

# RNA-seq data analysis at the gene and CDS levels provides a comprehensive view of transcriptome responses induced by 4-hydroxynonenal†

Cite this: *Mol. Biosyst.*, 2013, **9**, 3036

Qi Liu,<sup>\*ab</sup> Jody Ullery,<sup>c</sup> Jing Zhu,<sup>a</sup> Daniel C. Liebler,<sup>cde</sup> Lawrence J. Marnett<sup>c</sup> and Bing Zhang<sup>\*abde</sup>

Reactive electrophiles produced during oxidative stress, such as 4-hydroxynonenal (HNE), are increasingly recognized as contributing factors in a variety of degenerative and inflammatory diseases. Here we used the RNA-seq technology to characterize transcriptome responses in RKO cells induced by HNE at subcytotoxic and cytotoxic doses. RNA-seq analysis rediscovered most of the differentially expressed genes reported by microarray studies and also identified novel gene responses. Interestingly, differential expression detection at the coding DNA sequence (CDS) level helped to further improve the consistency between the two technologies, suggesting the utility and importance of the CDS level analysis. RNA-seq data analysis combining gene and CDS levels yielded an informative and comprehensive picture of gradually evolving response networks with increasing HNE doses, from cell protection against oxidative injury at low dose, initiation of cell apoptosis and DNA damage at intermediate dose to significant deregulation of cellular functions at high dose. These evolving dose-dependent pathway changes, which cannot be observed by the gene level analysis alone, clearly reveal the HNE cytotoxic effect and are supported by IC<sub>50</sub> experiments. Additionally, differential expression at the CDS level provides new insights into isoform regulation mechanisms. Taken together, our data demonstrate the power of RNA-seq to identify subtle transcriptome changes and to characterize effects induced by HNE through the generation of high-resolution data coupled with differential analysis at both gene and CDS levels.

Received 19th March 2013,  
Accepted 2nd September 2013

DOI: 10.1039/c3mb70114j

[www.rsc.org/molecularbiosystems](http://www.rsc.org/molecularbiosystems)

## Introduction

4-Hydroxynonenal (HNE), one of the major aldehydic products of lipid peroxidation, has been suggested to contribute to the development and progression of various diseases.<sup>1–6</sup> HNE reacts with a number of cellular molecules, including DNA, RNA and proteins, and has been shown to trigger multistep signal transduction cascades for suppression of cellular functions in a dose- and time-dependent manner.<sup>7–14</sup>

In a previous study, we used the microarray technology to examine the effects of HNE on gene expression in the RKO cell line.<sup>15</sup> Significant alterations were observed for genes involved in DNA damage and antioxidant, heat shock and ER stress responses. Integrating gene expression changes with protein adduction data further elucidated signaling and transcriptional regulatory mechanisms through which protein adduction triggers gene expression changes.<sup>16,17</sup> However, the datasets and integrative analysis represented only high micromolar treatment concentrations of HNE (*i.e.*, 60  $\mu$ M), as few gene expression changes were detected at lower concentrations using microarray technologies, making it difficult to study the dose-dependent response upon HNE treatment.

RNA-seq has become increasingly used to quantify expression of all genes with their alternative isoforms. Compared with microarray technology, the digital nature of RNA-seq enables a larger dynamic range, higher resolution and lower technical variance in measuring expression abundance, which makes RNA-seq more sensitive in capturing expression differences.<sup>18–22</sup> By properly assigning reads to each isoform, RNA-seq enables quantifying gene expression at an individual transcript level. Moreover, gene expression can also be quantified by grouping

<sup>a</sup> Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN 37232, USA. E-mail: [qi.liu@vanderbilt.edu](mailto:qi.liu@vanderbilt.edu), [bing.zhang@vanderbilt.edu](mailto:bing.zhang@vanderbilt.edu)

<sup>b</sup> Center for Quantitative Sciences, Vanderbilt Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

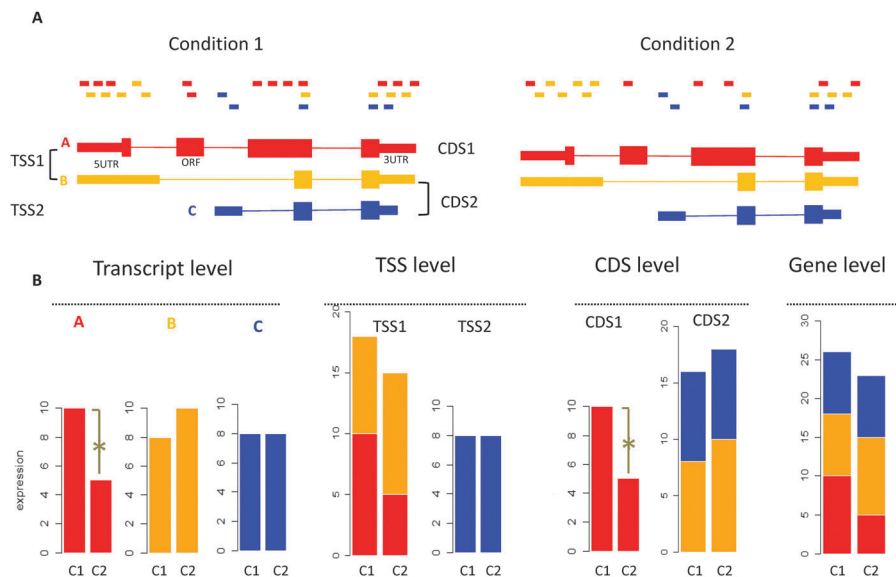
<sup>c</sup> Department of Biochemistry, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

<sup>d</sup> Jim Ayers Institute for Precancer Detection and Diagnosis, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

<sup>e</sup> Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3mb70114j





**Fig. 1** Significant changes detected at the high-resolution level but not at the low-resolution gene level by RNA-seq. (A) The gene produces three isoforms A, B and C at different abundances. TSS or CDS groups are formed by grouping isoforms sharing the same transcription start site (TSS) or coding the same protein sequences (CDS). For example, A and B are within the same TSS group, while B and C are within the same CDS group. (B) Analyzing expression difference at the transcript level, different isoform groups level and the gene level. At the transcript level, the expression of isoform A is significantly changed across conditions, while B and C are not. Adding expression values of A and B yields the expression value for the TSS1 group. At the TSS level, TSS1 and TSS2 groups are not significantly changed. Adding expression values of B and C yields the expression value for the CDS2 group. At the CDS level, the CDS1 group is significantly changed but the CDS2 group is not. Adding expression values of three isoforms yields the expression value of the gene, which is not significantly changed across conditions.

isoforms into biologically meaningful units. For example, isoforms with the same transcription start site (TSS) can be grouped together (Fig. 1, isoforms A and B). These isoforms are derived from the same pre-mRNA and differential expression at the TSS level suggests differential regulation of the pre-mRNA. Similarly, isoforms with the same coding DNA sequences (CDS) can be grouped together because they encode the same protein product (Fig. 1, isoforms B and C), and differential expression at the CDS level indicates potentially different protein outputs. Expression quantification at the transcript level and the intermediate functional unit level allows the detection of expression changes that may not be observable at the gene level. As shown in Fig. 1, RNA-seq reveals expression changes at the transcript level (isoform A) and the CDS level (CDS 1), although no significant change can be observed at the gene level. However, as many isoforms share exons, some reads cannot be accurately assigned to individual isoform. This read assignment uncertainty<sup>23</sup> and noisy splicing<sup>23,24</sup> make differential expression at the transcript level hard to detect and introduce false positives. To our knowledge, which level (gene, CDS group, TSS group, and transcript) is best suited for detecting differential expression has not been well studied.<sup>25</sup>

Here we applied RNA-seq to study the transcriptome changes in RKO cells in response to low, intermediate and high micromolar doses of HNE treatment. We first compared results from RNA-seq and microarrays at high HNE dose. Then we investigated whether the ability of RNA-seq to quantify expression of isoforms or isoform groups (CDS and TSS groups) could provide novel insights. We found that combining gene- and CDS-level analyses improved the consistency between RNA-seq and microarray and helped identify novel genes closely

related to HNE response, especially at low and intermediate HNE doses. This presented a clear picture of gradually evolving response networks with increasing HNE doses, from cell protection against oxidative injury, initiation of cell apoptosis and DNA damage to significant deregulation of cellular pathways. These dose-dependent pathway changes revealed the HNE cytotoxic effect and were supported by IC<sub>50</sub> experiments. Additionally, we discussed the relative contribution of transcriptional noise and isoform switching to the obscured expression changes at the gene level. Our study demonstrates that RNA-seq is a powerful tool to study dose-response relationships of altered pathways. Expression summarized at the CDS level complements gene-level analysis and provides novel and valuable information for characterizing molecular effects induced by HNE.

## Results

RNA-seq was conducted to explore transcriptional changes in RKO cells following treatment for 6 h with 15, 30, or 45  $\mu$ M HNE. Among the total of 1195 million reads, about 81% were aligned to the human genome and 75% were uniquely mapped. Although exons constitute less than 3% of the human genome, about 87% of reads were mapped to exons, suggesting that our poly(A)<sup>+</sup>-selected RNA samples were highly enriched with exonic sequences (Table S1, ESI<sup>†</sup>).

### Improved consistency between microarray and RNA-seq analyses

We compared the intraplatform and interplatform correlations of gene expression in RNA-seq and microarray analyses (Fig. 2)





**Fig. 2** Interplatform and intraplatform correlations of gene expression under control and 45  $\mu$ M HNE treatment between microarray and RNA-seq analyses.

after 45  $\mu$ M HNE treatment. Within each platform, a high reproducibility was observed among biological replicates (RNA-seq: Pearson correlation  $r = 0.98-1$ ; microarray:  $r = 0.98-1$ ). Both platforms detected the same correlation trend between samples: correlations between replicates of 45  $\mu$ M HNE treated samples ( $r = 0.99-1$ ) were slightly higher than those between replicates of controls (no HNE treatment,  $r = 0.98-1$ ), which were higher than those between 45  $\mu$ M HNE treated samples and controls ( $r = 0.95-0.98$ ). In contrast, expression correlations between platforms were much lower, with Pearson correlations ranging from 0.70 to 0.74. These results are in good agreement with previous works that reported high reproducibility among replicates within platforms and much lower correlation coefficients between platforms.<sup>19,21,22</sup>

To compare the capacities of two platforms to capture the expression changes, we also calculated the fold-change-based correlation. Interestingly, the cross-platform correlation was improved by using fold changes (Fig. 3A, Pearson correlation  $r = 0.76$ ). If only differentially expressed genes identified by either RNA-seq or microarray using the criteria of  $\text{abs}(\log_2 \text{FC}) > 1$  and  $\text{FDR} < 0.01$  were considered (91 genes), we obtained an even higher correlation (Fig. 3B, Pearson correlation  $r = 0.89$ ), suggesting that the two platforms were quite consistent in detecting differential expression. Among the 91 differentially expressed genes, 24 were identified by both platforms and 11 could not be detected by RNA-seq gene-level analysis. In contrast, differential expression of another 56 genes could not be captured by microarray analysis (Fig. 3C and Table S2, ESI†).

We used Cuffdiff to extend differential analysis from the gene level to higher resolution levels (transcript, CDS and TSS levels).<sup>26,27</sup> Among the 11 genes detected by microarray but missed by RNA-seq gene-level analysis, two upregulated (GCLM,

TXNRD1) and four downregulated genes (PPRC1, DOT1L, URB2, EGR3) could be rediscovered by the CDS level analysis (Fig. 3B). In contrast, only two upregulated (GCLM, TXNRD1) and two downregulated genes (EGR3, PPRC1) could be rediscovered by transcript or TSS-level analyses.

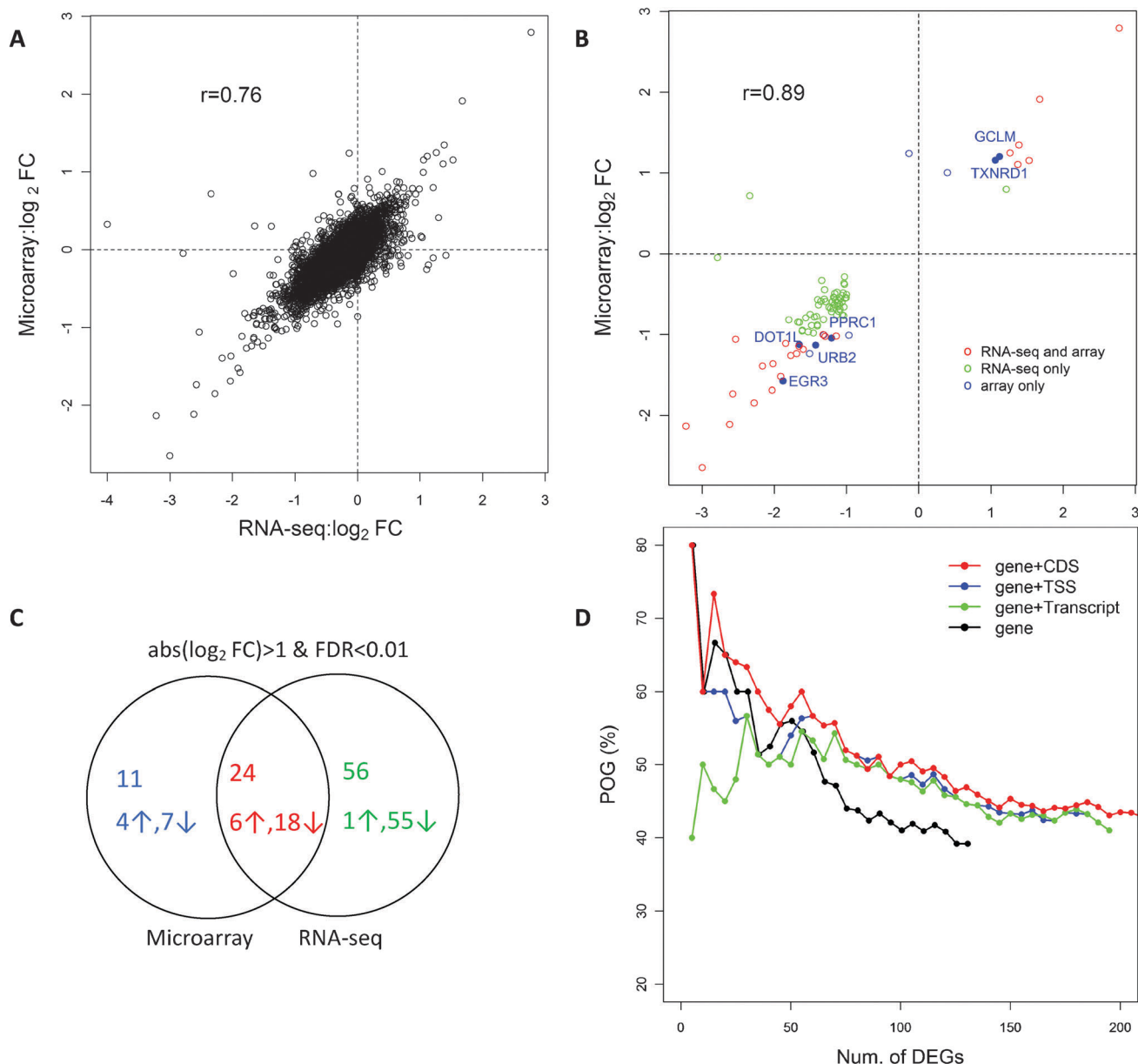
We further investigated the consistency in gene ranking between microarray and RNA-seq analyses at different levels. Using an FDR cutoff of 0.01, genes identified by microarray and RNA-seq through combining results from different levels were ranked by their fold change values, respectively, and were used to calculate the POG (percentage of overlapping genes). As shown in Fig. 3D, the POG between RNA-seq and microarray was improved when differential analysis at CDS, TSS or transcript levels was added to gene-level analysis. However, integrating TSS-level and transcript-level data introduced noise into highly changed genes, which led to lower POGs for the top ranked genes.

The six genes rediscovered by the CDS level analysis, including GCLM, TXNRD1, PPRC1, DOT1L, URB2, and EGR3, are important anti-oxidant genes or genes involved in DNA damage and cell proliferation, indicating their close relationship with HNE treatment. GCLM (the modifier subunit of glutamate cysteine ligase) is the first and the rate-limiting enzyme in the synthesis of GSH, a major player in cellular defense against oxidative stress.<sup>28</sup> TXNRD1 (thioredoxin reductase 1) reduces thioredoxin as well as other substrates and protects the cell from oxidative damage.<sup>29,30</sup> DOT1L (DOT1-like, histone H3 methyltransferase) has been reported to be involved in DNA damage response.<sup>31</sup> Furthermore, based on the assumption that functionally related genes have similar expression changes, we systematically evaluated the biological relevance of differentially expressed CDS using three protein-protein interaction (PPI) datasets (PPI HQ, PPI all and PrePPI, see the Materials and methods section for description).<sup>32,33</sup> Using the criteria of  $\text{FDR} < 0.05$  &  $\text{abs}(\log_2 \text{FC}) > 0.5$ , 297 genes were identified at the gene level and additional 195 genes were detected at the CDS level after 45  $\mu$ M HNE treatment (Fig. 4). The 195 genes detected only at the CDS level were more likely to interact with the 297 genes detected at the gene level than randomly selected genes ( $p = 3.5 \times 10^{-7}$  for PPI HQ,  $p = 2.2 \times 10^{-15}$  for PPI all,  $p = 2.4 \times 10^{-5}$  for PrePPI) (Table 1). These results suggest that differentially expressed CDS are highly likely to be involved in biological processes induced by HNE and differential analysis at the CDS level is a useful and appropriate complement to the gene level analysis.

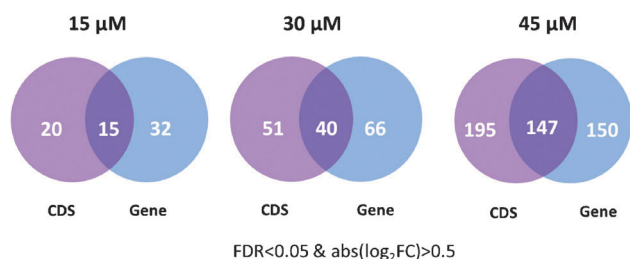
### Gradually evolving response networks presented by the combined level

Differential expressions at the CDS and gene levels were identified using Cuffdiff with  $\text{FDR} < 0.05$  &  $\text{abs}(\log_2 \text{FC}) > 0.5$  after 15, 30 and 45  $\mu$ M HNE treatment. CDS or genes were required to have  $\text{FPKM} > 1$  (Fragments Per Kilobase of transcript per Million fragments mapped) under at least one condition. The numbers of differentially expressed genes reported at CDS and gene levels at 15, 30, and 45  $\mu$ M HNE are illustrated in Venn diagrams (Fig. 4). In agreement with our previous study, the





**Fig. 3** Correlation of RNA-seq and microarray at the level of fold changes upon 45  $\mu\text{M}$  HNE treatment. (A) Correlation of fold change for all genes in microarray and RNA-seq. (B) Correlation of fold change of differentially expressed genes detected either by microarray or RNA-seq using the criteria of  $\text{abs}(\log_2 \text{FC}) > 1$  and  $\text{FDR} < 0.01$ . (C) A Venn diagram of the number of genes detected by microarray and RNA-seq. (D) POG values between the microarray and RNA-seq when the gene-level analysis was combined with higher resolution level analysis, CDS, TSS and transcript levels.



**Fig. 4** Differentially expressed genes detected at the CDS level and the gene level in HNE-treated RKO cells.

**Table 1** Relationships between differential expression detected at the gene level and only at the CDS level based on three PPI datasets. The table lists the observed and the expected number of differential expressions at the CDS level interacting with differentially expressed genes, and the probability to obtain at least the observed number at random

	Observed	Expected	P-value
PPI HQ	23	7.3	$3.5 \times 10^{-7}$
PPI all	86	39.1	$2.15 \times 10^{-15}$
PrePPI	69	45.8	$2.4 \times 10^{-5}$

PPI HQ (high quality protein-protein interaction dataset); PPI all (all protein-protein interaction dataset); PrePPI (protein-protein interaction dataset from PrePPI).

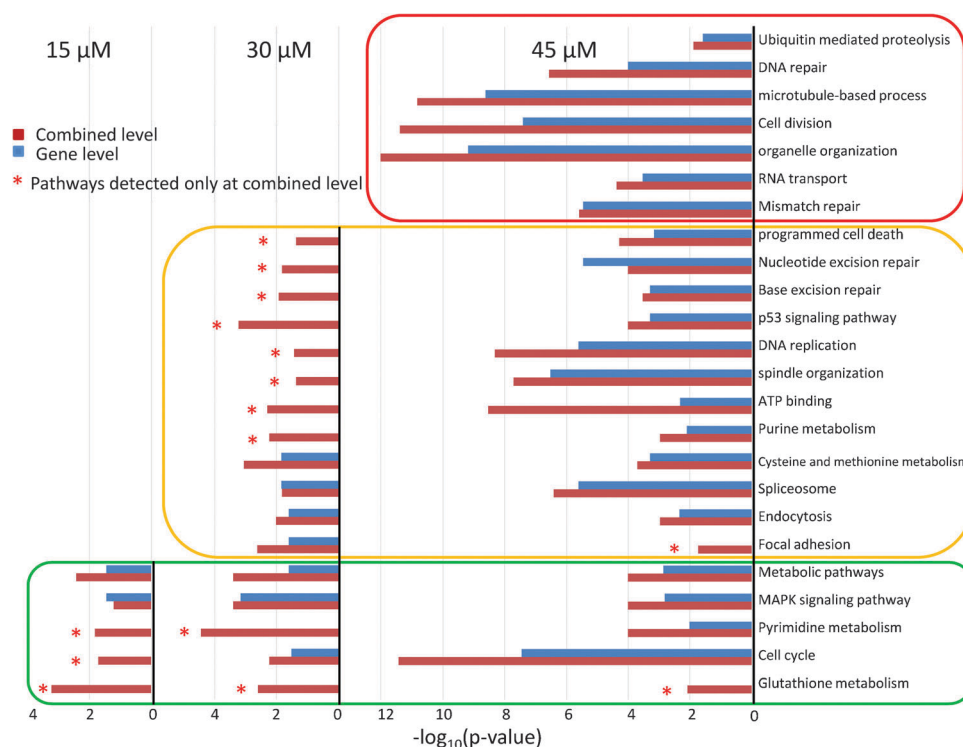


most pronounced changes in gene expression occurred in cells treated with the highest HNE concentrations.<sup>15</sup>

It should be noted that analysis at the CDS level detected a fraction of unique genes under each condition. Under 15  $\mu\text{M}$  HNE treatment, 15 genes were detected at both CDS and gene levels, whereas 20 genes were only captured at the CDS level including GCLM, RRM2, SLC1A5 and TXNRD1. GCLM and TXNRD1, showing a 1.8-fold and 2.4-fold increase at the CDS level respectively, have been reported to play a vital role in protecting cells from oxidative stress.<sup>28–30</sup> With 30  $\mu\text{M}$  HNE treatment, 40 genes were detected at both CDS and gene levels, whereas 51 genes were only captured at the CDS level, including RRM1, RRM2, CCND1, DKC1, BUB1B, POLE3 and GADD45A, which are involved in the cell cycle, DNA replication and glutathione metabolism. With 45  $\mu\text{M}$  HNE treatment, 147 genes were detected at both CDS and gene levels, whereas 195 genes were only captured at the CDS level, including many genes involved in the cell cycle (*e.g.*, EGFR, BUB1B, CCND1, and PPP5C), DNA replication (*e.g.*, HMGB1, MCM10, MCM5, MCM8, RRM1, and DKC1) and glutathione metabolism (*e.g.*, RRM1 and GCLM). Most of the unique genes detected at the CDS level closely related to the HNE response suggested that CDS level analysis is a useful complement to gene level analysis, which helps reveal important subtle biological changes.

Differentially expressed CDS and genes were further interpreted by functional enrichment analysis against Gene Ontology (GO) terms and KEGG pathways. Under 15  $\mu\text{M}$  HNE treatment, only the MAPK signaling pathway and the metabolic pathway were enriched in the differentially expressed genes.

Besides these two pathways, glutathione metabolism was observed at the combined level (combining differentially expressed CDS and genes, FDR = 0.0006) (Fig. 5 and Table S3, ESI†). Indeed, glutathione is a major intracellular antioxidant and glutathione synthesis is increased following HNE treatment to protect against oxidative injury.<sup>34,35</sup> Additionally, pyrimidine metabolism was also significantly represented at the combined level (FDR = 0.015) and pyrimidines have been reported to be a rich source for the synthesis of new antioxidant compounds.<sup>36–38</sup> With 30  $\mu\text{M}$  HNE treatment, additional pathways, such as focal adhesion, endocytosis, spliceosome and cysteine and methionine metabolism, were detected at both the gene level and the combined level. Interestingly, pathways associated with apoptosis and DNA repair were only revealed at the combined level, including programmed cell death (FDR = 0.044), nucleotide excision repair (FDR = 0.015), base excision repair (FDR = 0.012), p53 signaling pathways (FDR = 0.0006), DNA replication (FDR = 0.038), *etc.* (Fig. 5 and Table S4, ESI†). This is consistent with our previous studies, which reported that the  $\text{IC}_{50}$  value of HNE in RKO cells is 20  $\mu\text{M}$ <sup>39</sup> and a concentration equal to or greater than 30  $\mu\text{M}$  begins to induce apoptosis and cell cycle deregulation.<sup>12</sup> With 45  $\mu\text{M}$  HNE treatment, a number of additional pathways, such as ubiquitin mediated proteolysis, DNA repair, microtubule-based processes, and RNA transport, were affected, while most of these pathways showed a higher level of enrichment in combined level analysis compared to gene level analysis (Fig. 5 and Table S5, ESI†). For example, the FDR value for “DNA repair” is  $3.63 \times 10^{-7}$  at the combined level compared to  $1 \times 10^{-4}$  at the gene level.



**Fig. 5** Overrepresented pathways detected at the gene level and the combined level in HNE-treated RKO cells. Pathways observed only at the combined level are denoted by \*.



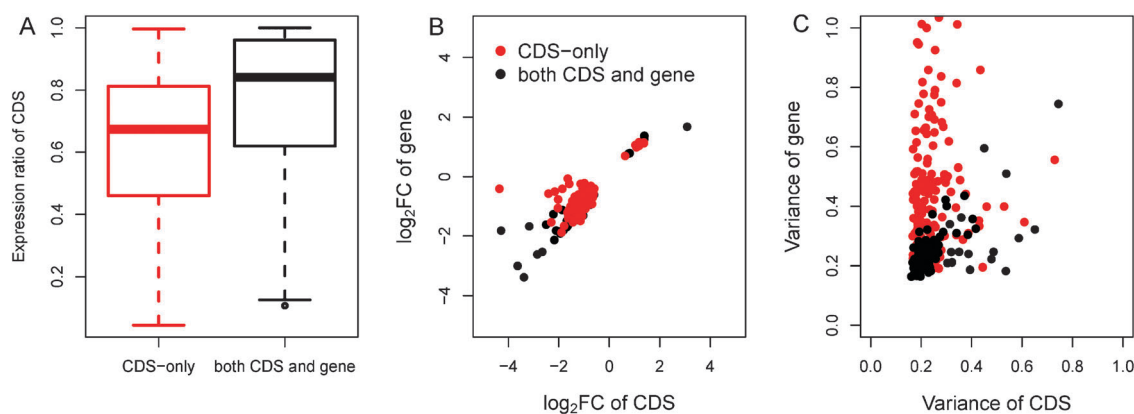
Taken together, the HNE cytotoxic effect was clearly shown by the dose-dependent pathway changes at the combined gene and CDS levels. At a low HNE concentration (15  $\mu\text{M}$ ), adaptive changes that protect cells against oxidative injury (*e.g.*, glutathione and pyrimidine metabolism) occurred. At the 30  $\mu\text{M}$  HNE concentration, repair of DNA damage was introduced along with an increase in the apoptotic response, which is consistent with  $\text{IC}_{50}$  experiments. Notably, cell protection against oxidative injury occurring at low dose and cell apoptosis initiated at intermediate dose were not identified by gene level analysis alone. At the 45  $\mu\text{M}$  concentration, HNE triggered many changes in signal transduction pathways that suppress cellular functions, which may lead to cell cycle arrest and apoptosis. Compared with gene level analysis combining gene and CDS levels helped reveal a gradual and continual involvement of biological pathways after low to high HNE dose treatment, which presents an informative and comprehensive picture of the dose-dependent cellular function changes.

## Discussion

RNA-seq provides the highest resolution of transcriptome information at the transcript level and the lowest resolution at the gene level. Our study is the first to estimate which level(s) are best suited to identify differential expression across conditions in terms of maximizing overlap with microarray data and providing biological relevance. At the gene level, differential expressions identified from RNA-seq and microarrays were quite consistent, with more genes identified by RNA-seq. At higher resolution, differential expression identified at the CDS level seemed to be a useful complement to gene-level analysis. Differential expression detected by the combined level (CDS and gene) achieved a higher overlap with microarray results and provided higher sensitivity in revealing biological insights into HNE dose-dependent responses than from gene-level analysis alone. The combined level analysis helped reveal gradually evolving response networks with increasing HNE

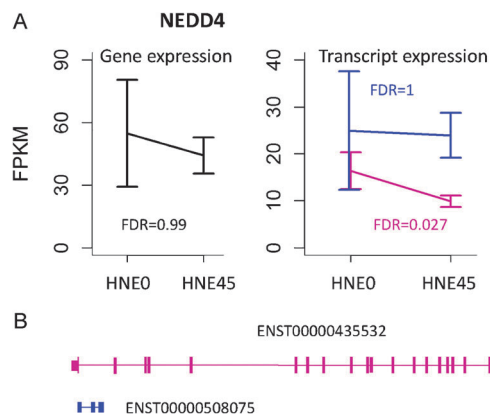
dose, from cell protection against oxidative stress (*e.g.*, glutathione metabolism) upon 15  $\mu\text{M}$  HNE treatment, initiation of apoptosis and the DNA damage response upon 30  $\mu\text{M}$  HNE treatment, to significant deregulation of cellular pathways upon 45  $\mu\text{M}$  HNE treatment.

Detection of differentially expressed CDS is technically more difficult than differentially expressed genes, due to greater uncertainty of read assignment and more stringent multiple test correction to account for a larger number of comparisons. There are two main possible explanations for differential expression detected at the CDS level but not at the gene level: transcriptional noise obscuring gene-level signal and isoform switching inducing differential splice variants without gene-level expression changes. To evaluate the relative contributions of these two factors to obscured gene expression changes, we compared the “CDS-only” group (differential expression detected only at the CDS level, 195 genes, Fig. 4) with the “both CDS and gene” group (differential expression detected at both CDS and gene levels, 147 genes, Fig. 4) after 45  $\mu\text{M}$  HNE treatment. These two groups have differentially expressed CDS but differ in gene-level expression changes. With the potential to change protein output, differentially expressed CDS is likely to function in HNE response and thus more informative. This informative CDS in the “CDS-only” group contributed less to the overall gene expression than those in the “both CDS and gene” group (Fig. 6A), suggesting higher background noise or splicing complexity in the “CDS-only” group. Furthermore, compared with genes in the “both CDS and gene” group, which exhibited similarity in both fold changes and expression variability (calculated by Cuffdiff,<sup>26</sup> including biological and technical variance) with their corresponding CDS, genes in the “CDS-only” group showed a similar fold change but with higher expression variances than their corresponding CDS (Fig. 6B and C). The high gene expression variability, resulting from transcriptional noise, obscures the gene level signal in the “CDS-only” group. Additionally, we only found one instance (SEPT6) out of 195 genes where isoform switching



**Fig. 6** (A) Cumulative distribution of the relative contributions of differentially expressed CDS to the genes in the “CDS-only” group and the “both CDS and gene” group. (B) Fold change of differentially expressed CDS vs. fold change of the corresponding genes in the “CDS-only” group and the “both CDS and gene” group upon 45  $\mu\text{M}$  HNE treatment. (C) Variances for differentially expressed CDS vs. variances of the corresponding genes in the “CDS-only” group and the “both CDS and gene” group under 45  $\mu\text{M}$  HNE treatment.





**Fig. 7** (A) Gene and transcript expression changes of NEDD4 in response to 45  $\mu$ M HNE treatment. (B) Transcript structure of differentially expressed CDS and non-differentially expressed transcripts. ENST00000435532 encodes differentially expressed CDS, while ENST00000508075 is a processed transcript.

led to differentially expressed CDS ( $\log_2$ FC =  $-1.58$ , FDR =  $0.004$ ) without detectable changes at the gene-level ( $\log_2$ FC =  $-0.24$ , FDR =  $1$ ) (Fig. S1–S6, ESI<sup>†</sup>). Thus transcriptional noise instead of isoform switching might be the main reason for the insignificant gene-level expression changes.

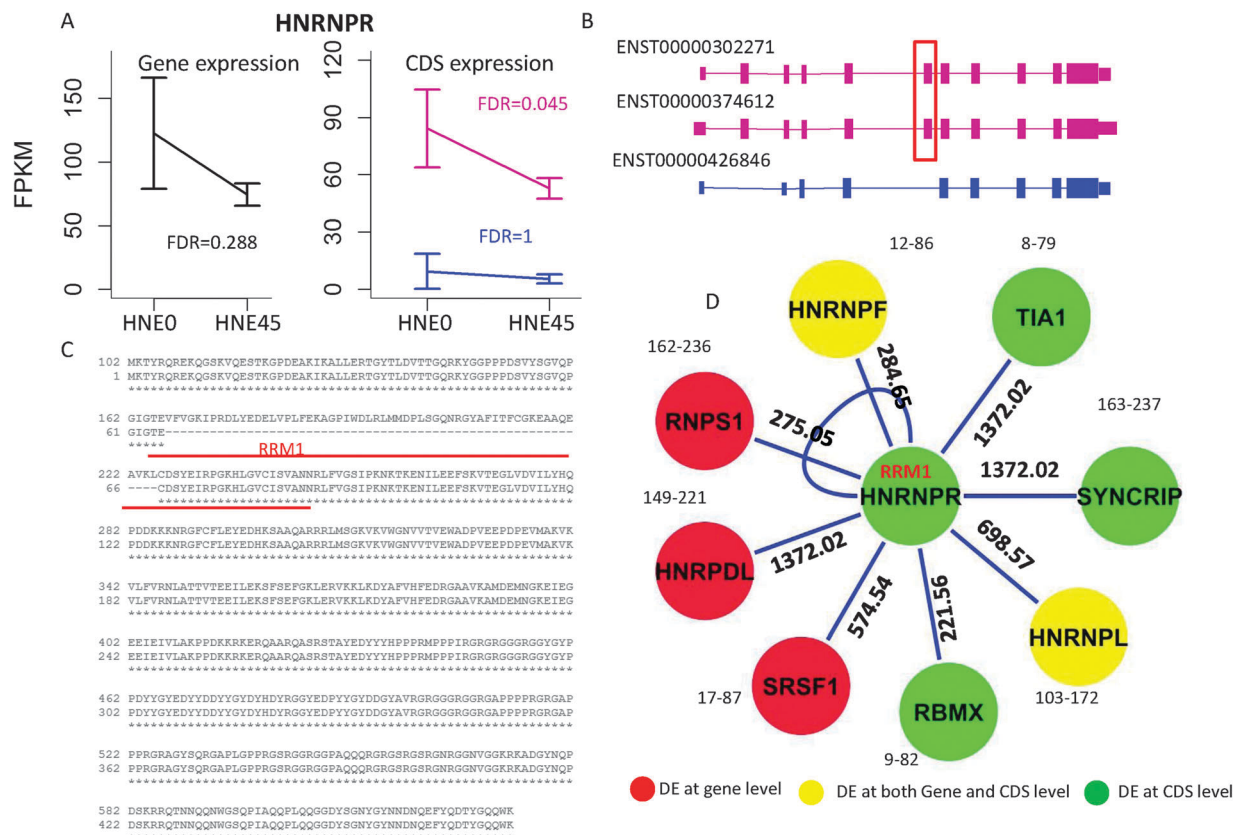
Transcriptional noise mainly stems from noncoding isoforms. Among 248 675 transcripts detected in the HNE transcriptome, 154 780 (62%) are noncoding isoforms. Noncoding isoforms, classified as retained intron or processed transcript, lack protein-coding capacity and do not contribute to protein output and thus may not be as functionally important as protein coding isoforms.<sup>40</sup> They are generally subject to less functional constraints on isoform abundance and have larger expression variances, which obscure the gene-level signal. For example, NEDD4 was identified to undergo significant expression changes at the CDS level (FDR <  $0.05$ ), but not at the gene level (FDR =  $0.99$ ) with 45  $\mu$ M HNE treatment (Fig. 7A). NEDD4 had two highly expressed transcripts, ENST00000435532 and ENST00000508075 (Fig. 7B). ENST00000435532 codes for a protein product and its expression was significantly changed (FDR =  $0.027$ ), whereas ENST00000508075 is a noncoding transcript whose expression varies a lot under two conditions and was not changed after 45  $\mu$ M HNE treatment (FDR =  $1$ ). Another possible source of transcriptional noise is from those coding isoforms lacking strong transcriptional control. Their expressions, to a large extent reflecting background transcription, make the gene-level signal hard to detect. For example, HNRNPR underwent significant expression changes at the CDS level (FDR <  $0.05$ ), but not at the gene level (FDR =  $0.29$ ) with 45  $\mu$ M HNE treatment (Fig. 8A). Besides the differentially expressed CDS (containing two isoforms, ENST00000302271 and ENST00000374612), HNRNPR had another CDS without significant expression change (ENST00000426846, FDR =  $1$ ), which obscured the gene-level signal (Fig. 8A). Comparing the transcript structure of these two CDS, we found that the significantly changed CDS has one more exon than the non-significant CDS (Fig. 8B). This exon encodes RNA recognition motif domain 1 (RRM1) (Fig. 8C), which is predicted to interact with many differentially expressed genes or

CDS by PrePPI, including HNRPDL, SRSF1, RNPS1, HNRNPF, HNRNPL, *etc.* (Fig. 8D). The CDS containing this important exon might be subjected to strong constraints on its expression, showing a higher transcriptional signal-to-noise ratio. In contrast, the non-informative CDS lacking the exon, subject to less functional constraints on isoform abundance, might undergo noisy splicing by erroneous splice site choice<sup>24</sup> and results in lower signal-to-noise ratio. This agrees with a previous study demonstrating that noise in gene expression is a biologically important variable and subject to natural selection.<sup>24</sup>

Additionally, differential expression observed at the CDS level but not at the gene level may present an opportunity for exploring potential post-transcriptional regulatory mechanisms to gain insights into isoform specific regulation. For example, the small expression variation of the functional transcripts within biological replicates suggests that their expression might be controlled by the coupling of transcription and splicing since RNA binding proteins usually have a low degree of transcriptional noise.<sup>41</sup> As another example, post-transcriptional regulation might be involved if only functional transcripts changed their abundance across conditions (*e.g.*, miRNA targeting of specific isoforms to induce mRNA decay). Analyzing the 3' UTR of genes with differentially expressed CDS is one way to find the miRNA involved in the process. For example, NEDD4 was found to be the target of several miRNAs from MSigDB (c3.mir.v3.1.symbols.gmt),<sup>42</sup> including miR-30, miR-27, miR-9 and miR-144. The binding sites are evolutionarily conserved and the miRNA–target relationships are also supported by other prediction algorithms (Table S6, ESI<sup>†</sup>). Consistently, miR-144 targets were highly enriched in differentially expressed gene sets, not only in those detected at the CDS level but also at the combined level (CDS and gene levels) (FDR <  $1 \times 10^{-6}$ , Table 2). Previous studies have found that the RKO cell line exhibits low expression levels of miR-144 and down regulation of miR-144 leads to colorectal cancer progression *via* activation of the mTOR signaling pathway.<sup>43</sup> Thus, miR-144 might be upregulated by HNE treatment, which leads to the down regulation of transcripts or genes and the inhibition of cell proliferation.

Differential CDS analysis can identify significant CDS abundance changes no matter gene expression changes or not, but this method is quite different from methods aimed at detecting differential spliced genes or differential exon usage, *e.g.*, MISO,<sup>44</sup> ALEXA-Seq,<sup>45</sup> DEXseq<sup>46</sup> and DSGseq.<sup>47</sup> The major difference is that if the gene's overall expression changes but the relative abundances of the different transcripts stay the same, the genes will be called significant by differential CDS analysis but not by methods focusing on differential splicing. Among 35 significantly changed genes detected by microarray analysis upon 45  $\mu$ M HNE treatment, 30 were identified by differential gene and CDS analysis from RNA-seq, but none of them were identified by DEXSeq. In addition, differential CDS analysis has several advantages. CDS is an important function unit, thus differential CDS analysis is biologically more meaningful and easier to interpret than differential exon usage. Furthermore, although exons are more sensitive and easier to calculate than CDS, the results based on the exon level are more prone to noise and will





**Fig. 8** (A) Gene and CDS expression changes of HNRNPR in response to 45  $\mu$ M HNE treatment. (B) Transcript structure of differentially expressed CDS and non-differentially expressed CDS. Two transcripts, ENST00000302271 and ENST00000374612 encode differentially expressed CDS, while ENST00000426846 encodes non-differentially expressed CDS, which lacks an exon situated in the RRM1 domain. (C) Comparison of protein sequences between differentially expressed and non-differentially expressed CDS. (D) The RRM1 domain of HNRNPR is predicted to interact with many genes by PrePI, whose expressions are significantly changed. The number on the edge denotes the likelihood ratio score based on the three-dimensional structural interaction. A score greater than 50 suggests the high probability of interaction between two proteins. The number beside the node shows the domain information. For example, the RRM1 domain of HNRNPR is predicted to interact with the domain 17–87 of SRSF1 with the score of 574.54.

**Table 2** miRNA targets enrichment analysis on differential expression at the CDS level and the combined level (CDS or gene)

miRNA	DEGs at CDS-level		DEGs at combined level	
	Number of targets	FDR	Number of targets	FDR
miR-144	13	$1.25 \times 10^{-7}$	15	$1.26 \times 10^{-7}$
miR-524	14	$6.07 \times 10^{-5}$	19	$2.20 \times 10^{-6}$
miR-518a-2	10	$6.07 \times 10^{-5}$	13	$2.23 \times 10^{-6}$
miR-101	10	$2 \times 10^{-4}$	14	$3.89 \times 10^{-6}$
miR-519a,b,c	13	$2 \times 10^{-4}$	16	$6.01 \times 10^{-5}$
miR-522	8	$2 \times 10^{-4}$	11	$6.82 \times 10^{-6}$
miR-204, miR-211	9	$3 \times 10^{-4}$	10	$3 \times 10^{-4}$
miR-181a,b,c,d	13	$4 \times 10^{-4}$	18	$1.55 \times 10^{-5}$
miR-324-3p	6	$5 \times 10^{-4}$	6	0.001
miR-30a-5p,30c,30d, 30b,30e-5p	14	$5 \times 10^{-4}$	24	$1.26 \times 10^{-7}$

be less robust and less stable. A large portion of differentially expressed/spliced genes at low dose is expected to be also significantly changed at high dose since HNE response networks will gradually evolve with the increasing dose. This expectation is better supported by differential CDS analysis than alternative splicing methods. 77% (23) of 30 significant CDS at 15  $\mu$ M were

found to be still significantly changed at 30  $\mu$ M, and 86% (78) of 91 significant CDS at 30  $\mu$ M were supported by 45  $\mu$ M. In contrast, only one (50%) of two exons detected by DEXSeq at 15  $\mu$ M found evidence of differential usage at 30  $\mu$ M, and 78% (18) of 23 exons detected at 30  $\mu$ M were re-identified at 45  $\mu$ M HNE. Even worse, MISO identified 29 exon skipping and 11 exon inclusion events at 15  $\mu$ M, but only one exon skipping and 2 inclusion events (8%) reappeared at 30  $\mu$ M. Among 23 exon skipping and 10 exon inclusion events detected at 30  $\mu$ M, only 4 skipping and 2 inclusion events (18%) were re-identified at 45  $\mu$ M (Fig. S7, ESI†).

Although RNA-seq offers high resolution transcriptome information, read assignment uncertainty remains a major challenge, especially for low abundance genes with many isoforms. Differential expression at the transcript level is the most difficult to detect, due to the largest read assignment uncertainty and the highest statistical significance required to account for the largest number of comparisons. Additionally, noisy splicing leads to false positives, especially when the number of replicates is small. Therefore, transcript level analysis did not help find more biologically relevant results in our experiments with only 3 replicates in each condition. Standard RNA-seq methods are not suited to



annotate the 5' start site, which may explain why differential expression detection at the TSS level was not as useful as expected. In contrast, each CDS group encompasses all transcripts coding for the same protein product, which reduces the read assignment ambiguity and the noise due to erroneous splice site choice. Thus, differential expression analysis at the CDS level is a useful complement to gene-level analysis. Combining CDS and gene levels revealed more subtle biological responses triggered by HNE treatment. In the future, adding more replicates, increasing sequencing depth, and using long pair-end reads will facilitate differential expression detection at the transcript level, which will create opportunities for regulation analysis with unprecedented scope and scale and allow researchers to better disentangle the complex interplay between transcriptional and post-transcriptional regulation.

## Materials and methods

### Cell culture and treatment

RKO human colorectal carcinoma cells were grown in McCoy's 5A medium supplemented with 10% fetal bovine serum, 2 mM L-glutamine, and antibiotics at 37 °C and 5% CO<sub>2</sub>. HNE was obtained from Cayman Chemical and was dissolved in MeOH as a 1000× stock solution. RKO cells were seeded and were treated with a vehicle or 15, 30, or 45 μM HNE for 6 h. Cell treatments were conducted three times for each condition. Experimental details have been described previously.<sup>15</sup>

### RNA sequencing

The twelve RNA samples were sequenced following the protocols recommended by the manufacturer (Illumina). Briefly, poly-A was purified and then fragmented into small pieces. Using reverse transcriptase and random primers, RNA fragments were used to synthesize the first and second strand cDNAs. Following end repair, addition of an "A" base, adapter ligation, size selection and amplification of cDNA templates, samples were sequenced in 5 lanes on the Illumina HiSeq 2000, generating about 70–110 million of 100 pair-end reads per sample (Table S1, ESI†).

### RNA-seq and microarray analyses

Reads were mapped to the human genome hg19 using TopHat version 1.4.0 with the reference annotation file (Homo\_sapiens.GRCh37.65.gtf).<sup>26,27,48</sup> Each sample obtained similar mapping quality, about 81% of the reads mapped to the genome, of which 87% were overlapping exons. The mapping results were summarized in Table S1 (ESI†). The aligned reads were assembled and transcript expression was quantified using FPKM (Fragments Per Kilobase of transcript per Million fragments mapped) by Cufflinks version 2.0.2, which uses a linear statistical model to compute the likelihood that the number of fragments would be observed given the proposed abundances on the transcripts.<sup>26</sup> Differential expression between four groups, HNE15 vs. HNE0, HNE30 vs. HNE0, and HNE45 vs. HNE0 was detected by Cuffdiff.<sup>26,27,49</sup> Genes, CDS, TSS or

transcripts with FPKM > 1 in any of four conditions were selected for further analysis.

Affymetrix cel files were normalized using the Robust MultiChip Analysis (RMA) algorithm<sup>50</sup> as implemented in Bioconductor.<sup>51</sup> Probe set identifiers (IDs) were mapped to gene symbols. Probe sets that mapped to multiple genes were eliminated. When multiple probe sets were mapped to the same gene, the probe set with the maximal IQR was used to represent the gene expression level. Differential expression analysis between HNE45 and HNE0 was performed using limma.<sup>52</sup>

A common set of genes shared by RNA-seq and microarray was used to compare gene expression between these two platforms. If genes had FDR < 0.01 at both gene-level and other levels (CDS, TSS or transcript), fold change values at the gene level were used. A fold change ranking with an FDR cutoff of 0.01 was applied separately to RNA-seq and microarray to calculate the percentage of overlapping genes (POG) using the equation  $POG = 100 \times (DD + UU)/2L$ , where DD and UU are the number of down- or up-regulated genes common in RNA-seq and microarray, respectively, and *L* is the number of selected genes ranked by fold change. Directionality of gene regulation is considered in POG calculations, that is, genes selected by two platforms but with different regulation directionalities are considered as discordant.<sup>53</sup>

### Functional interpretation

Three protein–protein interaction datasets, PPI HQ, PPI all and PrePPI, were downloaded from the PrePPI webserver (<http://bhapp.c2b2.columbia.edu/PrePPI/>).<sup>32,33</sup> PPI HQ contains 7409 interactions of at least two publication supports, involving 2976 proteins. PPI all includes 82 551 interactions between 12 104 proteins from HPRD, DIP, IntAct, BioGRID, and MINT. PrePPI comprises 317 813 high confidence interactions (LR > 600) for 11 219 proteins. For 492 genes whose differential expression was detected at the gene or the CDS level after 45 μM HNE treatment (Fig. 4), 154 were contained in PPI HQ, 405 were included in PPI all, and 217 were involved in PrePPI. The hypergeometric test was used to calculate the probability of differentially expressed CDS randomly connected to differentially expressed genes in the protein–protein network.

GO, KEGG and miRNA target enrichment analyses were performed using WebGestalt.<sup>54</sup> Functional categories or pathways containing no less than two differentially expressed CDS or genes with FDR < 0.05 were selected. Potential miRNAs targeting NEDD4 were obtained from MSigDB (c3.mir.v3.1.symbols.gmt),<sup>42</sup> which were further validated by evolutionary conservation and other miRNA target prediction algorithms, including TargetScan, DIANAmt, miRanda, miRDB, miRWalk, RNAhybrid, PICTAR4, PICTAR5, PITA, and RNA22.

## Acknowledgements

This work was supported by the National Institutes of Health grants P01 ES013125, P30 ES000267, U54CA126479 and R01GM088822. QL was partially supported by the State Key Program of National



Natural Science of China (31230058) and National Natural Science Foundation of China (31070746).

## References

- B. Hennig and C. K. Chow, *Free Radicals Biol. Med.*, 1988, **4**, 99–106.
- G. Jurgens, Q. Chen, H. Esterbauer, S. Mair, G. Ledinski and H. P. Dinges, *Arterioscler. Thromb.*, 1993, **13**, 1689–1699.
- A. Yoritaka, N. Hattori, K. Uchida, M. Tanaka, E. R. Stadtman and Y. Mizuno, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**, 2696–2701.
- N. Traverso, S. Menini, L. Cosso, P. Odetti, E. Albano, M. A. Pronzato and U. M. Marinari, *Diabetologia*, 1998, **41**, 265–270.
- X. Dou, S. Li, Z. Wang, D. Gu, C. Shen, T. Yao and Z. Song, *Am. J. Pathol.*, 2012, **181**, 1702–1710.
- W. C. Lee, H. Y. Wong, Y. Y. Chai, C. W. Shi, N. Amino, S. Kikuchi and S. H. Huang, *Biochem. Biophys. Res. Commun.*, 2012, **425**, 842–847.
- I. Nakashima, W. Liu, A. A. Akhand, K. Takeda, Y. Kawamoto, M. Kato and H. Suzuki, *Mol. Aspects Med.*, 2003, **24**, 231–238.
- W. Liu, M. Kato, A. A. Akhand, A. Hayakawa, H. Suzuki, T. Miyata, K. Kurokawa, Y. Hotta, N. Ishikawa and I. Nakashima, *J. Cell Sci.*, 2000, **113**(Pt 4), 635–641.
- D. Biswas, G. Sen and T. Biswas, *Toxicol. Appl. Pharmacol.*, 2010, **244**, 315–327.
- S. O. Abarikwu, A. B. Pant and E. O. Farombi, *Basic Clin. Pharmacol. Toxicol.*, 2012, **110**, 441–448.
- A. Sharma, R. Sharma, P. Chaudhary, R. Vatsyayan, V. Pearce, P. V. Jeyabal, P. Zimniak, S. Awasthi and Y. C. Awasthi, *Arch. Biochem. Biophys.*, 2008, **480**, 85–94.
- C. Ji, V. Amarnath, J. A. Pietenpol and L. J. Marnett, *Chem. Res. Toxicol.*, 2001, **14**, 1090–1096.
- J. Ruef, M. Moser, C. Bode, W. Kubler and M. S. Runge, *Basic Res. Cardiol.*, 2001, **96**, 143–150.
- U. Herbst, M. Toborek, S. Kaiser, M. P. Mattson and B. Hennig, *J. Cell. Physiol.*, 1999, **181**, 295–303.
- J. D. West and L. J. Marnett, *Chem. Res. Toxicol.*, 2005, **18**, 1642–1653.
- A. T. Jacobs and L. J. Marnett, *Acc. Chem. Res.*, 2010, **43**, 673–683.
- B. Zhang, Z. Shi, D. T. Duncan, N. Prodduturi, L. J. Marnett and D. C. Liebler, *Mol. Biosyst.*, 2011, **7**, 2118–2127.
- Z. Wang, M. Gerstein and M. Snyder, *Nat. Rev. Genet.*, 2009, **10**, 57–63.
- I. Nookaew, M. Papini, N. Pornputtapong, G. Scalcinati, L. Fagerberg, M. Uhlen and J. Nielsen, *Nucleic Acids Res.*, 2012, **40**, 10084–10097.
- Y. Xiong, X. Chen, Z. Chen, X. Wang, S. Shi, J. Zhang and X. He, *Nat. Genet.*, 2010, **42**, 1043–1047.
- Z. Su, Z. Li, T. Chen, Q. Z. Li, H. Fang, D. Ding, W. Ge, B. Ning, H. Hong, R. G. Perkins, W. Tong and L. Shi, *Chem. Res. Toxicol.*, 2011, **24**, 1486–1493.
- J. van Delft, S. Gaj, M. Lienhard, M. W. Albrecht, A. Kirpiy, K. Brauers, S. Claessen, D. Lizarraga, H. Lehrach, R. Herwig and J. Kleinjans, *Toxicol. Sci.*, 2012, **130**, 427–439.
- M. Garber, M. G. Grabherr, M. Guttman and C. Trapnell, *Nat. Methods*, 2011, **8**, 469–477.
- J. K. Pickrell, A. A. Pai, Y. Gilad and J. K. Pritchard, *PLoS Genet.*, 2010, **6**, e1001236.
- A. Oshlack, M. D. Robinson and M. D. Young, *Genome Biol.*, 2010, **11**, 220.
- C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn and L. Pachter, *Nat. Biotechnol.*, 2012, **31**, 46–53.
- C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter, *Nat. Protocols*, 2012, **7**, 562–578.
- T. B. Cole, G. Giordano, A. L. Co, I. Mohar, T. J. Kavanagh and L. G. Costa, *J. Toxicol.*, 2011, **2011**, 157687.
- S. Karimpour, J. Lou, L. L. Lin, L. M. Rene, L. Lagunas, X. Ma, S. Karra, C. M. Bradbury, S. Markovina, P. C. Goswami, D. R. Spitz, K. Hirota, D. V. Kalvakolanu, J. Yodoi and D. Gius, *Oncogene*, 2002, **21**, 6317–6327.
- D. K. Smart, K. L. Ortiz, D. Mattson, C. M. Bradbury, K. S. Bisht, L. K. Sieck, M. W. Brechbiel and D. Gius, *Cancer Res.*, 2004, **64**, 6716–6724.
- J. FitzGerald, S. Moureau, P. Drogaris, E. O'Connell, N. Abshiru, A. Verreault, P. Thibault, M. Grenon and N. F. Lowndes, *PLoS One*, 2011, **6**, e14714.
- Q. C. Zhang, D. Petrey, J. I. Garzon, L. Deng and B. Honig, *Nucleic Acids Res.*, 2013, **41**, D828–D833.
- Q. C. Zhang, D. Petrey, L. Deng, L. Qiang, Y. Shi, C. A. Thu, B. Bisikirska, C. Lefebvre, D. Accili, T. Hunter, T. Maniatis, A. Califano and B. Honig, *Nature*, 2012, **490**, 556–560.
- J. B. Schulz, J. Lindenau, J. Seyfried and J. Dichgans, *Eur. J. Biochem.*, 2000, **267**, 4904–4911.
- I. Kruman, A. J. Bruce-Keller, D. Bredesen, G. Waeg and M. P. Mattson, *J. Neurosci.*, 1997, **17**, 5089–5100.
- Y. Kotaiah, N. Harikrishna, K. Nagaraju and C. Venkata Rao, *Eur. J. Med. Chem.*, 2012, **58**, 340–345.
- S. S. Panda and P. V. Chowdary, *Indian J. Pharm. Sci.*, 2008, **70**, 208–215.
- T. Bano, N. Kumar and R. Dudhe, *Org. Med. Chem. Lett.*, 2012, **2**, 34.
- C. E. McGrath, K. A. Tallman, N. A. Porter and L. J. Marnett, *Chem. Res. Toxicol.*, 2011, **24**, 357–370.
- M. Gonzalez-Porta, A. Frankish, J. Rung, J. Harrow and A. Brazma, *Genome Biol.*, 2013, **14**, R70.
- A. R. Kornblihtt, *Adv. Exp. Med. Biol.*, 2007, **623**, 175–189.
- A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdottir, P. Tamayo and J. P. Mesirov, *Bioinformatics*, 2011, **27**, 1739–1740.
- T. Iwaya, T. Yokobori, N. Nishida, R. Kogo, T. Sudo, F. Tanaka, K. Shibata, G. Sawada, Y. Takahashi, M. Ishibashi, G. Wakabayashi, M. Mori and K. Mimori, *Carcinogenesis*, 2012, **33**, 2391–2397.
- Y. Katz, E. T. Wang, E. M. Airolidi and C. B. Burge, *Nat. Methods*, 2010, **7**, 1009–1015.
- M. Griffith, O. L. Griffith, J. Mwenifumbo, R. Goya, A. S. Morrissey, R. D. Morin, R. Corbett, M. J. Tang, Y. C. Hou, T. J. Pugh, G. Robertson, S. Chittaranjan, A. Ally, J. K. Asano, S. Y. Chan, H. I. Li, H. McDonald, K. Teague, Y. Zhao, T. Zeng, A. Delaney, M. Hirst,



- G. B. Morin, S. J. Jones, I. T. Tai and M. A. Marra, *Nat. Methods*, 2010, **7**, 843–847.
- 46 S. Anders, A. Reyes and W. Huber, *Genome Res.*, 2012, **22**, 2008–2017.
- 47 W. Wang, Z. Qin, Z. Feng, X. Wang and X. Zhang, *Gene*, 2013, **518**, 164–170.
- 48 C. Trapnell, L. Pachter and S. L. Salzberg, *Bioinformatics*, 2009, **25**, 1105–1111.
- 49 C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter, *Nat. Biotechnol.*, 2010, **28**, 511–515.
- 50 R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf and T. P. Speed, *Biostatistics*, 2003, **4**, 249–264.
- 51 M. Reimers and V. J. Carey, *Methods Enzymol.*, 2006, **411**, 119–134.
- 52 G. K. Smyth, *Stat. Appl. Genet. Mol. Biol.*, 2004, **3**, 3.
- 53 L. Shi, W. D. Jones, R. V. Jensen, S. C. Harris, R. G. Perkins, F. M. Goodsaid, L. Guo, L. J. Croner, C. Boysen, H. Fang, F. Qian, S. Amur, W. Bao, C. C. Barbacioru, V. Bertholet, X. M. Cao, T. M. Chu, P. J. Collins, X. H. Fan, F. W. Frueh, J. C. Fuscoe, X. Guo, J. Han, D. Herman, H. Hong, E. S. Kawasaki, Q. Z. Li, Y. Luo, Y. Ma, N. Mei, R. L. Peterson, R. K. Puri, R. Shippy, Z. Su, Y. A. Sun, H. Sun, B. Thorn, Y. Turpaz, C. Wang, S. J. Wang, J. A. Warrington, J. C. Willey, J. Wu, Q. Xie, L. Zhang, S. Zhong, R. D. Wolfinger and W. Tong, *BMC Bioinf.*, 2008, **9**(Suppl 9), S10.
- 54 B. Zhang, S. Kirov and J. Snoddy, *Nucleic Acids Res.*, 2005, **33**, W741–W748.

