RESEARCH

The impact of alternative splicing on RNA subcellular localization

Chao Zeng[1,2*] and Michiaki Hamada[1,2,3,4,5†]

*Correspondence: zeng.chao@aist.go.jp
[1] AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), 3-4-1, Okubo Shinjuku-ku, 169-8555 Tokyo, Japan
Full list of author information is available at the end of the article
†Correspondence may also be addressed to Michiaki Hamada. Email: mhamada@waseda.jp

Abstract

Background: Alternative splicing, a ubiquitous phenomenon in eukaryotes, provides a regulatory mechanism for the biological diversity of individual genes. Most studies have focused on the effects of alternative splicing for protein synthesis. However, the influence of alternative splicing on the RNA subcellular localization has rarely been studied.

Results: By analyzing RNA-seq data from subcellular fractions across thirteen human cell lines, we observed that splicing is apparent to promote cytoplasmic localization. We also discovered that intron retention is preferred by transcripts localized in the nucleus. Short and structurally stable introns show a positive correlation with nuclear localization. Such introns are predicted to be preferentially bound by MBNL1, an RNA-binding protein that contains two nuclear localization signals.

Conclusions: Our findings reveal that alternative splicing plays an important role in regulating RNA subcellular localization. This study provides valuable clues for understanding the biological mechanisms of alternative splicing.

Keywords: alternative splicing; localization; intron retention; RNA-binding protein

Introduction

Most eukaryotic genes consist of exons (encoding mature RNAs) and introns (removed by RNA splicing). Alternative splicing is known as a regulated process by which exons can be either included or excluded. Various RNAs (also called transcript variants) can be produced from a single gene through alternative splicing. Thus, alternative splicing enables a cell to express more RNA species with a limited number of genes, which increases the genome complexity [1]. In humans, for instance, about

95% of the multi-exonic genes undergoing alternative splicing were uncovered by high-throughput sequencing technology [2]. Exploring the functionality of alternative splicing is critical to our understanding of life mechanisms. Alternative splicing has been reported to be associated with protein functions, such as diversification, molecular binding properties, catalytic and signaling abilities, and stability. Such related studies have been reviewed elsewhere [3, 4]. Additionally, relationships between alternative splicing and disease [5] or cancer[6, 7] has received increasing attention. Understanding the pathogenesis associated with alternative splicing can shed light on diagnosis and therapy. With the emergence and rapid development of high-throughput technology, it has become possible to study the function and mechanism of alternative splicing genome-wide [8].

The location of an RNA in a cell can determine whether the RNA is translated, preserved, modified, or degraded [9, 10]. In other words, the subcellular location of an RNA is highly related to its biological function [9]. For example, the asymmetric distribution of RNA in cells can influence the expression of genes [9], the formation and interaction of protein complexes [11], the biosynthesis of ribosomes [12],the development of cells [13, 14], among other functions. Many techniques have been developed to investigate the subcellular localization of RNAs. RNA fluorescent in situ hybridization (RNA FISH) is a conventional method to track and visualize a specific RNA by hybridizing labeled probes to the target RNA molecule [15, 16]. Improved FISH methods using multiplexing probes to label multiple RNA molecules have been presented to expand the range of target RNA species [17, 18]. With the development of microarray and high-throughput sequencing technologies, approaches for the genome-wide investigation of RNA subcellular localizations have emerged [19]. Recently, a technology applying the ascorbate peroxidase APEX2 (APEX-seq) to detect proximity RNAs has been introduced [20, 21]. APEX-seq is expected to obtain unbiased, highly sensitive, and high-resolved RNA subcellular localization *in vivo*. Simultaneously, many related databases have been developed [22, 23], which integrate RNA localization information generated by the above methods and provide valuable resources for further studies of RNA functions.

Previous studies have reported that alternative splicing can regulate RNA/protein subcellular localization [24, 25, 26, 27]. However, to date, a limited number of genes have been examined. One approach to solve this problem involves the use of high-throughput sequencing. The goal of this research was to perform a comprehensive and genome-wide study of the impact of alternative splicing on RNA subcellular localization. Therefore, we analyzed RNA-sequencing (RNA-seq) data from subcellular (cytoplasmic and nuclear) fractions and investigated whether alternative splicing causes an imbalance of RNA expression between the cytoplasm and the nucleus. Briefly, we found that RNA splicing appeared to promote cytoplasmic localization. We also observed that transcripts with intron retentions preferred to localize in the nucleus. A further meta-analysis of retained introns indicated that short and structured intronic sequences were more likely to appear in nuclear transcripts. Notably, we also found that the CUG repeat sequence was enriched in the retained introns. Our analysis suggested that MBNL, an RNA-binding protein (RBP), recognized CUG-containing transcripts and led to nuclear localization by its nuclear signal sequence (NLS). The above results are consistently observed across thirteen cell lines, suggesting the reliability and robustness of the investigations. To our knowledge, this is the first genome-wide study to analyze and evaluate the effect of alternative splicing on RNA subcellular localization across multiple cell lines. We believe that this research may provide valuable clues for further understanding of the biological function and mechanism of alternative splicing.

Results

Thousands of transcript switches between cytoplasm and nucleus were identified across thirteen human cell lines

To assess the influence of alternative splicing on RNA subcellular localization, we analyzed RNA-seq datasets that cover the nuclear and cytoplasmic fractions in thirteen human cell lines (Additional file 1). Using these datasets, we calculated transcript usage changes (ΔTU) to identify transcript switches for genes between the cytoplasm and the nucleus. For a transcript, the transcript usage (TU) means the percentage of its expression in all transcript variants, and the ΔTU assesses the extent of differential transcript usage between two conditions (i.e., nuclear and cytoplasmic fractions). Thus, a transcript switch involves two transcript variants in a gene, one of which is predominantly expressed in the cytoplasm ($\Delta TU > 0$), while the other one is mainly expressed in the nucleus ($\Delta TU < 0$). The gene VPS4A, for instance, contains four transcript variants (Figure 1A, upper). In HeLa cells, we observed that the VPS4A-201 and VPS4A-204 (also termed as a transcript switch in this study) were prone to substantial expression in the cytoplasm and the nucleus (Figure 1A, bottom), respectively. In total, 2,073 transcript switches were detected in HeLa cells (Figure 1B). Specially, using a p-value of < 0.05, we further screened for the switches that are not significantly altered. For example, in HT1080, SK-MEL-5, and SK-N-DZ, we appropriately discarded transcripts that are more likely to exhibit changes in transcript usage because of low expression levels (Additional file 2).

On average, 1,650 pairs of transcript switches were identified across thirteen cell lines (Figure 1C). The smallest of 768 pairs and the largest of 2,766 pairs were both observed in brain tissue, in SK-N-SH cells, and SK-N-DZ cells, respectively. Next, we asked if there is a group of genes shared among cell lines that have similar biological functions through the inclusion of transcript switches. We observed a total of 8,720 genes containing transcript switches from these thirteen cell lines (hereafter referred to as "switching genes"). More than 93% (8,177 of 8,720) of these switching genes were found in fewer than six cell lines, indicating that most switching genes are cell-specific (Figure 1C and Additional file 3). Gene ontology (GO) analysis of the remaining genes revealed that these shared switching genes are associated with protein and RNA binding functions (Additional file 4).

To validate the transcripts defined by ΔTU, we first compared the ΔTU between protein-coding and non-coding transcripts over cell lines. Consistently, many cytoplasmic transcripts were observed to be protein-coding transcripts with positive ΔTU values, while non-coding transcripts prefer to locate in the nucleus and have more negative ΔTU values (Figure 2A). This result agrees with our understanding that the protein-coding transcript needs to be transported into the cytoplasm to produce proteins, while a large number of non-coding transcripts have been reported to localize in the nucleus to participate in transcriptional and post-transcriptional gene expression and chromatin organization, among other functions [28, 29]. Next, we compared the ribosome density on cytoplasmic transcripts ($\Delta TU > 0$) and nuclear transcripts ($\Delta TU < 0$) in HeLa cells. Ribosome density was calculated from ribosome profiling data (see "Methods"). Comparing this data with nuclear transcripts, we predicted that cytoplasmic transcripts interact more frequently with ribosomes resulting in a higher ribosome density. Since non-coding transcripts are considered to be a group of RNA molecules that are not involved in protein translation, they are suitable for the comparison of ribosome density. Indeed, we observed that those cytoplasmic non-coding transcripts tend to associate with more ribosomes relative to those of the nucleus (Figure 2B). However, due to the limited number of non-coding transcripts available for comparison, we did not observe a significant difference.

Taken together, we applied ΔTU to screen for a pair of transcripts for each gene, and the two transcripts were enriched in the cytoplasm (termed as "cytoplasmic transcripts") and the nucleus (termed as "nuclear transcripts"), respectively. Thus, we obtained a collection of transcript variants generated by alternative splicing, and the subcellular localization was different between them as well.

A transcript switch is associated with RNA splicing rather than RNA degradation

We first investigated the post-transcriptional influence of the Nonsense-Mediated RNA Decay (NMD) pathway on the transcript switch. Initially, the two transcripts in a transcript switch are equally distributed in the cytoplasm and nucleus. One reason for the transcript switch is that the rate of degradation of the two transcripts in the cytoplasm and nucleus is different. To test this hypothesis, we calculated the sensitivity of transcripts to NMD based on changes in RNA-seq data [30] before and after the knockout of UPF1, a core factor of NMD. However, after comparing the sensitivity of NMD to cytoplasmic and nuclear transcripts, we did not find significant differences (Figure 3A). This observation led us to conclude that NMD may not significantly affect transcript switches. The cause of the transcript switch should be mainly due to intrinsic (sequence features) and transcriptional (e.g., splicing) effects.

Considering that the longer the transcript the higher the splicing frequency (or exon number)should be, we next exterminated the association between the length of the transcript and its subcellular localization. We evaluated the relationship between length and subcellular localization by calculating the length ratio (logarithmic scale) between the cytoplasmic transcript and the nuclear transcript in each transcript switch. We divided the transcript switches into three categories based on the length ratio: positive ($ratio > 1$), negative ($ratio < -1$), and neutral (other). A positive category indicates that the longer the transcript, the more enriched in the cytoplasm, and vice versa. A neutral category means that there is no significant correlation between transcript length and its subcellular localization. We observed that the number of transcript switches in the positive category is higher than that in the negative category, which implies that the longer the transcript is, the more likely it is to be transported into the cytoplasm (Figure 3B). To verify whether the splicing frequency of the transcript is positively correlated with its cytoplasmic location, we further compared the distribution of exon numbers (i.e., splicing frequency) in the cytoplasmic transcripts and nuclear transcripts for the positive and negative categories, respectively (Figure 3B, inset). We found that the exon number of the cytoplasmic transcripts in the positive category is indeed higher than the nuclear transcripts. Based on the above observations, we speculate that there are significant differences in subcellular localization between transcripts with or without splicing events. To confirm this hypothesis, we divided the transcript switches into the mono-exonic (unspliced) and multi-exonic (spliced) groups and then compared the distribution of ΔTU values between them. As expected, the ΔTU value of multi-exonic transcripts was positive (indicating cytoplasmic localization) and was significantly and consistently higher than the negative ΔTU value of mono-exonic transcripts (representing nuclear localization) in all cell lines (Figure 3C).

In brief, we first compared the NMD sensitivity between cytoplasmic transcripts and nuclear transcripts, and found that transcript switches are not caused by NMD-induced imbalanced RNA levels between the cytoplasm and nucleus. Second, by comparing the length between cytoplasmic transcripts and nuclear transcripts, we found that longer transcripts with more splicing events are more likely to be enriched in the cytoplasm. Finally, by comparing the ΔTU values between spliced transcripts and unspliced transcripts, we found that those unspliced transcripts were enriched in the nucleus.

Enrichment and characterization of retained introns in the nuclear transcripts

Next, we asked whether there was a specific kind of splicing pattern associated with subcellular localization. Seven different splicing patterns were considered [31]. Comparing to an original transcript, each type of splicing pattern inserts, deletes, or replaces a partial sequence that may include sequence elements, which have important or decisive effects on subcellular localization (e.g., protein-binding sites, RNA structures, etc.). We used the $\Delta\Psi$, which ranges from -1 to 1, to measure the preference for a type of splicing pattern between the nucleus and the cytoplasm (see "Methods" for details). The smaller the $\Delta\Psi$, the more the corresponding splicing pattern prefers the nucleus. Interestingly, in HeLa cells, we observed that retained introns were obviously biased toward the nucleus, while other splicing patterns had no significant preference (Figure 4A). Consistent results were also observed in the other twelve cell lines (Additional file 5). In conclusion, we have observed in all thirteen cell lines that transcripts with retained introns prefer to be localized in the nucleus.

To further characterize the retained introns associated with nuclear localization, we first selected a set of retained introns enriched in the nucleus ($\Delta\Psi < 0$ and $p < 0.05$, termed as nuclear RIs). Compared with all retained introns, we should be able to observe some different features of the nuclear RIs, which are likely to be the determining factors for nuclear localization. We first investigated whether the nuclear RIs have a specific splicing signal. By applying this splicing signal, it is possible to regulate the intron retention specifically, thereby monitoring the subcellular localization of the transcript. Unfortunately, we found no significant difference in the splicing signal between the nuclear RIs and all RIs (Figure 4B). Then, we considered whether the structure of the nuclear RIs (including the length and RNA secondary structure) have a specific signature. Surprisingly, the length of the nuclear RIs is significantly shorter than the overall length level (Figure 4C, upper). We also found that the average stem probability of the nuclear RIs is higher than that of all RIs (Figure 4C, bottom). Thus, we conclude that nuclear RIs maintain a more compact and stronger structure when compared with the overall RIs.

The preferences of RNA-binding proteins on retained introns suggest their role in nuclear localization

We sought to further investigate whether the nuclear RIs mentioned above are associated with RBPs. We calculated the frequency of each dinucleotide in the nuclear RIs across thirteen cell lines and normalized it with the background frequency of all intronic sequences in the genome. The reason we examined the dinucleotide composition of nuclear nucleotides was due to the reported specific contact between amino acids and dinucleotides [32]. The normalized dinucleotide frequency represents the extent to which the corresponding dinucleotide is preferred in the nuclear RIs. We found the overexpression of the GC-rich sequence occurred in the nuclear RIs (Figure 5A). This observation also provides us an intuitive hypothesis of whether RBPs that preferentially bound to RNA containing GC-rich sequences directly or indirectly affect subcellular localization.

To further explore which RBPs preferentially interact with nuclear RIs, we predicted the binding preference between an RBP and an intronic sequence using RBPmap [33], a motif-based approach. We found that across thirteen cell lines, fourteen RBPs consistently (more than 80%) preferred attaching to the nuclear RIs (Figure 5B). These include serine/arginine-rich proteins (SRSF1, SRSF2, SRSF3, SRSF5, SRSF7), heterogeneous nuclear ribonucleoproteins (HNRNPH2, HNRNPF, PTBP1), CUG-binding proteins (CUG-BP, MBNL1, BRUNOL4, BRUNOL5), and others (TARDBP and NOVA1). Conversely, binding motifs of five RBPs (SART3, PABPC1, PABPC4, RBMS3, KHDRBS1) were consistently under-represented in the nuclear RIs. Considering a previous report showing that the repeat sequence drives the nuclear localization of lncRNA [34], we subsequently analyzed the enrichment of the repeat sequence in the nuclear RIs. In NHEK cells, we observed SINE (short interspersed nuclear element; consistent with the previous study [34]), LINE (long interspersed nuclear element), and DNA transposon enrichment in nuclear RIs (Figure 5C). However, such phenomena have not been seen in other cell lines, suggesting that retained introns may not be ubiquitous in guiding nuclear localization through repeat sequences. Additionally, we observed that an LTR (long terminal repeat) was relatively rare in the nuclear RIs in all cell lines, implying that an LTR is involved in the regulation of RNA subcellular localization.

Discussion

We emphasize that ΔTU measures the change in the proportion of gene expression for a transcript between the nucleus and the cytoplasm. An increased ΔTU does not imply that the transcript abundance in the cytoplasm is higher than that in the nucleus. Instead, a significantly increased ΔTU indicates that the transcript is stored dominantly for the corresponding gene in the cytoplasm, but is expressed as a minor isoform in the nucleus. Therefore, ΔTU measures the dynamic changes of a representative (dominantly expressed) transcript of a gene between the nucleus and the cytoplasm. By using ΔTU, we defined a transcript switch, which indicates that a single gene has two different representative transcripts in the nucleus and the cytoplasm, respectively. A gene containing a transcript switch is termed as a switching gene in this study.

We hypothesize that a switching gene can function separately in the nucleus and cytoplasm (hereafter "bifunctional gene"), respectively. Previous studies have discovered several bifunctional mRNAs, which generate coding and non-coding isoforms through alternative splicing. The coding isoform translates proteins in the cytoplasm while the non-coding isoform resides in the nucleus to function as a scaffolding agent [35, 36, 37, 38, 39], post-transcriptional component [40], or coactivator [41, 42, 43]. Furthermore, a controlling feedback loop between two such isoforms can be considered. This hypothesis is consistent with a phenomenon called genetic compensation response (GCR), which has been proposed and validated in recent years [44, 45, 46, 47]. In GCR, RNA sequence fragments obtained after mRNA degradation are imported to the nucleus to target genes based on sequence similarity and regulate their expression level. We argue that GCR-like feedback may be widely present in the switching genes. The reason is that parts of exonic sequences are usually shared (implying high sequence similarities) among transcript variants in the switching genes. Indeed, we have observed 768∼2,777 switching genes (Figure 1C) in thirteen cell lines, suggesting that bifunctional genes or GCR are more widespread in the human genome than expected. A total of 8,720 switching genes (Figure 1D) were identified in this study. We believe that this analysis promises to provide a valuable resource for studying bifunctional or GCR-associated genes. Additionally, we observed that most of the switching genes were cell-specific (Figure 1D), suggesting that such genes may be related to genetic adaption to the environmental changes (stress, development, disease, etc.).

Based on the observation that multi-exonic transcripts are preferentially exported to the cytoplasm, we argue that splicing supports cytoplasmic localization. One possible reason is that the exon-exon junction complex (EJC) may promote RNA export. The EJC containing or interacting with the export factor binds to the transcript during splicing, while the export factor promotes cytoplasmic localization. Previous studies have reported that the EJC component in *Xenopus* provides binding sites [48] for export factors (REF and/or TAP/p15). The coupling between a conserved RNA export machinery and RNA splicing has also been reviewed [49].

Additionally, we observed that transcript variants containing retained introns were more likely to reside in the nucleus. By further analyzing these retained introns, we found that transcripts, including short but structurally stable retained introns, were preferentially localized in the nucleus. We speculate that such introns function as platforms for attaching to localization shuttles (e.g. RBPs). Short introns appear to be favored by natural selection because of the low metabolic cost of transcription and other molecular processing activities [49]. The RNA structure context was discussed as being able to affect protein binding [50]. Interestingly, we also found that the retained introns of nucleus-localized transcripts favored CUG-binding sites for MBNL, which has been reported to contain two nuclear signals [51]. Previous studies have reported that the CUG repeat expansion caused nuclear aggregates and RNA toxicity, which in turn induces neurological diseases [52, 53]. Taken together, this study revealed a post-transcriptional regulation mechanism, in which alternative splicing regulates RNA subcellular localization by including or excluding those introns containing nuclear transporter binding sites.

Conclusions

This study explored whether alternative splicing can regulate RNA subcellular localization. The RNA-seq data derived from the nuclear and cytoplasmic fractions of thirteen cell lines were utilized to quantify transcript abundance. We applied ΔTU to define a pair of nuclear or cytoplasmic transcripts expressed from a single gene locus. By comparing nuclear and cytoplasmic transcripts, we observed that splicing appears to promote RNA export from the nucleus. Furthermore, we used $\Delta\Psi$ to analyze the effect of splicing patterns on RNA localization and found that intron retention was positively correlated with nuclear localization. Sequence analysis of the retained introns revealed that short and structurally stable introns were favored for nuclear transcripts. We argue that such intronic sequences provide a hotspot for interacting with nuclear proteins. Subsequently, we found the MBNL, an RBP that included two nuclear signal sequences, preferentially binds to the retained introns, driving nuclear localization. We argue that cells can regulate the subcellular localization and biological functions of RNAs through alternative splicing of introns, which contain localization elements (RNA structure context or sequence motifs).

Although we used an EM-based method [54] to predict the transcript abundance between overlapping transcripts, this transcript quantification is still very challenging. One possible solution is to obtain more accurate transcript structures and their expression level through long reads (such as nanopore sequencing [55, 56]). Studying the effects of alternative splicing on RNA localization in healthy tissues or other species would be another challenging work. In summary, this research provides important clues for studying the mechanism of alternative splicing on gene expression regulation. We also believe that the switching genes defined in this study (Additional file 3) will provide a valuable resource for studying gene functions.

Methods

RNA-seq data and bioinformatics

We used the RNA-seq data from the ENCODE [57] subcellular (nuclear and cytoplasmic) fractions of 13 human cell lines (A549, GM12878, HeLa-S3, HepG2, HT1080, HUVEC, IMR-90, MCF-7, NHEK, SK-MEL-5, SK-N-DZ, SK-N-SH, and K562) to quantify the localization of the transcriptome. In brief, cell lysates were divided into the nuclear and cytoplasmic fractions by centrifugal separation, filtered for RNA with poly-A tails to include mature and processed transcript sequences, and then subjected to high-throughput RNA sequencing. H1-hESC and NCI-H460 samples without replicates were discarded. See additional file 1 for accessions of RNA-seq data.

The human genome (GRCh38) and comprehensive gene annotation were obtained from GENCODE (v29) [58]. RNA-seq reads were mapped with STAR (2.7.1a) [59] and quantified with RSEM (v1.3.1) [54] using the default parameters. Finally, we utilized SUPPA2 [60] to calculate the differential usage (change of transcript usage, ΔTU [61]) and the differential splicing (change of splicing inclusion, $\Delta\Psi$ [62]) of transcripts.

ΔTU and $\Delta\Psi$

To investigate the inconsistency of subcellular localization of transcript variants, we calculated the ΔTU from the RNA-seq data to quantify this bias. For each transcript variant, ΔTU indicates the change in the proportion of expression level in a gene, which is represented as

$$\Delta TU = T_i/G_i - T_j/G_j, \tag{1}$$

where i and j represent different subcellular fractions (i.e. cytoplasmic and nuclear, respectively). T and G are the expression abundance (transcripts per million, TPM) of the transcript and the corresponding gene, respectively. Expression abundances are estimated from RNA-seq data using the RSEM mentioned above. A high ΔTU value of a transcript indicates that it is enriched in the cytoplasm, and a low value indicates that it is enriched in the nucleus. We utilized SUPPA2 to compute ΔTU and assess the significance of differential transcript usage between the cytoplasm and the nucleus. For a gene, we selected a pair of transcript variants T_c and T_n that

$$T_c := max\{\Delta TU_i | \Delta TU_i > 0.2, P_i < 0.05\}, \tag{2}$$

$$T_n := min\{\Delta TU_i | \Delta TU_i < -0.2, P_i < 0.05\}, \tag{3}$$

where P_i is the significance level of the i-th transcript variant.

To study whether the different splicing patterns affect the subcellular localization of transcript, we used $\Delta\Psi$ to quantify this effect. Seven types of splicing patterns – alternative $3'$ splice-site (A3), alternative $5'$ splice-site (A5), alternative first exon (AF), alternative last exon (AL), mutually exclusive exon (MX), retained intron (RI), and skipping exon (SE) – were considered. These splicing patterns were well-defined in previous studies [60, 63]. For each alternative splicing site (e.g. RI),

$$\Delta\Psi = \sum TPM_{include}/(\sum TPM_{include} + \sum TPM_{exclude}), \tag{4}$$

where $TPM_{include}$ and $TPM_{exclude}$ represent the expression values of transcripts included and excluded, respectively, from the splicing site.

NMD sensitivity

NMD sensitivity was defined by the fold-change of transcript expression before and after knockdown of UPF1(a main factor in NMD). The larger the value, the higher the degradation rate by NMD. We used the RNA-seq data (GSE86148) [30] to calculate the NMD sensitivity.

Ribosome density

Ribosome density was estimated from the ribosome profiling (also termed Ribo-seq) data. For each transcript i, the Ribo-seq and RNA-seq data were used to measure the number of binding ribosomes ($Ribo_i$) and RNA abundance (RNA_i), respectively.

$$Ribosome\ density\ (i) = log(Ribo_i/RNA_i). \tag{5}$$

We obtained the ribosome densities of transcripts in HeLa cells from our previous work [64].

RBP-binding prediction and repeat sequence analysis

We applied the stand-alone version of RBPmap[33] to predict RBP binding sites. All RBPmap human/mouse stored motifs were used, and other parameters used default values. We used the repeat library (built on 20140131) that mapped to human (hg38) from Repeatmasker[65].

Abbreviations

RNA FISH, RNA fluorescent in situ hybridization; RNA-seq, RNA sequencing; RBP, RNA-binding protein; NLS, nuclear signal sequence; ΔTU, change of transcript usage; GO, Gene Ontology; NMD, nonsense-mediated decay; $\Delta\Psi$, change of splicing inclusion; RI, retained intron; SINE, Short interspersed nuclear element; LINE, Long interspersed nuclear element; LTR, Long terminal repeat; GCR, genetic compensation response; EJC, exon-exon junction complex; TPM, transcripts per million; A3, alternative $3'$ splice-site; A5, alternative $5'$ splice-site; AF, alternative first exon; AL, alternative last exon; MX, mutually exclusive exon; SE, skipping exon; Ribo-seq, ribosome profiling;

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets and materials supporting the findings of this article are included within Additional files 1–5.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (KAKENHI) [grant numbers JP17K20032, JP16H05879, JP16H01318, and JP16H02484 to MH].

Author's contributions

MH and CZ conceived this study and contributed to the analysis and interpretation of the data. CZ performed all the experiments and wrote the draft, MH revised it critically. All authors read and approved the final manuscript.

Acknowledgements

CZ and MH are grateful to Martin Frith, Yutaka Saito, and Masahiro Onoguchi for valuable discussions. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), 3-4-1, Okubo Shinjuku-ku, 169-8555 Tokyo, Japan. [2]Faculty of Science and Engineering, Waseda University, 55N-06-10, 3-4-1 Okubo Shinjuku-ku, 169-8555 Tokyo, Japan. [3]Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6 Aomi, Koto-ku, 135-0064 Tokyo, Japan. [4]Institute for Medical-oriented Structural Biology, Waseda University, 2-2, Wakamatsu-cho Shinjuku-ku, 162-8480 Tokyo, Japan. [5]Graduate School of Medicine, Nippon Medical School, 1-1-5, Sendagi, Bunkyo-ku, 113-8602 Tokyo, Japan.

References

1. Brett, D., Pospisil, H., Valcárcel, J., Reich, J., Bork, P.: Alternative splicing and genome complexity. Nature Genetics **30**(1), 29–30 (2002)
2. Pan, Q., Shai, O., Lee, L.J., Frey, B.J., Blencowe, B.J.: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics **40**(12), 1413–1415 (2008)
3. Graveley, B.R.: Alternative splicing: increasing diversity in the proteomic world. Trends in Genetics **17**(2), 100–107 (2001)
4. Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T., Soreq, H.: Function of alternative splicing. Gene **344**, 1–20 (2005)
5. Cieply, B., Carstens, R.P.: Functional roles of alternative splicing factors in human disease. Wiley Interdisciplinary Reviews: RNA **6**(3), 311–326 (2015)
6. David, C.J., Manley, J.L.: Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. Genes & Development **24**(21), 2343–2364 (2010)
7. Di, C., Zhang, Q., Chen, Y., Wang, Y., Zhang, X., Liu, Y., Sun, C., Zhang, H., Hoheisel, J.D., *et al.*: Function, clinical application, and strategies of Pre-mRNA splicing in cancer. Cell Death & Differentiation **26**(7), 1181–1194 (2019)
8. Ule, J., Blencowe, B.J.: Alternative splicing regulatory networks: Functions, mechanisms, and evolution. Molecular Cell **76**(2), 329–345 (2019)
9. Buxbaum, A.R., Haimovich, G., Singer, R.H.: In the right place at the right time: visualizing and understanding mRNA localization. Nature Reviews Molecular Cell Biology **16**(2), 95–109 (2015)
10. Fasken, M.B., Corbett, A.H.: Mechanisms of nuclear mRNA quality control. RNA Biology **6**(3), 237–241 (2009)
11. Nevo-Dinur, K., Govindarajan, S., Amster-Choder, O.: Subcellular localization of RNA and proteins in prokaryotes. Trends in Genetics **28**(7), 314–322 (2012)
12. Dermit, M., Dodel, M., Lee, F.C., Azman, M.S., Schwenzer, H., Jones, J.L., Blagden, S.P., Ule, J., Mardakheh, F.K.: Subcellular mRNA localization regulates ribosome biogenesis in migrating cells. bioRxiv, 829739 (2019)
13. Bashirullah, A., Cooperstock, R.L., Lipshitz, H.D.: RNA localization in development. Annual Review of Biochemistry **67**(1), 335–394 (1998)
14. Zhou, Y., King, M.L.: Sending RNAs into the future: RNA localization and germ cell fate. IUBMB life **56**(1), 19–27 (2004)
15. Lawrence, J.B., Singer, R.H.: Intracellular localization of messenger RNAs for cytoskeletal proteins. Cell **45**(3), 407–415 (1986)
16. Femino, A.M., Fay, F.S., Fogarty, K., Singer, R.H.: Visualization of single RNA transcripts in situ. Science **280**(5363), 585–590 (1998)
17. Raj, A., Van Den Bogaard, P., Rifkin, S.A., Van Oudenaarden, A., Tyagi, S.: Imaging individual mRNA molecules using multiple singly labeled probes. Nature Methods **5**(10), 877–879 (2008)
18. Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Yang, J.L., Ferrante, T.C., Terry, R., Jeanty, S.S.F., Li, C., Amamoto, R., Peters, D.T., Turczyk, B.M., Marblestone, A.H., Inverso, S.A., Bernard, A., Mali, P., Rios,

X., Aach, J., Church, G.M.: Highly multiplexed subcellular RNA sequencing in situ. Science **343**(6177), 1360–1363 (2014)

19. Taliaferro, J.M., Wang, E.T., Burge, C.B.: Genomic analysis of RNA localization. RNA Biology **11**(8), 1040–1050 (2014)
20. Fazal, F.M., Han, S., Parker, K.R., Kaewsapsak, P., Xu, J., Boettiger, A.N., Chang, H.Y., Ting, A.Y.: Atlas of subcellular RNA localization revealed by APEX-seq. Cell **178**(2), 473–490 (2019)
21. Padron, A., Iwasaki, S., Ingolia, N.T.: Proximity RNA labeling by APEX-seq reveals the organization of translation initiation complexes and repressive RNA granules. Molecular Cell **75**(4), 875–887 (2019)
22. Zhang, T., Tan, P., Wang, L., Jin, N., Li, Y., Zhang, L., Yang, H., Hu, Z., Zhang, L., Hu, C., Li, C., Qian, K., Zhang, C., Huang, Y., Li, K., Lin, H., Wang, D.: RNALocate: a resource for RNA subcellular localizations. Nucleic Acids Research **45**(D1), 135–138 (2016)
23. Mas-Ponte, D., Carlevaro-Fita, J., Palumbo, E., Pulido, T.H., Guigo, R., Johnson, R.: LncATLAS database for subcellular localization of long noncoding RNAs. RNA **23**(7), 1080–1087 (2017)
24. Zhou, W., Liu, Z., Wu, J., Liu, J.-h., Hyder, S.M., Antoniou, E., Lubahn, D.B.: Identification and characterization of two novel splicing isoforms of human estrogen-related receptor β. The Journal of Clinical Endocrinology & Metabolism **91**(2), 569–579 (2006)
25. Hakre, S., Tussie-Luna, M.I., Ashworth, T., Novina, C.D., Settleman, J., Sharp, P.A., Roy, A.L.: Opposing functions of TFII-I spliced isoforms in growth factor-induced gene expression. Molecular Cell **24**(2), 301–308 (2006)
26. Wiesenthal, A., Hoffmeister, M., Siddique, M., Kovacevic, I., Oess, S., Müller-Esterl, W., Siehoff-Icking, A.: NOSTRINβ–A shortened NOSTRIN variant with a role in transcriptional regulation. Traffic **10**(1), 26–34 (2009)
27. Bombail, V., Collins, F., Brown, P., Saunders, P.T.: Modulation of ERα transcriptional activity by the orphan nuclear receptor ERRβ and evidence for differential effects of long-and short-form splice variants. Molecular and cellular endocrinology **314**(1), 53–61 (2010)
28. Quinn, J.J., Chang, H.Y.: Unique features of long non-coding RNA biogenesis and function. Nature Reviews Genetics **17**(1), 47 (2016)
29. Marchese, F.P., Raimondi, I., Huarte, M.: The multidimensional mechanisms of long noncoding RNA function. Genome Biology **18**(1), 206 (2017)
30. Colombo, M., Karousis, E.D., Bourquin, J., Bruggmann, R., Mühlemann, O.: Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6- and SMG7-mediated degradation pathways. RNA **23**(2), 189–201 (2017)
31. Alamancos, G.P., Pagès, A., Trincado, J.L., Bellora, N., Eyras, E.: Leveraging transcript quantification for fast computation of alternative splicing profiles. RNA **21**(9), 1521–1531 (2015)
32. Fernandez, M., Kumagai, Y., Standley, D.M., Sarai, A., Mizuguchi, K., Ahmad, S.: Prediction of dinucleotide-specific rna-binding sites in proteins. BMC Bioinformatics **12**(13), 5 (2011)
33. Paz, I., Kosti, I., Ares Jr, M., Cline, M., Mandel-Gutfreund, Y.: RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Research **42**(W1), 361–367 (2014)
34. Lubelsky, Y., Ulitsky, I.: Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. Nature **555**(7694), 107–111 (2018)
35. Kloc, M., Wilk, K., Vargas, D., Shirato, Y., Bilinski, S., Etkin, L.D.: Potential structural role of non-coding and coding RNAs in the organization of the cytoskeleton at the vegetal cortex of Xenopus oocytes. Development **132**(15), 3445–3457 (2005)
36. Jenny, A., Hachet, O., Závorszky, P., Cyrklaff, A., Weston, M.D., St Johnston, D., Erdélyi, M., Ephrussi, A.: A translation-independent role of oskar RNA in early Drosophila oogenesis. Development **133**(15), 2827–2833 (2006)
37. Lim, S., Kumari, P., Gilligan, P., Quach, H.N.B., Mathavan, S., Sampath, K.: Dorsal activity of maternal squint is mediated by a non-coding function of the RNA. Development **139**(16), 2903–2915 (2012)
38. Shevtsov, S.P., Dundr, M.: Nucleation of nuclear bodies by RNA. Nature Cell Biology **13**(2), 167–173 (2011)
39. Jansen, G., Groenen, P.J., Bächner, D., Jap, P.H., Coerwinkel, M., Oerlemans, F., van den Broek, W., Gohlsch, B., Pette, D., Plomp, J.J., Molenaar, P.C., Nederhoff, M.G., van Echteld, C.J., Dekker, M., Berns, A.,

Hameister, H., Wieringa, B.: Abnormal myotonic dystrophy protein kinase levels produce only mild myopathy in mice. Nature Genetics **13**(3), 316 (1996)

40. Gong, C., Maquat, L.E.: lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature **470**(7333), 284–288 (2011)
41. Chooniedass-Kothari, S., Emberley, E., Hamedani, M., Troup, S., Wang, X., Czosnek, A., Hube, F., Mutawe, M., Watson, P., Leygue, E.: The steroid receptor RNA activator is the first functional RNA encoding a protein. FEBS Letters **566**(1-3), 43–47 (2004)
42. Hubé, F., Velasco, G., Rollin, J., Furling, D., Francastel, C.: Steroid receptor RNA activator protein binds to and counteracts SRA RNA-mediated activation of MyoD and muscle differentiation. Nucleic Acids Research **39**(2), 513–525 (2010)
43. Lanz, R.B., McKenna, N.J., Onate, S.A., Albrecht, U., Wong, J., Tsai, S.Y., Tsai, M.-J., O'Malley, B.W.: A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. Cell **97**(1), 17–27 (1999)
44. Rossi, A., Kontarakis, Z., Gerri, C., Nolte, H., Hölper, S., Krüger, M., Stainier, D.Y.: Genetic compensation induced by deleterious mutations but not gene knockdowns. Nature **524**(7564), 230 (2015)
45. El-Brolosy, M.A., Stainier, D.Y.: Genetic compensation: A phenomenon in search of mechanisms. PLoS Genetics **13**(7), 1006780 (2017)
46. El-Brolosy, M.A., Kontarakis, Z., Rossi, A., Kuenne, C., Günther, S., Fukuda, N., Kikhi, K., Boezio, G.L.M., Takacs, C.M., Lai, S.-L., Fukuda, R., Gerri, C., Giraldez, A.J., Stainier, D.Y.R.: Genetic compensation triggered by mutant mRNA degradation. Nature **568**(7751), 193–197 (2019)
47. Ma, Z., Zhu, P., Shi, H., Guo, L., Zhang, Q., Chen, Y., Chen, S., Zhang, Z., Peng, J., Chen, J.: PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. Nature **568**(7751), 259–263 (2019)
48. Le Hir, H., Gatfield, D., Izaurralde, E., Moore, M.J.: The exon–exon junction complex provides a binding platform for factors involved in mrna export and nonsense-mediated mrna decay. The EMBO Journal **20**(17), 4987–4997 (2001)
49. Reed, R., Hurt, E.: A conserved mRNA export machinery coupled to pre-mRNA splicing. Cell **108**(4), 523–531 (2002)
50. Wan, Y., Kertesz, M., Spitale, R.C., Segal, E., Chang, H.Y.: Understanding the transcriptome through RNA structure. Nature Reviews Genetics **12**(9), 641–655 (2011)
51. Taylor, K., Sznajder, Ł.J., Cywoniuk, P., Thomas, J.D., Swanson, M.S., Sobczak, K.: MBNL splicing activity depends on RNA binding site structural context. Nucleic Acids Research **46**(17), 9119–9133 (2018)
52. Goodwin, M., Mohan, A., Batra, R., Lee, K.-Y., Charizanis, K., Gómez, F.J.F., Eddarkaoui, S., Sergeant, N., Buée, L., Kimura, T., *et al.*: MBNL sequestration by toxic RNAs and RNA misprocessing in the myotonic dystrophy brain. Cell Reports **12**(7), 1159–1168 (2015)
53. Wang, P.-Y., Chang, K.-T., Lin, Y.-M., Kuo, T.-Y., Wang, G.-S.: Ubiquitination of mbnl1 is required for its cytoplasmic localization and function in promoting neurite outgrowth. Cell Reports **22**(9), 2294–2306 (2018)
54. Li, B., Dewey, C.N.: RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics **12**, 323 (2011)
55. Soneson, C., Yao, Y., Bratus-Neuenschwander, A., Patrignani, A., Robinson, M.D., Hussain, S.: A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. Nature Communications **10**, 3359 (2019)
56. Workman, R.E., Tang, A.D., Tang, P.S., Jain, M., Tyson, J.R., Razaghi, R., Zuzarte, P.C., Gilpatrick, T., Payne, A., Quick, J., Sadowski, N., Holmes, N., de Jesus, J.G., Jones, K.L., Soulette, C.M., Snutch, T.P., Loman, N., Paten, B., Loose, M., Simpson, J.T., Olsen, H.E., Brooks, A.N., Akeson, M., Timp, W.: Nanopore native RNA sequencing of a human poly(a) transcriptome. Nature Methods **16**, 1297–1305 (2019)
57. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H.,

Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., Gingeras, T.R.: Landscape of transcription in human cells. Nature **489**(7414), 101–108 (2012)

58. Frankish, A., Diekhans, M., Ferreira, A.M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I.T., García Girón, C., Gonzalez, J.M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O.G., Lagarde, J., Martin, F.J., Martínez, L., Mohanan, S., Muir, P., Navarro, F.C.P., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B.M., Stapleton, E., Suner, M.M., Sycheva, I., Uszczynska-Ratajczak, B., Xu, J., Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J.S., Gerstein, M., Guigó, R., Hubbard, T.J.P., Kellis, M., Paten, B., Reymond, A., Tress, M.L., Flicek, P.: GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Research **47**(D1), 766–773 (2019)
59. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R.: STAR: Ultrafast universal RNA-seq aligner. Bioinformatics **29**(1), 15–21 (2013)
60. Trincado, J.L., Entizne, J.C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D.J., Eyras, E.: SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biology **19**(1), 1–11 (2018)
61. Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., Pachter, L.: Differential analysis of gene regulation at transcript resolution with rna-seq. Nature biotechnology **31**(1), 46 (2013)
62. Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., Burge, C.B.: Alternative isoform regulation in human tissue transcriptomes. Nature **456**(7221), 470 (2008)
63. Shen, S., Park, J.W., Lu, Z.-x., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., Xing, Y.: rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences of the United States of America **111**(51), 5593–601 (2014)
64. Zeng, C., Fukunaga, T., Hamada, M.: Identification and analysis of ribosome-associated lncRNAs using ribosome profiling data. BMC Genomics **19**, 414 (2018)
65. Smit, A., Hubley, R., Green, P.: 2013–2015. RepeatMasker Open-4.0 (2013). `http://www.repeatmasker.org` Accessed 1 May 2019
66. Mi, H., Muruganujan, A., Ebert, D., Huang, X., Thomas, P.D.: PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research **47**(D1), 419–426 (2018)

Figures

Figure 1 Transcript switches between cytoplasm and nucleus. (A) The upper shows 4 transcript variants of the gene VPS4A. The bottom indicates an example for Identifying a transcript variant switch in the gene VPS4A by using ΔTU. VPS4A-201 and VPS41-204 are observed to be localized dominantly in the cytoplasm and the nucleus, respectively, in HeLa-S3 cells. (B) Genome-wide identification of transcript variant switches in HeLa-S3 cells. We applied $|\Delta TU| > 0.25$ and $p < 0.05$ to filter out cytoplasmic ($|\Delta TU| > 0$, blue) and nuclear ($|\Delta TU| < 0$, red) transcript variants. Unknown (black) are transcripts filtered as no significant. (C) Number of transcript variant switches across 13 cell lines. (D) Number of genes containing transcript variant switches shared across 13 cell lines. For example, 4015 genes are identified as cell specific.

Figure 2 Validation of the $\boldsymbol{\Delta TU}$ metric. (A) Comparison of ΔTU between protein-coding transcripts (red) and non-coding transcripts (black). Protein-coding transcripts possess higher ΔTU scores, which is consistent with our understanding that the transcript encoding protein is more prone to be located in the cytoplasm. (B) Comparison of ΔTU and ribosome density in HeLa cells. For non-coding transcripts (solid lines), the ΔTU shows a weak positive correlation with the ribosome density. However, no significant difference was observed in the protein-coding transcripts (dotted lines). ΔTU^{+} and ΔTU^{-} represent transcripts with $\Delta TU > 0$ (cytoplasmic) or $\Delta TU < 0$ (nuclear), respectively.

Figure 3 Characterization of cytoplasmic and nuclear transcripts. (A) NMD does not explain transcript variant switches between the cytoplasm and the nucleus (ns indicates not significant). (B) Comparison of transcript length between cytoplasmic (ΔTU^{+}) and nuclear (ΔTU^{-}) transcripts. For each pair of transcript switches, the ratio of transcript length between the cytoplasm and the nucleus is defined as a metric. $Ratio > 1$ indicates a group of transcripts where the length of the cytoplasmic transcripts is greater than that of the nuclear transcripts. In the group, splicing frequency (or exon number) is over-expressed in the cytoplasmic transcripts (inset, upper), suggesting splicing can promote cytoplasmic localization. While no significant difference in splicing frequency was observed in the group if the transcripts $Ratio < -1$ (inset, bottom). (C) Comparison of ΔTU between mono-exonic (black) and multi-exonic (red) transcripts across 13 cell lines. ΔTU shows a significant positive correlation with splicing, indicating that splicing appears to be a dominant factor for RNA export from the nucleus.

Figure 4 Comparison of alternative splicing patterns between cytoplasmic and nuclear transcripts. (A) Retained introns are enriched in the nuclear transcripts (HeLa-S3). $\Delta\Psi$ indicates the inclusion level for a specific alternative splicing pattern. A3: alternative $3'$ splice-site; A5: alternative $5'$ splice-site; AF: alternative first exon; AL: alternative last exon; MX: mutually exclusive exon; RI: retained intron; SE: skipping exon. Comparison of (B) splice sites, (C) Length (top) and RNA secondary structure (bottom) between all introns and nuclear RIs ($\Delta\Psi < 0$ and $p < 0.05$). Sequence logos show intron-exon (the left) and exon-intron (the right) splice boundaries.

Figure 5 Preference of dinucleotide, RBPs, and TEs on nuclear RIs. The heatmap represents the relative density of a specific (A) dinucleotide, (B) RBP-binding site, or (C) TE (row) on nuclear RIs across multiple cell lines (columns) compared with that of controlled introns (Ctr.).

Additional Files

Additional file 1 — RNA-seq accession numbers for subcellular localization.

Additional file 2 — Genome-wide identification of transcript switches in twelve other cell lines.

Additional file 3 — List of switching genes identified in this study.

Additional file 4 — GO analysis of switching genes (shared $\geq$ 7 cell lines) with PANTHER[66].

Additional file 5 — Alternative splicing analyses in twelve other cell lines.

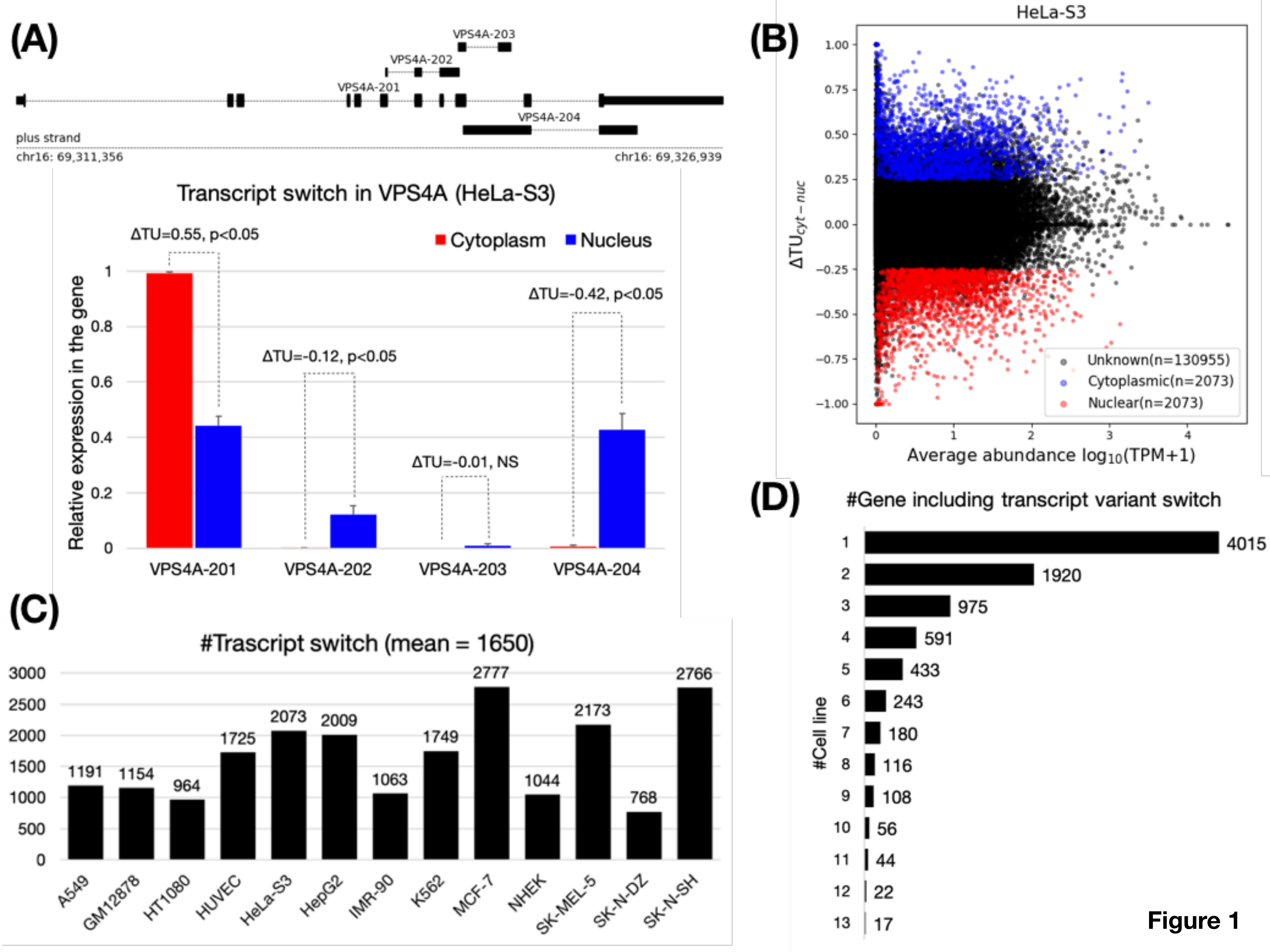

(A)
VPS4A-203
VPS4A-202
VPS4A-201
VPS4A-204
plus strand
chr16: 69,311,356
chr16: 69,326,939
Transcript switch in VPS4A (HeLa-S3)
Cytoplasm
Nucleus
ΔTU=0.55, p<0.05
ΔTU=-0.12, p<0.05
ΔTU=-0.01, NS
ΔTU=-0.42, p<0.05
Relative expression in the gene
0
0.2
0.4
0.6
0.8
1
VPS4A-201
VPS4A-202
VPS4A-203
VPS4A-204
(B)
HeLa-S3
ΔTU$_{cyt-nuc}$
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
1
2
3
4
Average abundance log$_{10}$(TPM+1)
Unknown(n=130955)
Cytoplasmic(n=2073)
Nuclear(n=2073)
(C)
#Trascript switch (mean = 1650)
3000
2500
2000
1500
1000
500
0
1191
1154
964
1725
2073
2009
1063
1749
2777
1044
2173
768
2766
A549
GM12878
HT1080
HUVEC
HeLa-S3
HepG2
IMR-90
K562
MCF-7
NHEK
SK-MEL-5
SK-N-DZ
SK-N-SH
(D)
#Gene including transcript variant switch
#Cell line
1
2
3
4
5
6
7
8
9
10
11
12
13
4015
1920
975
591
433
243
180
116
108
56
44
22
17

Figure 1

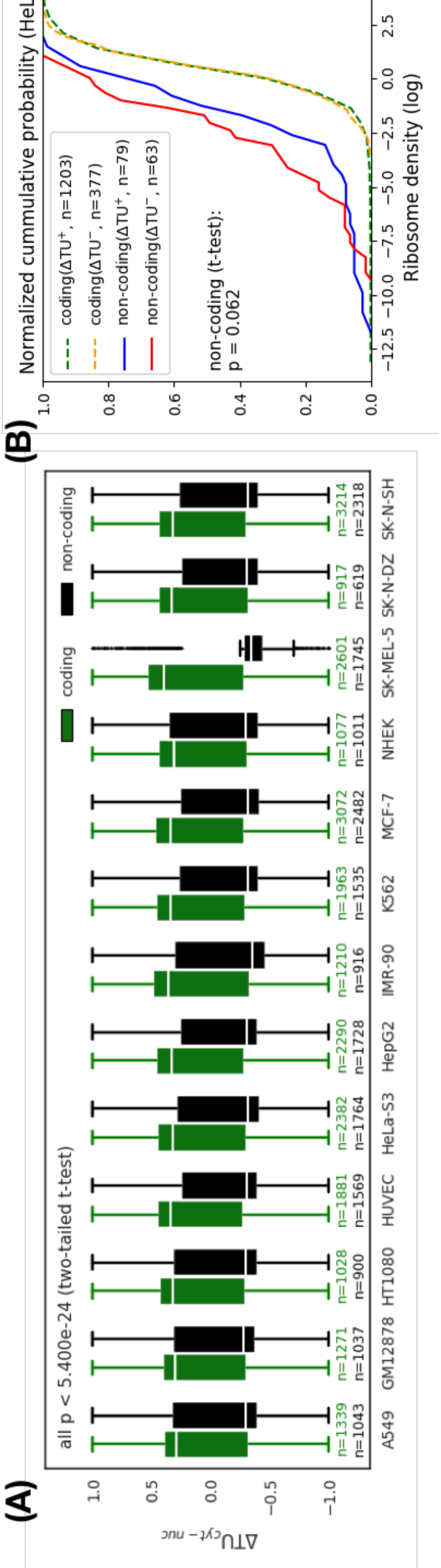

(A)
$\Delta TU_{cyt-nuc}$
1.0
0.5
0.0
−0.5
−1.0
all p < 5.400e-24 (two-tailed t-test)
🟩 coding
■ non-coding
n=1339
n=1043
A549
n=1271
n=1037
GM12878
n=1028
n=900
HT1080
n=1881
n=1569
HUVEC
n=2382
n=1764
HeLa-S3
n=2290
n=1728
HepG2
n=1210
n=916
IMR-90
n=1963
n=1535
K562
n=3072
n=2482
MCF-7
n=1077
n=1011
NHEK
n=2601
n=1745
SK-MEL-5
n=917
n=619
SK-N-DZ
n=3214
n=2318
SK-N-SH
(B)
Normalized cummulative probability (HeLa)
1.0
0.8
0.6
0.4
0.2
0.0
coding(ΔTU⁺, n=1203)
coding(ΔTU⁻, n=377)
non-coding(ΔTU⁺, n=79)
non-coding(ΔTU⁻, n=63)
non-coding (t-test):
p = 0.062
−12.5
−10.0
−7.5
−5.0
−2.5
0.0
2.5
Ribosome density (log)

Figure 2

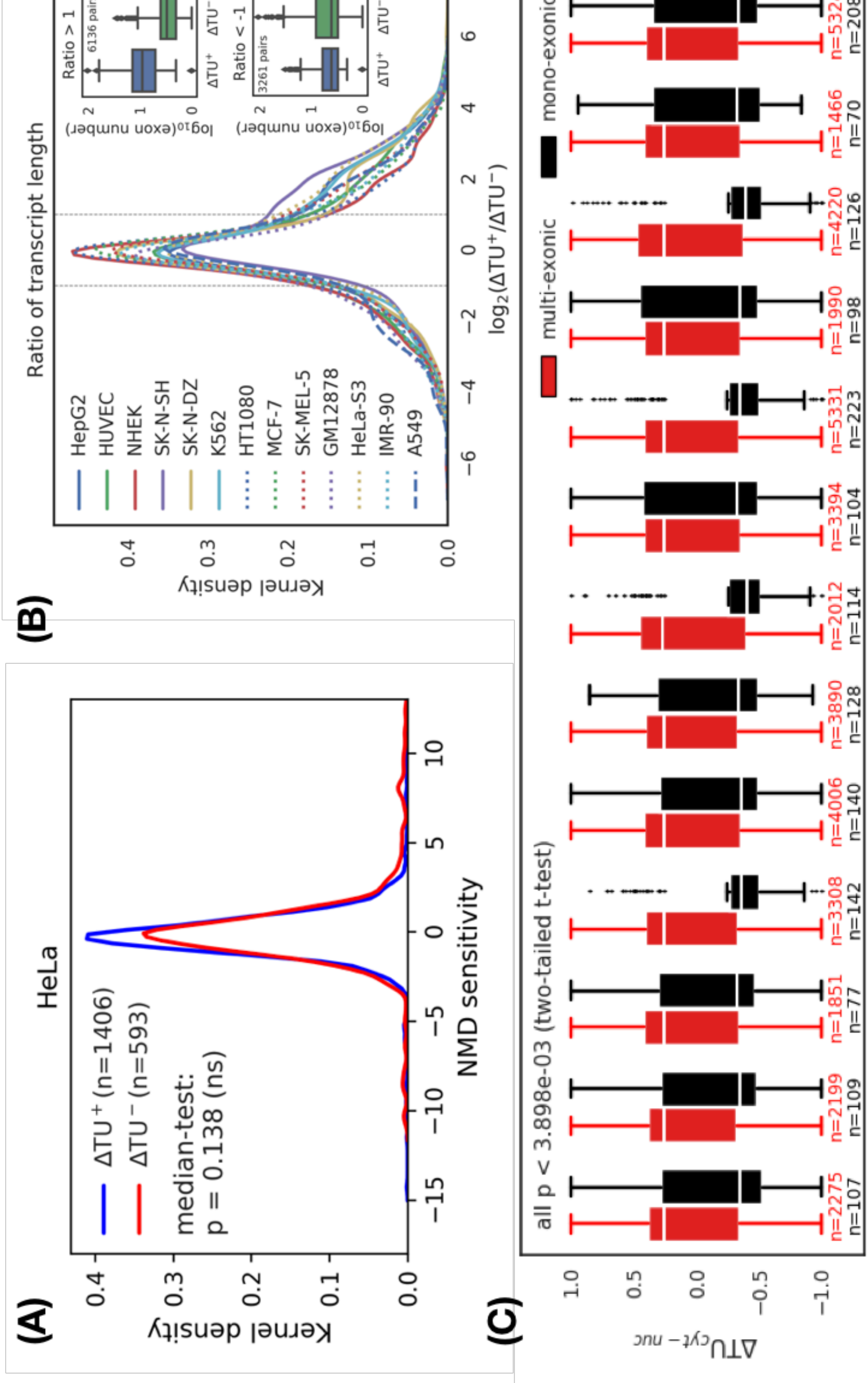
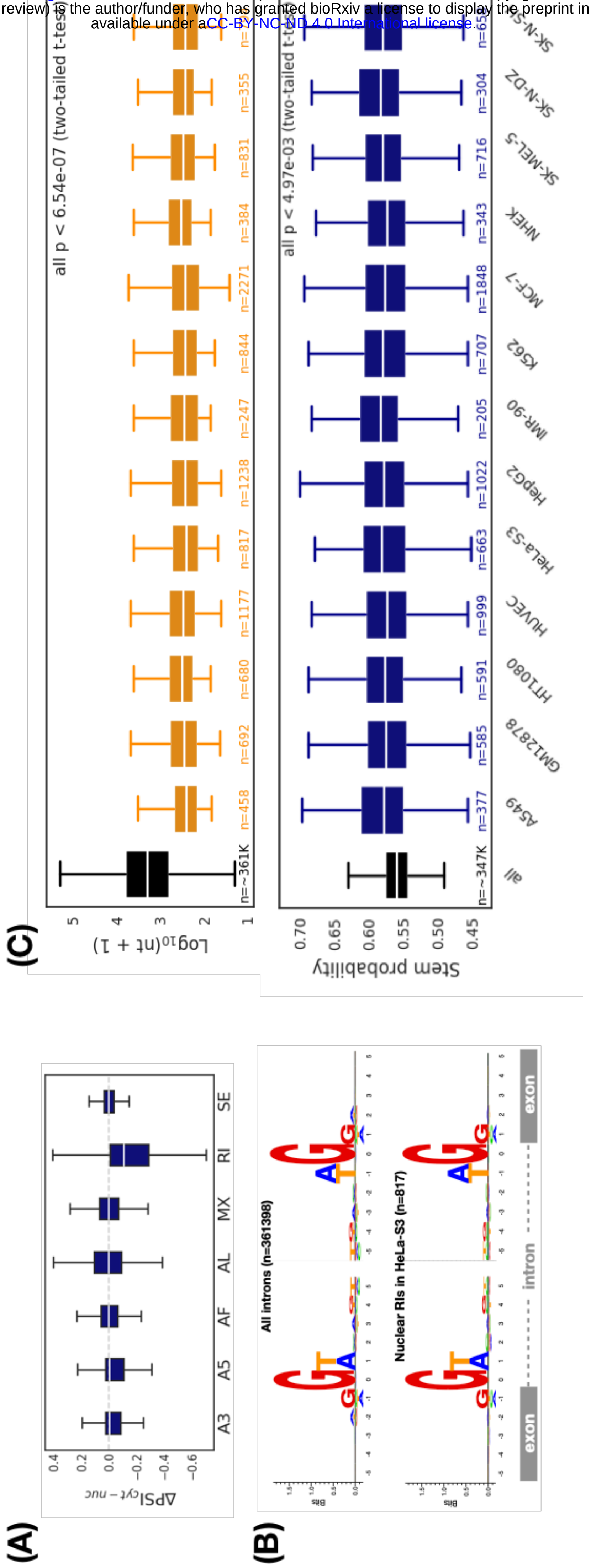
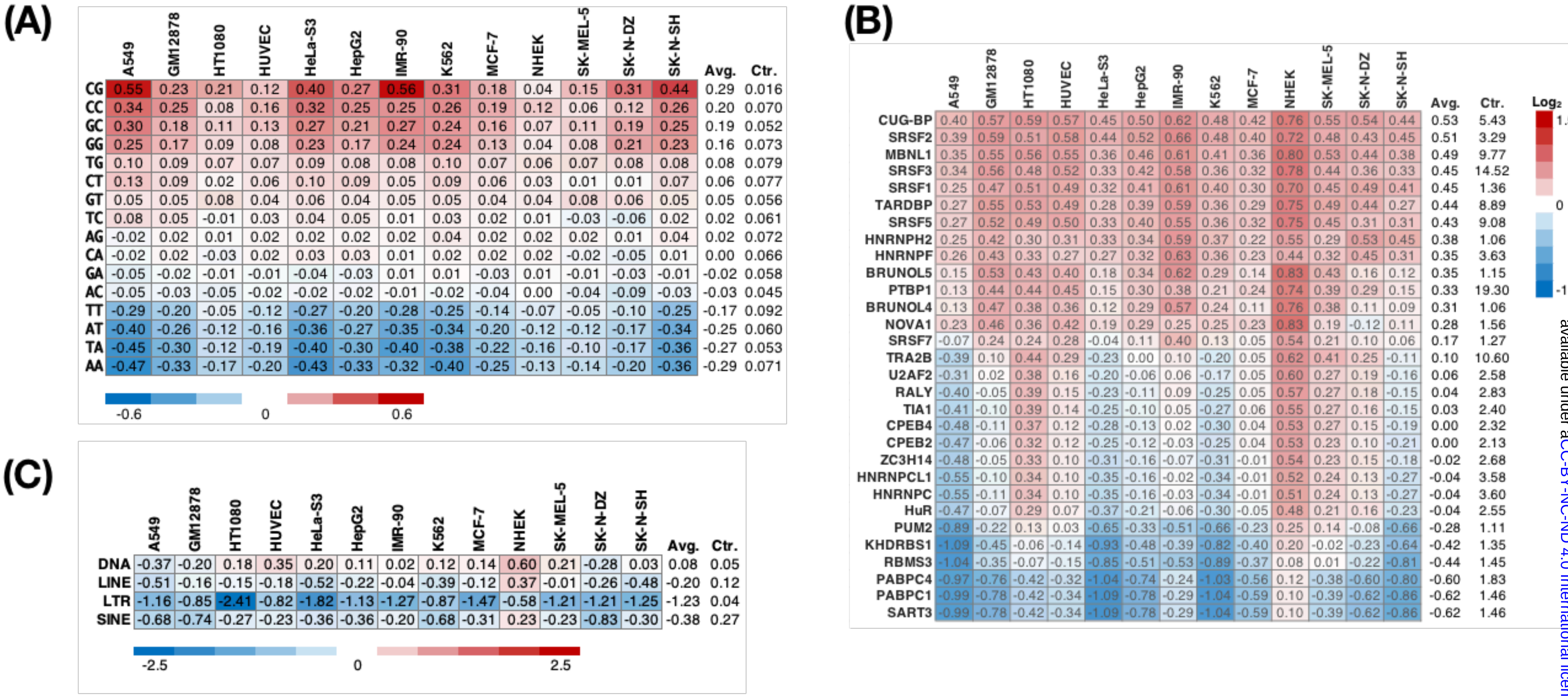

(A)
HeLa
ΔTU⁺ (n=1406)
ΔTU⁻ (n=593)
median-test:
p = 0.138 (ns)
Kernel density
0.0
0.1
0.2
0.3
0.4
−15
−10
−5
0
5
10
NMD sensitivity
(B)
Ratio of transcript length
HepG2
HUVEC
NHEK
SK-N-SH
SK-N-DZ
K562
HT1080
MCF-7
SK-MEL-5
GM12878
HeLa-S3
IMR-90
A549
Kernel density
0.0
0.1
0.2
0.3
0.4
−6
−4
−2
0
2
4
6
log₂(ΔTU⁺/ΔTU⁻)
Ratio > 1
6136 pairs
Ratio < -1
3261 pairs
log₁₀(exon number)
0
1
2
ΔTU⁺
ΔTU⁻
(C)
all p < 3.898e-03 (two-tailed t-test)
multi-exonic
mono-exonic
ΔTU cyt − nuc
1.0
0.5
0.0
−0.5
−1.0
n=2275
n=107
A549
n=2199
n=109
GM12878
n=1851
n=77
HT1080
n=3308
n=142
HUVEC
n=4006
n=140
HeLa-S3
n=3890
n=128
HepG2
n=2012
n=114
IMR-90
n=3394
n=104
K562
n=5331
n=223
MCF-7
n=1990
n=98
NHEK
n=4220
n=126
SK-MEL-5
n=1466
n=70
SK-N-DZ
n=5324
n=208
SK-N-SH

Figure 3

Figure 4


(A)
ΔPSI$_{cyt-nuc}$
0.4
0.2
0.0
−0.2
−0.4
−0.6
A3
A5
AF
AL
MX
RI
SE
(B)
All introns (n=361398)
Nuclear RIs in HeLa-S3 (n=817)
Bits
exon
intron
exon
(C)
all p < 6.54e-07 (two-tailed t-test)
Log$_{10}$(nt + 1)
5
4
3
2
1
n=~361K
n=458
n=692
n=680
n=1177
n=817
n=1238
n=247
n=844
n=2271
n=384
n=831
n=355
n=65
all p < 4.97e-03 (two-tailed t-test)
Stem probability
0.70
0.65
0.60
0.55
0.50
0.45
n=~347K
n=377
n=585
n=591
n=999
n=663
n=1022
n=205
n=707
n=1848
n=343
n=716
n=304
n=65
all
A549
GM12878
HT1080
HUVEC
HeLa-S3
HepG2
IMR-90
K562
MCF-7
NHEK
SK-MEL-5
SK-N-DZ
SK-N-SH



(A)

	A549	GM12878	HT1080	HUVEC	HeLa-S3	HepG2	IMR-90	K562	MCF-7	NHEK	SK-MEL-5	SK-N-DZ	SK-N-SH	Avg.	Ctr.
CG	0.55	0.23	0.21	0.12	0.40	0.27	0.56	0.31	0.18	0.04	0.15	0.31	0.44	0.29	0.016
CC	0.34	0.25	0.08	0.16	0.32	0.25	0.25	0.26	0.19	0.12	0.06	0.12	0.26	0.20	0.070
GC	0.30	0.18	0.11	0.13	0.27	0.21	0.27	0.24	0.16	0.07	0.11	0.19	0.25	0.19	0.052
GG	0.25	0.17	0.09	0.08	0.23	0.17	0.24	0.24	0.13	0.04	0.08	0.21	0.23	0.16	0.073
TG	0.10	0.09	0.07	0.07	0.09	0.08	0.08	0.10	0.07	0.06	0.07	0.08	0.08	0.08	0.079
CT	0.13	0.09	0.02	0.06	0.10	0.09	0.05	0.09	0.06	0.03	0.01	0.01	0.07	0.06	0.077
GT	0.05	0.05	0.08	0.04	0.06	0.04	0.05	0.05	0.04	0.04	0.08	0.06	0.05	0.05	0.056
TC	0.08	0.05	-0.01	0.03	0.04	0.05	0.01	0.03	0.02	0.01	-0.03	-0.06	0.02	0.02	0.061
AG	-0.02	0.02	0.01	0.02	0.02	0.02	0.02	0.04	0.02	0.02	0.02	0.01	0.04	0.02	0.072
CA	-0.02	0.02	-0.03	0.02	0.03	0.03	0.01	0.02	0.02	0.02	-0.02	-0.05	0.01	0.00	0.066
GA	-0.05	-0.02	-0.01	-0.01	-0.04	-0.03	0.01	0.01	-0.03	0.01	-0.01	-0.03	-0.01	-0.02	0.058
AC	-0.05	-0.03	-0.05	-0.02	-0.02	-0.02	-0.01	-0.02	-0.04	0.00	-0.04	-0.09	-0.03	-0.03	0.045
TT	-0.29	-0.20	-0.05	-0.12	-0.27	-0.20	-0.28	-0.25	-0.14	-0.07	-0.05	-0.10	-0.25	-0.17	0.092
AT	-0.40	-0.26	-0.12	-0.16	-0.36	-0.27	-0.35	-0.34	-0.20	-0.12	-0.12	-0.17	-0.34	-0.25	0.060
TA	-0.45	-0.30	-0.12	-0.19	-0.40	-0.30	-0.40	-0.38	-0.22	-0.16	-0.10	-0.17	-0.36	-0.27	0.053
AA	-0.47	-0.33	-0.17	-0.20	-0.43	-0.33	-0.32	-0.40	-0.25	-0.13	-0.14	-0.20	-0.36	-0.29	0.071

(B)

	A549	GM12878	HT1080	HUVEC	HeLa-S3	HepG2	IMR-90	K562	MCF-7	NHEK	SK-MEL-5	SK-N-DZ	SK-N-SH	Avg.	Ctr.
CUG-BP	0.40	0.57	0.59	0.57	0.45	0.50	0.62	0.48	0.42	0.76	0.55	0.54	0.44	0.53	5.43
SRSF2	0.39	0.59	0.51	0.58	0.44	0.52	0.66	0.48	0.40	0.72	0.48	0.43	0.45	0.51	3.29
MBNL1	0.35	0.55	0.56	0.55	0.36	0.46	0.61	0.41	0.36	0.80	0.53	0.44	0.38	0.49	9.77
SRSF3	0.34	0.56	0.48	0.52	0.33	0.42	0.58	0.36	0.32	0.78	0.44	0.36	0.33	0.45	14.52
SRSF1	0.25	0.47	0.51	0.49	0.32	0.41	0.61	0.40	0.30	0.70	0.45	0.49	0.41	0.45	1.36
TARDBP	0.27	0.55	0.53	0.49	0.28	0.39	0.59	0.36	0.29	0.75	0.49	0.44	0.27	0.44	8.89
SRSF5	0.27	0.52	0.49	0.50	0.33	0.40	0.55	0.36	0.32	0.75	0.45	0.31	0.31	0.43	9.08
HNRNPH2	0.25	0.42	0.30	0.31	0.33	0.34	0.59	0.37	0.22	0.55	0.29	0.53	0.45	0.38	1.06
HNRNPF	0.26	0.43	0.33	0.27	0.27	0.32	0.63	0.36	0.23	0.44	0.32	0.45	0.31	0.35	3.63
BRUNOL5	0.15	0.53	0.43	0.40	0.18	0.34	0.62	0.29	0.14	0.83	0.43	0.16	0.12	0.35	1.15
PTBP1	0.13	0.44	0.44	0.45	0.15	0.30	0.38	0.21	0.24	0.74	0.39	0.29	0.15	0.33	19.30
BRUNOL4	0.13	0.47	0.38	0.36	0.12	0.29	0.57	0.24	0.11	0.76	0.38	0.11	0.09	0.31	1.06
NOVA1	0.23	0.46	0.36	0.42	0.19	0.29	0.25	0.25	0.23	0.83	0.19	-0.12	0.11	0.28	1.56
SRSF7	-0.07	0.24	0.24	0.28	-0.04	0.11	0.40	0.13	0.05	0.54	0.21	0.10	0.06	0.17	1.27
TRA2B	-0.39	0.10	0.44	0.29	-0.23	0.00	0.10	-0.20	0.05	0.62	0.41	0.25	-0.11	0.10	10.60
U2AF2	-0.31	0.02	0.38	0.16	-0.20	-0.06	0.06	-0.17	0.05	0.60	0.27	0.19	-0.16	0.06	2.58
RALY	-0.40	-0.05	0.39	0.15	-0.23	-0.11	0.09	-0.25	0.05	0.57	0.27	0.18	-0.15	0.04	2.83
TIA1	-0.41	-0.10	0.39	0.14	-0.25	-0.10	0.05	-0.27	0.06	0.55	0.27	0.16	-0.15	0.03	2.40
CPEB4	-0.48	-0.11	0.37	0.12	-0.28	-0.13	0.02	-0.30	0.04	0.53	0.27	0.15	-0.19	0.00	2.32
CPEB2	-0.47	-0.06	0.32	0.12	-0.25	-0.12	-0.03	-0.25	0.04	0.53	0.23	0.10	-0.21	0.00	2.13
ZC3H14	-0.48	-0.05	0.33	0.10	-0.31	-0.16	-0.07	-0.31	-0.01	0.54	0.23	0.15	-0.18	-0.02	2.68
HNRNPCL1	-0.55	-0.10	0.34	0.10	-0.35	-0.16	-0.02	-0.34	-0.01	0.52	0.24	0.13	-0.27	-0.04	3.58
HNRNPC	-0.55	-0.11	0.34	0.10	-0.35	-0.16	-0.03	-0.34	-0.01	0.51	0.24	0.13	-0.27	-0.04	3.60
HuR	-0.47	-0.07	0.29	0.07	-0.37	-0.21	-0.06	-0.30	-0.05	0.48	0.21	0.16	-0.23	-0.04	2.55
PUM2	-0.89	-0.22	0.13	0.03	-0.65	-0.33	-0.51	-0.66	-0.23	0.25	0.14	-0.08	-0.66	-0.28	1.11
KHDRBS1	-1.09	-0.45	-0.06	-0.14	-0.93	-0.48	-0.39	-0.82	-0.40	0.20	-0.02	-0.23	-0.64	-0.42	1.35
RBMS3	-1.04	-0.35	-0.07	-0.15	-0.85	-0.51	-0.53	-0.89	-0.37	0.08	0.01	-0.22	-0.81	-0.44	1.45
PABPC4	-0.97	-0.76	-0.42	-0.32	-1.04	-0.74	-0.24	-1.03	-0.56	0.12	-0.38	-0.60	-0.80	-0.60	1.83
PABPC1	-0.99	-0.78	-0.42	-0.34	-1.09	-0.78	-0.29	-1.04	-0.59	0.10	-0.39	-0.62	-0.86	-0.62	1.46
SART3	-0.99	-0.78	-0.42	-0.34	-1.09	-0.78	-0.29	-1.04	-0.59	0.10	-0.39	-0.62	-0.86	-0.62	1.46

(C)

	A549	GM12878	HT1080	HUVEC	HeLa-S3	HepG2	IMR-90	K562	MCF-7	NHEK	SK-MEL-5	SK-N-DZ	SK-N-SH	Avg.	Ctr.
DNA	-0.37	-0.20	0.18	0.35	0.20	0.11	0.02	0.12	0.14	0.60	0.21	-0.28	0.03	0.08	0.05
LINE	-0.51	-0.16	-0.15	-0.18	-0.52	-0.22	-0.04	-0.39	-0.12	0.37	-0.01	-0.26	-0.48	-0.20	0.12
LTR	-1.16	-0.85	-2.41	-0.82	-1.82	-1.13	-1.27	-0.87	-1.47	-0.58	-1.21	-1.21	-1.25	-1.23	0.04
SINE	-0.68	-0.74	-0.27	-0.23	-0.36	-0.36	-0.20	-0.68	-0.31	0.23	-0.23	-0.83	-0.30	-0.38	0.27

Figure 5