

# SCIENTIFIC REPORTS



OPEN

## A Multi-Parameter Analysis of Cellular Coordination of Major Transcriptome Regulation Mechanisms

Wen Jiang<sup>1,2</sup>, Zhanyong Guo<sup>4</sup>, Nuno Lages<sup>5</sup>, W. Jim Zheng<sup>6</sup>, Denis Feliers<sup>7</sup>, Fangyuan Zhang<sup>3</sup> & Degeng Wang<sup>1,2</sup>

To understand cellular coordination of multiple transcriptome regulation mechanisms, we simultaneously measured transcription rate (TR), mRNA abundance (RA) and translation activity (TA). This revealed multiple insights. First, the three parameters displayed systematic statistical differences. Sequentially more genes exhibited extreme (low or high) expression values from TR to RA, and then to TA; that is, cellular coordination of multiple transcriptome regulatory mechanisms leads to sequentially enhanced gene expression selectivity as the genetic information flow from the genome to the proteome. Second, contribution of the stabilization-by-translation regulatory mechanism to the cellular coordination process was assessed. The data enabled an estimation of mRNA stability, revealing a moderate but significant positive correlation between mRNA stability and translation activity. Third, the proportion of mRNA occupied by un-translated regions (UTR) exhibited a negative relationship with the level of this correlation, and was thus a major determinant of the mode of regulation of the mRNA. High-UTR-proportion mRNAs tend to defy the stabilization-by-translation regulatory mechanism, staying out of the polysome but remaining stable; mRNAs with little UTRs largely followed this regulation. In summary, we quantitatively delineated the relationship among multiple transcriptome regulation parameters, *i.e.*, cellular coordination of corresponding regulatory mechanisms.

Genomic sequences are readily available for a large and ever-increasing number of species. These sequences, like English literature, represent static strings of symbols/alphabets (A, T, C, and G). Hence, genomic sequences are often analogized to the “book” of life. As an extension of the analogy, the cell can be considered as the “reader” of the genomic “book”, and the multi-stepped gene expression process as the “reading” process<sup>1–4</sup>. Through the gene expression process, the seemingly simplistic genomic alphabetical strings are selectively and dynamically transcribed into transcriptome sequences, which are in turn translated into amino acid sequences in the proteome – the main machinery that controls biochemical reactions and processes in support of cellular functions. This process is integral to essentially all cellular activity and entails multiple regulatory mechanisms. A complex picture of the relationship among these regulatory mechanisms has recently emerged<sup>5–11</sup>.

In some omics experiments, mRNA and protein abundance are measured simultaneously. One lesson we learned is that correlation between the two is not always satisfactory enough for mRNA abundance to be a reliable predictor of protein abundance. This discrepancy had been observed prior to the genomic era<sup>12,13</sup>. It was confirmed in the yeast *S. cerevisiae* by one of the first simultaneous transcriptome and proteomic measurements<sup>14</sup>, and then observed in many other high-throughput studies<sup>5,9,10,15–21</sup>.

<sup>1</sup>Department of Environmental Toxicology, Texas Tech University, Lubbock, TX, 79409, USA. <sup>2</sup>The Institute of Environmental and Human Health (TIEHH), Texas Tech University, Lubbock, TX, 79409, USA. <sup>3</sup>Department of Mathematics & Statistics, Texas Tech University, Lubbock, TX, 79409, USA. <sup>4</sup>National Key Laboratory of Wheat and Maize Crop Science, College of Agronomy, Henan Agricultural University, Zhengzhou, Henan, 450002, P. R. China. <sup>5</sup>Instituto Superior Técnico, Universidade de Lisboa, IST-Taguspark, 2744-016, Porto Salvo, Portugal. <sup>6</sup>School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin Street, Suite 600, Houston, TX, 77030, USA. <sup>7</sup>Department of Medicine/Renal Diseases, The University of Texas Health Science Center at San Antonio, San Antonio, TX, USA. Correspondence and requests for materials should be addressed to D.W. (email: [degeng.wang@ttu.edu](mailto:degeng.wang@ttu.edu))

Transcriptome analysis techniques have also been coupled to conventional experimental protocols to measure other gene expression parameters. Initially microarray<sup>22–25</sup>, and then next generation sequencing (NGS)<sup>26,27</sup>, were coupled to the nuclear run-on technique for genome-wide transcription rate measurement. Additionally, NGS was coupled to metabolic labeling of nascent transcripts to measure transcription rates<sup>10,28–31</sup>. It was also coupled to RNA polymerase II chromatin immunoprecipitation (ChIP) for the same purpose. These strategies enabled simultaneous transcription rate and mRNA abundance measurement. Once again, some levels of discrepancy were observed in that mRNA abundance was not always a good predictor of transcription rate.

These observed discrepancies among gene expression parameters were a reflection of the complexity of the gene expression process<sup>32</sup>, and should be informative for us to unravel the complexity. At the same time when nascent RNA and protein are produced, existing RNA and protein are being selectively degraded. The abundance of protein and mRNA represent the balance of the respective production and degradation. Discrepancy among gene expression parameters is considered evidence for some levels of decoupling among transcription, translation, mRNA degradation and protein degradation; that is, the gene expression parameters can be divergently regulated. Given the technical feasibility, multi-parameter approaches are being used to study the discrepancy and glean out fundamental gene expression regulation principles. Such studies will potentially lead to more efficient gene expression analysis strategies that generate more informative data.

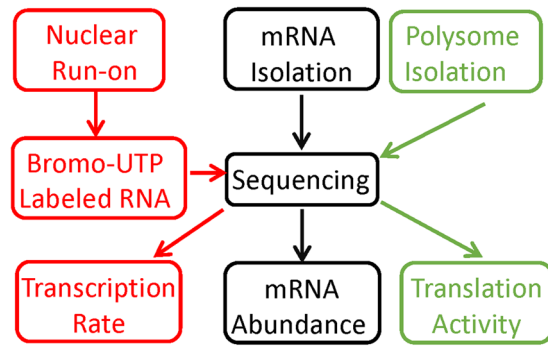
Such multi-parameter approaches should be especially fruitful for transcriptome analysis. The polysome profiling analysis utilizes NGS to quantify polysome-associated mRNAs, i.e., actively translating mRNAs, thus enabling genome-wide analysis of translation activity. Similarly, the ribosome profiling analysis utilize NGS to quantify mRNA fragments protected from RNase digestion by the ribosome<sup>33,34</sup>. Thus, all techniques are in place for genome-wide integration of transcription rate (i.e., GRO-seq), mRNA abundance (RNA-seq) and mRNA translation activity (i.e., polysome profiling). This will generate an integrative view of the transcriptome and its dynamic regulation, i.e., how the multiple transcriptome regulatory mechanisms are coordinated.

Additionally, such data is needed as a platform to study mediation of post-transcriptional regulation by mRNA untranslated regions (UTR), where regulatory signals for post-transcriptional regulation are embedded. It is well documented that mRNA UTRs are responsible for mRNA stability and translation control. They contain binding sites for microRNA and many regulatory RNA-binding proteins. They are common in mammalian mRNAs; e.g., human mRNAs, on average, have ~1000 nucleotide long UTRs (~800 nucleotide 3'- and ~200 nucleotide 5'-UTRs). Systematic functional study of the UTRs, however, awaits multi-parameter datasets that enables simultaneous study of mRNA stability and translation activity.

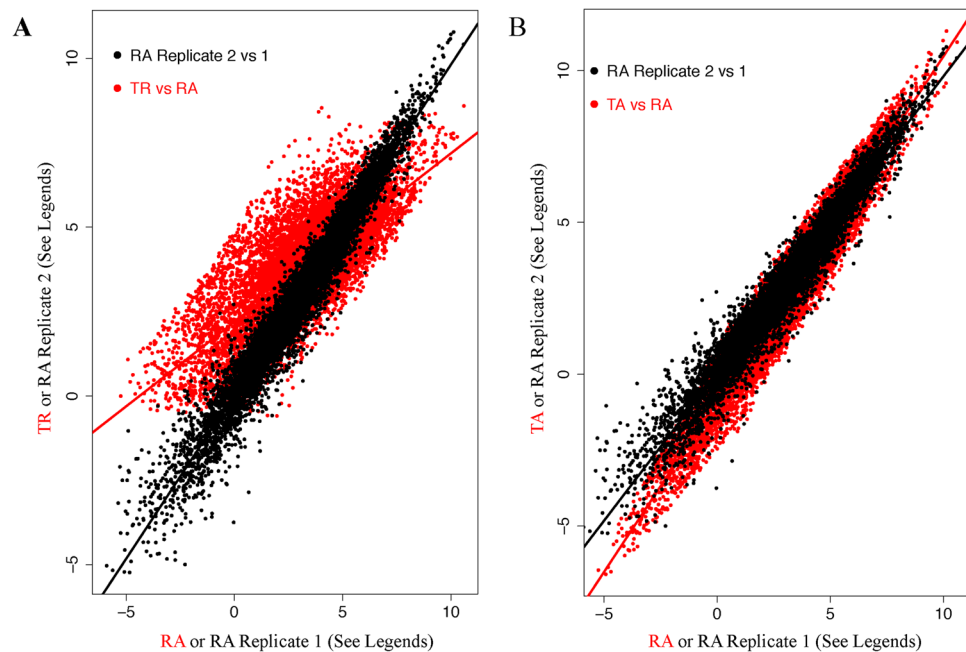
Thus, we generated a multi-parameter snapshot of the transcriptome by simultaneous genome-wide measurement of transcription rate (TR), mRNA abundance (RA) and translation activity (TA); we also estimated mRNA stability/degradation by the RA to TR ratio. The data enabled a quantitative delineation of cellular coordination of these regulatory mechanisms. Briefly, the data revealed functional consequences of the cellular coordination activity. We assessed, for the first time, the contribution of the mRNA-stabilization-by-translation regulatory mechanism to transcriptome regulation, as it is known that actively translating mRNA is protected from degradation<sup>35,36</sup>. Analysis of the data in conjunction with mRNA UTRs revealed further insights into, and the roles of UTRs in, cellular coordination of these transcriptome regulatory mechanisms.

## Results

**Simultaneous measurement of TR, RA and TA.** Previously, we analyzed publicly available genomic datasets, in which two gene expression parameters were simultaneously measured<sup>37,38</sup>. In those studies, we attempted to explain the discrepancy among gene expression parameters, which seemed mysterious then to most scientists, from the perspectives of biochemical pathway/network control and cellular operations. Though the genome-wide measurement techniques have since greatly advanced and many datasets have recently been published<sup>39</sup>, we have not seen a dataset that integrate TR, RA, TA and mRNA stability. The translational data in such studies are mass-spectrometry-based, and thus the coverage is not nearly genome-wide. In the present work, we took advantage of the genome-wide analysis power of NGS and its versatility through successful coupling to a variety of conventional experimental protocols. Our goal was to obtain a genome-wide multi-parameter snapshot of the transcriptome through simultaneous measurements of TR, RA and TA in the HCT116 human colon cancer cells. We chose HCT116 due to its popularity and its near diploid. Lots of data are available and many useful genetic modifications have been done. This makes our dataset more valuable to the research community, and also makes it feasible to facilitate our future studies with the HCT116 genetic variants and these public datasets. The experimental strategy is illustrated in Fig. 1. The experiments were performed with HCT116 cells in exponential growth (log) phase (see Materials and Methods for details). RA was measured using standard RNA-seq method. Simultaneously, we measured genome-wide TR and TA, using the GRO-seq technique and the polysome profiling technique, respectively. We chose GRO-seq for two reasons: first, its higher sensitivity than the RNA Polymerase II ChIP-seq (Pol-II ChIP-seq) analysis; and, second, to avoid the need for label-time calibration associated with metabolic labeling based methods (i.e., 4sU-seq). Nevertheless, it has been shown that 4sU-seq data agree well with Pol-II ChIP-seq, and thus GRO-seq, data<sup>29</sup>. Therefore, our selection of TR measurement method should not matter. Polysome profiling was performed as previously described (see Methods for detail)<sup>40,41</sup>, with a polysome profile illustrated in Figure S1. The NGS reads were aligned to the human genome with TopHat<sup>42</sup> and the read counts for expressed genes were calculated with the HTSeq-count software<sup>43</sup>. The read counts were then converted into Reads Per Kilo-base Per Million Mapped Reads (RPKM) values. With a cut-off of 1 RPKM for at least one of the three parameters, 12921 genes were found expressed in the HCT116 cells. As expected, the experimental results are highly repeatable. For all three parameters, respective biological replicates are extremely consistent with each other, with linear regression R-squared values of at least 0.94 (see Methods for detail). The dataset provides a unique opportunity to generate mechanistic insight into cellular coordination of major transcriptome regulatory mechanisms.



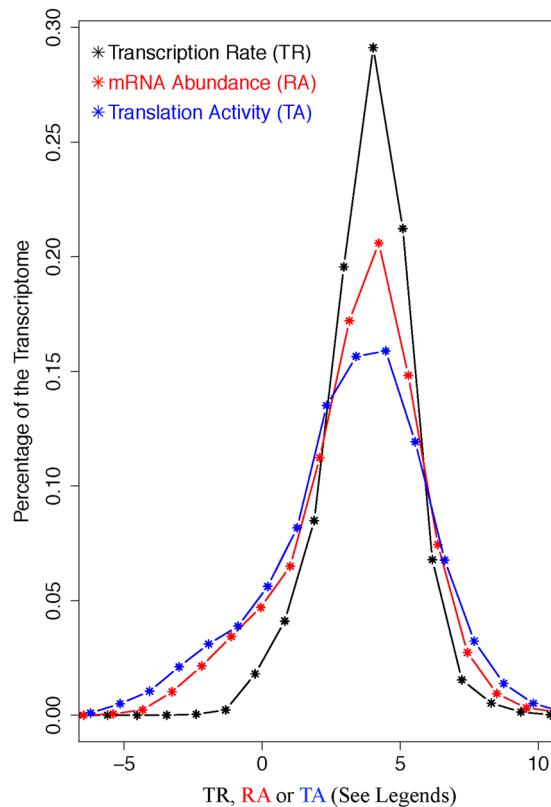
**Figure 1.** Experimental strategy. Log-phase HCT116 cells were split up into three parts. One part was used to extract total mRNA for RNA-seq analysis to measure steady-state mRNA abundance (RA) (black texts and arrows). One part was used to perform nuclear run-on to generate bromo-UTP labeled nascent RNA for sequencing, that is, GRO-seq analysis to measure transcription rate (TR) (red texts and arrows). The last part was used to isolate and quantify polysome associated mRNA to measure translation activity (TA) (green texts and arrows).



**Figure 2.** Comparison of TR, RA and TA to illustrate the discrepancy among the three parameters. **(A)** Scatter plot of TR (y-axis) versus RA (x-axis) (red) and RA biological replicate 2 (Y-axis) versus RA biological replicate 1 (x-axis) (black). **(B)** Scatter plot of TA (y-axis) versus RA (x-axis) (red) and of two RA biological replicates (black). The same two RA biological replicates are used in **(A)** and **(B)**. The linear regression lines are also shown.

### Comparative analysis of the three gene expression parameters, revealing sequentially higher levels of selectivity from transcription to translation.

Comparative analysis of the three gene expression parameters revealed extensive differences among them. Individual pairwise comparison resulted in, as expected, a general trend of good correlation; that is, an association of a high value of one parameter with high values of other parameters. However, regression analysis revealed quite dramatic differences, which are beyond intrinsic experimental noises, among the three parameters (Fig. 2). In Fig. 2A, TR and RA are plotted as red data points, with TR as y-axis and RA as x-axis; in the same scatter plot, one RA biological replicate (RA2) and another RA replicate (RA1) is plotted as black data points, with RA2 as y-axis and RA1 as x-axis; the two RA replicates illustrate the level of the intrinsic experimental noises. The two RA replicates agree with each other, with a linear regression slope of about 1 and a low level of dispersion along the regression line. However, the TR-RA regression is dramatically different. The slope of the regression line is only 0.5, suggesting systematic difference between the two parameters. Additionally, in Fig. 2B, TA and RA are plotted as red data points, with TR as y-axis and RA as x-axis; RA2 and RA1 are plotted in the same manner as in Fig. 2A. The TA-RA regression line is different from the RA2-RA1 line. The change in the slope of the regression line, an increase to 1.11, is not as dramatic. But, statistically, it is highly significant, with a p-value of less than  $1E-200$  – essentially zero (see Methods for detail). Thus, systematic discrepancies exist among the three key transcriptome parameters.

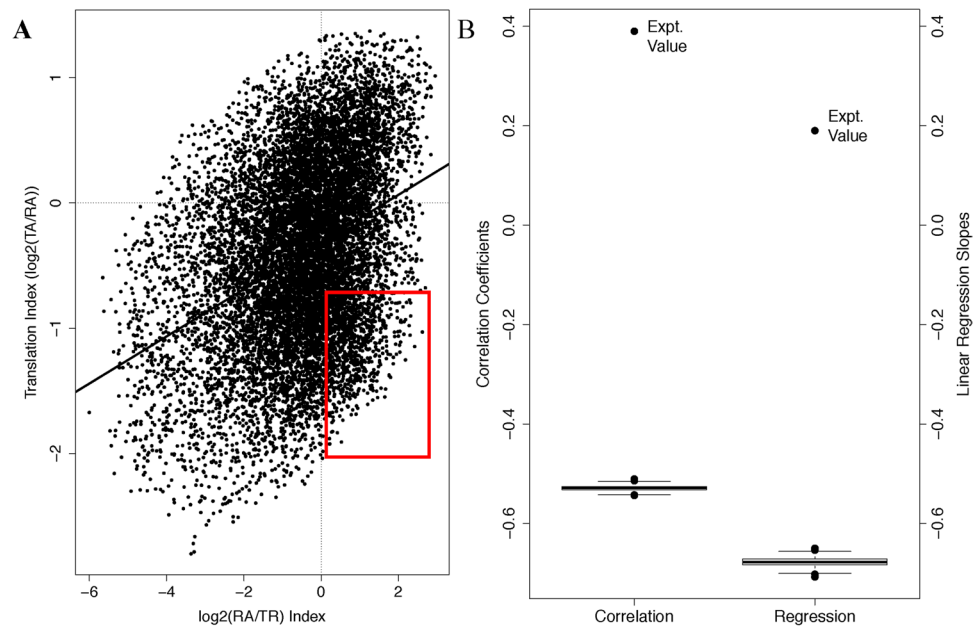


**Figure 3.** Comparison of the profiles/histograms of the three transcriptome parameters – TR, RA and TA. Histograms of TR (black), RA (red) and TA (blue) are shown as specified by the legends inside the figure. The RA and TA histograms are shifted a little bit so that the three histograms have their peaks in the same x-axis range, in order to better display the increased levels of dispersion from TR and RA and then to TA.

A systematic trend was observed when the statistical features of the three parameters were compared. The levels of dispersion of the three profiles increase from transcription rate to mRNA abundance, and then to translation activity. Schematically, the trend is illustrated by a comparison of the histograms of the three parameters shown in Fig. 3; quantitatively, this trend is illustrated by the sequential increase of two standard statistical parameters – the standard deviation and the value range of the three profiles (Table S1). This trend is consistent with the observations in Fig. 2, where the two regression lines in Fig. 2A indicate that RA tends to have higher values than TR for high expression level genes and lower values than TR for low expression level genes, and thus has higher dispersion than TR; similarly, the two regression lines in Fig. 2B indicate TA has higher dispersion than RA.

Thus, both regression analysis and statistical parameters of the distributions demonstrate that more and more genes display extreme (either low or high) parameter values from TR to RA, and then to TA. That is, the functional consequence of cellular coordination of the major transcriptome regulatory mechanisms is sequentially enhanced gene expression selectivity as the genetic information flow in the direction dictated by the molecular biology Central Dogma (from genome to mRNA, then to protein). We are aware that high-throughput analysis suffers from potential batch effect<sup>44</sup>. Batch effect tends to be randomly distributed and enhance experimental noise, and becomes a significant technical problem when it somehow correlates with key analysis parameter. Since our experiment uses three methodologies, batch effect might occur. However, this is unlikely the cause of observed discrepancy among the three parameters due to: (1) the histogram of mRNA and protein abundance in Fig. 2b of the work of Schwanhausser *et al.* was generated with different methodologies and independently support this notion of enhanced gene expression selectivity from transcriptome to proteome, though the authors did not notice it in the paper<sup>10</sup>; (2) potential batch effect is unlikely aligned so-well with the direction of the molecular biology central dogma (transcription to mRNA abundance, and then to translation); (3) the function/pathway-specific pattern and the effect of mRNA UTR proportion on the mRNA stabilization-by-translation regulatory mechanism, to be described later, further exclude batch effects, as batch effect cannot possibly correlate with all these parameters. Next, we further dissected our dataset to decipher the underlying mechanisms that give rise to the observed discrepancies and lead to the sequential enhancement of gene expression selectivity.

**Contribution of mRNA-stabilization-by-translation to the sequential enhancement of gene expression selectivity.** We tested whether, and to what extent, the TA-RA and the RA-TR discrepancy are related with each other. The TA-RA discrepancy is a reflection of mRNA translation regulation; enrichment of mRNA species with high translation activity in polysome complexes leads to higher TA values than RA values – and *vice versa*. In the case of TR and RA, the discrepancy mostly results from regulation of both mRNA degradation and, to a much lesser extent, from



**Figure 4.** Overall positive correlation between the stability index ( $\log_2(\text{RA/TR})$ ) and the translation index ( $\log_2(\text{TA/RA})$ ). **(A)** Scatter plot of the stability index versus the translation index. The correlation coefficient and the linear regression line between the two indices are also shown. The red rectangle identifies mRNAs with high stability but low translation activity to be further analyzed later (see text and Fig. 7). **(B)** Box plot of the correlation coefficient and the slope of linear regression lines between the two indexes upon randomization of the experimental dataset. Values from 1000 randomization were used to generate the boxplot. Experimental values are also shown and denoted.

RNA processing. RNA processing is extensively coupled to transcription<sup>45</sup>, and transcription was shown to be on average more than three folds slower than RNA processing<sup>6</sup>. Transcription should be, in most cases, the rate limiting step in mRNA production. Thus, TR, as has been reported, closely correlates with mRNA production rate<sup>6</sup>. By extension, TR-RA discrepancies should be mostly accounted for by mRNA stability control; unstable mRNA (or high degradation rate) leads to lower RA values than mRNA production rate and, in turn, TR values – and *vice versa*. Furthermore, it is known that active translation shields mRNAs from degradation, thus stabilizing the mRNA molecules and contributing to the discrepancy between mRNA production rate and RA<sup>35</sup>. In other words, translation activity should be a major determinant of how RA deviates from mRNA production rate and, in light of close correlation between TR and mRNA production rate, the TR-RA discrepancy. Our data provide a unique opportunity, to our knowledge for the first time, for a genome-wide and quantitative assessment of the contribution of this stabilization-by-translation regulatory mechanism to transcriptome regulation. For this purpose, we used the  $\log_2(\text{TA/RA})$  and  $\log_2(\text{RA/TR})$  log ratios as translation index and RA-TR discrepancy index, respectively. The former is the log ratio between actively translated mRNA abundance and total mRNA abundance, thus a measurement of mRNA translation activity normalized against RA; the latter is the log ratio between total mRNA abundance and transcription rate, and can be operationally considered as a close estimate of mRNA stability.

We hypothesized that the stabilization-by-translation regulatory mechanism should exert a significant effect and lead to positive correlation between the two indices. Indeed, our experimental results support this hypothesis. As shown in Fig. 4A, a dot-plot of the two indices illustrates an overall positive relationship, with a correlation coefficient of 0.39; the linear regression line is also shown to quantify the relationship, with a slope of 0.19. In other words, an overall positive correlation was observed. If our hypothesis is wrong, the two indices should have been negatively correlated, since RA is the numerator in the RA-TR index ( $\log_2(\text{RA/TR})$ ) and the denominator in the translation index ( $\log_2(\text{TA/RA})$ ).

To examine the significance of this observation, we randomly permuted the TR and TA parameters simultaneously to generate a statistical background model. As expected, this led to negative correlation coefficients, and negative slopes of the linear regression line, between the two indexes. We performed the randomization for 1000 times. This generated 1000 correlation coefficients and 1000 slopes of the corresponding linear regression lines, the boxplots of both of which were shown in Fig. 4B. Out of the 1000 randomization, not a single positive correlation was observed – both values were always negative. Figure 4B also shows the experimentally determined positive values of the correlation coefficient and the linear regression line slope, demonstrating a sharp contrast with the respective randomly generated values. This contrast illustrates the magnitude of the effect of the stabilization-by-translation regulation mechanism on the relationship between the two indices. Thus, consistent with our hypothesis, the stabilization-by-translation regulatory mechanism renders the relationship into an overall positive one. Moreover, it should be pointed out that the analysis likely under-estimates the significance of the stabilization-by-translation regulatory mechanism, due to the lack of RNA processing information in the RA-TR index.

To examine a potential causal relationship from this stabilization-by-translation regulatory mechanism to the enhancement of gene expression selectivity from transcription to translation, another statistical experiment was performed. We standardized the three genome profiles. This statistical procedure – mean subtraction followed by division by the standard deviation – is routinely used to eliminate statistical differences between distributions. The transformation should, in this case, reduce the numerical differences among the three parameters for many genes shown in Fig. 2A,B. For instance, a gene with higher TA and lower TR values than its RA value, for instance, should have similar values for all three parameters in the transformed dataset. That is, the statistical differences among the three genomic profiles shown in Fig. 3 and Table S1 were eliminated by the standardization procedure. If the stabilization-by-translation regulatory mechanism, that is, selective degradation of translationally inactive mRNA, is a major cause of the discrepancy among the three parameters, the positive correlation between the translation and the stability indices must also be significantly reduced in the transformed data. This is, indeed, the case. The correlation coefficient was reduced from 0.39 to 0.12, supporting selective degradation of translationally inactive mRNA as a major cause for the enhancement of gene expression selectivity from transcription to translation. Additionally, many functional genomic normalization procedures rely on standardization of the datasets; and many others, such as the rank and quantile based methods, have similar effects. Our results raised a technical issue that they may not be good choices for multi-parameter gene expression studies, as valuable information will be lost.

Summarily, all three analysis (correlation, permutation and standardization) support mRNA stabilization-by-translation as a major contributor to enhancement of gene expression selectivity from transcription to translation, which is an important functional aspect of transcriptome regulation. To our knowledge, this is the first quantitative and genome-wide examination of the contribution of this regulatory mechanism.

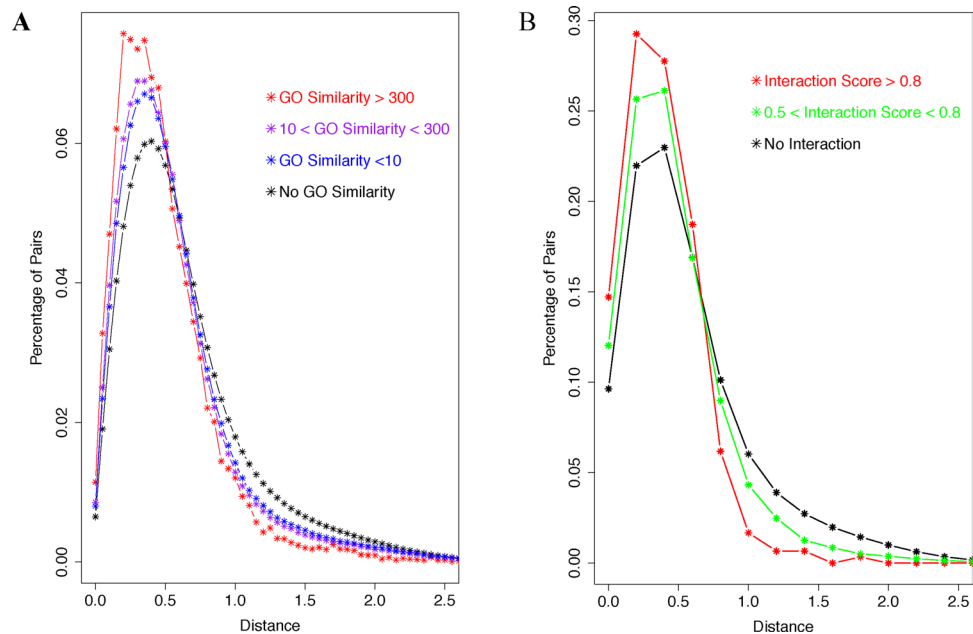
### Pathway/function specific pattern of correlation between the stability and the translation indices.

The correlation between the two indices, on the other hand, is not nearly unequivocal; too many genes deviate significantly from the overall trend – the linear regression line (Fig. 4A). We asked whether this is due to function specific patterns of gene expression regulation, as genes involved in the same biological process have been shown to share a similar pattern in other datasets. To answer this question, we performed two systematic analyses. First, we calculated the distances between the coordinates of each gene pair in Fig. 4A. We then created the histograms of the pairwise distances between gene pairs associated with similar sets of gene ontology (GO) terms/fingerprints (see Methods for detail), and also a histogram for the distances between gene pairs with no significant GO similarity. The distance between genes associated with similar GO terms tend to be smaller than those between genes with no significant similarity in their GO association (Fig. 5A). The trend is correlated with the GO fingerprint similarity score; the higher the score, the more the histogram shifts toward short distance range. Second, we performed this comparison of distances between gene pairs whose proteins interact with each other versus gene pairs whose proteins have not been found to interact with each other. This was done with protein-protein interaction data, which was, as previously described<sup>46</sup>, downloaded from the IntAct database<sup>47,48</sup>. As shown in Fig. 5B, the interacting pairs exhibit shorter distance than non-interacting pairs. And the trend is correlated with the confidence score assigned to the protein pairs in the IntAct database. Since the protein-protein interaction datasets are generally considered noisy, the confidence score quantifies the reliability of the interaction. As shown in Fig. 5B, the more reliable the interaction, the more the histogram shifts toward short distance range.

This function specific pattern is illustrated by the distinct patterns of two exemplary functional groups of genes – the genes for the proteasome subunit (PMS) proteins (PMSA1 to 7 and PMSB1 to 7) and the like-Sm (LSM) genes (LSM1 to 8) (Figure S2). The PMS genes code for proteins that constitute the proteasome 20S core structure<sup>49</sup>. Their mRNAs share a pattern of high levels in both indices. The LSM genes code for subunits of two single-stranded-RNA-binding hetero-heptameric ring structures – one cytoplasmic and the other nucleus<sup>50</sup>. Subunits LSM1 to 7 form the heptamer that is part of the P-body and functions during mRNA degradation in the cytoplasm. Consistently, LSM1-7 mRNAs share a common pattern. However, the pattern is strikingly different from the pattern shared by the PMS mRNAs. While the LSM1-7 mRNAs exhibit relatively high TR-RA index values, unlike PMS mRNA, they exhibit largely lower than average translation activity. The LSM8 subunit interacts with, and nucleus-retains, LSM2 to 7 subunits to form the nucleus heptameric ring structure<sup>50</sup>. That is, it replaces the LSM1 subunit to form the nucleus heptameric structure. This heptameric structure binds to the U6 snRNA and U8 small nucleolar RNA (snRNA), and thus functions during general RNA maturation in the nucleus. Consistent with this unique LSM8 function, the LSM8 mRNA does not follow the pattern shared by LSM1 to 7 mRNAs, in that it is relatively unstable (Figure S2).

Thus, we observed a function/pathway specific pattern of variation in how much individual mRNA species are regulated by the stabilization-by-translation regulatory mechanism, *i.e.*, the level of correlation between the two indexes; some mRNAs defy this regulation completely in that they stay out of polysome but retain high RA-TR index values and, likely, high stability. As discussed earlier, this function/pathway-specific pattern further exclude batch effects as the cause of discrepancies among the gene expression parameters shown in Figs 2 and 3, as potential batch effect cannot possibly correlate with all these GO annotation and biochemical pathways. Next, we tried to gain mechanistic insight into this variation, and turned our attention to other post-transcription regulatory mechanisms and the untranslated region (UTR) of mRNAs.

**mRNA UTR proportion is a major determinant of whether a mRNA obeys or defies the stabilization-by-translation regulatory mechanism.** Besides stabilization-by-translation regulation, many other mechanisms exist in multi-cellular eukaryotic species, but have not been accounted for in our analysis. For instance, the miRNAs/siRNAs target and regulate a large portion of the transcriptome. Essentially all regulatory signals for such regulation are embedded in mRNA UTR sequences. Consistently, UTRs are abundant in



**Figure 5.** Pathway/function specific pattern of the correlation between the stability and the translation indexes. **(A)** Histograms of the distances between their mRNAs' coordinates in Fig. 3A for gene pairs with different levels of GO similarity. **(B)** Histograms of the distances for gene pairs whose proteins were shown to mutually interact with different levels of interaction confidence.

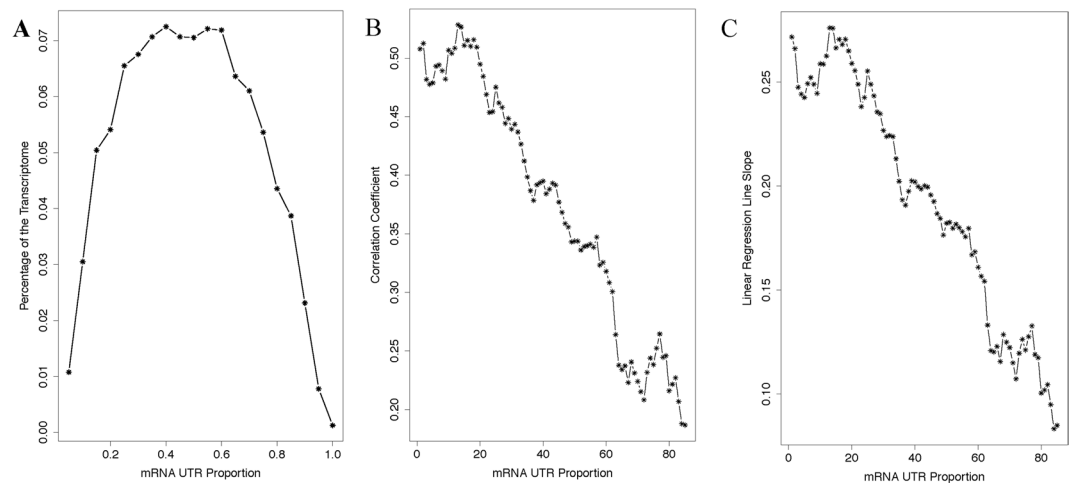
multi-cellular transcriptomes. This is especially true in human. As shown in Fig. 6A, on average, the ORF occupies only ~50% of a human mRNA; the other half is devoted to the UTRs. In many mRNAs, the UTR occupies more than 90% of the total length; for instance, the mRNAs of the all-important CREB1 (cyclic AMP-responsive element-binding protein 1) gene. Since the regulatory signals for mRNA post-transcription regulation are mostly embedded in the UTR sequences, the proportion of an mRNA that is occupied by the UTRs should serve as a good measure of the degree to which the mRNA is controlled by these regulatory mechanisms. Thus, we hypothesized that this proportion should be a major explanatory factor for the high level of variation in the correlation between the two indexes. Indeed, our results support this hypothesis. First, the correlation coefficient between the two indices is optimal at ~20% UTR, but steadily decreases as this proportion further increases (Fig. 6B); as well as the slope of the linear regression line between the two indices (Fig. 6C). Second, the mRNAs that defy the stabilization-by-translation regulatory mechanism, those that show low translation activity but relatively high RA-TR index values as identified by the red rectangle in Fig. 4A, display higher proportion of UTRs. The histogram of their UTR proportions, when compared with that of the whole human transcriptome, shifts clearly toward high proportion ranges (Fig. 7). As mentioned earlier, the patterns illustrated in Fig. 6B,C further help to exclude batch effects as the cause of discrepancies among the gene expression parameters shown in Figs 2 and 3, as potential batch effect should be randomly distributed among the genes and cannot possibly correlate so well with mRNA UTR proportions.

To exemplify this observation, we once again used the PMS and the LSM functional groups of genes shown in Figure S2 (Table S2). The mRNAs of the PMS group of genes have high levels of both RA-TR index and translation activity, and thus exemplify mRNAs controlled by the stabilization-by-translation mechanism. Consistently, as shown in Table S2, they have less-than-average UTR proportions (an average of 35.3% and a median of 26.4%). The mRNAs for the LSM genes, on the other hand, have low levels of translation activity but relatively high RA-TR index values, and thus exemplify mRNAs defying the stabilization-by-translation regulatory mechanism. Not surprisingly, they have higher-than-average UTR proportions (a mean of 69.6% and a median of 68.6%). The difference between the UTR proportions of the mRNAs of the two functional groups of genes is statistically significant; it has, according to a t-test, a p-value of 0.0004 (Table S2).

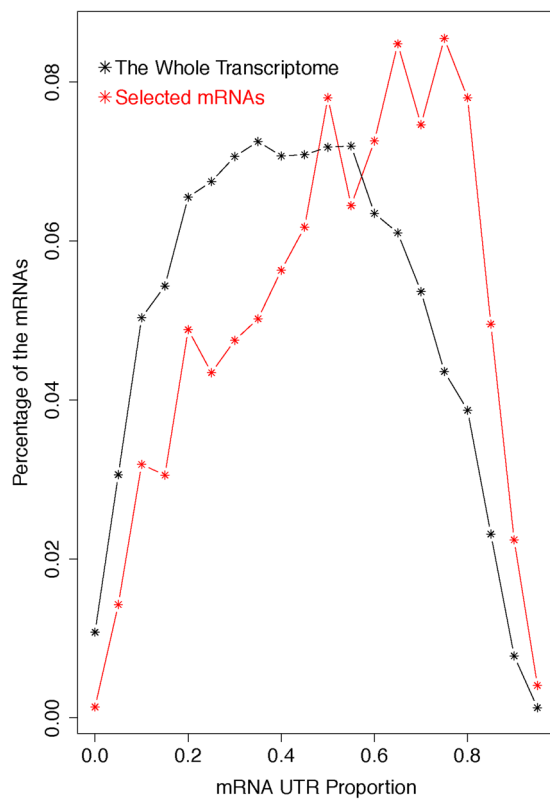
Thus, variation in the correlation between the two indices can be partially explained by post-transcription regulatory mechanisms mediated by mRNA UTRs. To put it another way, multiple regulatory mechanisms control the transcriptome. Multi-parameter approaches, such as ours, have the urgently needed power to dissect cellular co-ordination of these mechanisms and functionally characterize UTR sequences.

## Discussion

Regulation of the transcriptome is a major underpinning of cellular operation. It involves, in addition to transcription, many post-transcriptional processes. Multiple parameters, such as TR, RA and mRNA degradation rate, are relevant to this multi-faceted process. Many multi-parameter studies have reported significant discrepancies among the parameters, such as those between TR and RA and those between RA and protein abundance, prompting an appreciation for the complexity of transcriptome regulation.



**Figure 6.** The proportion of a mRNA that is occupied by the UTRs is a determinant of the level of the correlation between the stability and the translation indices. **(A)** Histogram of the UTR proportions of human mRNAs. **(B and C)** The correlation coefficient **(B)** and the slope of the linear regression line **(C)** between the stability and the translation indices decrease as the mRNA UTR proportion increases.



**Figure 7.** Histogram of the UTR proportions of the mRNAs that have relative high stability but lower-than-average translation activity in comparison with that of the whole human transcriptome. The set of mRNAs with high stability but low translation activity was selected with the red rectangle in Fig. 4A.

We have previously participated in the study of the complexity of transcriptome regulation, with a desire for a mechanistic understanding of the discrepancies among TR, RA and protein abundance as well as potential operational advantages the cells gain from them. In this study, we took advantages of the power and versatility of NGS analysis through its coupling to traditional experimental protocols. We simultaneously measured three transcriptome regulation parameters: TR, RA and TA. Based on the close correlation between TR and mRNA production rate, we also indirectly estimate mRNA stability (or degradation rate) by the log ratio of RA and TR ( $\log_2(RA/TR)$ ). Given the importance of mRNA UTRs in post-transcriptional regulation, the data was analyzed in conjunction with an individual mRNAs UTR proportion. To put it another way, we broke open the “blackbox” of



transcriptome regulation and peeked inside for mechanistic insight into how the cells co-ordinate multiple factors that regulate the transcriptome. In the present paper, we publish, to our knowledge, the first genome-wide dataset that enables integrative analysis of TR, RA, TA and, to some degree, mRNA stability.

It is well known that actively translating mRNA are likely protected from degradation, and thus stabilized. In bacteria, this is considered the primary mechanism for mRNA stability regulation. In eukaryotes, more post-transcriptional regulatory mechanisms, such as micro-RNA control, evolutionarily emerged, giving rise to a more complicated scheme of mRNA stability regulation. However, it is certain that the stabilization-by-translation mechanism still plays a prominent role in eukaryotic transcriptome regulation. We provide a quantitative analysis of the impact of translation activity on transcriptome regulation, by showing a moderate but significant positive correlation between the mRNA translation index and the RA-TR index.

This correlation has some explanatory power over the discrepancy between TR and RA. High translation activity protects a mRNA species from degradation, while other less translated mRNAs are being actively degraded and removed out of the transcriptome. This leads to enrichment of the mRNA species, resulting in higher steady-state abundance level than implied by its production rate and, by extension, TR. Conversely, low translation activity makes a mRNA species more susceptible to the degradation process, leading to situations where the steady-state abundance level is lower than that implied by the production rate. The operational advantages the cells gained by implementing this regulatory scheme remain to be elucidated.

Our multi-parameter approach represents a feasible option to enable the much-needed systematic analysis of mRNA UTRs. UTRs are much more abundant in the human transcriptome than in any other transcriptome. Their functions in post-transcriptional regulation are well documented. Essentially, all signals for post-transcriptional regulation reside in the UTRs; for instance, microRNA and siRNA target sites, ARE, IRE-IRP etc. But systematic study of mRNA UTRs has been lacking, and our knowledge about their functions remains fragmentary at best. This is perhaps due to a lack of relevant genomic experimental approaches and datasets. mRNA abundance measurements alone are ill-suited for the study of post-transcription regulatory mechanisms and functional analyses of mRNA UTRs. Additionally, though microRNAs/siRNAs target mRNA UTRs and are major regulators of both translation and mRNA degradation, to our knowledge, microRNA/siRNA study has not been integrated with simultaneous genome-wide measurement of translation activity and mRNA stability. Thus, our integrative multi-parameter analysis represents a novel functional genomic approach to mRNA UTR analysis. It is able to reveal that the UTRs potentially play an important role in maintaining the stability of translationally inactive mRNA species, thus conferring to human cells the capacity for a post-transcriptional regulatory mechanism that is absent in prokaryotic species and mostly in uni-cellular species such as the yeast *S. cerevisiae*. It should be noted that our results represent only a single time-point snap-shot of actively growing human cells. More power of this analysis approach, we believe, is yet to be relished in analyzing dynamic changes of the three parameters during physiological processes, and when it is coupled to more accurate mRNA production rate measurement techniques such as nascent-RNA-metabolic-labeling based approaches.

Additionally, computational analysis of mRNA UTRs for key regulatory signals embedded in the UTR sequences remains technically challenging. This is due to low signal-to-noise ratio and the lack of a general guiding principle. For instance, a typical mammalian microRNA target site is no more than 8 nucleotide long. Our approach provides a way to classify the mRNAs based on their patterns in the generated datasets, i.e., their behavior in the multi-faceted transcriptome regulation process. Key regulatory signals shall be shared by the UTRs of similarly classified mRNAs, and thus can be computationally extracted from them – a much easier approach than *de novo* computational analysis of mRNA UTR sequences. That is, datasets generated through this approach should provide a functional context for enhancing the signal-to-noise ratio in computational analysis of mRNA UTR sequences.

We also quantitatively describe the trend of sequentially higher levels of selectivity as the genetic information flow from the genome to the proteome in the gene expression process. In other words, the gene expression machinery focuses its resources on less and less genes, so that only mission critical proteins are expressed in the proteome. The multi-stepped gene expression process can be considered as, to some degree, a selective amplification process. Transcription selectively amplifies the genomic sequences into multiple copies of mRNA sequences. Translation, in turn, selectively amplifies individual mRNA molecules into multiple copies of protein sequences. The selectivity of this process is further enhanced by selective mRNA degradation. Even though obvious from the results in previous publications, this trend of sequentially higher levels of selectivity in the gene expression process has not received much attention, and was never explicitly stated in these reports. In this study, we quantitatively described this trend by comparing the dispersions of the genomic profiles of the three gene expression parameters. Our results also suggest that mRNA degradation plays perhaps the biggest role in this trend, as the jump in selectivity from transcription rate to mRNA abundance is much bigger than the increase from mRNA abundance to translation activity. That is, selective degradation of those mRNAs, which are not protected from degradation by active translation or other processes mediated by their UTRs, play an important role in shaping up the transcriptome and priming it for efficient production of mission-critical proteins.

## Conclusions

In summary, we present a quantitative delineation of cellular coordination of transcription, mRNA abundance, mRNA translation activity, to some degree mRNA stability as well as mechanistic involvement of mRNA UTRs in the coordination process. As a consequence of the coordination activity, the cells exhibit sequentially higher level of gene expression selectivity from transcription to mRNA abundance, and then to translation activity. The results contribute to our understanding of the complexity of the multi-stepped gene expression process, through which the cells “read” the genomic “book” of seemingly simplistic string of nucleotides and “translate” information embedded in the sequences into cellular operations, that is, dynamic control of the biochemical flow through biochemical reactions, pathways and networks<sup>1–4</sup>.

## Methods

**Tissue Culture and mRNA Isolation for RNA-seq Analysis.** The human HCT116 cells were cultured in a serum-free medium (McCoy's 5A (Sigma) with pyruvate, vitamins, amino acids and antibiotics) supplemented with 10 ng/ml epidermal growth factor, 20 µg/ml insulin and 4 µg/ml transferrin. Cells were maintained at 37 °C in a humidified incubator with 5% CO<sub>2</sub>.

To extract mRNA for RNA-seq analysis, RNeasy kit (Qiagen) was used to extract total RNA from the HCT116 cells according to manufacturer's specification. GeneRead Pure mRNA Kit (Qiagen) was then used to isolate mRNA from the total RNA for Illumina NGS sequencing according to manufacturer's specification.

**GRO-seq Analysis.** Global run-on was done as previously described<sup>26,51,52</sup>. Briefly, two 100 cm plates of HCT116 cells were washed 3 times with cold PBS buffer. Cells were then swelled in swelling buffer (10 mM Tris-pH 7.5, 2 mM MgCl<sub>2</sub>, 3 mM CaCl<sub>2</sub>) for 5 min on ice. Harvested cells were re-suspended in 1 ml of the lysis buffer (swelling buffer with 0.5% IGEPAL and 10% glycerol) with gentle vortex and brought to 10 ml with the same buffer for nuclei extraction. Nuclei were washed with 10 ml of lysis buffer and re-suspended in 1 ml of freezing buffer (50 mM Tris-pH 8.3, 40% glycerol, 5 mM MgCl<sub>2</sub>, 0.1 mM EDTA), pelleted down again, and finally re-suspended in 100 µl of freezing buffer.

For the nuclear run-on step, re-suspended nuclei were mixed with an equal volume of reaction buffer (10 mM Tris-pH 8.0, 5 mM MgCl<sub>2</sub>, 1 mM DTT, 300 mM KCl, 20 units of SUPERase-In, 1% Sarkosyl, 500 µM ATP, GTP, and Br-UTP, 2 µM CTP) and incubated for 5 min at 30 °C. Nuclei RNA was extracted with TRIzol LS reagent (Invitrogen) following manufacturer's instructions, and was resuspended in 20 µl of DEPC-water. RNA was then purified through a p-30 RNase-free spin column (BioRad), according to the manufacturer's instructions and treated with 6.7 µl of DNase buffer and 10 µl of RQ1 RNase-free DNase (Promega), purified again through a p-30 column. A volume of 8.5 µl 10 × antarctic phosphatase buffer, 1 µl of SUPERase-In, and 5 µl of antarctic phosphatase was added to the run-on RNA and treated for 1 hr at 37 °C. Before proceeding to immuno-purification, RNA was heated to 65 °C for 5 min and kept on ice.

Anti-BrdU argarose beads (Santa Cruz Biotech) were blocked in blocking buffer (0.5 × SSPE, 1 mM EDTA, 0.05% Tween-20, 0.1% PVP, and 1 mg/ml BSA) for 1 hr at 4 °C. Heated run-on RNA (~85 µl) was added to 60 µl beads in 500 µl binding buffer (0.5 × SSPE, 1 mM EDTA, 0.05% Tween-20) and allowed to bind for 1 hr at 4 °C with rotation. After binding, beads were washed once in low salt buffer (0.2 × SSPE, 1 mM EDTA, 0.05% Tween-20), twice in high salt buffer (0.5% SSPE, 1 mM EDTA, 0.05% Tween-20, 150 mM NaCl), and twice in TET buffer (TE pH 7.4, 0.05% Tween-20). BrdU-incorporated RNA was eluted with 4 × 125 µl elution buffer (20 mM DTT, 300 mM NaCl, 5 mM Tris-pH 7.5, 1 mM EDTA, and 0.1% SDS). RNA was then extracted with acidic phenol/chloroform once, chloroform once and precipitated with ethanol overnight. The precipitated RNA was re-suspended in 50 µl reaction (45 µl of DEPC water, 5.2 µl of T4 PNK buffer, 1 µl of SUPERase-In and 1 µl of T4 PNK (NEB)) and incubated at 37 °C for 1 hr. The RNA was extracted and precipitated again as above before being processed for Illumina NGS sequencing.

**Polysome Isolation and mRNA extraction.** Polysome was isolated as previously described<sup>40,41</sup>. Briefly, the HCT116 cells were incubated with 100 µg/ml cycloheximide for 15 minutes, washed three times with PBS, scraped off into PBS, and then pelleted by micro-centrifugation. Cell pellet was homogenized in a hypertonic re-suspension buffer (10 mM Tris (pH 7.5), 250 mM KCl, 2 mM MgCl<sub>2</sub> and 0.5% Triton X100) with RNase inhibitor and a protease cocktail. Homogenates were centrifuged for 10 min at 12,000 g to pellet the nuclei. The post-nuclear supernatants were laid on top of a 10–50% (w/v) sucrose gradient, followed by centrifugation for 90 min at 200,000 g. The polysomal fractions were identified by OD<sub>254</sub> and collected. RNeasy kit (Qiagen) was used to extract RNA from the polysome fractions according to manufacturer's specification. GeneRead Pure mRNA Kit (Qiagen) was then used to isolate mRNA for Illumina NGS sequencing from the RNA according to manufacturer's specification.

**Illumina NGS Sequencing.** Sequencing libraries were generated with the Illumina TruSeq RNA Sample Preparation Kit. Briefly, RNA molecules were fragmented into small pieces using divalent cations under elevated temperature. The cleaved RNA fragments are copied into first strand cDNA synthesis using reverse transcriptase and random primers. This was followed by second strand cDNA synthesis using DNA Polymerase I and RNase H. These cDNA fragments were end-repaired using T4 DNA polymerase, Klenow polymerase and T4 polynucleotide kinase. The resulting blunt-ended fragments were A-tailed using a 3'-5' exonuclease-deficient Klenow fragment and ligated to Illumina adaptor oligonucleotides in a 'TA' ligation. The ligation mixture was further size-selected by AMPure beads and enriched by PCR amplification following Illumina TruSeq DNA Sample Preparation protocol. The resulting library is attached and amplified on a flow-cell by cBot Cluster Generation System.

The sequencing was done with an Illumina HiSeq2000 sequencer. Multiplexing was used to pool 4 samples into one sequencing lane. After each sequencing run, the raw reads were pro-processed to filter out low quality reads and to remove the multiplexing barcode sequences. The dataset is available through the NCBI GEO database (accession number GSE111222).

**NGS Data Analysis.** The sequencing reads were mapped to the UCSC hg19 human genome sequences with the TopHat software, using the default input parameter values. For each sample, at least 80% of the reads were successfully mapped. For the sake of consistency across the three transcriptome regulation parameters, we counted the reads for each gene for the exon regions only. The counting was performed with the HTSeq-count software, and the counts were then transformed into Reads Per Kilo-base Per Million Mapped Reads (RPKM) values. 12921 genes have a minimal RPKM value of 1 for at least one of the three parameters, and were considered expressed in

the HCT116 cells. Linear regression of log transformed data was used to examine consistence between biological replicate samples.

**Statistical Analysis.** The R open source statistical software (version 3.3.1) installed on a Mac Pro desktop computer was used for statistical analysis. Outlier identification, student t-test, standard deviation calculation, correlation coefficient calculation, linear regression and other statistical procedure are all done with this R software.

The procedure for comparing the linear regression slopes/coefficients shown in Fig. 2B is described as follows. We first applied the following linear regression models to the data:

$$\begin{aligned}\log_2(\text{TA}) &= \mu_1 + \beta_1 \log_2(\text{RA}) + \varepsilon_1, \\ \log_2(\text{RA2}) &= \mu_2 + \beta_2 \log_2(\text{RA1}) + \varepsilon_2,\end{aligned}$$

and  $\varepsilon_1$  and  $\varepsilon_2$  follow normal distribution. It is estimated that

$$\begin{aligned}\widehat{\beta}_1 &= 1.11, \quad \widehat{\sigma}(\widehat{\beta}_1) = 0.003058, \quad \text{and } \widehat{\beta}_1/\widehat{\sigma}(\widehat{\beta}_1) \sim T_{12920}, \\ \widehat{\beta}_2 &= 0.99, \quad \widehat{\sigma}(\widehat{\beta}_2) = 0.001879, \quad \text{and } \widehat{\beta}_2/\widehat{\sigma}(\widehat{\beta}_2) \sim T_{12920},\end{aligned}$$

Therefore, the 97.5% confidence interval for  $\beta_1$  is

$$\beta_1 \pm t_{0.0125, 12920} \widehat{\sigma}(\widehat{\beta}_1) = (1.101403, 1.115113).$$

The 97.5% confidence interval for  $\beta_2$  is

$$\beta_2 \pm t_{0.0125, 12920} \widehat{\sigma}(\widehat{\beta}_2) = (0.9819329, 0.9903571).$$

These two confidence intervals do not overlap, implying that, at significant level 0.05, the two regression coefficients are different. In addition, because T distribution with degrees of freedom of 12920 is very close to standard normal distribution, the t-score, calculated as below,

$$\widehat{\beta}_2 - \widehat{\beta}_1 / \sqrt{\widehat{\sigma}^2(\widehat{\beta}_2) + \widehat{\sigma}^2(\widehat{\beta}_1)}$$

approximately follow standard normal distribution. This allows p-value calculation. The p-value is essentially 0 (smaller than 1E-200).

**Gene Ontology (GO) Similarity Analysis.** Pairwise GO similarity scores between human genes were computed as previously described<sup>53–55</sup>. Briefly, for each gene, we first generated GO fingerprint – a set of ontology terms enriched in the PubMed abstracts linked to the gene, along with the adjusted p-value reflecting the degree of enrichment of each term. The GO similarity score quantifies similarity between the GO fingerprints of corresponding gene pair. For detail about GO fingerprint generation and similarity calculation, please see description in previous publications<sup>54</sup>.

**Data Availability.** The dataset is deposited into the NCBI GEO database, with an accession number GSE111222.

## References

- Searls, D. B. Linguistic approaches to biological sequences. *Comput Appl Biosci* **13**, 333–344 (1997).
- Searls, D. B. Reading the book of life. *Bioinformatics* **17**, 579–580 (2001).
- Searls, D. B. The language of genes. *Nature* **420**, 211–217 (2002).
- Wang, D. G. “Molecular gene”: Interpretation in the right context. *Biol Philos* **20**, 453–464, <https://doi.org/10.1007/s10539-005-5584-6> (2005).
- Jovanovic, M. *et al.* Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science* **347**, 1259038, <https://doi.org/10.1126/science.1259038> (2015).
- Rabani, M. *et al.* High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. *Cell* **159**, 1698–1710, <https://doi.org/10.1016/j.cell.2014.11.015> (2014).
- Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2**, e270, <https://doi.org/10.7717/peerj.270> (2014).
- Liu, Y. & Aebersold, R. The interdependence of transcript and protein abundance: new data–new complexities. *Mol Syst Biol* **12**, 856, <https://doi.org/10.15252/msb.20156720> (2016).
- McManus, J., Cheng, Z. & Vogel, C. Next-generation analysis of gene expression regulation—comparing the roles of synthesis and degradation. *Mol Biosyst* **11**, 2680–2689, <https://doi.org/10.1039/c5mb00310e> (2015).
- Schwanhauser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342, <https://doi.org/10.1038/nature10098> (2011).
- Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**, 227–232, <https://doi.org/10.1038/nrg3185> (2012).
- Anderson, L. & Seilhamer, J. A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis* **18**, 533–537 (1997).
- Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol* **19**, 1720–1730 (1999).
- Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
- Flory, M. R. *et al.* Quantitative proteomic analysis of the budding yeast cell cycle using acid-cleavable isotope-coded affinity tag reagents. *Proteomics* **6**, 6146–6157 (2006).

16. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425**, 737–741, <https://doi.org/10.1038/nature02046> (2003).
17. Griffin, T. J. *et al.* Complementary profiling of gene expression at the transcriptome and proteome levels in *Saccharomyces cerevisiae*. *Mol Cell Proteomics* **1**, 323–333 (2002).
18. Le Roch, K. G. *et al.* Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome Res* **14**, 2308–2318, <https://doi.org/10.1101/gr.2523904> (2004).
19. Tian, Q. *et al.* Integrated genomic and proteomic analyses of gene expression in Mammalian cells. *Mol Cell Proteomics* **3**, 960–969, <https://doi.org/10.1074/mcp.M400055-MCP200> (2004).
20. Washburn, M. P. *et al.* Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* **100**, 3107–3112 (2003).
21. Cheng, Z. *et al.* Differential dynamics of the mammalian mRNA and protein expression response to misfolding stress. *Mol Syst Biol* **12**, 855, <https://doi.org/10.15252/msb.20156423> (2016).
22. Garcia-Martinez, J., Aranda, A. & Perez-Ortin, J. E. Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. *Mol Cell* **15**, 303–313 (2004).
23. Marin-Navarro, J. *et al.* Global estimation of mRNA stability in yeast. *Methods Mol Biol* **734**, 3–23, [https://doi.org/10.1007/978-1-61779-086-7\\_1](https://doi.org/10.1007/978-1-61779-086-7_1) (2011).
24. Molina-Navarro, M. M. *et al.* Comprehensive transcriptional analysis of the oxidative response in yeast. *J Biol Chem* **283**, 17908–17918 (2008).
25. Romero-Santacreu, L., Moreno, J., Perez-Ortin, J. E. & Alepuz, P. Specific and global regulation of mRNA stability during osmotic stress in *Saccharomyces cerevisiae*. *RNA* **15**, 1110–1120 (2009).
26. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848, <https://doi.org/10.1126/science.1162228> (2008).
27. Hah, N. *et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145**, 622–634, <https://doi.org/10.1016/j.cell.2011.03.042> (2011).
28. Eser, P. *et al.* Periodic mRNA synthesis and degradation co-operate during cell cycle gene expression. *Mol Syst Biol* **10**, 717, <https://doi.org/10.1002/msb.134886> (2014).
29. Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat Biotechnol* **29**, 436–442, <https://doi.org/10.1038/nbt.1861> (2011).
30. Dolken, L. *et al.* High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* **14**, 1959–1972, <https://doi.org/10.1261/rna.1136108> (2008).
31. Friedel, C. C., Dolken, L., Ruzsics, Z., Koszinowski, U. H. & Zimmer, R. Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. *Nucleic Acids Res* **37**, e115, <https://doi.org/10.1093/nar/gkp542> (2009).
32. Greenbaum, D., Colangelo, C., Williams, K. & Gerstein, M. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* **4**, 117–117 (2003).
33. Brar, G. A. *et al.* High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335**, 552–557, <https://doi.org/10.1126/science.1215110> (2012).
34. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. & Weissman, J. S. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223, <https://doi.org/10.1126/science.1168978> (2009).
35. Coldwell, M. J., Gray, N. K. & Brook, M. Cytoplasmic mRNA: move it, use it or lose it! *Biochem Soc Trans* **38**, 1495–1499, <https://doi.org/10.1042/BST0381495> (2010).
36. Morozov, I. Y. *et al.* mRNA 3' tagging is induced by nonsense-mediated decay and promotes ribosome dissociation. *Mol Cell Biol* **32**, 2585–2595, <https://doi.org/10.1128/MCB.00316-12> (2012).
37. Hayles, B., Yellaboina, S. & Wang, D. Comparing transcription rate and mRNA abundance as parameters for biochemical pathway and network analysis. *PLoS One* **5**, e9908, <https://doi.org/10.1371/journal.pone.0009908> (2010).
38. Wang, D. Discrepancy between mRNA and protein abundance: insight from information retrieval process in computers. *Comput Biol Chem* **32**, 462–468, <https://doi.org/10.1016/j.compbiolchem.2008.07.014> (2008).
39. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**, 333–351, <https://doi.org/10.1038/nrg.2016.49> (2016).
40. Day, R. T. *et al.* Acute hyperglycemia rapidly stimulates VEGF mRNA translation in the kidney. Role of angiotensin type 2 receptor (AT2). *Cell Signal* **22**, 1849–1857, <https://doi.org/10.1016/j.cellsig.2010.07.012> (2010).
41. Feliars, D., Duraisamy, S., Barnes, J. L., Ghosh-Choudhury, G. & Kasinath, B. S. Translational regulation of vascular endothelial growth factor expression in renal epithelial cells by angiotensin II. *Am J Physiol Renal Physiol* **288**, F521–529, <https://doi.org/10.1152/ajprenal.00271.2004> (2005).
42. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36, <https://doi.org/10.1186/gb-2013-14-4-r36> (2013).
43. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169, <https://doi.org/10.1093/bioinformatics/btu638> (2015).
44. Taub, M. A., Corrada Bravo, H. & Irizarry, R. A. Overcoming bias and systematic errors in next generation sequencing data. *Genome Med* **2**, 87, <https://doi.org/10.1186/gm208> (2010).
45. Pandit, S., Wang, D. & Fu, X. D. Functional integration of transcriptional and RNA processing machineries. *Curr Opin Cell Biol* **20**, 260–265, <https://doi.org/10.1016/j.ceb.2008.03.001> (2008).
46. Guo, Z., Jiang, W., Lages, N., Borchers, W. & Wang, D. Relationship between gene duplicability and diversifiability in the topology of biochemical networks. *BMC Genomics* **15**, 577, <https://doi.org/10.1186/1471-2164-15-577> (2014).
47. Orchard, S. *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358–363, <https://doi.org/10.1093/nar/gkt1115> (2014).
48. Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res* **40**, D841–846, <https://doi.org/10.1093/nar/gkr1088> (2012).
49. Kish-Trier, E. & Hill, C. P. Structural biology of the proteasome. *Annu Rev Biophys* **42**, 29–49, <https://doi.org/10.1146/annurev-biophys-083012-130417> (2013).
50. Khusial, P., Plaag, R. & Zieve, G. W. LSm proteins form heptameric rings that bind to RNA via repeating motifs. *Trends Biochem Sci* **30**, 522–528, <https://doi.org/10.1016/j.tibs.2005.07.006> (2005).
51. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294, <https://doi.org/10.1038/nature12644> (2013).
52. Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* **474**, 390–394, <https://doi.org/10.1038/nature10006> (2011).
53. Qin, T. *et al.* Finding pathway-modulating genes from a novel Ontology Fingerprint-derived gene network. *Nucleic Acids Res* **42**, e138, <https://doi.org/10.1093/nar/gku678> (2014).
54. Qin, T., Tsoi, L. C., Sims, K. J., Lu, X. & Zheng, W. J. Signaling network prediction by the Ontology Fingerprint enhanced Bayesian network. *BMC Syst Biol* **6**(Suppl 3), S3, <https://doi.org/10.1186/1752-0509-6-S3-S3> (2012).
55. Tsoi, L. C., Boehnke, M., Klein, R. L. & Zheng, W. J. Evaluation of genome-wide association study results through development of ontology fingerprints. *Bioinformatics* **25**, 1314–1320, <https://doi.org/10.1093/bioinformatics/btp158> (2009).

## Acknowledgements

We would like to acknowledge Dr. Jordan Crago of Texas Tech University for proofreading this manuscript. This work is supported by grants from the National Institute of Health (R01LM010212 and R15GM122006) to DW and a grant from Cancer Prevention and Research Institute of Texas (CPRIT) (RP170668) to WJZ.

## Author Contributions

W.J. performed molecular and cellular experimental portion of the project. Z.G., N.L., W.J.Z. and F.Z. performed computational and statistical analysis. D.F. performed polysome isolation. W.J. and D.G. designed the project and drafted the manuscript. All authors participated in critical revision of the manuscript and have given final approval for publication.

## Additional Information

**Supplementary information** accompanies this paper at <https://doi.org/10.1038/s41598-018-24039-1>.

**Competing Interests:** The authors declare no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018