected 18 DEGs based on the aflatoxin b1 (training set, DNA damage MOA) treatment. These genes were identified as DEGs by either RNA-seq only, microarray only or the overlapping of both RNA-seq and microarray (i.e., six genes in each category) and all of them were randomly selected among those that were annotated in cancer-related pathways to ensure their biological relevance (**Supplementary Table 5**). We further examined the expression of these 18 genes across the training and test sets and randomly selected five to eight genes for each of the following chemicals: ifosfamide and *N*-nitrosodimethylamine (DNA damage, training set); miconazole and fluconazole (CAR/PXR, test set); clofibric acid, rosiglitazone and gemfibrozil (PPARA, test set). These chemicals were selected to ensure that the validation results reflect the objectives of the overall study design: (i) they were taken from both training and test sets, which involved all three separated library preparation/sequencing batches for RNA-seq; (ii) they involved the most studied MOAs (PPARA, CAR/PXR and DNA damage) and elicited a wide range of transcriptional response (the number of DEGs ranging from 500 to 5,000 out of the common set;) and (iii) they also covered all three types of vehicle controls used in this study. A total of 65 chemical-gene pairs were assayed by qPCR using TaqMan Gene Expression Assays. cDNA was synthesized using High Capacity RNA-to-cDNA Kit according to the manufacturer's protocol. The qPCR was performed on a ViiA 7 Real-Time PCR System using TaqMan MGB probe and primers specific for each of the above gene, with cDNA synthesized from 25 ng RNA and TaqMan Universal Master Mix II (Life Technologies, Grand Island, NY). After enzyme activation at 95 °C for 5 min, PCR was carried out at 95 °C for 15 s and 60 °C for 30 s for 40 cycles. Comparative Ct method (delta delta Ct method) was used to calculate the fold differences between chemically treated and control groups with housekeeping gene *GAPDH* as endogenous control. To be consistent in comparison of qPCR against RNA-seq and microarray, both microarray and RNA-seq data were normalized by housekeeping gene *GAPDH*. For the qPCR data, a gene is called DEG when the P-value < 0.05 and the absolute fold change is greater than 1.5. For RNA-seq and microarray, their respective DEG validation rate by qPCR is calculated as the percentage of DEGs in agreement with the results of qPCR.

Mapping Affymetrix probe sets to RNA-seq genes. A master table was generated to facilitate the comparison across platforms, which contained the Affymetrix GeneChip Rat Genome 230 2.0 array probe sequences mapped to all six pipelines (**Supplementary Table 14**). Probe sets with at least eight probes mapping unambiguously (maximum, 1 mismatch) to a gene or to a compact region of the genome were considered matched. The number of common genes between microarray and RNA-seq for all six RNA-seq pipelines was listed in **Figure 1b**.

Statistical analysis. *DEG analysis.* DEGs were determined from a treated versus control comparison of $\log_2$-transformed expression measurements (RMA or MAS5 intensities for microarray, normalized expression values for RNA-seq) using the limma[29] package within Bioconductor (R statistical computing environment) with the absolute fold change cutoff at 1.5 and uncorrected $P < 0.05$ as determined by standards set forth by previous MAQC efforts[4]. Results from the Levene's Test[42] for homogeneity of variance across groups showed that the percentage of genes with unequal variance in the DEG lists is below 5% (for all but one chemical, which had one sequenced sample removed owing to mislabeling), which is less than the expected false-discovery rate (FDR), and thus supported the applicability of limma with the equal variance assumption to this data set.

Cross-platform concordance analysis. Concordance was computed as:

$$\frac{2 \times intersect(DEGs_{microarray}, DEGs_{RNA-Seq})}{DEGs_{microarray} + DEGs_{RNA-Seq}}$$

with agreement in directionality of fold change.

doi:10.1038/nbt.3001

The concordance is adjusted by random chance before the computation. Below is the description of the adjustment approach. By treating the relative area of a set to the whole set as a probability, the intersection of two independent sets is equal to the product of their individual probabilities. Letting N be the number of items in the whole set, n_1 and n_2 the number of items in two independent sets, then the number of items in the intersection between these two sets is $\left(\frac{n_1}{N}\right)\times\left(\frac{n_2}{N}\right)*N=\frac{n_1\times n_2}{N}$. This background intersection increases as the sets get enlarged. Using this property, the equation below was used to estimate the background-corrected intersection between two sets, which may not be independent.

$$x+\frac{(n_1-x)\times(n_2-x)}{N-x}=n_0$$

In the above equation, n_0 is the number of items in the observed intersection and x is the background-corrected result. (n_1-x) and (n_2-x) can be interpreted as the numbers of items in two independent sets drawn from a new whole set with $(N-x)$ items. The observed intersection is thus made of two parts: the true intersection and the background. The solution for the background-corrected intersection is $\frac{n_0\times N-n_1\times n_2}{n_0+N-n_1-n_2}$. For two independent sets, this solution equals 0 since $n_0=n_1\times\frac{n_2}{N}$.

Root mean squared distance (RMSD) *analysis*. RMSD was used to measure the distance between pairs of samples. Genes were first ranked from high to low expression by the sample average per gene. A moving bin of 10% of all expressed genes, with a step size of 5%, was then selected along the expression rank as set x to calculate the RMSD between sample i and j as:

$$\mathrm{RMSD}_{ij}=\sqrt{\frac{\sum_g (I_{ig}-I_{jg})^2}{N_g}}$$

where g is a gene in set x, I_g is the $\log_2$ transformed expression level of gene g in the corresponding sample and N_g is the number of genes in set x. For each chemical, there were three rats treated and the RMSD was averaged for each bin among all three pairs of samples. This mean RMSD could be viewed as a measure of the difference at a certain expression level (i.e., bin) among the three biological replicate rats treated by the chemical. To compare the two profiling platforms and their data analysis methods (i.e., six data analysis pipelines for RNA-seq and two data normalization methods for microarray data), the RMSD was finally averaged over all 15 chemicals in the training set and plotted for all bins.

Pathway analyses. Two pathway analysis strategies were applied in this study.

First, Entrez gene IDs for DEGs were mapped to MetaCore canonical pathway maps (CPMs), a database of manually inferred and curated gene, protein and transcript regulator interactions from primary scientific literature. The CPMs are an ontology of experimentally confirmed signaling and metabolic multistep pathways in human, mouse and rat (http://thomsonreuters.com/metacore/). For the purposes of this study, 693 pathway maps were selected that covered signaling and metabolic processes as well as mechanisms associated with toxic pathologies. The enrichment analysis was done using a hypergeometric test to find over-representation of CPMs. Only statistically significant pathways with $P < 0.05$ were considered. The Benjamini-Hochberg FDR method was used to adjust the P-values for multiple testing. In addition, Entrez gene IDs were mapped onto the canonical pathway collection of the Ingenuity Knowledge Base (http://www.ingenuity.com/), which is a constructed and manually curated database of information on genes, proteins, chemicals, drugs and their molecular interaction and regulation. Significance of the enrichment of the gene list for a particular pathway was calculated and reported as a right-tailed Fisher's exact test P-value. Only pathways with a P-value < 0.05 (by Ingenuity Pathway Analysis) are reported.

Second, to determine if there is systematic bias in pathway enrichment between the two platforms, we classified each enriched pathway for all test set treatments into 1 of 3 classes: enriched in RNA-seq only, enriched in microarray only, or enriched in both. The enriched pathway counts (n) for each chemical were then compiled and a rank score was calculated to determine the distribution of pathway enrichment across the two platforms (M for microarray, R for RNA-seq). The rank score

$$RS=\left(\frac{n_M-n_R}{n_M+n_R+n_{M\cap R}}\right)\times n_M$$

was calculated where n_M is the number of times a pathway is solely enriched in microarray, n_R is the number of times a pathway is solely enriched in RNA-seq and $n_{M\cap R}$ is the number of times a pathway is enriched in both platforms. The more positive the RS, the more a pathway is selectively enriched in microarray and the more negative, the more often it is selectively enriched by RNA-seq.

Signal transduction pathway profiling. Using the rat Rn4 October 15, 2010 release of GenBank (version 480), the Affymetrix GeneChip array probe sets were collapsed into 27,146 UniGene transcript clusters and the filtered RNA-seq transcripts that mapped to the Affymetrix array were collapsed into 21,598 UniGene transcript clusters. The gene expression data from the same UniGene cluster were averaged resulting in 21,199 UniGene transcripts in microarray and 14,831 in RNA-seq. The UniGene downstream targets (DSTs) of transcript-regulators (TRs: transcription factors (TFs), miRNAs, cofactors and complexes) were obtained from the March 31, 2011 release (version 2011.1) of the TRANSFAC database[43,44]. Significance of signal transduction pathway profiling was determined as previously described[45]. Briefly, for each population individually, significant TRs were based on a Group Correlation Score

$$GCS=\sum_{i\neq j} r_{i,j}^2$$

defined as the sum of the squares of the Pearson correlations (r) among all pairs of genes i and j determined to be DSTs of the TR. The P-value for a score was determined from a nonparametric distribution of correlation scores obtained from random cases (B = 10,000 reshuffles of the genes) and the number of times (n) one of these permuted scores is greater than the observed correlation score. Thus, $P = n/B$. The null hypothesis keeps the structure and overlap of all signaling interactions fixed, but changes the identity of the genes.

Reconstruction of gene co-expression networks. Consistency between RNA-seq and microarray data on gene co-expression networks reconstructed for the five MOAs in the training data set was evaluated on a unified signature built among the 13,079 common gene set between RNA-seq and microarray. The signature was derived from a GGSSL (semi-supervised) classifier, with ReliefF and KNN used for feature ranking, on 400 features (200 most discriminant for RNA-seq and 200 for microarray), achieving 69% accuracy on validation data (see Subsection 'Classifier development – methodology 5' in **Supplementary Note 2**). We selected the top 15 features for RNA-seq and the top 15 for microarray, finding five common features, thus obtaining a set of 25 genes to use as nodes. MOA-specific networks were built by computing TOM correlation (TOMsimilarityFromExpr function in the WGCNA R package, with default parameters) between pairs of nodes only on the corresponding MOA samples. For each MOA, binarization was performed by considering only the top 10% of TOM values. To identify common co-correlation patterns across MOAs, we built a unified weighted network with edge weight k between two nodes i and j if the edge was present in exactly k MOA-specific binary networks. Visualization of the network was based on igraph R package with parameters charge = 0.1 and spring.length = 100 in the layout.graphopt function.

Ranked-based correlation. Rank-based enrichment analysis was performed to compute correlation scores of the RNA-seq signatures with all the DrugMatrix Affymetrix Liver Signatures contained in the NextBio database[33]. There are 657 DrugMatrix-based Affymetrix Liver Signatures derived from 197 different chemical treatments over a variety of doses and durations of exposure contained in NextBio. The RNA-seq signatures employed for the analysis were the same used for Ingenuity and MetaCore pathway enrichment analysis. Only gene identifiers recognized by NextBio were considered in the analysis.

P-values associated with rank-based enrichment analysis were calculated using the Running Fisher algorithm as described elsewhere[33].

Detecting genes, transcripts and exon-exon junctions. For gene detection, the threshold was four read pairs or eight reads mapped depending on whether or not reads are mapped in pairs by the RNA-seq data analysis pipeline. The counts data for genes and transcripts from pipeline P6 were then used for differential expression test by edgeR. Based on Ensembl gene annotation (Release r58) that was adopted by pipeline P6, RNAs were further separated into two categories (protein coding genes and noncoding RNAs). Detection of exon-exon junctions and coverage quantification were done through TopHat[46] with the subsequent differential expression test conducted through edgeR. For both count-based differential expression test methods (edgeR and DESeq), the absolute fold change cutoff at 1.5 and uncorrected $P < 0.05$ were also adopted as in the case for limma.

Based on spliced junctions mapped by TopHat within transcripts (pipeline P4), the quantification of the reads for each sample (PHE, PIR and matched control each had their replicate's mapped reads merged) and each RefSeq transcript was determined using the mixture of Isoforms (MISO) approach[35] with default settings but also with the 'paired-end' argument using mean insert size = 164 and s.d. (SD) = 40. The mean insert size and s.d. were determined using a set of long (≥ 1,000 bps) constitutive exons such that both read pairs map within the exon and not outside of it. The reads from a pair that were on the same strand or had no pair mate were discarded. In addition, insert lengths 2-SDs outside of the mean were discarded. MISO uses a Bayesian inference to compute the probability that a read originated from a particular isoform. To find differentially expressed isoforms, we compared treated (PHE or PIR) versus the control sample. A Bayes factor is computed for the exon, which represents the weight of the evidence in the data in favor of differential expression versus not. For example, a Bayes factor of 2 would mean that the isoform/exon is two times more likely to be differentially expressed than not. Statistical significance of a differentially expressed isoform between the treated sample and control was determined using a Fisher's exact test based on the inferred assignment of reads to each isoform and where the corresponding *P*-values of the Fisher's exact tests were computed from the hypergeometric distribution. We restrict the statistical analysis to only those isoforms of a gene where each isoform had at least five reads supporting it. An FDR *q*-value < 0.05 was used to account for multiple comparisons and as threshold for the significance of the differentially expressed isoforms.

Detection of shortened 3′ UTRs. RNA-seq read alignments by TopHat (pipeline P4) were analyzed by a Poisson hidden Markov model (PHMM) to detect 3′ UTR shortening[34]. Briefly, all the terminal exons located within the 3′ UTR region of the RefSeq genes were collected. Only exons with unique genomic coordinates and with length (l) > 600 bps were retained. Next, a sliding window of k base-pairs (bp) was applied to each terminal exon, where the number of reads mapped to each sliding window was recorded and where

$$k = \begin{cases} 100 & \text{if } l < 2 \text{ kb} \\ 200 & \text{if } l < 4 \text{ kb} \\ 400 & \text{if } l < 8 \text{ kb} \\ 800 & \text{if } l \geq 8 \text{ kb} \end{cases}$$

In order to identify the potential shortened 3′ UTR of a gene transcript, we applied the PHMM to the sequences of read counts obtained above. If the total read counts (from all windows) is less than 10, we do not perform the model fitting due to low coverage.

We fit the sequence with a two-state PHMM, i.e., $S_x = 1{,}2$, with parameters estimated by the Expectation and Maximization (EM) method using the depmixS4 R package. Transcripts with potential shortened 3′ UTRs (i.e., having two or more states) were selected based on the Bayesian information criterion (*BIC*) of the model fits as follows:

$$S_x = \begin{cases} 1 & \text{if } BIC_1 < BIC_2 \\ 2 & \text{if } BIC_1 \geq BIC_2 \end{cases}$$

where BIC_1 is the Bayesian information criterion from the 1-state model and BIC_2 is the Bayesian information criterion from the 2-state model. If a 2-state model is preferred, we only select transcripts with transitions from high-expression state to the low-expression state (as here we only focus on 3′ UTR shortening). Statistical significance of a shortened 3′ UTR for a transcript between control and treated samples is determined using a Fisher's exact test. The mean of the read counts as integers for the high expression state and the low expression state (padded by 5 to account for zero counts) are used for comparing the proportions between the two samples. For the genes with PHMM modeled, fitted parameters, the corresponding *P*-values of the Fisher's exact test are computed from the hypergeometric distribution. An FDR *q*-value ≤ 0.01 is used as a threshold for the significance of the difference of 3′ UTR shortening between the two samples.

Prediction methodology. A variety of modeling methods were used to fully explore the predictive potentials of both platforms. The modeling process usually consisted of the following four steps: (i) pre-processing and ratio data generation; (ii) discriminative feature selection; (iii) classifier training; and (iv) application to the test set. Several dimensions had been explored in each step. For Step 1, there were two options (RMA and MAS5) tested for Affymetrix microarray data processing and six different pipelines for RNA-seq. When ratio data were generated, some models applied some filtering based on expression levels and variance across control samples. For some approaches, the feature selection step was embedded in the classifier training step. Cross-validation during the training step was not performed owing to the small sample size. These were the commonly used classifier approaches: K-nearest neighbors (KNN), Support Vector Machines (SVM), Principal Vector Algorithm (PVA), graph-based semi-supervised learning, and a novel method based on covariance. During the last step, each model assigned each sample in the test set to one of the following three MOAs: PPARA, CAR/PXR, or unknown. Details about each modeling method were included in the **Supplementary Note 2**.

38. Ganter, B. *et al.* Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *J. Biotechnol.* **119**, 219–244 (2005).
39. Liu, W.M. *et al.* Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* **18**, 1593–1599 (2002).
40. Affymetrix Technical Note. *Guide to Probe Logarithmic Intensity Error (PLIER) Estimation* (http://www.affymetrix.com/support/technical/technotes/plier_technote.pdf) (2005).
41. Wu, Z., Irizarry, R.A., Gentleman, R., Martinez-Murillo, F. & Spencer, F. A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.* **99**, 909–917 (2004).
42. Fox, J. & Weisberg, S. *An R Companion to Applied Regression* (Sage, Thousand Oaks, CA, 2011).
43. Wingender, E. *et al.* The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.* **29**, 281–283 (2001).
44. Wingender, E. *et al.* TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* **28**, 316–319 (2000).
45. Breslin, T., Krogh, M., Peterson, C. & Troein, C. Signal transduction pathway profiling of individual tumor samples. *BMC Bioinformatics* **6**, 163 (2005).
46. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

doi:10.1038/nbt.3001