TOXICOLOGICAL SCIENCES **130(2)**, 427–439 (2012)
doi:10.1093/toxsci/kfs250
Advance Access publication August 13, 2012

RNA-Seq Provides New Insights in the Transcriptome Responses Induced by the Carcinogen Benzo[a]pyrene

Joost van Delft,*,†,[1],[2] Stan Gaj,*,†,[2] Matthias Lienhard,‡ Marcus W. Albrecht,‡ Alexander Kirpiy,‡ Karen Brauers,* Sandra Claessen,* Daneida Lizarraga,*,† Hans Lehrach,‡ Ralf Herwig,‡ and Jos Kleinjans*,†

**Department of Toxicogenomics, Maastricht University, Maastricht, The Netherlands; †Netherlands Toxicogenomics Centre, Maastricht, The Netherlands; and ‡Department Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Berlin, Germany*

[1] To whom correspondence should be addressed at Department of Toxicogenomics, Faculty of Health, Medicine and Life Sciences, Maastricht University, Universiteitssingel 50, 6229 ER Maastricht, The Netherlands. Fax: +31 (43) 3884146. E-mail: j.vandelft@maastrichtuniversity.nl.
[2] These authors contributed equally to this study.

Received March 12, 2012; accepted August 6, 2012

Whole-genome transcriptome measurements are pivotal for characterizing molecular mechanisms of chemicals and predicting toxic classes, such as genotoxicity and carcinogenicity, from *in vitro* and *in vivo* assays. In recent years, deep sequencing technologies have been developed that hold the promise of measuring the transcriptome in a more complete and unbiased manner than DNA microarrays. Here, we applied this RNA-seq technology for the characterization of the transcriptomic responses in HepG2 cells upon exposure to benzo[a]pyrene (BaP), a well-known DNA damaging human carcinogen. Based on EnsEMBL genes, we demonstrate that RNA-seq detects ca 20% more genes than microarray-based technology but almost threefold more significantly differentially expressed genes. Functional enrichment analyses show that RNA-seq yields more insight into the biology and mechanisms related to the toxic effects caused by BaP, i.e., two- to fivefold more affected pathways and biological processes. Additionally, we demonstrate that RNA-seq allows detecting alternative isoform expression in many genes, including regulators of cell death and DNA repair such as TP53, BCL2 and XPA, which are relevant for genotoxic responses. Moreover, potentially novel isoforms were found, such as fragments of known transcripts, transcripts with additional exons, intron retention or exon-skipping events. The biological function(s) of these isoforms remain for the time being unknown. Finally, we demonstrate that RNA-seq enables the investigation of allele-specific gene expression, although no changes could be observed. Our results provide evidence that RNA-seq is a powerful tool for toxicology, which, compared with microarrays, is capable of generating novel and valuable information at the transcriptome level for characterizing deleterious effects caused by chemicals.

Key Words: **RNA-seq; chemical carcinogenesis; gene expression profiling; microarrays; DNA-reactive agents.**

One of today’s main challenges in pharmaceutical, chemical, and cosmetic industries is to accurately assess the toxic properties and mechanisms of new and known chemical entities, such as on carcinogenicity and genotoxicity, preferably without the use of animal experiments. From the early days of microarray-based genomics technologies, these have been embraced to this aim by the toxicological research community (Aardema and MacGregor, 2002; Decristofaro and Daniels, 2008; Waters and Fostel, 2004). Gene expression profiling is currently widely applied for unravelling mechanisms that underlie toxic properties of chemicals as well as for predicting toxicity of chemical compounds (Kim *et al.*, 2005; Mathijs *et al.*, 2009; McHale *et al.*, 2010; Paules *et al.*, 2011). Now, the emerging next-generation DNA-sequencing technologies and their application for gene expression analyses have the potential to advance transcriptomics-based risk assessment.

Over the last decade, several microarray-based platforms have been developed for whole-genome gene expression profiling, the most prominent ones being provided by Affymetrix, Agilent, and Illumina. For the purpose of enhancing the acceptance of the genomic profiling technologies for chemical risk assessment by regulatory authorities, both the microarray technology as well as their associated data analysis approaches have been thoroughly validated (MicroArray Quality Control I and II [MAQC I and II]) (Arikawa *et al.*, 2008; Canales *et al.*, 2006; Chen *et al.*, 2007; Liu *et al.*, 2009; Shi *et al.*, 2006).

Although all this is promising, microarray technology suffers from some important limitations. First, the platforms are inflexible and incomplete, since hybridization probes are only present for “known” or predicted transcripts, therefore, by default the unknowns are missing. Information on splice and other variants is hardly available, although exon arrays can partly fill that gap (Chen *et al.*, 2011; Laajala *et al.*, 2009). Secondly, microarray scanners have a limited dynamic range and sensitivity that hampers the detection of many low-expressed genes and are at best semiquantitative, as demonstrated by MAQC I (Shi *et al.*, 2006). Due to this, the so-called whole-genome arrays do not provide complete and accurate information on the actual cellular transcriptome and on how that can be affected by toxic compounds. Risk assessment of chemical exposure, however, requires quantitative knowledge of dose-response relationships, and it is obvious that because of their limited utility, microarray technologies cannot completely meet that goal.

The arrival of deep sequencing applications for transcriptome analyses, RNA-seq, may circumvent these disadvantages of microarray platforms. Digital counting of all transcripts that are actually present in the cell, holds the potential to provide unbiased and complete measurements of all small and large RNA molecules present in a cell (Wang *et al.*, 2009). This implies that no *a priori* knowledge of the transcriptome is required, that splice variants and other isoforms can be measured, and that the dynamic range is unlimited (Sultan *et al.*, 2008). However, RNA-seq results are still difficult to interpret because of the huge amounts of data generated, which in turn results in new challenges for the processing of sequence alignments as well as for subsequent statistical analyses. Also, the restricted *a priori* knowledge of the transcriptome hampers functional interpretation. Furthermore, although new and improved algorithms/tools are constantly designed, their standardization is still limited (Auer and Doerge, 2010; Bullard *et al*., 2010).

To date, within the toxicology community, the application of RNA-seq has hardly been evaluated. To our knowledge, there are only two publications available that describes this application. One focuses on RNA-seq analysis of kidney tissue taken from rats treated with aristolochic acid (Su *et al.*, 2011). In this study, RNA-seq appeared to generate a consistent biological interpretation compared with traditional microarray platforms while simultaneously generating more sensitive results. However, alternate splicing and other sequence variations were not investigated. The second study deals with RNA-seq analyses of human A549 cells exposed to $NiCl_2$ (Tchou-Wong *et al*., 2011). In this article, no comparison was made with microarray technology and their RNA-seq analysis did not cover transcript isoforms.

Thus, as there is a clear need for further toxicological investigations to better understand the advantages and limitations of RNA-seq, the aim of this article is to explore the performance of RNA-seq by comparing its results with those from classical microarray analysis. Information on genes, transcripts, splice variants, and allele-specific expression has been gathered and the functional interpretation of these data at the level of pathways and biological processes performed. The input to this study is RNA from a human hepatoma (HepG2) liver cell line, exposed for 12 h and 24 h to benzo[a]pyrene (BaP), a ubiquitously present and potent DNA-damaging carcinogen, that is also considered to be a human carcinogen by the International Agency for Research on Cancer (IARC) (http://monographs.iarc.fr/ENG/Monographs/vol100F/mono100F-14.pdf). BaP is especially interesting for transcriptomic studies, as it has both genotoxic and nongenotoxic properties. The genotoxic properties are due to biotransformation of BaP to DNA-reactive metabolites thus also generating reactive oxygen species (Burczynski *et al*., 1999; Cavalieri and Rogan, 1995; Cheng *et al*., 1989). The nongenotoxic properties of polycyclic aromatic hydrocarbons (PAH) are assumed to result from its capability of activating the transcription factor *AhR*, the aromatic hydrocarbon receptor, which controls the transcription of many biotransformation genes (Ma, 2001; Nebert *et al*., 2004). In addition, BaP-mediated oxidative stress leads via the activation of transcription factor *Nrf2* to the induction of multiple phase I and II biotransformation enzymes (Wang *et al*., 2007). Recently, an extensive time series study on BaP-induced transcriptome changes in HepG2 demonstrated that a network of transcription factors may regulate the effects on functional gene sets (van Delft *et al.*, 2010). The mRNA profiles obtained from these networks at 12 h and 24 h exposure time points were clearly different, making these attractive time points for the RNA-seq study.

Several previous microarray studies have shown that HepG2 cells are capable of metabolizing BaP, which in turn alters the expression of numerous genes, such as those related to the DNA damage response, apoptosis, cell cycle, DNA repair, metabolism, etc. (Hockley *et al.*, 2006, 2009; Lin and Yang, 2007). These findings indicate that the HepG2 cells are a good model for this study, although we acknowledge that because HepG2 cells are already transformed. As a consequence some gene changes may be missed. For comparison with microarray-based methods, the same samples were also investigated using the Affymetrix whole-genome HGU133Plus 2.0 GeneChip.

MATERIALS AND METHODS

Chemicals. BaP, purity 96%, CAS no. 50-32-8 was obtained from Sigma-Aldrich (Zwijndrecht, The Netherlands). The chemical was dissolved in dimethyl sulfoxide (DMSO).

Cell culture and treatment. Human hepatocellular carcinoma HepG2 cells (ATCC HB-6065) were used in all experiments. HepG2 cells were maintained as a monolayer culture in 95% humidity, an atmosphere with 5% of CO_2, and at 37°C. HepG2 cells were passaged at preconfluent densities with the use of a trypsin-EDTA solution. Cells were cultured and passaged in a minimal essential medium with 10% of fetal bovine serum, 1% penicillin/streptomycin, 1% sodium-pyruvate, and 1% nonessential amino acids. All medium compounds were obtained from Gibco BRL (Breda, The Netherlands). Three milliliter of cells (1×10^5/ml) were seeded into each well of a six-well microtitre plate. When the cells were 80% confluent, the medium was replaced with fresh medium containing either 2μM BaP or the vehicle control (0.5% DMSO) during 12 h and 24 h in two independent experiments (biological replicates). More details have been provided elsewhere (Jennen *et al.*, 2010).

RNA isolation. Total RNA was isolated after 12 h and 24 h of incubation with BaP or DMSO in HepG2 using the miRNeasy mini kit (Qiagen Westburg BV, Leusden, The Netherlands) according to manufacturer's instructions, followed by a DNase I (Qiagen Inc) treatment. RNA quantity was measured on a spectrophotometer and the quality was determined on the BioAnalyzer system (Agilent Technologies, Breda, The Netherlands). All RNA samples showed clear 18S and 28S rRNA peaks and demonstrated a RNA integrity number (RIN) level higher than 8.

RNA-seq: Library preparation and sequencing. Poly-A RNA was purified with the Dynabeads mRNA purification kit (Invitrogen) following the manufacturer's instructions and was treated for 30 min at 37°C with TURBO DNase (Ambion; 0.2 units/1 µg of RNA). First- and second-strand synthesis followed the manufacturer's protocol. About 500 ng of double-stranded cDNA was fragmented by sonication with a UTR200 (Hielscher Ultrasonics GmbH, Germany). DNA was parceled in 2% high-resolution agarose gel and 230–270 bp fragments were cut. Library amplification consisted of the following two steps: (1) real-time PCR was performed on StepOne Real-Time PCR System with SYBR Green core reagents (Applied Biosystems) according to the manufacturer's instructions and (2) large scale PCR was performed on PTS-100 Programmable Thermal Controller (MJ Research Inc.) with 2×Phusion HF master mix (NEB) and primers for amplification of the mate-paired library ("PE_PCR primer 1.0" and "PE_PCR primer 2.0," Illumina). Amplified material was loaded to a flow cell at a concentration of 8 pM. Sequencing was carried out on the Illumina Genome Analyzer by running 51 paired-end cycles according to the manufacturer's instructions. For real-time monitoring of run quality and parameters Illumina's Sequencing Control Software and for image analysis and base-calling Illumina Genome Analyzer Pipeline v1.5.0 was used. Resulting FASTQ files were quality checked using FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/).

RNA-seq: Analysis. RNA-seq reads were mapped to the EnsEMBL v58 cDNA sequences using bowtie v0.12.7 (Turro *et al.*, 2011). On top of this alignment for uniquely mappable reads, RNA-seq by expectation-maximization (RSEM) v1.1.7 has been used to assign the ambiguous reads to different transcript isoforms using an expectation-maximization algorithm for estimation of each gene's isoform distribution (Langmead *et al.*, 2009). After the alignment step, RSEM provides read quantification at the individual transcript and gene level.

Both lists were separately imported into R for further statistical analysis using the Bioconductor package edgeR (Robinson and Oshlack, 2010). Differentially expressed (DE) genes/transcripts were identified using the following criteria: (1) absolute fold-change ≥ 1.5, (2) q value (false discovery rate [FDR]) ≤ 0.05, and (3) average read count in at least one experimental group ≥ 10.

Affymetrix microarray analysis. Sample preparation, hybridization, washing, staining, and scanning of the Affymetrix Human Genome U133 Plus 2.0 GeneChip arrays were conducted according to the manufacturer's manual as previously described (Jennen *et al.*, 2010). Quality controls were within acceptable limits. Hybridization controls were called present on all arrays and yielded the expected increases in intensities. The data discussed in this publication have been deposited in Gene Expression Omnibus (Barrett and Edgar, 2006) under accession number GSE36244.

The Affymetrix CEL files were imported in R 2.12.2 under BioConductor v2.7 (Gentleman *et al.*, 2004). Probe sets were reannotated using the BrainArray custom Chip Definition File (CDF) v13.0 annotations for both EnsEMBL genes and transcripts, generating two separate datasets. These custom CDFs were based on EnsEMBL v58. Next, both datasets were separately log transformed, Robust Multi-Average (RMA) normalized, and further evaluated through an extensive quality control pipeline consisting of box plots, Probe Level Model fitting (fitPLM), Normalized Unscaled Standard Error (NUSE), Relative Log Expression (RLE), clustering/heat maps, Principal Component Analysis (PCA), and correlation plots. There were no technically deviating arrays. The limma package (Smyth, 2005) was applied for computing a linear model using time and dose as main components. The resulting *p* values were familywise error rate (FWER) corrected using the false discovery rate (FDR) approach. Genes were considered to be DE when their (1) absolute fold change ≥ 1.5, (2) q value (FDR) < than 0.05, and (3) average expression in at least one of the two experimental groups > 6 (log2-scale).

Functional interpretation. The biological outcome of the experiment was assessed using a combined strategy consisting of interpreting and visualization of the modulated genes at the biological pathway level (PathVisio; van Iersel *et al.*, 2008) as well as of a gene ontology (GO) classification method as performed by GO-Elite (Zambon *et al.*, 2012).

For PathVisio, all human pathways were obtained from WikiPathways (Pico *et al.*, 2008). Also, a gene database (HS_Derby_20100601.gdb) was used to map the genes on pathways. The annotations present in this gene database were based on EnsEMBL v58. For each time point, a separate gene list was imported using EnsEMBL gene identifiers. After the data were imported, pathway statistics were calculated based on the DE genes by means of a Z-score. A pathway is considered significantly altered when (1) Z-score ≥ 1.96, (2) the minimum amount of DE genes was three, and (3) permutation *p* value ≤ 0.05 (*n* = 2.000).

GO-Elite v1.21 was used for functional classification using GO terminology. In preparation for the analysis, the EnsEMBL v58 database was downloaded in combination with its associated GO OBO database (release date: 26 July 2009). Gene identifiers of the DE genes were placed in the input file and the identifiers of the detected genes were placed in the denominator file. For a more thorough interpretation of the results, we only focused on GO terms belonging to the biological processes (P) class. A GO term was considered to be differentially altered when (1) Z-score > 1.96, (2) there were at least three DE genes present in that process, and (3) permuted *p* value ≤ 0.05 (*n* =2.000). Large GO-processes (> 300 genes) were filtered out to minimize process redundancy.

Expression of alternative isoforms. In order to assess differences in isoform expression upon BaP treatment, we have used an entropy argument (Cover and Thomas, 2006). Short reads were counted for each isoform *i* of a gene separately using RSEM. An isoform was called expressed if at least 10 reads were annotated for the respective gene sequence and had at least one read for the specific isoform. Let $P_1,\ldots, P_N$ be the different relative isoform counts of a particular gene, i.e., P_1 1 … + P_N + 1, then we defined isoform entropy (IE) as

$$\mathrm{IE}{=}\sum_i \log_2\left(\mathrm{P}_i\right)P_i \quad \text{where } i = 1,\ldots,N.$$

IE is a measure for the uniformity of expression of the different isoforms. It is high if many of the gene's isoforms are expressed at similar levels and low if only one or few isoforms are predominantly expressed. In order to identify differences in isoform entropy between the BaP-treated and untreated HepG2 cells, we computed for each gene the log2-ratio of IEs computed from the two states and considered IEs to be different if their absolute log2-ratio was ≥1.

Expression of alleles. Allele-specific gene expression can be measured if a heterozygous single nucleotide polymorphism (SNP) is present in the transcript.

Raw sequence reads of untreated HepG2 cells were aligned using bowtie (v0.12.7) against UCSC human genome (hg19) reference. Only reads that aligned to a unique position within the reference were used for the detection of SNPs. Further, potential PCR artifacts were filtered using rmdup (from samtools v0.1.17). Overall, 78,510,329 reads were used as input to the detection of gene polymorphisms, which yields on average 37× coverage of the HepG2 transcriptome. SNP calling was incorporated with samtools (mpileup, bcfutils, vcfutils) using standard parameters except for the maximum coverage, which was artificially set to 10,000 and the minimum mutational penetrance, which was set to 20. Statistical tests for each treatment/control pair at each SNP were performed to investigate changes in the allelic ratio using a binomial test.

RESULTS

Differential Expression in HepG2 Cells Upon BaP Treatment Using RNA-Seq

In total, eight lanes were sequenced resulting in a total of 110.7 million reads. For each lane, the percentage of total reads uniquely mapped to the EnsEMBL v58 genome was between 57.7 and 61.6%. Applying the RSEM algorithm increased these mapping numbers to 74.8–78.4%. All alignment statistics are summarized in Supplementary table 1.

TABLE 1
Interpretation of BaP-Induced Gene Expression Changes Measured by RNA-Seq In HepG2 at the Level of Biological Pathways (WikiPathways)

12 h pathway name	Z-score	24 h pathway name	Z-score
BaP metabolism	6.33#	BaP metabolism	7.38#
Oxidative stress	5.09#	Cholesterol biosynthesis	6.08
Keap1-Nrf2	4.63#	Codeine and morphine metabolism	4.16
GPCRs, class A rhodopsin-like	3.67	Keap1-Nrf2	3.85#
Metapathway biotransformation	3.47#	Urea cycle and metabolism of amino groups	3.66
Myometrial relaxation and contraction	3.12#	Metapathway biotransformation	3.65#
Nucleotide GPCRs	3.06	Oxidative stress	3.34#
Hypertrophy model	2.96	Statin pathway	3.26
Estrogen metabolism	2.94#	Estrogen metabolism	3.13#
Focal adhesion	2.88	Tryptophan metabolism	2.74#
DNA damage response	2.82	Adipogenesis	2.31
Blood clotting cascade	2.77	Inflammatory response pathway	2.30
Biogenic amine synthesis	2.70	Myometrial relaxation and contraction	2.18#
Calcium regulation in the cardiac cell	2.64		
DNA damage response (ATM dependent)	2.32		
Glutathione metabolism	2.28		
Wnt signaling pathway	2.18		
Ovarian infertility genes	2.14		
Type II diabetes mellitus	2.11		
G-protein signaling pathways	2.07		
Tryptophan metabolism	1.98#		

# Changed at both 12 h and 24 h.

Applying the filtering criteria to each time point resulted in the detection of 13,385 (12 h) and 12,918 (24 h) EnsEMBL-annotated genes, which were subsequently used for downstream analysis. Within these gene lists, 1080 (12 h) and 704 (24 h) genes were found to be DE between BaP-treated and untreated cells (see Materials and Methods). These gene sets were then further interpreted with regard to their biological context (pathways) as well as their function (GO).

Pathway analysis revealed 21 (at 12 h) and 13 (at 24 h) significantly altered pathways (Table 1). The majority of these changed pathways were related to relevant toxicological processes like BaP metabolism, biotransformation, and oxidative stress. Furthermore, after 12 h of exposure, BaP induced large changes in both the ATM-dependent and non–ATM-dependent DNA damage response.

Analyses of GO terms belonging to the biological processes (P) class indicated that 387 processes were affected at 12 h and 214 processes at 24 h, with 97 overlapping processes. A full list of all regulated processes is provided in Supplementary table 2. Many of the processes affected at both time points are associated with responses to xenobiotic compounds and stress responses. Also, at 12 h modifications of apoptosis-related pathways were found, whereas at 24 h effects on sterol metabolism and DNA damage response were observed.

Alterations of Isoform Expression Upon BaP Treatment

An advantage of RNA-seq compared with microarrays is that sequencing allows the quantification of gene expression at the isoform level, such as splice variants and polyadenylation variants. After alignment of the short reads to EnsEMBL transcripts, we observed a large number of genes that expressed two or more isoforms regardless of exposure time and treatment. Approximately, 57% of the genes expressed more than two isoforms over any of the experimental conditions, indicating that alternative splicing is a common event in HepG2 cells (Fig. 1A).

In order to measure differences in isoform expression distribution upon BaP treatment, we applied an entropy criterion (see Materials and Methods). Figure 1B displays the histogram of the log2-ratios (treatment vs. control) of isoform entropy. The vast majority of genes did not change isoform entropy upon BaP treatment, which implies that isoform distribution was not altered and thus that alternative splicing is not affected by BaP. However, after treatment, 839 genes showed a change in isoform entropy of more than twofold meaning that there is either a loss or a gain of isoforms upon BaP treatment. Interestingly, these genes contained, among others, 13 apoptosis genes (out of 86 genes from the Kyoto Encyclopedia of Genes and Genomes pathway, which represents a significant overrepresentation, Fisher's $P = 0.000113$, $Q = 0.0394$). Several of these genes that are known to be associated with carcinogenic processes, i.e., *TP53*, *BCL2*, and *XPA* and with oxidative stress response such as *AKR1B10*, are depicted in Fig. 1C–F. Additionally, changes of isoform entropy were observed across the full range of expression strength. For example, *TP53* is highly expressed in the cells with 1194–1675 short read assignments (reads per kilobase per million [RPKM]: 14.40–18.97), whereas *BCL2* is expressed at a very low level with 11–30 short reads (RPKM: 0.05–0.14).

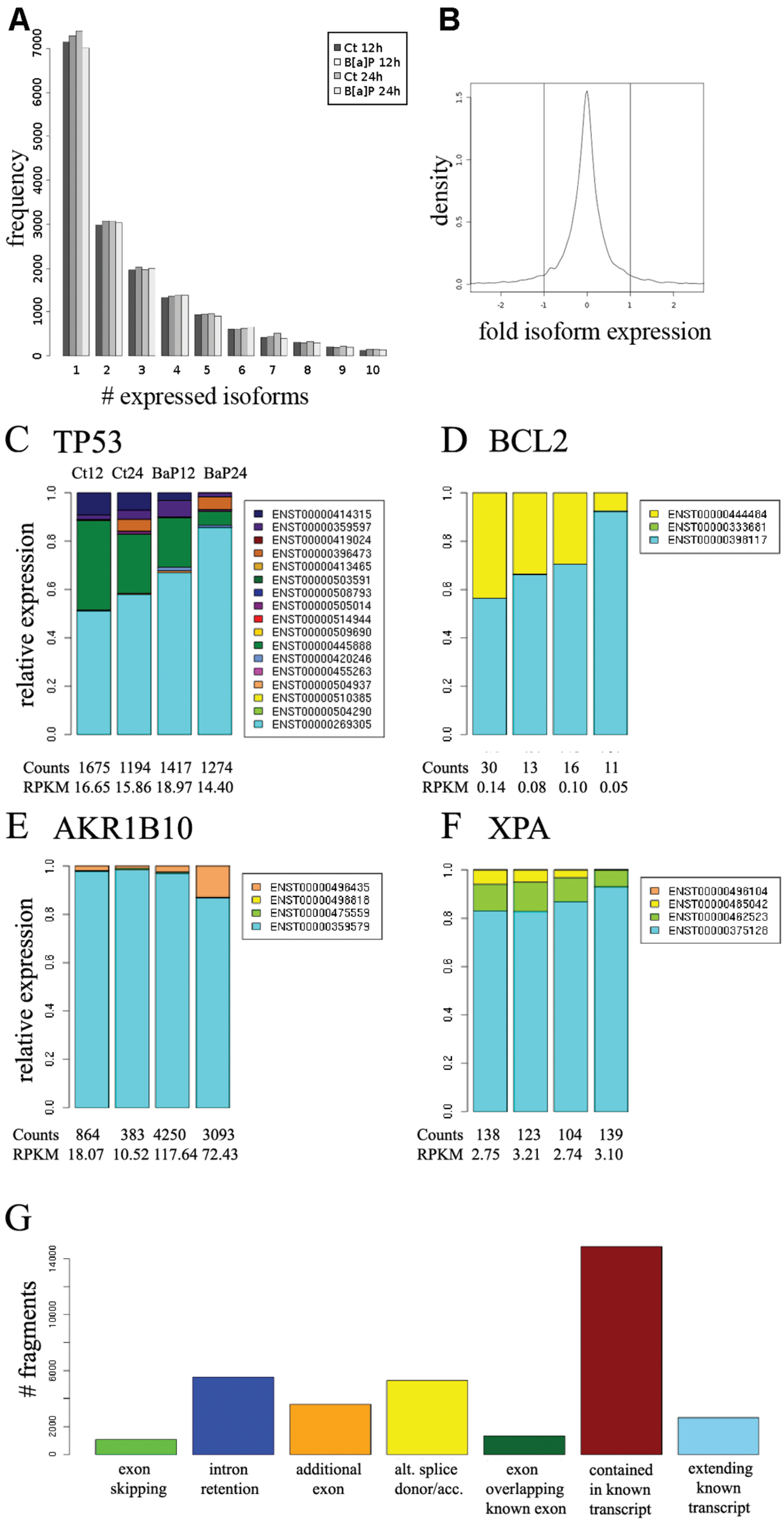

FIG. 1. (A) Histogram showing the number of expressed isoforms per gene in the four experimental settings. (B) Density plot of fold-changes (log2-scale) of isoform entropy (IE) in BaP-treated cells (24 h exposure time) versus untreated cells. (C) *TP53* isoforms. Bars show relative proportions of reads mapped to the different isoforms with respect to the four experimental conditions. Isoforms are color coded with a legend given in the box. Total reads and RPKM values taking into account the library size and the gene length are shown below each bar. (D) *BCL2* isoforms, (E) *AKR1B10* isoforms, and (F) *XPA* isoforms.

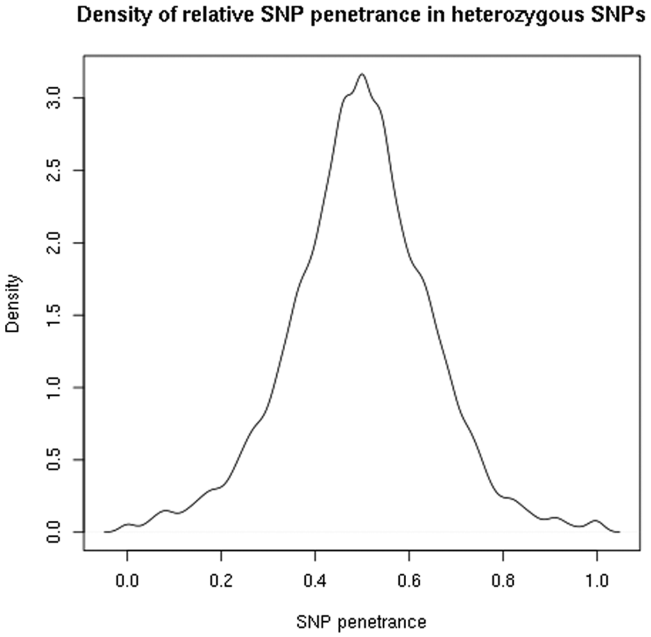

FIG. 2. SNP penetrance density plot demonstrating that most genes have balanced expression of both alleles.

Characterizing Alterations of Allele-Specific Expression Upon BaP Treatment

Because HepG2 is a human hepatocarcinoma cell line, the abundant presence of heterozygous SNPs enables us to investigate the allele-specific transcriptional landscape, which might help in explaining eminent features with respect to *in vitro* toxicology and pharmacology in future studies. In order to identify heterozygous SNPs in HepG2 cells, we processed all uniquely mapped short reads. In total, 7351 heterozygous SNPs were assigned to coding regions of which 3749 and 2471 were sufficiently covered in treatment and control samples (≥ 20 reads) in the 12-h and 24-h samples, respectively. Most SNPs refer to known variants across human populations (dbSNP entries) (Sherry *et al.*, 2001), whereas 241 had no match and appear specific for HepG2 cells.

The SNP penetrance density plot has a peak at 0.5, hence most genes show a balanced expression of both alleles (Fig. 2). The binomial test found 38 genes in the 12-h samples and 27 genes in the 24-h samples to have changed allelic ratios between treatment and control samples ($p < 0.05$), however, following correction for multiple testing, no gene was significant in either experiment. Therefore, we conclude that there is no evidence that BaP influences allele-specific gene expression in HepG2 cells.

Detection of Novel Isoforms

Another advantage of RNA-seq over microarray technology is the possibility to detect novel isoforms. We used the combination of TopHat/Cufflinks in order to detect splice junctions and compared the assembled sequence reads against EnsEMBL annotation (Supplementary fig. 1). In total, 60,403 expressed transcript assemblies were identified, of which 28,966 (48%) were classified as "potentially novel isoforms." In order to distinguish between the unknown alternative splicing events detected by cufflinks, the class of "potentially novel isoforms" was further examined comparing the transcript boundaries with known annotation. Specific types of alternative splicing included unknown exons, exon skipping, intron retention, alternative splice donor, or acceptor sites. In 23,487 out of 28,966 transcripts that were tagged as "potentially novel isoforms," we found evidence for at least one of the analyzed alternative splicing events. The majority of these transcripts (14,872) were fragments of known transcripts, whereas 3588 isoforms had additional exons that did not overlap with known exons, most of them in front or after the known gene structure, leading to "extended" transcripts (2636). In total, 5305 transcripts had alternative splice sites, not present in EnsEMBL 58, whereas 5530 transcripts showed evidence for intron retention. Novel exon skipping events where detected in 1080 cases (Fig. 1G).

Comparison of RNA-Seq and Microarray Results

The outcomes of the Illumina RNA-seq and Affymetrix platforms were obviously evaluated at the gene level because the microarray did not allow us to measure isoform-specific expression levels. For each platform, a baseline of detectable genes that were mapped against EnsEMBL genes was established and summarized in Supplementary table 3. On the Affymetrix platform, 11,435 (12 h) and 11,383 (24 h) reannotated genes passed the detection criteria. For RNA-seq, these numbers were slightly higher, i.e., 13,385 (12 h) and 12,918 (24 h). In each dataset, 90–95% of all genes were annotated with a meaningful EnsEMBL description and/or gene name. Within each platform, the log ratios of the biological replicates showed a high correlation (Affymetrix: R = 0.87; RNA-seq: R = 0.72; Supplementary figs. 2A and 2B), whereas those between platforms were slightly lower (Replicate 1: R = 0.71; Replicate 2: R = 0.72; Supplementary figs. 2C and 2D). Across all platforms and time points, there was a large overlap in detected genes (9846), as illustrated in Fig. 3A. Additionally, each platform detected a fraction of unique genes, which could not be confirmed by the other platform. With respect to Affymetrix readouts, about ~65% of these additional genes were also picked up by the RNA-seq system but showed low counts. Similarly, with respect to the RNA-seq platform, about ~35% of these additional genes were measured by means of the Affymetrix technology and corresponded with a very low intensity. All other additional RNA-seq genes were not present on the Affymetrix chip.

The changes reported by RNA-seq were mostly larger than measured on the Affymetrix chip, examples are as follows: (1) *CYP1A1* that showed a 93-fold (12 h) and 79-fold (24 h) increase on Affymetrix, whereas RNA-seq reported a 199-fold (12 h) and 214-fold (24 h) increase and (2) *ALDH3A1* that showed a 6.3-fold (12 h) and 6.1-fold (24 h) increase on Affymetrix compared with the 32-fold (12 h) and 36-fold (24 h) increase reported by RNA-seq.

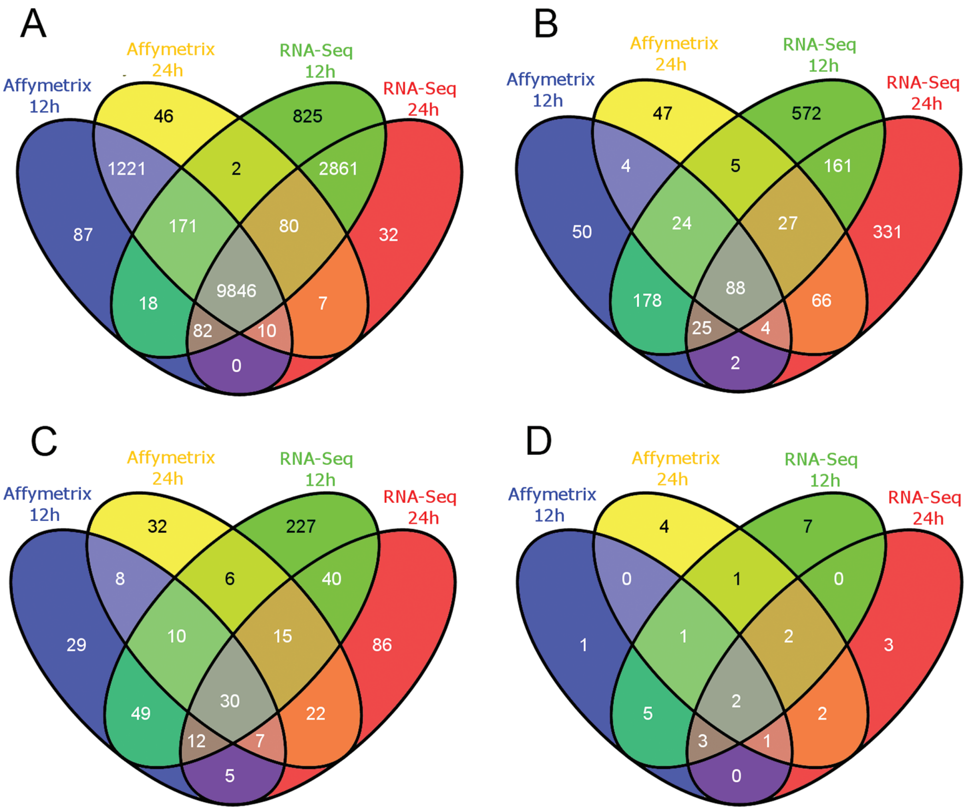

FIG. 3. Between-platform comparison of total number of (A) detectable genes, (B) DE genes, (C) significantly altered biological processes (GO), and (D) significantly altered biological pathways (WikiPathways).

Our next focus was on the total number of DE genes per time point. Figure 3B shows that the majority of the DE genes found by Affymetrix were also demonstrated to be DE by the RNA-seq platform (12 h: 315; 24 h: 185). However, RNA-seq analysis resulted in a much higher number of DE genes, many of which were not found by microarray technology. Gene expressions that are expected to be affected by BaP were indeed identified on both platforms, such as *CYP1A1*, which is regulated by *AHR*, *NQO1*, and *AKR1B10*, which are regulated by *NRF2*, and *CDKN1A*, which is regulated by *TP53*. Furthermore, the proto-oncogenes *FAS* and *MDM2* were also demonstrated to be upregulated by both platforms, whereas expressions of the proto-oncogenes *FOS* and *JUN* and the tumor suppressor *EXT1* were only retrieved by RNA-seq. At both time points, 161 genes were uniquely identified by RNA-seq as DE. These included the *NRF2*-regulated gene *OSGIN1*, whereas there were only four DE genes that were only found by the microarrays.

Next, these lists of DE genes were interpreted in a functional context (GO biological process annotations; Fig. 3C and Supplementary tables 3 and 5) as well as in a biological pathway context (WikiPathways; Fig. 3D and Supplementary fig. 4). Not only was the RNA-seq approach able to cover a larger part of the genes in nearly all relevant biological pathways and GO processes (Fig. 4 and Supplementary fig. 3), but it also provided a larger number of biological processes to be significantly affected compared with the microarray analyses. For pathways, at both exposure periods, there appears a considerable overlap between both platforms, with RNA-seq identifying more pathways. A between-platform comparison of the most important pathways is presented in Fig. 5, a similar presentation for all pathways is provided in Supplementary fig. 4. Two pathways were observed under all four conditions (both time points and both platforms), namely "benzo(a) pyrene metabolism" and "metapathway biotransformation." Some are only observed at 12 h of treatment, e.g., "glutathione metabolism," or at 24 h, such as "cholesterol biosynthesis." Moreover, 357 processes and 10 pathways were uniquely identified by RNA-seq, whereas with microarray analysis, these numbers were lower (69 and 5, respectively). More platform-specific pathways are observed for RNA-seq than for microarrays, such as "Wnt signaling pathway," which at 12 h of BaP exposure is only observed for RNA-seq, or "Keap1-Nrf2" and "oxidative stress," which at 24 h are only observed for RNA-seq. The DNA damage response pathway can be considered the prototypical effect for a genotoxic carcinogen like BaP. A part of this pathway that focuses on the regulation of TP53 is shown in Fig. 6, where the affected genes under all experimental conditions and platforms are also visualized. The complete DNA damage response pathway (and gene visualization) can be found in Supplementary fig. 5.

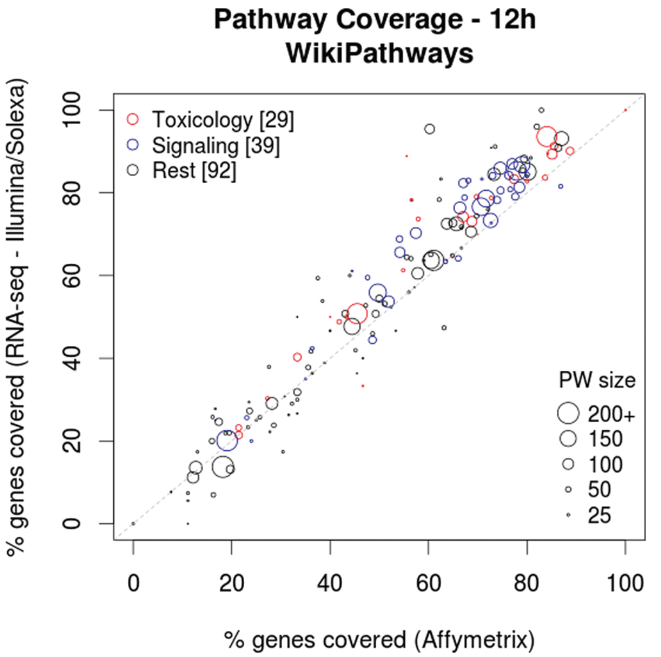

FIG. 4. Coverage plot comparing the relative number of detected genes in WikiPathways between each platform after 12 h BaP exposure showing that the majority of the relevant pathways (i.e., [coverage] >50%) show an increase of measurable genes using the RNA-seq approach. The circle size corresponds with the total number of genes in that pathway. A distinction was also made between toxicology-related pathways (red), signaling pathways (blue), and all other Wikipathways pathways (black).

DISCUSSION

Because the development of next-generation high-throughput DNA sequencing technologies several years ago, gene expression analysis based on "digital counting" of the transcripts (RNA-seq) has come within reach. Here, we explored the performance of RNA-seq within a toxicogenomics setting. The gene expression changes induced in HepG2 cells by the well-studied carcinogen BaP have been investigated. Information on genes, transcripts, splice variants, as well as the functional interpretation of these data at the level of pathways and biological processes was gathered. For anchoring and comparing with microarray-based methods, the same samples were also investigated using the Affymetrix whole-genome HGU133Plus 2.0 GeneChips.

The major advantages of RNA-seq are that it provides insight into all transcripts and their variants present in the cell as it is not biased to the "known transcripts," and that it is quantitative over a large dynamic range thanks to digital counting of the transcripts (Sultan *et al.*, 2008; Wang *et al.*, 2009). These added values are confirmed by our study. Unfortunately, interpreting the transcript data in a functional and biological pathway context requires the mapping of the reads to the "known transcripts," thereby losing information on "unknown" transcripts. Despite this, we demonstrated that RNA-seq results in more enriched biological processes and pathways being identified as significantly affected compared with microarray analyses (Fig. 3C and D). Also more platform-specific pathways are found by RNA-seq than by microarrays. This implies that RNA-seq provides better insight into the biology and mechanisms related to the toxic effects caused by BaP.

Both at the level of genes as well as that of biological processes and pathways, profound differences are observed between the 12 h and 24 h exposure periods (Fig. 3B–D, Supplementary table 1, and Supplementary fig. 4), which is in agreement with previous publications (Hockley *et al.*, 2006; van Delft *et al.*, 2010). For instance, in an extensive time series study using 20k genes Agilent arrays, temporal profiles for functional gene sets demonstrated both early and late effects in up- and downregulation of relevant gene sets involved in cell cycle, apoptosis, DNA repair, and metabolism of amino acids and lipids. Pathway analyses of the RNA-seq data indicate that at 12 h specifically G-protein-coupled receptor pathways and DNA damage response pathways are affected, whereas at 24 h this was "cholesterol biosynthesis" and "codeine and morphine metabolism." Most of these time-dependent differences were also observed by the microarray analyses. Also in the previous study, it was shown that genotoxic responses occur before effects on metabolism pathways (van Delft *et al.*, 2010). Visualization of the effects in the "DNA damage response" pathway demonstrates that especially the network of genes around TP53 is upregulated (Fig. 6). This is irrespective of the transcriptomics platform used. It also shows for several genes that the effects at 12 h are larger than at 24 h, which agrees with the pathway analyses.

Also by microarray analyses, unique DE genes and biological processes and pathways were found, which suggests that this technology may reveal biology that is missed by RNA-seq but may also represent false-positive observations. Obviously, neither can be excluded. Examples are the pathways "cytochrome P450" and "nuclear receptors in lipid metabolism and toxicity." However, because methods based on quantitative hybridization are sensitive to mismatches, cross-hybridization with other mRNAs may occur, resulting in reporting DEs that actually are not affected. It is noteworthy to mention that microarray platforms report in general smaller gene expression changes than RT-PCR-based methods, as was shown in the MAQC I project (Canales *et al.*, 2006; Shi *et al.*, 2006). Furthermore, because RNA-seq detects more genes, the background list of unaffected genes also increases. In overrepresentation analysis tools like PathVisio and GO-Elite, these background lists are required for correct statistical analyses. When a background list increases, significant changes may become less significant. For instance, the biological process "GO:0042770 DNA damage response, signal transduction" was significant for microarrays with 7 DE out of 74 measurable genes (permuted p value = 0.01), whereas for RNA-seq, 10 DE among 80 measurable genes was not significant (permuted p value = 0.214). Nevertheless, it is undisputable that the RNA-seq analysis represents a more complete and thus a more correct picture of the biology.

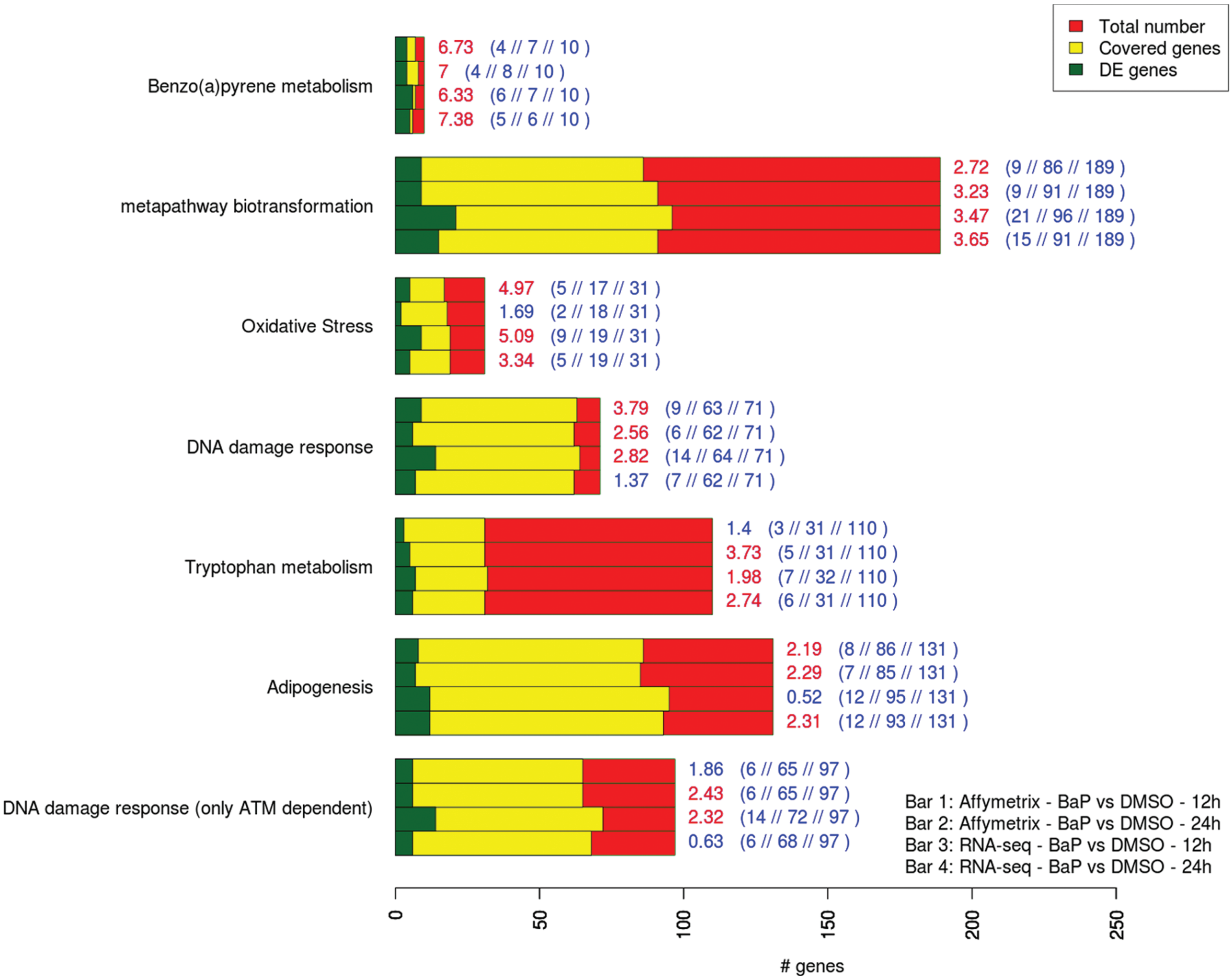

FIG. 5. Between-platform comparisons of the pathways affected in HepG2 cells by BaP treatment. A summary is displayed of some relevant toxicological pathways that were significantly altered across most experimental conditions. The length of the bar corresponds with the pathway size (red), number of genes in the pathway detected by the platform (yellow), and the number of DE genes found in that pathway (green). The Z-score as well as the numerical values are displayed next to each bar. This Z-score is colored red when that pathway was significantly altered at that time point.

Besides the detection of DE genes and known transcript isoforms, RNA-seq also offers the potential to detect novel isoforms. Although the majority of transcript assemblies overlapped with current Ensembl annotation, we detected 1080 transcripts with novel exon-skipping events that correspond to 735 different genes. These genes span a wide range of functional annotations including cancer and genotoxic response pathways, e.g., p53 signaling (*CHEK1*, *DDB2*, *ABL1*), apoptosis (*PLEKHG2*), and cell cycle (*CDC25*, *RBL1*).

We in particular showed that RNA-seq can provide information on BaP-induced gene expression changes at the level of splice variants and allows investigating allele-specific effects, which to our knowledge has never before been described in the context of a toxicogenomics study. It has been shown previously that DNA damage induces apoptosis and regulates alternative splicing (Muñoz *et al.*, 2009). Where BaP treatment effects on alternative splicing have not yet been globally investigated, however, previous work has selectively identified transcribed genes with activated alternative splicing after BaP treatment (Yan *et al.*, 2010). Furthermore, BaP has been found to induce alternative splicing of *MDM2* in lung cancer cell lines and that this process is associated with lung cancer (Weng *et al.*, 2005).

For 839 gene expressions, we observed isoform changes, meaning that either a loss or a gain of isoforms occurred following BaP treatment. Interestingly, several of these genes are known to be relevant for chemical carcinogenesis, such as *TP53*, *BCL2*, and XPA or chemical oxidative stress response, such as *AKR1B10* (Fig. 1D–F). *TP53* and *BCL2* are both tumor suppressor genes involved in cell cycle and/or apoptosis of cells following exposure to DNA damaging carcinogens like BaP. In fact, *TP53* is a transcription factor and has a key function in the regulation of expression of many genes following the DNA damage formation, also in HepG2 cells (Hockley *et al.*, 2006; van Delft *et al.*, 2010). Especially at 24 h of BaP treatment,

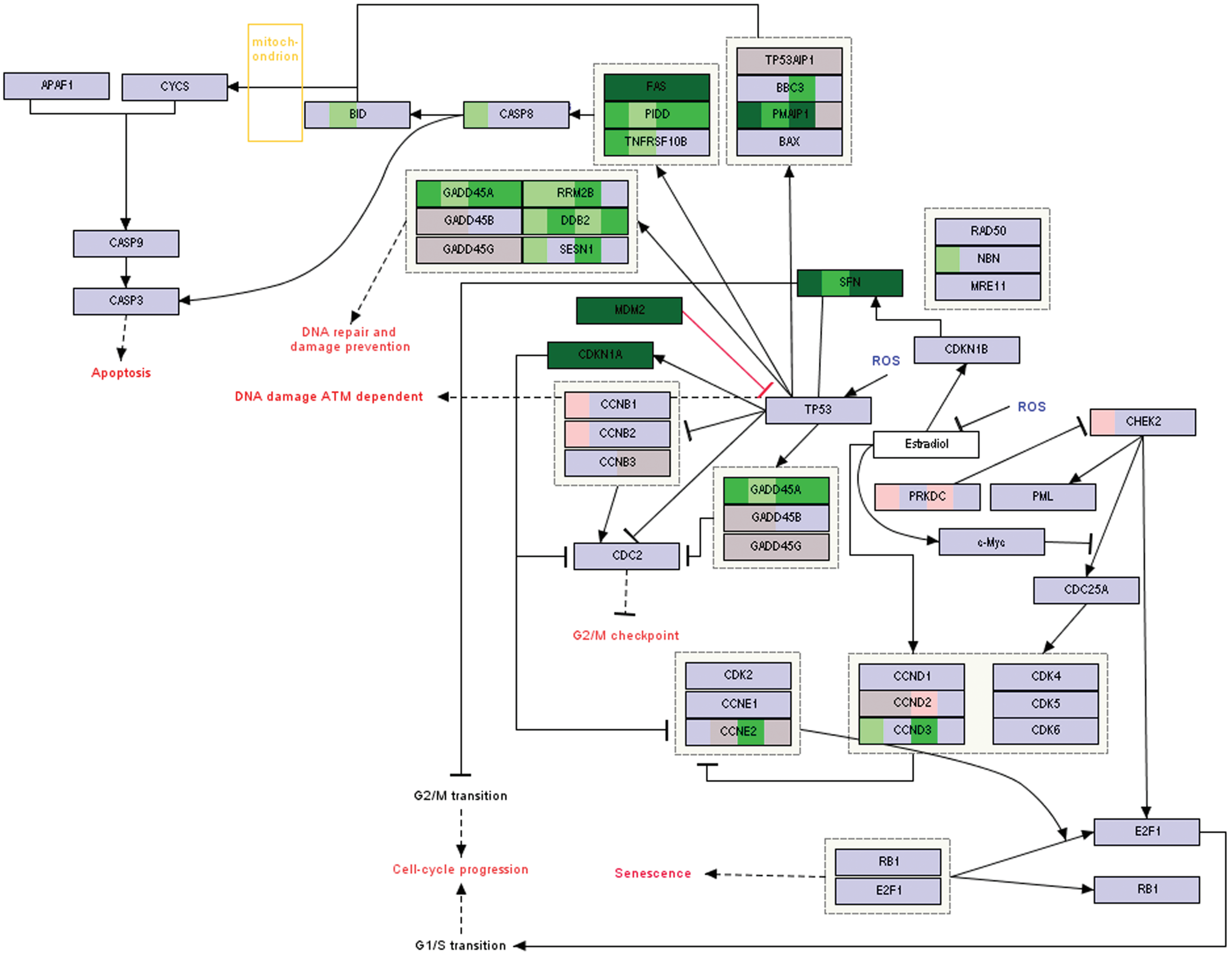

FIG. 6. Part of the "DNA damage response" pathway, thereby focussing on the network around TP53 and visualizing gene expression changes according to analyses by RNA-seq and Affymetrix arrays. Every block represents a gene present in the DNA damage response pathway. Every block is divided in four subblocks, corresponding from left to right with: "Affymetrix 12h," "Affymetrix 24h," "RNA-seq 12h," and "RNA-seq 24h." The color code is based on the following: (1) grey for genes not passing the detection criteria, (2) light-blue for not DE genes, (3) light-green or light-red (up/downregulated) for genes that are statistically altered but did not meet our FC criteria, and (4) dark-green/dark-red color for up/downregulated genes that are DE.

a large shift in transcript isoforms occurs, mainly due to the decrease of ENST0000044588 and ENST00000414315 and increase of the main transcript ENST00000269305 (Fig. 1C). ENST00000414315 codes for a truncated protein, whose function is unclear. A total of 10 human *TP53* isoforms are currently characterized described in literature that are produced by alternative splicing, use of alternative translation site, or alternative promoter (reviewed in Marcel and Hainaut, 2009 and http://www-p53.iarc.fr/p53isoforms.html), whereas in EnsEMBL, even 16 isoforms are mentioned. Current knowledge on the role and activities of these isoforms is still limited. Also for *BCL2*, a splice variant that codes for a truncated protein decreases following BaP treatment (Fig. 1D). To date, no information is available about the possible role of this variant protein. *AKR1B10* is one of the most induced genes by BaP, already giving an almost 10-fold induction at early 12h. This has been reported previously as well and agrees with activation of the transcription factor *NRF2* by reactive metabolites and/or oxidative stress (Nishinaka *et al.*, 2011; van Delft *et al.*, 2010). Interestingly, at 24h, the level of splice variant ENST00000496435 is increased even further (Fig. 1E), although this variant does not appear to code for a protein product. *XPA* is involved in nucleotide excision repair of DNA damages, among others that caused by BaP (de Vries *et al.*, 1997a, 1997b). Following BaP treatment, transcript ENST00000485042 is almost completely lost at 24h (Fig. 1F). The relevance of this transcript is unknown, as it appears not to code for a protein product. All together, these data show that exposure to genotoxic compounds causes differential expression of splice variants. The biological and mechanistic relevance of these are to-date unknown but difficult to ignore because isoform variation in many relevant genes occurs. To understand the significance of these, more in-depth studies will be needed, e.g., by overexpressing specific splice variants and the analyses of protein variants using proteomics-based technologies.

Furthermore, we measured isoform expression by RT-PCR in order to validate differential expression, which appeared difficult. Although we could confirm the relative expression of the different isoforms in general and also could confirm the induction upon BaP treatment, validating the changes in relative expression of the different isoforms between treatment and control samples was not quite successful (results not shown). Validating isoform expression levels is a difficult task because it is typically dependant on multiple and different factors (Fang and Cui, 2011) such as annotation differences of the transcripts between different genome databases (i.e., EnsEMBL and RefSeq) and the numerical methods that quantify transcript-specific expression. Additionally, for detecting isoform expression differences upon chemical treatment a further increase in sequencing depth might still be needed in order to resolve the expression levels of only moderately expressed isoforms.

HepG2 cells are human by origin and thus heterozygous for their chromosome pairs, with each cell having maternal and paternal chromosomes. Furthermore, allelic imbalance in gene expression is a phenomenon known to occur in man (Ju *et al.*, 2011; Zhang *et al.*, 2009). Due to the presence of heterozygous SNPs, it is possible to investigate allele-specific transcriptional changes by RNA-seq, but so far this has not been exploited in a toxicological context. Despite that we were technologically able to investigate this, we did not find any evidence that BaP influences allele-specific gene expression changes in HepG2 cells.

To date, there are a few studies that also investigated the relationship between microarray and RNA-seq gene expression data (Beane *et al.*, 2011; Bottomly *et al.*, 2011; Liu *et al.*, 2011; Marioni *et al.*, 2008; Su *et al.*, 2011). Marioni *et al.* (2008) illustrated that the data between the microarrays and RNA-seq, both Illumina platforms, were highly reproducible. After comparison of DE genes, they concluded that there is a high correlation for gene expressions that had on average a large number of reads (> 250), whereas this correlation decreased when the number of reads decreased. These results are in concordance with our own (see Supplementary fig. 2). Bottomly *et al.* (2011) investigated the expression levels between two different mouse strains and reported that the Illumina RNA-seq platform was in high concordance with the output obtained from the Illumina and Affymetrix microarray platforms. As with our results, the RNA-seq platform was able to detect more DE genes. Beane *et al.* (2011) compared different RNA-seq protocols, sequencing types (paired end/single end) in combination with results obtained from Affymetrix exon and whole-genome arrays. Among all DE genes detected in the RNA-seq experiment, only 21% was also DE on either the exon or whole-genome chip. Thus, our study is in agreement with these previous findings.

In conclusion, RNA-seq data appear to provide a more complete picture at the level both of DE genes and of affected pathways than microarrays. Additionally, we highlight the potential of RNA-seq for characterizing mechanisms related to alternative splicing and other transcript isoforms and to allele-specific expression, and therewith gathering new information. Our results give evidence that RNA-seq is a powerful tool for toxicogenomics that, compared with microarrays, will add valuable information at the transcriptome level for characterizing toxic effects caused by chemicals, as evidenced from the large numbers of processes and pathways that are missed by microarrays. Clearly, more in-depth studies are needed to further substantiate our findings. Among others, studies should be directed on verification of splice variants and the differential expression for these.

SUPPLEMENTARY DATA

Supplementary data are available online at http://toxsci.oxfordjournals.org/.

ACKNOWLEDGMENTS

The authors would like to express their thanks to Bernd Timmermann, Marc Sultan, and Marie-Laure Yaspo for support in the next-generation sequencing and to Rachel Cavill for critical reviewing the manuscript.

FUNDING

The Netherlands Genomics Initiative for The Netherlands Toxicogenomics Center [grant number 050-06-510], the European Commission under its 6th framework project carcinoGENOMICS (PL037712), the 7th framework project diXa (283775), the German Ministry for Education and Research under its MedSys program (PREDICT 0315428A), and the Max Planck Society.

REFERENCES

Aardema, M. J., and MacGregor, J. T. (2002). Toxicology and genetic toxicology in the new era of "toxicogenomics": Impact of "-omics" technologies. *Mutat. Res.* **499,** 13–25.

Arikawa, E., Sun, Y., Wang, J., Zhou, Q., Ning, B., Dial, S. L., Guo, L., and Yang, J. (2008). Cross-platform comparison of SYBR Green real-time PCR with TaqMan PCR, microarrays and other gene expression measurement technologies evaluated in the MicroArray Quality Control (MAQC) study. *BMC Genomics* **9,** 328.

Auer, P. L., and Doerge, R. W. (2010). Statistical design and analysis of RNA sequencing data. *Genetics* **185,** 405–416.

Barrett, T., and Edgar, R. (2006). Gene expression omnibus: Microarray data storage, submission, retrieval, and analysis. *Meth. Enzymol.* **411,** 352–369.

Beane, J., Vick, J., Schembri, F., Anderlind, C., Gower, A., Campbell, J., Luo, L., Zhang, X. H., Xiao, J., Alekseyev, Y. O., *et al.* (2011). Characterizing the impact of smoking and lung cancer on the airway transcriptome using RNA-Seq. *Cancer Prev. Res. (Phila).* **4,** 803–817.

Bottomly, D., Walter, N. A., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., Searles, R. P., Mooney, M., McWeeney, S. K., and Hitzemann, R. (2011). Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLOS ONE* **6,** e17820.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11,** 94.

Burczynski, M. E., Lin, H. K., and Penning, T. M. (1999). Isoform-specific induction of a human aldo-keto reductase by polycyclic aromatic hydrocarbons (PAHs), electrophiles, and oxidative stress: Implications for the alternative pathway of PAH activation catalyzed by human dihydrodiol dehydrogenase. *Cancer Res.* **59,** 607–614.

Canales, R. D., Luo, Y., Willey, J. C., Austermiller, B., Barbacioru, C. C., Boysen, C., Hunkapiller, K., Jensen, R. V., Knight, C. R., Lee, K. Y., *et al.* (2006). Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.* **24,** 1115–1122.

Cavalieri, E. L., and Rogan, E. G. (1995). Central role of radical cations in metabolic activation of polycyclic aromatic hydrocarbons. *Xenobiotica* **25,** 677–688.

Chen, J. J., Hsueh, H. M., Delongchamp, R. R., Lin, C. J., and Tsai, C. A. (2007). Reproducibility of microarray data: A further analysis of microarray quality control (MAQC) data. *BMC Bioinformatics* **8,** 412.

Chen, P., Lepikhova, T., Hu, Y., Monni, O., and Hautaniemi, S. (2011). Comprehensive exon array data processing method for quantitative analysis of alternative spliced variants. *Nucleic Acids Res.* **39,** e123.

Cheng, S. C., Hilton, B. D., Roman, J. M., and Dipple, A. (1989). DNA adducts from carcinogenic and noncarcinogenic enantiomers of benzo[a]pyrene dihydrodiol epoxide. *Chem. Res. Toxicol.* **2,** 334–340.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wily-Interscience, Hoboken, NJ.

de Vries, A., Dollé, M. E., Broekhof, J. L., Muller, J. J., Kroese, E. D., van Kreijl, C. F., Capel, P. J., Vijg, J., and van Steeg, H. (1997a). Induction of DNA adducts and mutations in spleen, liver and lung of XPA-deficient/lacZ transgenic mice after oral treatment with benzo[a]pyrene: Correlation with tumour development. *Carcinogenesis* **18,** 2327–2332.

de Vries, A., van Oostrom, C. T., Dortant, P. M., Beems, R. B., van Kreijl, C. F., Capel, P. J., and van Steeg, H. (1997b). Spontaneous liver tumors and benzo[a]pyrene-induced lymphomas in XPA-deficient mice. *Mol. Carcinog.* **19,** 46–53.

Decristofaro, M. F., and Daniels, K. K. (2008). Toxicogenomics in biomarker discovery. *Methods Mol. Biol.* **460,** 185–194.

Fang, Z., and Cui, X. (2011). Design and validation issues in RNA-seq experiments. *Brief. Bioinformatics* **12,** 280–287.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., *et al.* (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80.

Hockley, S. L., Arlt, V. M., Brewer, D., Giddings, I., and Phillips, D. H. (2006). Time- and concentration-dependent changes in gene expression induced by benzo(a)pyrene in two human cell lines, MCF-7 and HepG2. *BMC Genomics* **7,** 260.

Hockley, S. L., Mathijs, K., Staal, Y. C., Brewer, D., Giddings, I., van Delft, J. H., and Phillips, D. H. (2009). Interlaboratory and interplatform comparison of microarray gene expression analysis of HepG2 cells exposed to benzo(a) pyrene. *OMICS* **13,** 115–125.

Jennen, D. G., Magkoufopoulou, C., Ketelslegers, H. B., van Herwijnen, M. H., Kleinjans, J. C., and van Delft, J. H. (2010). Comparison of HepG2 and HepaRG by whole-genome gene expression analysis for the purpose of chemical hazard identification. *Toxicol. Sci.* **115,** 66–79.

Ju, Y. S., Kim, J. I., Kim, S., Hong, D., Park, H., Shin, J. Y., Lee, S., Lee, W. C., Kim, S., Yu, S. B., *et al.* (2011). Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nat. Genet.* **43,** 745–752.

Kim, J. Y., Kwon, J., Kim, J. E., Koh, W. S., Chung, M. K., Yoon, S., Song, C. W., and Lee, M. (2005). Identification of potential biomarkers of genotoxicity and carcinogenicity in L5178Y mouse lymphoma cells by cDNA microarray analysis. *Environ. Mol. Mutagen.* **45,** 80–89.

Laajala, E., Aittokallio, T., Lahesmaa, R., and Elo, L. L. (2009). Probe-level estimation improves the detection of differential splicing in Affymetrix exon array studies. *Genome Biol.* **10,** R77.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25.

Lin, T., and Yang, M. S. (2007). Benzo[a]pyrene-induced elevation of GSH level protects against oxidative stress and enhances xenobiotic detoxification in human HepG2 cells. *Toxicology* **235,** 1–10.

Liu, Q., Sung, A. H., Chen, Z., Liu, J., Huang, X., and Deng, Y. (2009). Feature selection and classification of MAQC-II breast cancer and multiple myeloma microarray gene expression data. *PLOS ONE* **4,** e8250.

Liu, S., Lin, L., Jiang, P., Wang, D., and Xing, Y. (2011). A comparison of RNA-Seq and high-density exon array for detecting differential gene expression between closely related species. *Nucleic Acids Res.* **39,** 578–588.

Ma, Q. (2001). Induction of CYP1A1. The AhR/DRE paradigm: Transcription, receptor regulation, and expanding biological roles. *Curr. Drug Metab.* **2,** 149–164.

Marcel, V., and Hainaut, P. (2009). p53 isoforms—a conspiracy to kidnap p53 tumor suppressor activity? *Cell. Mol. Life Sci.* **66,** 391–406.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18,** 1509–1517.

Mathijs, K., Brauers, K. J., Jennen, D. G., Boorsma, A., van Herwijnen, M. H., Gottschalk, R. W., Kleinjans, J. C., and van Delft, J. H. (2009). Discrimination for genotoxic and nongenotoxic carcinogens by gene expression profiling in primary mouse hepatocytes improves with exposure time. *Toxicol. Sci.* **112,** 374–384.

McHale, C. M., Zhang, L., Hubbard, A. E., and Smith, M. T. (2010). Toxicogenomic profiling of chemically exposed humans in risk assessment. *Mutat. Res.* **705,** 172–183.

Muñoz, M. J., Pérez Santangelo, M. S., Paronetto, M. P., de la Mata, M., Pelisch, F., Boireau, S., Glover-Cutter, K., Ben-Dov, C., Blaustein, M., Lozano, J. J., *et al.* (2009). DNA damage regulates alternative splicing through inhibition of RNA polymerase II elongation. *Cell* **137,** 708–720.

Nebert, D. W., Dalton, T. P., Okey, A. B., and Gonzalez, F. J. (2004). Role of aryl hydrocarbon receptor-mediated induction of the CYP1 enzymes in environmental toxicity and cancer. *J. Biol. Chem.* **279,** 23847–23850.

Nishinaka, T., Miura, T., Okumura, M., Nakao, F., Nakamura, H., and Terada, T. (2011). Regulation of aldo-keto reductase AKR1B10 gene expression: Involvement of transcription factor Nrf2. *Chem. Biol. Interact.* **191,** 185–191.

Paules, R. S., Aubrecht, J., Corvi, R., Garthoff, B., and Kleinjans, J. C. (2011). Moving forward in human cancer risk assessment. *Environ. Health Perspect.* **119,** 739–743.

Pico, A. R., Kelder, T., van Iersel, M. P., Hanspers, K., Conklin, B. R., and Evelo, C. (2008). WikiPathways: Pathway editing for the people. *PLOS Biol.* **6,** e184.

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11,** R25.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311.

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., de Longueville, F., Kawasaki, E. S., Lee, K. Y., *et al.*; MAQC

Consortium. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24,** 1151–1161.

Smyth, G. (2005). Limma: Linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, 397–420.

Su, Z., Li, Z., Chen, T., Li, Q. Z., Fang, H., Ding, D., Ge, W., Ning, B., Hong, H., Perkins, R. G., *et al.* (2011). Comparing next-generation sequencing and microarray technologies in a toxicological study of the effects of aristolochic acid on rat kidneys. *Chem. Res. Toxicol.* **24,** 1486–1493.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., *et al.* (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321,** 956–960.

Tchou-Wong, K. M., Kiok, K., Tang, Z., Kluz, T., Arita, A., Smith, P. R., Brown, S., and Costa, M. (2011). Effects of nickel treatment on H3K4 trimethylation and gene expression. *PLOS ONE* **6,** e17728.

Turro, E., Su, S. Y., Gonçalves, Â., Coin, L. J., Richardson, S., and Lewin, A. (2011). Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol.* **12,** R13.

van Delft, J. H., Mathijs, K., Staal, Y. C., van Herwijnen, M. H., Brauers, K. J., Boorsma, A., and Kleinjans, J. C. (2010). Time series analysis of benzo[a]pyrene-induced transcriptome changes suggests that a network of transcription factors regulates the effects on functional gene sets. *Toxicol. Sci.* **117,** 381–392.

van Iersel, M. P., Kelder, T., Pico, A. R., Hanspers, K., Coort, S., Conklin, B. R., and Evelo, C. (2008). Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics* **9,** 399.

Wang, W., Kwok, A. M., and Chan, J. Y. (2007). The p65 isoform of Nrf1 is a dominant negative inhibitor of ARE-mediated transcription. *J. Biol. Chem.* **282,** 24670–24678.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10,** 57–63.

Waters, M. D., and Fostel, J. M. (2004). Toxicogenomics and systems toxicology: Aims and prospects. *Nat. Rev. Genet.* **5,** 936–948.

Weng, M. W., Lai, J. C., Hsu, C. P., Yu, K. Y., Chen, C. Y., Lin, T. S., Lai, W. W., Lee, H., and Ko, J. L. (2005). Alternative splicing of MDM2 mRNA in lung carcinomas and lung cell lines. *Environ. Mol. Mutagen.* **46,** 1–11.

Yan, C., Wu, W., Li, H., Zhang, G., Duerksen-Hughes, P. J., Zhu, X., and Yang, J. (2010). Benzo[a]pyrene treatment leads to changes in nuclear protein expression and alternative splicing. *Mutat. Res.* **686,** 47–56.

Zambon, A. C., Gaj, S., Ho, I., Hanspers, K., Vranizan, K., Evelo, C. T., Conklin, B. R., Pico, A. R., and Salomonis, N. (2012). GO-Elite: A flexible solution for pathway and ontology over-representation. *Bioinformatics* **28,** 2209–2210.

Zhang, K., Li, J. B., Gao, Y., Egli, D., Xie, B., Deng, J., Li, Z., Lee, J. H., Aach, J., Leproust, E. M., *et al.* (2009). Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat. Methods* **6,** 613–618.