RNA-Seq Analysis Reveals Localization-Associated Alternative Splicing across 13 Cell Lines

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

[1]AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), Tokyo, 169-8555, Japan
[2]Faculty of Science and Engineering, Waseda University, Tokyo, 169-8555, Japan
[3]Institute for Medical-oriented Structural Biology, Waseda University, Tokyo, 162-8480, Japan
[4]Graduate School of Medicine, Nippon Medical School, Tokyo, 113-8602, Japan
[*]CZ (zeng.chao@aist.go.jp) or MH (mhamada@waseda.jp)

ABSTRACT

Alternative splicing, a ubiquitous phenomenon in eukaryotes, is 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 transcriptome-wide influence of alternative splicing on RNA subcellular localization has rarely been studied. By analyzing RNA-seq data obtained from subcellular fractions across 13 human cell lines, we identified 8720 switching genes between the cytoplasm and the nucleus. Consistent with previous reports, intron retention was observed to be enriched in the nuclear transcript variants. Interestingly, we found that short and structurally stable introns were positively correlated with nuclear localization. Motif analysis reveals that fourteen RNA-binding protein (RBPs) are prone to be preferentially bound with such introns. To our knowledge, this is the first transcriptome-wide study to analyze and evaluate the effect of alternative splicing on RNA subcellular localization. Our findings reveal that alternative splicing plays a promising role in regulating RNA subcellular localization.

1 Introduction

Most eukaryotic genes consist of exons that are processed to produce mature RNAs and introns that are removed by RNA splicing. Alternative splicing is a regulated process by which exons can be either included or excluded in the mature RNA. 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 than the number of genes it has, which increases the genome complexity[1]. In humans, for instance, approximately 95% of the multi-exonic genes undergoing alternative splicing have been revealed by high-throughput sequencing technology[2]. Exploring the functionality of alternative splicing is critical to our understanding of life mechanisms. Alternative splicing is 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] have received increasing attention. Understanding the pathogenesis associated with alternative splicing can shed light on diagnosis and therapy. With the rapid development of high-throughput technology, it has become possible to study the transcriptome-wide function and mechanism of alternative splicing[8].

The localization of 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 strongly related to its biological function[9]. For example, the asymmetric distribution of RNA in cells can influence the expression of genes[9], formation and interaction of protein complexes[11], biosynthesis of ribosomes[12], and 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, and have expanded the range of target RNA species[17,18]. With the development of microarray and high-throughput sequencing technologies, approaches for the transcriptome-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.

Alternative splicing can regulate RNA/protein subcellular localization[24–27]. However, to date, alternative splicing in only a limited number of genes has 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 transcriptome-wide study of the impact of alternative splicing on RNA subcellular localization. Therefore, we analyzed RNA-sequencing (RNA-seq) data obtained from subcellular (cytoplasmic and nuclear) fractions and investigated whether alternative splicing causes an imbalance in 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. Additionally, motif analysis revealed that fourteen RBPs were enriched in the retained introns that are associated with nuclear localization. The above results were consistently observed across 13 cell lines, suggesting the reliability and robustness of the investigations. To our knowledge, this is the first transcriptome-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.

2 Materials and Methods

2.1 RNA-Seq Data and Bioinformatics

We used the RNA-seq data from the ENCODE[28] 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 (>200 bp in size), and then subjected to high-throughput RNA sequencing. Ribosomal RNA molecules (rRNA) were removed with RiboMinus™ Kit. RNA-seq libraries of strand-specific paired-end reads were sequenced by Illumina HiSeq 2000. See Supplementary Data 1 for accessions of RNA-seq data. The human genome (GRCh38) and comprehensive gene annotation were obtained from GENCODE (v29)[29]. RNA-seq reads were mapped with STAR (2.7.1a)[30] and quantified with RSEM (v1.3.1)[31] using the default parameters. Finally, we utilized SUPPA2[32] to calculate the differential usage (change of transcript usage, ΔTU[33]) and the differential splicing (change of splicing inclusion, $\Delta\Psi$[34]) of transcripts.

2.2 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 were estimated from RNA-seq data using the aforementioned RSEM. 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. SUPPA2 calculates the significance level by comparing the empirical distribution of ΔTU between replicates and conditions (see[32] for details). 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[32,35]. 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.

2.3 Gene Ontology (GO) Enrichment Analysis

GO enrichment analysis was performed with g:Profiler (Benjamini–Hochberg false discovery rate (FDR) < 0.001)[36]. g:Profiler applies the hypergeometric test to detect statistically significantly enriched molecular function (MF), biological processes

(BP), and cellular component (CC). In addition to GO enrichment analysis of switching genes in each cell line, we also defined the shared switching genes (detected as switching genes $\geq$ 5 cell lines in this study, n = 1219) and conducted GO enrichment analysis.

2.4 Nonsense-Mediated RNA Decay (NMD) Sensitivity

NMD sensitivity was defined by the fold-change of transcript expression before and after knockdown of *UPF1*, *SMG6*, and *SMG7* (main factors in NMD) in HeLa-S3 cells. The larger the value, the higher the degradation rate by NMD. We used the RNA-seq data (GSE86148)[37] to calculate the NMD sensitivity. RNA-seq reads were mapped with STAR (2.7.1a)[30] and quantified with RSEM (v1.3.1)[31] using the default parameters and GENCODE (v29) comprehensive gene annotation. EBseq (v1.20.0)[38] was used to calculate the fold-change of transcript expression.

2.5 RNA Structure Context Analysis

We used CapR[39] with default parameters to predict secondary structures for RNA sequences. We extracted the stem probabilities of RNA sequence to represent the stability of the RNA secondary structure. The higher the stem probability, the more stable the secondary structure of RNA.

2.6 RBP-Binding Prediction and Repeat Sequence Analysis

We applied the stand-alone version of RBPmap[40] 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 31 January 2014) that mapped to human (hg38) from Repeatmasker[41]. Bedtools[42] was used to obtain overlaps between these repeat features (TEs) and the target sequences (i.e., retained introns).

3 Results

3.1 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 13 human cell lines (Supplementary Data 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 ($\Delta TU > 0$, termed "cytoplasmic transcripts"), while the other one is mainly expressed in the nucleus ($\Delta TU < 0$, termed "nuclear transcripts"). For instance, the *C16orf91* gene contains two transcript variants (Figure 1A, upper). In HeLa-S3 cells, we observed that the *C16orf91-201* and *C16orf91-202* (also termed a transcript switch in this study) were prone to substantial expression in the cytoplasm and the nucleus, respectively (see Figure 1A, lower and Supplementary Data 2). In total, 2073 transcript switches were detected in HeLa-S3 cells (Figure 1B). Specially, using a p-value of <0.05, we further screened for the switches that were not significantly altered. For example, in HT1080, SK-MEL-5, and SK-N-DZ, we appropriately discarded transcripts that were more likely to exhibit changes in transcript usage because of low expression levels (Supplementary Data 3).

On average, 1650 pairs of transcript switches were identified across the 13 cell lines (Figure 1C). The smallest (768 pairs) and the largest (2766 pairs) were both observed in brain tissue, SK-N-DZ and SK-N-SH cells, respectively. Next, we assessed if there is a group of genes shared among cell lines with similar biological functions through the inclusion of transcript switches. We observed a total of 8720 genes containing transcript switches from these 13 cell lines (hereafter referred to as "switching genes"). More than 93% (8177 of 8720) of these switching genes were found in fewer than six cell lines, indicating that most switching genes were cell-specific (Figure 1D and Supplementary Data 4, some may also caused by the cell specificity of the gene expression). GO enrichment analysis of the remaining genes revealed that these shared switching genes (detected as switching genes $\geq$ 5 cell lines, n = 1219) are associated with receptor/transducer activities (adjusted $p < 1 \times 10^{-10}$, GO:0004888, GO:0060089, GO:0038023, and GO:0004930), multicellular organismal process and G protein-coupled receptor signaling pathway (adjusted $p < 1 \times 10^{-10}$, GO:0032501 and GO:0007186), plasma membrane and cell periphery (adjusted $p < 1 \times 10^{-10}$, GO:0005886 and GO:0071944). Although the GO enrichment analysis results of switching genes in each cell line are somewhat different, they are also roughly similar. (see Supplementary Data 5).

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 preferred to locate in the nucleus and displayed more negative ΔTU values (all $p < 5.4 \times 10^{-24}$, Supplementary Data 6). This result agreed 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[43,44].

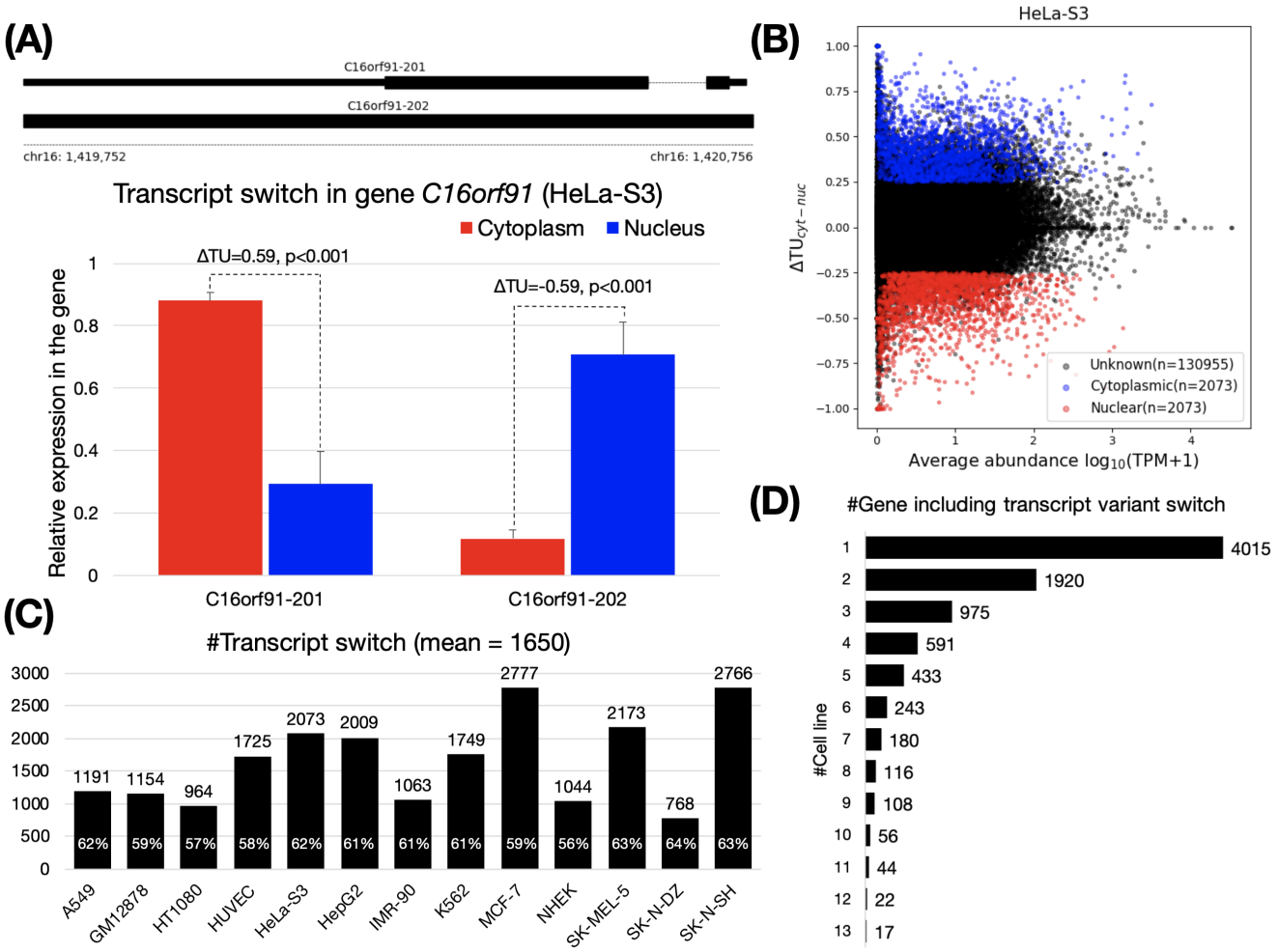


Figure 1. Transcript switches between cytoplasm and nucleus. (**A**) The upper panel shows two transcript variants of the gene *C16orf91*. The lower panel is an example for the identification of a transcript variant switch in the gene *C16orf91* using ΔTU. C16orf91-201 and C16orf91-202 were mainly localized in the cytoplasm and the nucleus, respectively, in HeLa-S3 cells (see Supplementary Data 2 for read coverage on *C16orf91* gene). P values are calculated with SUPPA2[32] (see Materials and Methods). (**B**) Transcriptome-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 the 13 cell lines. The percentage on the histogram indicates the content of protein-coding transcript. (**D**) Number of genes containing transcript variant switches shared across the 13 cell lines. For example, 4015 genes were identified as being cell specific.

3.2 A Transcript Switch Is Associated with RNA Splicing rather than RNA Degradation

We first investigated the post-transcriptional influence of the NMD pathway on the transcript switch. Initially, the two transcripts in a transcript switch are equally distributed in the cytoplasm and nucleus. We consider one possible 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 in HeLa-S3 cells based on changes in RNA-seq data[37] before and after the knockout of *UPF1*, *SMG6*, and *SMG7* (core factors of NMD). However, after comparing the sensitivity of NMD to cytoplasmic and nuclear transcripts, we did not find significant differences (Figure 2A). 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 is, the higher the splicing frequency (or exon number) should be[45], we next

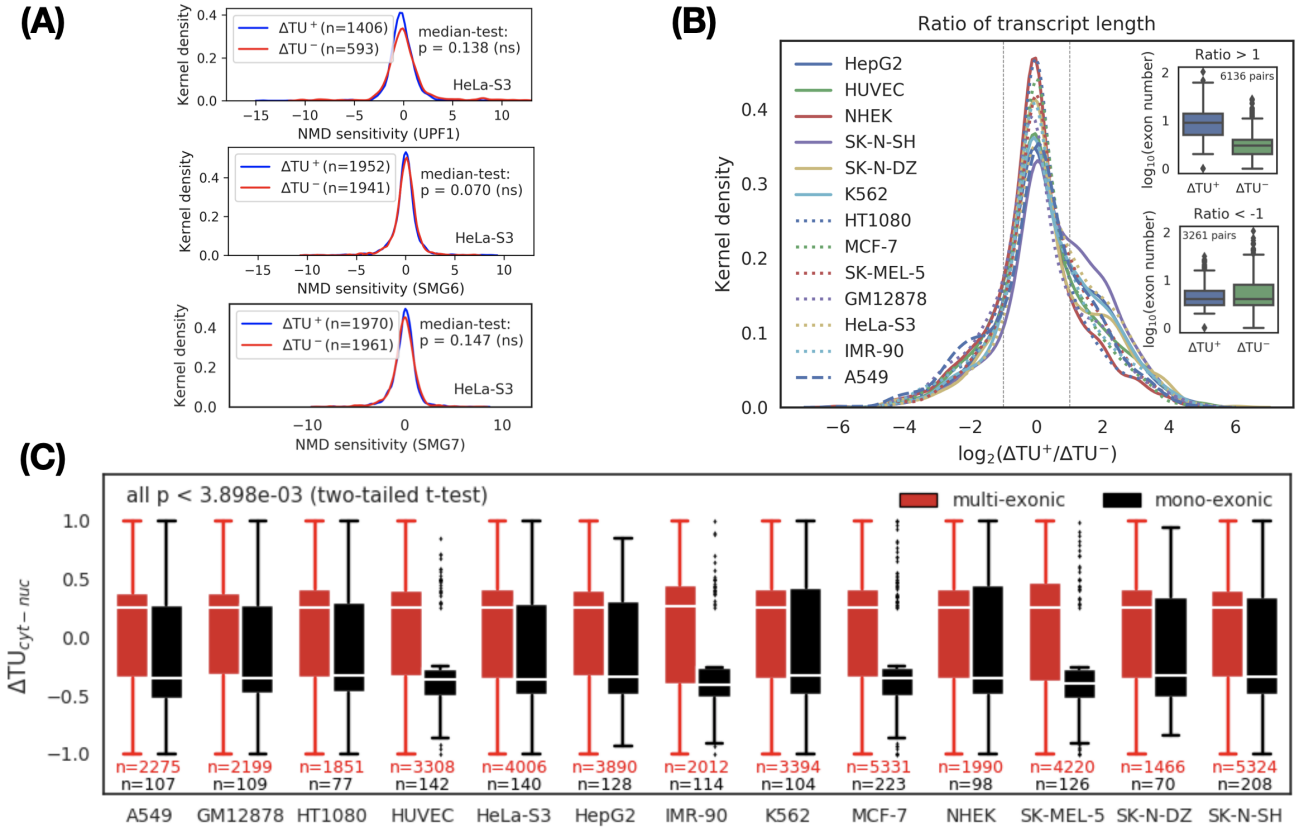


Figure 2. Characterization of cytoplasmic and nuclear transcripts. (**A**) NMD does not explain transcript variant switches between the cytoplasm and the nucleus. NMD sensitivity was defined by the fold-change of transcript expression before and after knockdown of NMD factor *UPF1* (upper), *SMG6* (middle), and *SMG7* (lower) in HeLa-S3 cells. P values: Mood's median test. ns: 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, $p < 2.2 \times 10^{-308}$ Mood's median test), suggesting splicing can promote cytoplasmic localization. No significant difference in splicing frequency was observed in the group if the transcripts $Ratio < -1$ (inset, bottom, $p < 4.8 \times 10^{-12}$ Mood's median test). (**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. Based on two-tailed t-test, we calculated the significant level of difference between ΔTU in mono-exonic and multi-exonic transcripts in each cell line. The largest p value is 3.898×10^{-3}.

determined 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 weak or no correlation ($|ratio| \leq 1$) between transcript length and its subcellular localization. We observed that the number of transcript switches in the positive category was higher than that in the negative category ($p < 1 \times 10^{-6}$, Binomial test), which implied that the longer the transcript is, the more likely it is to be transported into the cytoplasm (Figure 2B). 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 2B, inset). We found that the exon number of the cytoplasmic transcripts in the positive category was higher than the nuclear transcripts (see Figure 2B, inset upper, $p < 2.2 \times 10^{-308}$ Mood's median test). Based on the above observations, we speculated 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 2C and Supplementary Data 7).

3.3 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 (A3, A5, AF, AL, MX, RI, and SE) were considered[46]. 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 "Materials and Methods" for details). The smaller the $\Delta\Psi$, the more the corresponding splicing pattern prefers the nucleus. Interestingly, in HeLa-S3 cells, we observed that retained introns were biased toward the nucleus, while other splicing patterns had no preference (Figure 3A). Consistent results were also observed in the other twelve cell lines (Supplementary Datas 8 and 9).

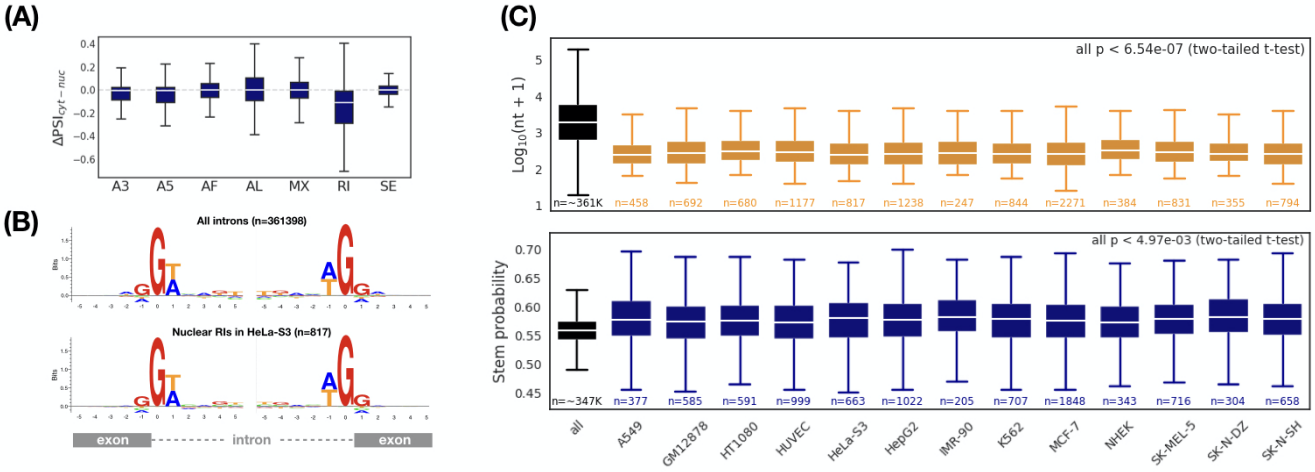


Figure 3. Comparison of alternative splicing patterns between cytoplasmic and nuclear transcripts. (**A**) Retained introns are enriched in the nuclear transcripts (HeLa-S3). $\Delta\Psi$ indicates the inclusion level for a specific alternative splicing pattern. A3: alternative 3′ splice-site; A5: alternative 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. p values: two-tailed t-test between all introns and nuclear RIs. The largest p value is 6.54×10^{-7} and 4.97×10^{-3} for length and RNA secondary structure, respectively.

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 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 differences in the splicing signal between the nuclear RIs and all RIs (Figure 3B). We then 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 was significantly shorter than the overall length level (Figure 3C, upper (all $p < 6.64 \times 10^{-7}$) and Supplementary Data 10). We also found that the average stem probability of the nuclear RIs was higher than that of all RIs (Figure 3C, bottom and Supplementary Data 11). Thus, we concluded that nuclear RIs maintain a more compact and stronger structure when compared with the overall RIs.

3.4 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 the 13 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[47]. The normalized dinucleotide frequency represents the extent to which the corresponding dinucleotide is preferred in the nuclear RIs. Overexpression of the GC-rich sequence occurred in the nuclear RIs (Figure 4A). This observation also provided an intuitive hypothesis of whether RBPs that preferentially bound to RNA containing GC-rich sequences directly or indirectly affect subcellular localization.

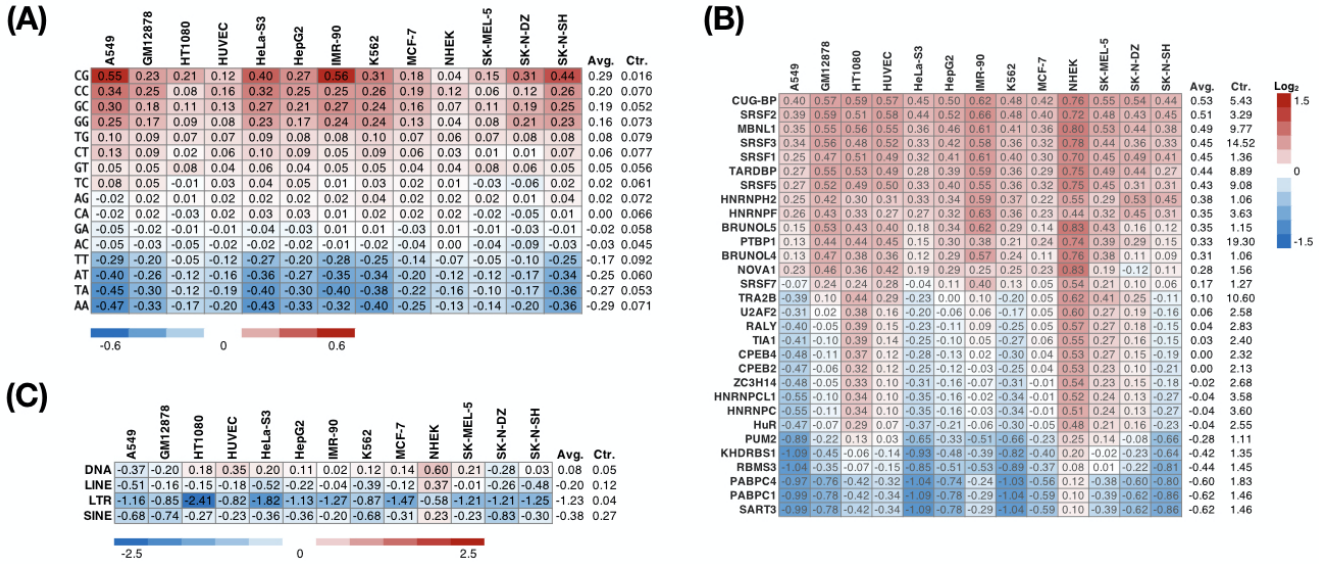


Figure 4. Preference of dinucleotide, RBPs, and TEs on nuclear RIs. The heatmap represents the relative density of a specific (**A**) dinucleotide, (**B**) RBP-binding site, or (**C**) TE (row) on nuclear RIs across multiple cell lines (columns) compared with that of all introns (Ctr.).

To further explore which RBPs preferentially interact with nuclear RIs, we predicted the binding preference between an RBP and an intronic sequence using RBPmap[40], a motif-based approach. We found that across the 13 cell lines, 14 of 30 RBPs tested consistently (more than 80%) preferred attaching to the nuclear RIs (Figure 4B). 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[48], 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[48]), LINE (long interspersed nuclear element), and DNA transposon enrichment in nuclear RIs (Figure 4C). 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.

4 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. Due to the limitation of available experimental data, we calculated ΔTU and performed alternative splicing analysis based on short-read RNA-seq. However, long-read RNA-seq and direct RNA-seq can provide more clear and convincing splicing events[49–52].

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[53–57], post-transcriptional component[58], or coactivator[59–61]. 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[62–65]. 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∼2777 switching genes (Figure 1C) in the 13 cell lines, suggesting that bifunctional genes or GCR are more widespread in the human genome than expected. A total of 8720 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[66] for export factors (REF and/or TAP/p15). The coupling between a conserved RNA export machinery and RNA splicing has also been reviewed[67].

Additionally, we observed that transcript variants containing retained introns were more likely to reside in the nucleus, which is consistent with the previous studies[68,69]. By further analyzing these retained introns, we found that transcripts, including short but structurally stable retained introns, were preferentially localized in the nucleus. It is worth noting that the above observations are still robust after we distinguish between protein coding and noncoding genes (Supplementary Datas 7, 10 and 11). 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[67]. The RNA structure context was discussed as being able to affect protein binding[70]. Motif analysis of the retained introns of nucleus-localized transcripts detetced 14 enriched RBPs, in which MBNL1 has been reported to bind RNA in the CUG repeats[71] and contain two nuclear signals[72]. Previous studies have reported that the CUG repeat expansion caused nuclear aggregates and RNA toxicity, which in turn induces neurological diseases[73,74]. Taken together, the data reveal a post-transcriptional regulation mechanism in which alternative splicing regulates RNA subcellular localization by including or excluding those introns containing nuclear transporter binding sites.

5 Conclusions

This study explored whether alternative splicing can regulate RNA subcellular localization. The RNA-seq data derived from the nuclear and cytoplasmic fractions of 13 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 consistently observed that splicing appears to promote RNA export from the nucleus[75]. 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, fourteen RBPs are consistently identified to prone to associate with retained introns across 13 cell lines. 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). These findings can help us better understand

the regulatory mechanisms of RNA localization, and thus, for example, predict RNA localization more precisely from sequence characteristics[76].

Although we used an EM-based method[31] 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[77,78]). 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 (Supplementary Data 4) will provide a valuable resource for studying gene functions.

References

1. Brett, D., Pospisil, H., Valcárcel, J., Reich, J. & Bork, P. Alternative splicing and genome complexity. *Nat. Genet.* **30**, 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. *Nat. Genet.* **40**, 1413–1415 (2008).
3. Graveley, B. R. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* **17**, 100–107 (2001).
4. Stamm, S. *et al.* 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 Interdiscip. Rev. RNA* **6**, 311–326 (2015).
6. David, C. J. & Manley, J. L. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes & Dev.* **24**, 2343–2364 (2010).
7. Di, C. *et al.* Function, clinical application, and strategies of Pre-mRNA splicing in cancer. *Cell Death & Differ.* **26**, 1181–1194 (2019).
8. Ule, J. & Blencowe, B. J. Alternative splicing regulatory networks: Functions, mechanisms, and evolution. *Mol. Cell* **76**, 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. *Nat. Rev. Mol. Cell Biol.* **16**, 95–109 (2015).
10. Fasken, M. B. & Corbett, A. H. Mechanisms of nuclear mRNA quality control. *RNA Biol.* **6**, 237–241 (2009).
11. Nevo-Dinur, K., Govindarajan, S. & Amster-Choder, O. Subcellular localization of RNA and proteins in prokaryotes. *Trends Genet.* **28**, 314–322 (2012).
12. Dermit, M. *et al.* 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. *Annu. Rev. Biochem.* **67**, 335–394 (1998).
14. Zhou, Y. & King, M. L. Sending RNAs into the future: RNA localization and germ cell fate. *IUBMB life* **56**, 19–27 (2004).
15. Lawrence, J. B. & Singer, R. H. Intracellular localization of messenger RNAs for cytoskeletal proteins. *Cell* **45**, 407–415 (1986).
16. Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H. Visualization of single RNA transcripts in situ. *Science* **280**, 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. *Nat. Methods* **5**, 877–879 (2008).
18. Lee, J. H. *et al.* Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
19. Taliaferro, J. M., Wang, E. T. & Burge, C. B. Genomic analysis of RNA localization. *RNA Biol.* **11**, 1040–1050 (2014).
20. Fazal, F. M. *et al.* Atlas of subcellular RNA localization revealed by APEX-seq. *Cell* **178**, 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. *Mol. Cell* **75**, 875–887 (2019).
22. Zhang, T. *et al.* RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res.* **45**, D135–D138 (2016).
23. Mas-Ponte, D. *et al.* LncATLAS database for subcellular localization of long noncoding RNAs. *RNA* **23**, 1080–1087 (2017).

24. Zhou, W. *et al.* Identification and characterization of two novel splicing isoforms of human estrogen-related receptor β. *The J. Clin. Endocrinol. & Metab.* **91**, 569–579 (2006).

25. Hakre, S. *et al.* Opposing functions of TFII-I spliced isoforms in growth factor-induced gene expression. *Mol. Cell* **24**, 301–308 (2006).

26. Wiesenthal, A. *et al.* NOSTRINβ–A shortened NOSTRIN variant with a role in transcriptional regulation. *Traffic* **10**, 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. *Mol. cellular endocrinology* **314**, 53–61 (2010).

28. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

29. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

30. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

31. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinforma.* **12**, 323 (2011).

32. Trincado, J. L. *et al.* SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19**, 1–11 (2018).

33. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with rna-seq. *Nat. biotechnology* **31**, 46 (2013).

34. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470 (2008).

35. Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. United States Am.* **111**, E5593–601 (2014).

36. Raudvere, U. *et al.* g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research* **47**, W191–W198 (2019).

37. 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**, 189–201 (2017).

38. Leng, N. *et al.* Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics* **29**, 1035–1043 (2013).

39. Fukunaga, T. *et al.* Capr: revealing structural specificities of rna-binding protein target recognition using clip-seq data. *Genome biology* **15**, R16 (2014).

40. 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 Res.* **42**, W361–W367 (2014).

41. Smit, A., Hubley, R. & Green, P. 2013–2015. repeatmasker open-4.0 (2013).

42. Quinlan, A. R. & Hall, I. M. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

43. Quinn, J. J. & Chang, H. Y. Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* **17**, 47 (2016).

44. Marchese, F. P., Raimondi, I. & Huarte, M. The multidimensional mechanisms of long noncoding RNA function. *Genome Biol.* **18**, 206 (2017).

45. Regulapati, R., Bhasi, A., Singh, C. K. & Senapathy, P. Origination of the split structure of spliceosomal genes from random genetic sequences. *PloS one* **3**, e3456 (2008).

46. Alamancos, G. P., Pagès, A., Trincado, J. L., Bellora, N. & Eyras, E. Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* **21**, 1521–1531 (2015).

47. Fernandez, M. *et al.* Prediction of dinucleotide-specific rna-binding sites in proteins. *BMC Bioinforma.* **12**, S5 (2011).

48. Lubelsky, Y. & Ulitsky, I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* **555**, 107–111 (2018).

49. Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The expanding landscape of alternative splicing variation in human populations. *The Am. J. Hum. Genet.* **102**, 11–26 (2018).

50. Wen, W. X., Mead, A. J. & Thongjuea, S. Technological advances and computational approaches for alternative splicing analysis in single cells. *Comput. structural biotechnology journal* **18**, 332–343 (2020).

51. Parker, M. T. *et al.* Nanopore direct rna sequencing maps the complexity of arabidopsis mrna processing and m6a modification. *Elife* **9**, e49658 (2020).

52. Tang, A. D. *et al.* Full-length transcript characterization of sf3b1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat. communications* **11**, 1–12 (2020).

53. Kloc, M. *et al.* Potential structural role of non-coding and coding RNAs in the organization of the cytoskeleton at the vegetal cortex of Xenopus oocytes. *Development* **132**, 3445–3457 (2005).

54. Jenny, A. *et al.* A translation-independent role of oskar RNA in early Drosophila oogenesis. *Development* **133**, 2827–2833 (2006).

55. Lim, S. *et al.* Dorsal activity of maternal squint is mediated by a non-coding function of the RNA. *Development* **139**, 2903–2915 (2012).

56. Shevtsov, S. P. & Dundr, M. Nucleation of nuclear bodies by RNA. *Nat. Cell Biol.* **13**, 167–173 (2011).

57. Jansen, G. *et al.* Abnormal myotonic dystrophy protein kinase levels produce only mild myopathy in mice. *Nat. Genet.* **13**, 316 (1996).

58. Gong, C. & Maquat, L. E. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature* **470**, 284–288 (2011).

59. Chooniedass-Kothari, S. *et al.* The steroid receptor RNA activator is the first functional RNA encoding a protein. *FEBS Lett.* **566**, 43–47 (2004).

60. 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 Res.* **39**, 513–525 (2010).

61. Lanz, R. B. *et al.* A steroid receptor coactivator, SRA, functions as an RNA and is present in an SRC-1 complex. *Cell* **97**, 17–27 (1999).

62. Rossi, A. *et al.* Genetic compensation induced by deleterious mutations but not gene knockdowns. *Nature* **524**, 230 (2015).

63. El-Brolosy, M. A. & Stainier, D. Y. Genetic compensation: A phenomenon in search of mechanisms. *PLoS Genet.* **13**, e1006780 (2017).

64. El-Brolosy, M. A. *et al.* Genetic compensation triggered by mutant mRNA degradation. *Nature* **568**, 193–197 (2019).

65. Ma, Z. *et al.* PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. *Nature* **568**, 259–263 (2019).

66. 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 J.* **20**, 4987–4997 (2001).

67. Reed, R. & Hurt, E. A conserved mRNA export machinery coupled to pre-mRNA splicing. *Cell* **108**, 523–531 (2002).

68. Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J. & Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome biology* **14**, R70 (2013).

69. Boutz, P. L., Bhutkar, A. & Sharp, P. A. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes & development* **29**, 63–80 (2015).

70. Wan, Y., Kertesz, M., Spitale, R. C., Segal, E. & Chang, H. Y. Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.* **12**, 641–655 (2011).

71. Warf, M. B. & Berglund, J. A. Mbnl binds similar rna structures in the cug repeats of myotonic dystrophy and its pre-mrna substrate cardiac troponin t. *Rna* **13**, 2238–2251 (2007).

72. Taylor, K. *et al.* MBNL splicing activity depends on RNA binding site structural context. *Nucleic Acids Res.* **46**, 9119–9133 (2018).

73. Goodwin, M. *et al.* MBNL sequestration by toxic RNAs and RNA misprocessing in the myotonic dystrophy brain. *Cell Reports* **12**, 1159–1168 (2015).

74. 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**, 2294–2306 (2018).

75. Valencia, P., Dias, A. P. & Reed, R. Splicing promotes rapid and efficient mrna export in mammalian cells. *Proc. Natl. Acad. Sci.* **105**, 3386–3391 (2008).

76. Zuckerman, B. & Ulitsky, I. Predictive models of subcellular localization of long rnas. *RNA* **25**, 557–572 (2019).

77. Soneson, C. *et al.* A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. *Nat. Commun.* **10**, 3359 (2019).

78. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(a) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).

Acknowledgements

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

Author contributions

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

Competing interests

The authors declare that they have no competing interests.

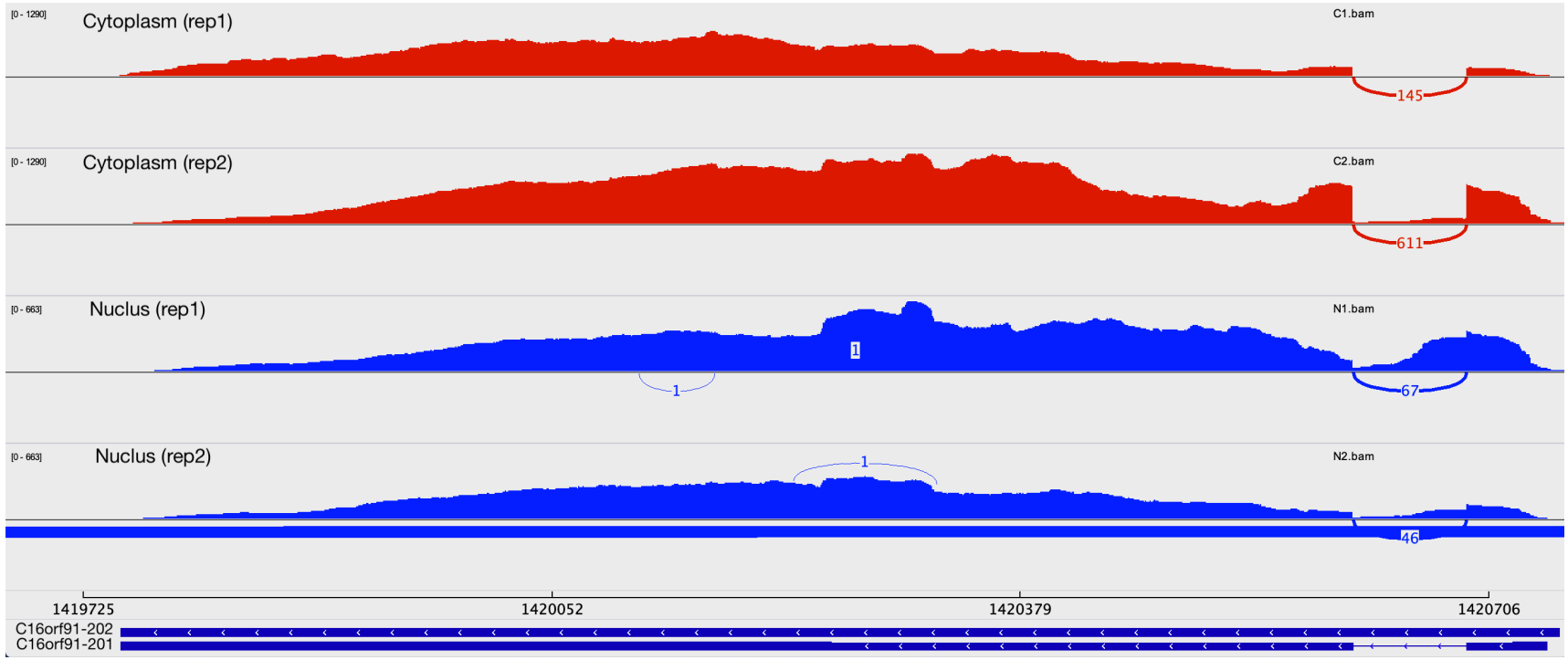
Additional information

- Supplementary Data 1 — RNA-seq data used in this study.
- Supplementary Data 2 — Read coverage of gene *C16orf91* in HeLa-S3.
- Supplementary Data 3 — Transcriptome-wide identification of transcript variant switches in 12 human cell lines.
- Supplementary Data 4 — List of switching genes identified across 13 cell lines in this study. Gene Ontology enrichment analysis for switch genes.
- Supplementary Data 5 — Gene Ontology enrichment analysis for switch genes.
- Supplementary Data 6 — Comparison of ΔTU between protein-coding transcripts and non-coding transcripts.
- Supplementary Data 7 — Comparison of ΔTU between mono-exonic and multi-exonic transcripts across 13 cell lines.
- Supplementary Data 8 — Comparison of alternative splicing patterns between cytoplasmic and nuclear transcripts (all genes).
- Supplementary Data 9 — Comparison of alternative splicing patterns between cytoplasmic and nuclear transcripts (protein-coding and noncoding genes).
- Supplementary Data 10 — Comparison of length between all introns and nuclear RIs.
- Supplementary Data 11 — Comparison of RNA secondary structure (stem probability) between all introns and nuclear RIs.

Supplementary Data 1. RNA-seq data used in this study

Cell line	Tissue	Cytosolic_rep1 (C1)	Cytosolic_rep2 (C2)	Nuclear_rep1 (N1)	Nuclear_rep2 (N2)
A549	Lung	ENCFF000EKW	ENCFF000ELC	ENCFF000EMA	ENCFF000ELQ
		ENCFF000EKX	ENCFF000ELF	ENCFF000EMD	ENCFF000EMB
GM12878	Blood	ENCFF000EZH	ENCFF000EZF	ENCFF000FCG	ENCFF000FCH
		ENCFF000EZR	ENCFF000EZT	ENCFF000FCI	ENCFF000FCU
HeLa-S3	Cervix	ENCFF000FQK	ENCFF000FQW	ENCFF000FSU	ENCFF000FTB
		ENCFF000FQX	ENCFF000FQV	ENCFF000FTA	ENCFF000FTJ
HepG2	Liver	ENCFF000FXU	ENCFF000FYG	ENCFF000GAF	ENCFF000GAH
		ENCFF000FYF	ENCFF000FYH	ENCFF000GAT	ENCFF000GAS
HT1080	Connective tissue	ENCFF447UEE	ENCFF331FNA	ENCFF028JQU	ENCFF365ZYO
		ENCFF058MHC	ENCFF503XDE	ENCFF381YLG	ENCFF087HWA
HUVEC	Umbilical vein	ENCFF000GTX	ENCFF000GUJ	ENCFF000GWF	ENCFF000GWS
		ENCFF000GUH	ENCFF000GUI	ENCFF000GWT	ENCFF000GWV
IMR-90	Lung	ENCFF000HCB	ENCFF000HCI	ENCFF000HDD	ENCFF000HDE
		ENCFF000HCC	ENCFF000HCJ	ENCFF000HDF	ENCFF000HDH
MCF-7	Breast	ENCFF000HSB	ENCFF000HSC	ENCFF000HTE	ENCFF000HTD
		ENCFF000HSI	ENCFF000HSK	ENCFF000HVN	ENCFF000HVM
NHEK	Skin	ENCFF000IDC	ENCFF000IDF	ENCFF000IFF	ENCFF000IFI
		ENCFF000IDI	ENCFF000IDH	ENCFF000IFH	ENCFF000IFW
SK-MEL-5	Skin	ENCFF470ZTS	ENCFF687XKK	ENCFF005NLJ	ENCFF394ODW
		ENCFF299CXA	ENCFF570YTU	ENCFF635CQM	ENCFF365MLV
SK-N-DZ	Brain	ENCFF960DLP	ENCFF138HQU	ENCFF482SFO	ENCFF993CAL
		ENCFF288VWD	ENCFF822LCB	ENCFF691TRA	ENCFF689LXP
SK-N-SH	Brain	ENCFF000INM	ENCFF000INN	ENCFF000IOD	ENCFF000IOP
		ENCFF000INT	ENCFF000INV	ENCFF000IOQ	ENCFF000IPB
K562	Bone marrow	ENCFF000HJP	ENCFF000HJF	ENCFF000HOC	ENCFF000HOD
		ENCFF000HJW	ENCFF000HJX	ENCFF000HOE	ENCFF000HOQ

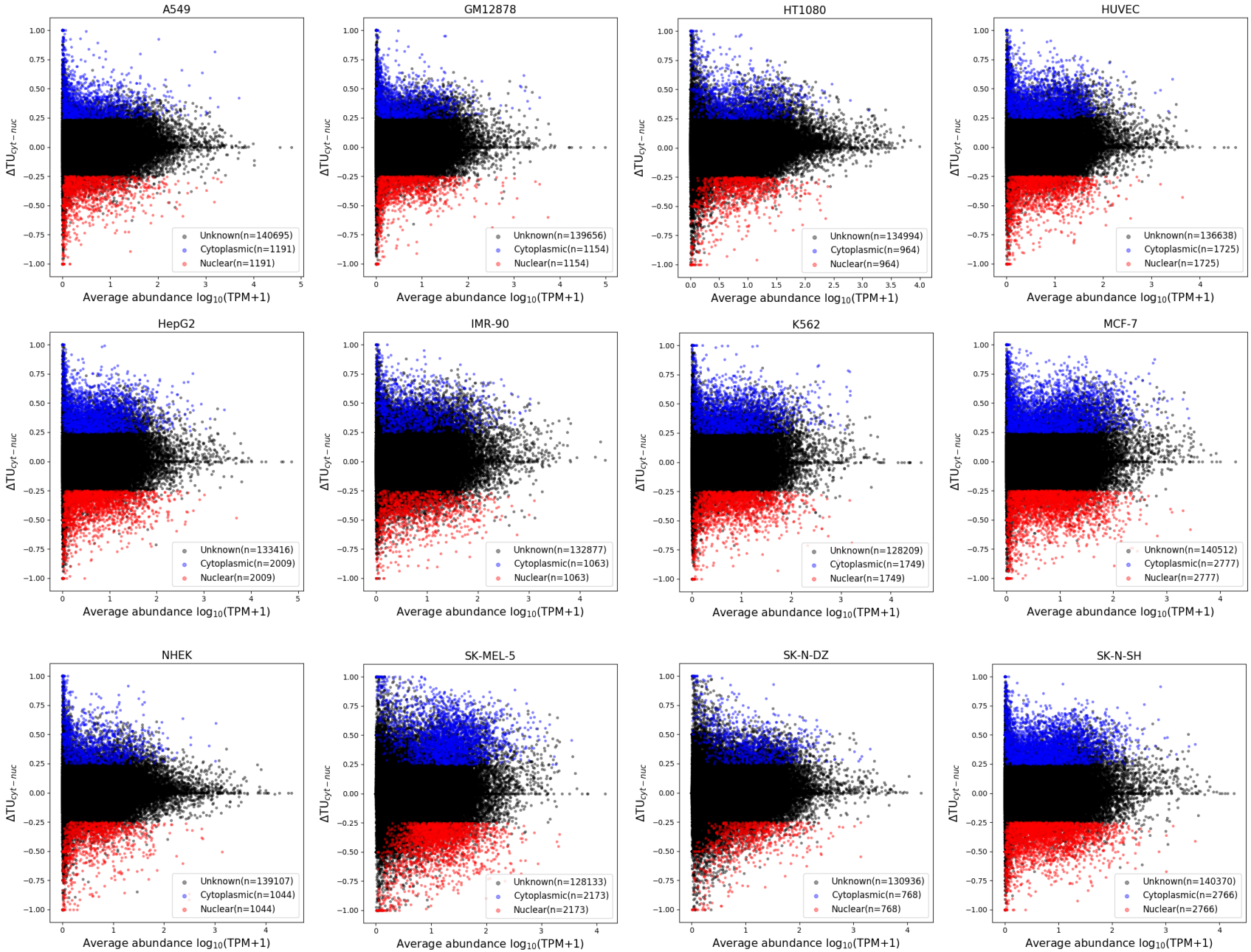
- The libraries were constructed from rRNA-depleted Ploy-A+ RNA (> 200bp in size)
- Strand-specific, PE101nt or PE76nt



Supplementary Data 2. Read coverage of gene *C16orf91* in HeLa-S3. IGV (Robinson et al 2011) was used to visualize the read coverage.

Reference:

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P., 2011. Integrative genomics viewer. Nature biotechnology, 29(1), pp.24-26.



Supplementary Data 3. Transcriptome-wide identification of transcript variant switches in 12 human cell lines. We applied |ΔTU|>0.25 and p<0.05 to filter out cytoplasmic(|ΔTU|>0, blue) and nuclear (|ΔTU|<0, red) transcript variants. Unknown (black) are transcripts filtered as no significant.

Supplementary Data 4. List of switching genes identified across 13 cell lines in this study.

Online resource: https://doi.org/10.6084/m9.figshare.11999658

Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (sheared, continued)

source	term_name	term_id	adjusted_p_value	negative_log10_of_adjusted_p_value
GO:MF	transmembrane signaling receptor activity	GO:0004888	1.55E-16	
GO:MF	molecular transducer activity	GO:0060089	5.59E-16	
GO:MF	signaling receptor activity	GO:0038023	5.68E-16	
GO:MF	G protein-coupled receptor activity	GO:0004930	6.56E-12	
GO:MF	passive transmembrane transporter activity	GO:0022803	0.000493696	
GO:MF	channel activity	GO:0015267	0.000493696	
GO:BP	multicellular organismal process	GO:0032501	3.31E-12	
GO:BP	G protein-coupled receptor signaling pathway	GO:0007186	2.10E-11	
GO:BP	system process	GO:0003008	3.32E-09	
GO:BP	sensory perception of chemical stimulus	GO:0007606	1.85E-07	
GO:BP	nervous system process	GO:0050877	2.73E-07	
GO:BP	detection of stimulus involved in sensory perception	GO:0050906	1.36162E-06	
GO:BP	detection of chemical stimulus involved in sensory perception	GO:0050907	1.61549E-06	
GO:BP	signaling	GO:0023052	2.52304E-06	
GO:BP	signal transduction	GO:0007165	1.66204E-05	
GO:BP	detection of chemical stimulus	GO:0009593	2.20905E-05	
GO:BP	sensory perception	GO:0007600	2.65966E-05	
GO:BP	cell communication	GO:0007154	2.81869E-05	
GO:BP	defense response	GO:0006952	6.05683E-05	
GO:BP	detection of stimulus	GO:0051606	6.68613E-05	
GO:BP	response to external stimulus	GO:0009605	0.000100343	
GO:BP	multicellular organism development	GO:0007275	0.000100343	
GO:BP	adaptive immune response	GO:0002250	0.000171355	
GO:BP	cell differentiation	GO:0030154	0.000179869	
GO:BP	system development	GO:0048731	0.000207557	
GO:BP	ion transport	GO:0006811	0.000405413	
GO:BP	positive regulation of immune system process	GO:0002684	0.000439648	
GO:BP	anatomical structure development	GO:0048856	0.000518076	
GO:BP	regulation of multicellular organismal process	GO:0051239	0.000537973	
GO:BP	defense response to bacterium	GO:0042742	0.000820237	
GO:CC	plasma membrane	GO:0005886	3.50E-13	
GO:CC	cell periphery	GO:0071944	4.22E-13	
GO:CC	extracellular region	GO:0005576	1.17977E-05	
GO:CC	plasma membrane region	GO:0098590	0.000637715	
GO:CC	intrinsic component of membrane	GO:0031224	0.000637715	

GO(Gene Ontology) enrichment analysis for 1219 sheared switching genes (detected as switch gene ≥ 5 cell lines in this study) was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

* MF, BP, CC represent Molecular Function, Biological Process, Cellular Component, respectively.

Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (A549, continued)

source	term_name	term_id	adjusted_p_value	negative_log10_of_adjusted_p_value
GO:MF	transmembrane signaling receptor activity	GO:0004888	2.58E-09	
GO:MF	signaling receptor activity	GO:0038023	9.26E-08	
GO:MF	molecular transducer activity	GO:0060089	9.26E-08	
GO:MF	G protein-coupled receptor activity	GO:0004930	1.27132E-05	
GO:BP	G protein-coupled receptor signaling pathway	GO:0007186	1.535E-05	
GO:BP	sensory perception of chemical stimulus	GO:0007606	2.3181E-05	
GO:BP	detection of chemical stimulus involved in sensory perception	GO:0050907	2.3181E-05	
GO:BP	sensory perception of smell	GO:0007608	3.06052E-05	
GO:BP	detection of chemical stimulus	GO:0009593	3.06052E-05	
GO:BP	detection of stimulus involved in sensory perception	GO:0050906	8.11614E-05	
GO:CC	no significant results			

GO(Gene Ontology) enrichment analysis for 1191 switching genes in A549 was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

* MF, BP, CC represent Molecular Function, Biological Process, Cellular Component, respectively.

Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (GM12878, continued)

source	term_name	term_id	adjusted_p_value	negative_log10_of_adjusted_p_value
GO:MF	G protein-coupled receptor activity	GO:0004930	4.27792E-05	
GO:BP	G protein-coupled receptor signaling pathway	GO:0007186	3.80158E-05	
GO:BP	system process	GO:0003008	0.000132801	
GO:CC	no significant results			

GO(Gene Ontology) enrichment analysis for 1154 switching genes in GM12878 was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (HT1080, continued)

source	term_name	term_id	adjusted_p_value	negative_log10_of_adjusted_p_value
GO:MF	transmembrane signaling receptor activity	GO:0004888	6.49781E-05	
GO:MF	G protein-coupled receptor activity	GO:0004930	0.000178858	
GO:MF	signaling receptor activity	GO:0038023	0.000211021	
GO:MF	molecular transducer activity	GO:0060089	0.000211021	
GO:BP	signal transduction	GO:0007165	4.50381E-06	
GO:BP	multicellular organismal process	GO:0032501	4.50381E-06	
GO:BP	G protein-coupled receptor signaling pathway	GO:0007186	5.06491E-06	
GO:BP	signaling	GO:0023052	5.06491E-06	
GO:BP	cell communication	GO:0007154	2.71825E-05	
GO:CC	extracellular region	GO:0005576	4.51E-12	
GO:CC	extracellular space	GO:0005615	5.10E-10	
GO:CC	cell periphery	GO:0071944	1.51052E-06	
GO:CC	plasma membrane	GO:0005886	2.08264E-06	
GO:CC	extracellular organelle	GO:0043230	6.65978E-05	
GO:CC	extracellular vesicle	GO:1903561	6.65978E-05	
GO:CC	extracellular exosome	GO:0070062	9.1387E-05	

GO(Gene Ontology) enrichment analysis for 964 switching genes in HT1080 was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (HUVEC, continued)

source	term_name	term_id	adjusted_p_value	negative_log10_of_adjusted_p_value
GO:MF	transmembrane signaling receptor activity	GO:0004888	1.04E-08	
GO:MF	G protein-coupled receptor activity	GO:0004930	1.04E-08	
GO:MF	signaling receptor activity	GO:0038023	1.73E-07	
GO:MF	molecular transducer activity	GO:0060089	1.96E-07	
GO:BP	detection of stimulus involved in sensory perception	GO:0050906	5.37E-10	
GO:BP	detection of chemical stimulus involved in sensory perception	GO:0050907	9.09E-10	
GO:BP	detection of stimulus	GO:0051606	5.62E-09	
GO:BP	sensory perception of chemical stimulus	GO:0007606	2.26E-08	
GO:BP	sensory perception of smell	GO:0007608	2.26E-08	
GO:BP	multicellular organismal process	GO:0032501	1.40E-07	
GO:BP	system process	GO:0003008	2.72E-07	
GO:BP	detection of chemical stimulus	GO:0009593	2.90E-07	
GO:BP	sensory perception	GO:0007600	4.30E-07	
GO:BP	nervous system process	GO:0050877	1.37468E-05	
GO:BP	G protein-coupled receptor signaling pathway	GO:0007186	7.79881E-05	
GO:BP	cell differentiation	GO:0030154	0.000238532	
GO:CC	extracellular region	GO:0005576	2.7509E-06	
GO:CC	extracellular space	GO:0005615	0.000209584	

GO(Gene Ontology) enrichment analysis for 1725 switching genes in HUVEC was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

* MF, BP, CC represent Molecular Function, Biological Process, Cellular Component, respectively.

Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (HeLa-S3, continued)

source	term_name	term_id	adjusted_p_value	-log10 (adjusted_p_value)
GO:MF	transmembrane signaling receptor activity	GO:0004888	6.36E-20	
GO:MF	signaling receptor activity	GO:0038023	3.32E-18	
GO:MF	molecular transducer activity	GO:0060089	4.04E-18	
GO:MF	G protein-coupled receptor activity	GO:0004930	1.43E-16	
GO:MF	gated channel activity	GO:0022836	0.00014682	
GO:BP	G protein-coupled receptor signaling pathway	GO:0007186	4.22E-14	
GO:BP	detection of chemical stimulus involved in sensory perception	GO:0050907	4.22E-13	
GO:BP	detection of stimulus involved in sensory perception	GO:0050906	5.95E-13	
GO:BP	sensory perception of chemical stimulus	GO:0007606	6.90E-13	
GO:BP	multicellular organismal process	GO:0032501	7.57E-13	
GO:BP	nervous system process	GO:0050877	1.73E-12	
GO:BP	sensory perception	GO:0007600	4.13E-12	
GO:BP	sensory perception of smell	GO:0007608	1.13E-11	
GO:BP	detection of chemical stimulus	GO:0009593	1.59E-11	
GO:BP	system process	GO:0003008	4.10E-11	
GO:BP	detection of stimulus	GO:0051606	2.85E-07	
GO:BP	adaptive immune response	GO:0002250	4.41341E-05	
GO:BP	signaling	GO:0023052	0.000978609	
GO:CC	plasma membrane	GO:0005886	3.22E-20	
GO:CC	cell periphery	GO:0071944	6.49E-20	
GO:CC	integral component of plasma membrane	GO:0005887	1.44E-09	
GO:CC	integral component of membrane	GO:0016021	1.84E-09	
GO:CC	intrinsic component of membrane	GO:0031224	3.40E-09	
GO:CC	intrinsic component of plasma membrane	GO:0031226	6.90E-09	
GO:CC	receptor complex	GO:0043235	3.00457E-06	
GO:CC	plasma membrane region	GO:0098590	2.56347E-05	
GO:CC	extracellular matrix	GO:0031012	3.55608E-05	
GO:CC	plasma membrane protein complex	GO:0098797	0.000121469	
GO:CC	plasma membrane signaling receptor complex	GO:0098802	0.000238214	
GO:CC	collagen-containing extracellular matrix	GO:0062023	0.000679446	
GO:CC	extracellular region	GO:0005576	0.000679446	
GO:CC	synaptic membrane	GO:0097060	0.000679446	

GO(Gene Ontology) enrichment analysis for 2073 switching genes in HeLa-S3 was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

* MF, BP, CC represent Molecular Function, Biological Process, Cellular Component, respectively.

Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (HepG2, continued)

source	term_name	term_id	adjusted_p_value	negative_log10_of_adjusted_p_value
GO:MF	transmembrane signaling receptor activity	GO:0004888	3.54E-15	
GO:MF	G protein-coupled receptor activity	GO:0004930	9.45E-13	
GO:MF	olfactory receptor activity	GO:0004984	1.15E-12	
GO:MF	signaling receptor activity	GO:0038023	1.15E-12	
GO:MF	molecular transducer activity	GO:0060089	1.15E-12	
GO:BP	G protein-coupled receptor signaling pathway	GO:0007186	5.33E-14	
GO:BP	sensory perception of chemical stimulus	GO:0007606	9.52E-14	
GO:BP	sensory perception of smell	GO:0007608	1.60E-12	
GO:BP	detection of chemical stimulus involved in sensory perception	GO:0050907	5.18E-12	
GO:BP	detection of chemical stimulus involved in sensory perception of smell	GO:0050911	1.43E-11	
GO:BP	detection of chemical stimulus	GO:0009593	2.53E-10	
GO:BP	system process	GO:0003008	4.56E-09	
GO:BP	multicellular organismal process	GO:0032501	7.88E-09	
GO:BP	detection of stimulus involved in sensory perception	GO:0050906	9.69E-09	
GO:BP	sensory perception	GO:0007600	1.23E-08	
GO:BP	nervous system process	GO:0050877	2.19E-08	
GO:BP	detection of stimulus	GO:0051606	8.61E-08	
GO:BP	signaling	GO:0023052	6.64E-07	
GO:BP	signal transduction	GO:0007165	9.84594E-06	
GO:BP	cell communication	GO:0007154	1.09238E-05	
GO:BP	keratinization	GO:0031424	0.000190632	
GO:BP	regulation of response to stimulus	GO:0048583	0.000408547	
GO:BP	humoral immune response	GO:0006959	0.000504249	
GO:BP	keratinocyte differentiation	GO:0030216	0.000527629	
GO:BP	gene silencing by miRNA	GO:0035195	0.000527629	
GO:CC	cell periphery	GO:0071944	4.18E-14	
GO:CC	plasma membrane	GO:0005886	1.57E-13	
GO:CC	integral component of membrane	GO:0016021	1.44303E-06	
GO:CC	intrinsic component of membrane	GO:0031224	2.02693E-06	
GO:CC	integral component of plasma membrane	GO:0005887	0.000194967	
GO:CC	intrinsic component of plasma membrane	GO:0031226	0.00034656	
GO:CC	plasma membrane protein complex	GO:0098797	0.000698663	

GO(Gene Ontology) enrichment analysis for 2009 switching genes in HepG2 was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

* MF, BP, CC represent Molecular Function, Biological Process, Cellular Component, respectively.

Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (IMR-90, continued)

source	term_name	term_id	adjusted_p_value	negative_log10_of_adjusted_p_value
GO:MF	transmembrane signaling receptor activity	GO:0004888	5.05E-11	
GO:MF	signaling receptor activity	GO:0038023	7.05E-11	
GO:MF	molecular transducer activity	GO:0060089	1.38E-10	
GO:MF	G protein-coupled receptor activity	GO:0004930	3.27E-09	
GO:MF	olfactory receptor activity	GO:0004984	5.7643E-05	
GO:BP	multicellular organismal process	GO:0032501	3.2905E-06	
GO:BP	sensory perception	GO:0007600	4.58447E-05	
GO:BP	system process	GO:0003008	6.72981E-05	
GO:BP	nervous system process	GO:0050877	0.000186145	
GO:BP	G protein-coupled receptor signaling pathway	GO:0007186	0.000239019	
GO:BP	detection of stimulus involved in sensory perception	GO:0050906	0.000612089	
GO:BP	detection of chemical stimulus involved in sensory perception of smell	GO:0050911	0.000612089	
GO:BP	sensory perception of chemical stimulus	GO:0007606	0.000618172	
GO:BP	detection of chemical stimulus involved in sensory perception	GO:0050907	0.000618172	
GO:CC	plasma membrane	GO:0005886	1.24E-07	
GO:CC	cell periphery	GO:0071944	1.24E-07	

GO(Gene Ontology) enrichment analysis for 1063 switching genes in IMR-90 was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

* MF, BP, CC represent Molecular Function, Biological Process, Cellular Component, respectively.

Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (K562, continued)

source	term_name	term_id	adjusted_p_value	negative_log10_of_adjusted_p_value
GO:MF	G protein-coupled receptor activity	GO:0004930	6.83E-11	
GO:MF	transmembrane signaling receptor activity	GO:0004888	3.65E-09	
GO:MF	signaling receptor activity	GO:0038023	2.52E-08	
GO:MF	molecular transducer activity	GO:0060089	2.89E-08	
GO:BP	G protein-coupled receptor signaling pathway	GO:0007186	1.94E-15	
GO:BP	system process	GO:0003008	4.80E-12	
GO:BP	nervous system process	GO:0050877	1.01E-10	
GO:BP	detection of chemical stimulus	GO:0009593	1.18E-09	
GO:BP	detection of chemical stimulus involved in sensory perception	GO:0050907	1.18E-09	
GO:BP	multicellular organismal process	GO:0032501	9.58E-09	
GO:BP	detection of stimulus involved in sensory perception	GO:0050906	1.36E-08	
GO:BP	sensory perception of smell	GO:0007608	3.31E-08	
GO:BP	sensory perception	GO:0007600	2.87E-07	
GO:BP	sensory perception of chemical stimulus	GO:0007606	5.46E-07	
GO:BP	detection of stimulus	GO:0051606	1.31782E-06	
GO:BP	adaptive immune response	GO:0002250	0.000259623	
GO:CC	cell periphery	GO:0071944	3.90E-09	
GO:CC	plasma membrane	GO:0005886	4.12E-09	
GO:CC	extracellular region	GO:0005576	0.000163539	
GO:CC	intrinsic component of membrane	GO:0031224	0.000205003	
GO:CC	extracellular matrix	GO:0031012	0.000216974	
GO:CC	integral component of membrane	GO:0016021	0.000295277	

GO(Gene Ontology) enrichment analysis for 1749 switching genes in K562 was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

* MF, BP, CC represent Molecular Function, Biological Process, Cellular Component, respectively.

Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (MCF-7, continued)

source	term_name	term_id	adjusted_p_value	negative_log10_of_adjusted_p_value
GO:MF	transmembrane signaling receptor activity	GO:0004888	3.29E-22	
GO:MF	signaling receptor activity	GO:0038023	9.46E-20	
GO:MF	molecular transducer activity	GO:0060089	9.46E-20	
GO:MF	G protein-coupled receptor activity	GO:0004930	1.26E-17	
GO:MF	receptor regulator activity	GO:0030545	0.000673771	
GO:BP	detection of chemical stimulus involved in sensory perception	GO:0050907	9.20E-19	
GO:BP	detection of stimulus involved in sensory perception	GO:0050906	4.21E-17	
GO:BP	detection of chemical stimulus	GO:0009593	4.21E-17	
GO:BP	multicellular organismal process	GO:0032501	1.97E-15	
GO:BP	G protein-coupled receptor signaling pathway	GO:0007186	8.82E-15	
GO:BP	sensory perception of chemical stimulus	GO:0007606	3.28E-14	
GO:BP	sensory perception of smell	GO:0007608	1.02E-12	
GO:BP	detection of stimulus	GO:0051606	5.08E-12	
GO:BP	sensory perception	GO:0007600	5.29E-12	
GO:BP	system process	GO:0003008	1.22E-11	
GO:BP	nervous system process	GO:0050877	2.33E-11	
GO:BP	gene silencing by miRNA	GO:0035195	2.77583E-05	
GO:BP	posttranscriptional gene silencing	GO:0016441	3.61883E-05	
GO:BP	posttranscriptional gene silencing by RNA	GO:0035194	3.66099E-05	
GO:BP	response to external stimulus	GO:0009605	7.15327E-05	
GO:BP	signaling	GO:0023052	7.15327E-05	
GO:BP	adaptive immune response	GO:0002250	9.03152E-05	
GO:BP	regulation of multicellular organismal process	GO:0051239	0.000169194	
GO:BP	cell communication	GO:0007154	0.000223032	
GO:BP	humoral immune response	GO:0006959	0.000229014	
GO:BP	gene silencing by RNA	GO:0031047	0.000233736	
GO:BP	keratinocyte differentiation	GO:0030216	0.000373816	
GO:BP	keratinization	GO:0031424	0.000471737	
GO:BP	system development	GO:0048731	0.000493893	
GO:BP	regulation of developmental process	GO:0050793	0.000493893	
GO:BP	defense response	GO:0006952	0.000572092	
GO:BP	multicellular organism development	GO:0007275	0.000586089	
GO:BP	signal transduction	GO:0007165	0.000639678	
GO:BP	positive regulation of developmental process	GO:0051094	0.000859708	
GO:BP	defense response to bacterium	GO:0042742	0.000886349	
GO:BP	G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger	GO:0007187	0.000886349	
GO:BP	anatomical structure development	GO:0048856	0.000946523	
GO:CC	plasma membrane	GO:0005886	2.37E-12	
GO:CC	cell periphery	GO:0071944	2.37E-12	
GO:CC	receptor complex	GO:0043235	1.57E-07	
GO:CC	intrinsic component of membrane	GO:0031224	1.81786E-05	
GO:CC	integral component of plasma membrane	GO:0005887	2.17165E-05	
GO:CC	extracellular region	GO:0005576	2.17165E-05	
GO:CC	integral component of membrane	GO:0016021	2.30812E-05	
GO:CC	intrinsic component of plasma membrane	GO:0031226	2.30812E-05	
GO:CC	intermediate filament	GO:0005882	6.0376E-05	
GO:CC	plasma membrane protein complex	GO:0098797	0.000166003	
GO:CC	plasma membrane region	GO:0098590	0.000387874	

GO(Gene Ontology) enrichment analysis for 2777 switching genes in MCF-7 was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

* MF, BP, CC represent Molecular Function, Biological Process, Cellular Component, respectively.

Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (NHEK, continued)

source	term_name	term_id	adjusted_p_value	negative_log10_of_adjusted_p_value
GO:MF	transmembrane signaling receptor activity	GO:0004888	0.000223642	
GO:MF	signaling receptor activity	GO:0038023	0.000223642	
GO:MF	molecular transducer activity	GO:0060089	0.000223642	
GO:MF	olfactory receptor activity	GO:0004984	0.000669489	
GO:BP	multicellular organismal process	GO:0032501	0.000809884	
GO:CC	extracellular region	GO:0005576	0.000461703	
GO:CC	plasma membrane	GO:0005886	0.00049703	
GO:CC	cell periphery	GO:0071944	0.000666079	

GO(Gene Ontology) enrichment analysis for 2777 switching genes in NHEK was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (SK-MEL-5, continued)

source	term_name	term_id	adjusted_p_value	negative_log10_of_adjusted_p_value
GO:MF	transmembrane signaling receptor activity	GO:0004888	2.50E-20	
GO:MF	G protein-coupled receptor activity	GO:0004930	2.50E-20	
GO:MF	signaling receptor activity	GO:0038023	2.50E-20	
GO:MF	molecular transducer activity	GO:0060089	2.50E-20	
GO:MF	receptor regulator activity	GO:0030545	4.34223E-05	
GO:MF	passive transmembrane transporter activity	GO:0022803	0.000322256	
GO:MF	channel activity	GO:0015267	0.000322256	
GO:MF	signaling receptor activator activity	GO:0030546	0.000481617	
GO:MF	receptor ligand activity	GO:0048018	0.000631058	
GO:MF	DNA-binding transcription factor activity	GO:0003700	0.000739906	
GO:BP	G protein-coupled receptor signaling pathway	GO:0007186	3.38E-17	
GO:BP	detection of stimulus involved in sensory perception	GO:0050906	3.39E-17	
GO:BP	detection of chemical stimulus involved in sensory perception	GO:0050907	3.73E-16	
GO:BP	multicellular organismal process	GO:0032501	1.60E-15	
GO:BP	system process	GO:0003008	1.60E-15	
GO:BP	detection of chemical stimulus	GO:0009593	2.10E-15	
GO:BP	sensory perception	GO:0007600	5.11E-15	
GO:BP	sensory perception of chemical stimulus	GO:0007606	4.57E-14	
GO:BP	nervous system process	GO:0050877	1.96E-13	
GO:BP	sensory perception of smell	GO:0007608	3.80E-12	
GO:BP	detection of stimulus	GO:0051606	4.66E-11	
GO:BP	adaptive immune response	GO:0002250	1.89E-08	
GO:BP	signaling	GO:0023052	1.21971E-06	
GO:BP	humoral immune response	GO:0006959	9.91038E-06	
GO:BP	defense response to bacterium	GO:0042742	1.30477E-05	
GO:BP	cell communication	GO:0007154	1.33747E-05	
GO:BP	signal transduction	GO:0007165	4.62854E-05	
GO:BP	response to bacterium	GO:0009617	8.22478E-05	
GO:BP	cell differentiation	GO:0030154	0.000132096	
GO:BP	cell adhesion	GO:0007155	0.000194183	
GO:BP	biological adhesion	GO:0022610	0.000230205	
GO:BP	cellular developmental process	GO:0048869	0.000259699	
GO:BP	keratinization	GO:0031424	0.000331074	
GO:BP	epidermal cell differentiation	GO:0009913	0.000362886	
GO:BP	G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger	GO:0007187	0.000362886	
GO:BP	regulation of transmembrane transport	GO:0034762	0.000362886	
GO:BP	biological regulation	GO:0065007	0.000400538	
GO:BP	inflammatory response	GO:0006954	0.000400538	
GO:BP	humoral immune response mediated by circulating immunoglobulin	GO:0002455	0.000400538	
GO:BP	regulation of ion transport	GO:0043269	0.000451217	
GO:BP	keratinocyte differentiation	GO:0030216	0.000664627	
GO:BP	regulation of ion transmembrane transport	GO:0034765	0.000944702	
GO:BP	regulation of cellular process	GO:0050794	0.000998044	
GO:BP	response to stimulus	GO:0050896	0.000998044	
GO:CC	plasma membrane	GO:0005886	2.21E-24	
GO:CC	cell periphery	GO:0071944	3.52E-24	
GO:CC	intrinsic component of plasma membrane	GO:0031226	6.58E-09	
GO:CC	integral component of plasma membrane	GO:0005887	3.39E-08	
GO:CC	intrinsic component of membrane	GO:0031224	1.36001E-06	
GO:CC	plasma membrane protein complex	GO:0098797	1.73042E-06	
GO:CC	integral component of membrane	GO:0016021	8.63394E-06	
GO:CC	extracellular region	GO:0005576	0.00012106	
GO:CC	receptor complex	GO:0043235	0.000138826	
GO:CC	intermediate filament	GO:0005882	0.000227515	
GO:CC	plasma membrane region	GO:0098590	0.000400338	

GO(Gene Ontology) enrichment analysis for 2173 switching genes in SK-MEL-5 was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

* MF, BP, CC represent Molecular Function, Biological Process, Cellular Component, respectively.

Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (SK-N-DZ, continued)

source	term_name	term_id	adjusted_p_value	negative_log10_of_adjusted_p_value
GO:MF	transmembrane signaling receptor activity	GO:0004888	0.000446565	
GO:BP	multicellular organismal process	GO:0032501	2.27025E-05	
GO:CC	no significant results			

GO(Gene Ontology) enrichment analysis for 768 switching genes in SK-N-DZ was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

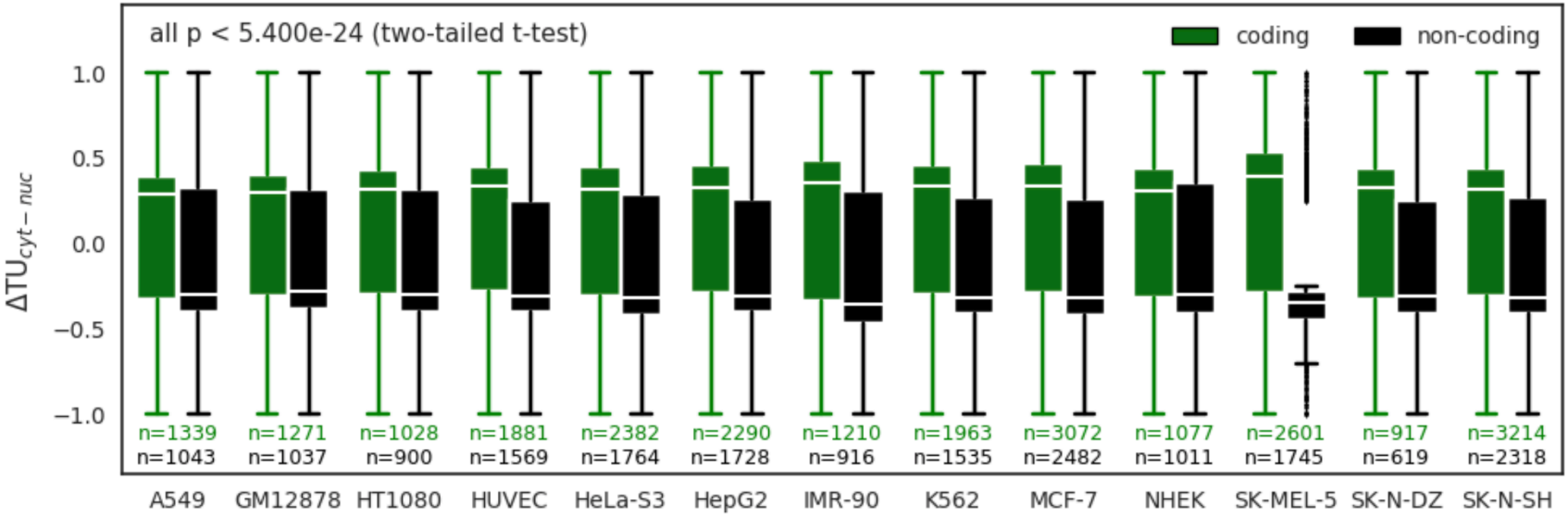
Supplementary Data 5. Gene Ontology enrichment analysis for switch genes (SK-N-SH)

source	term_name	term_id	adjusted_p_value	negative_log10_of_adjusted_p_value
GO:MF	transmembrane signaling receptor activity	GO:0004888	3.19E-16	
GO:MF	signaling receptor activity	GO:0038023	3.19E-16	
GO:MF	G protein-coupled receptor activity	GO:0004930	5.19E-16	
GO:MF	molecular transducer activity	GO:0060089	5.19E-16	
GO:MF	olfactory receptor activity	GO:0004984	2.89E-15	
GO:MF	antigen binding	GO:0003823	0.000131859	
GO:BP	detection of chemical stimulus	GO:0009593	3.47E-15	
GO:BP	detection of chemical stimulus involved in sensory perception	GO:0050907	8.52E-15	
GO:BP	G protein-coupled receptor signaling pathway	GO:0007186	8.77E-15	
GO:BP	sensory perception of chemical stimulus	GO:0007606	1.30E-14	
GO:BP	detection of chemical stimulus involved in sensory perception of smell	GO:0050911	6.08E-14	
GO:BP	sensory perception of smell	GO:0007608	2.32E-12	
GO:BP	nervous system process	GO:0050877	5.53E-12	
GO:BP	detection of stimulus involved in sensory perception	GO:0050906	8.87E-12	
GO:BP	system process	GO:0003008	1.37E-10	
GO:BP	sensory perception	GO:0007600	2.82E-10	
GO:BP	multicellular organismal process	GO:0032501	1.60E-09	
GO:BP	detection of stimulus	GO:0051606	2.96E-07	
GO:BP	keratinization	GO:0031424	5.33503E-06	
GO:BP	defense response to bacterium	GO:0042742	2.54298E-05	
GO:BP	adaptive immune response	GO:0002250	8.74689E-05	
GO:BP	signaling	GO:0023052	0.000829655	
GO:BP	humoral immune response mediated by circulating immunoglobulin	GO:0002455	0.000881868	
GO:CC	plasma membrane	GO:0005886	2.24E-19	
GO:CC	cell periphery	GO:0071944	2.11E-18	
GO:CC	extracellular region	GO:0005576	1.40E-08	
GO:CC	intrinsic component of membrane	GO:0031224	6.73E-08	
GO:CC	integral component of membrane	GO:0016021	7.91E-08	
GO:CC	membrane	GO:0016020	2.26958E-05	
GO:CC	plasma membrane region	GO:0098590	8.71218E-05	
GO:CC	receptor complex	GO:0043235	8.71218E-05	
GO:CC	extracellular space	GO:0005615	8.71218E-05	
GO:CC	cell surface	GO:0009986	8.72698E-05	
GO:CC	intrinsic component of plasma membrane	GO:0031226	0.000264325	
GO:CC	integral component of plasma membrane	GO:0005887	0.000455522	

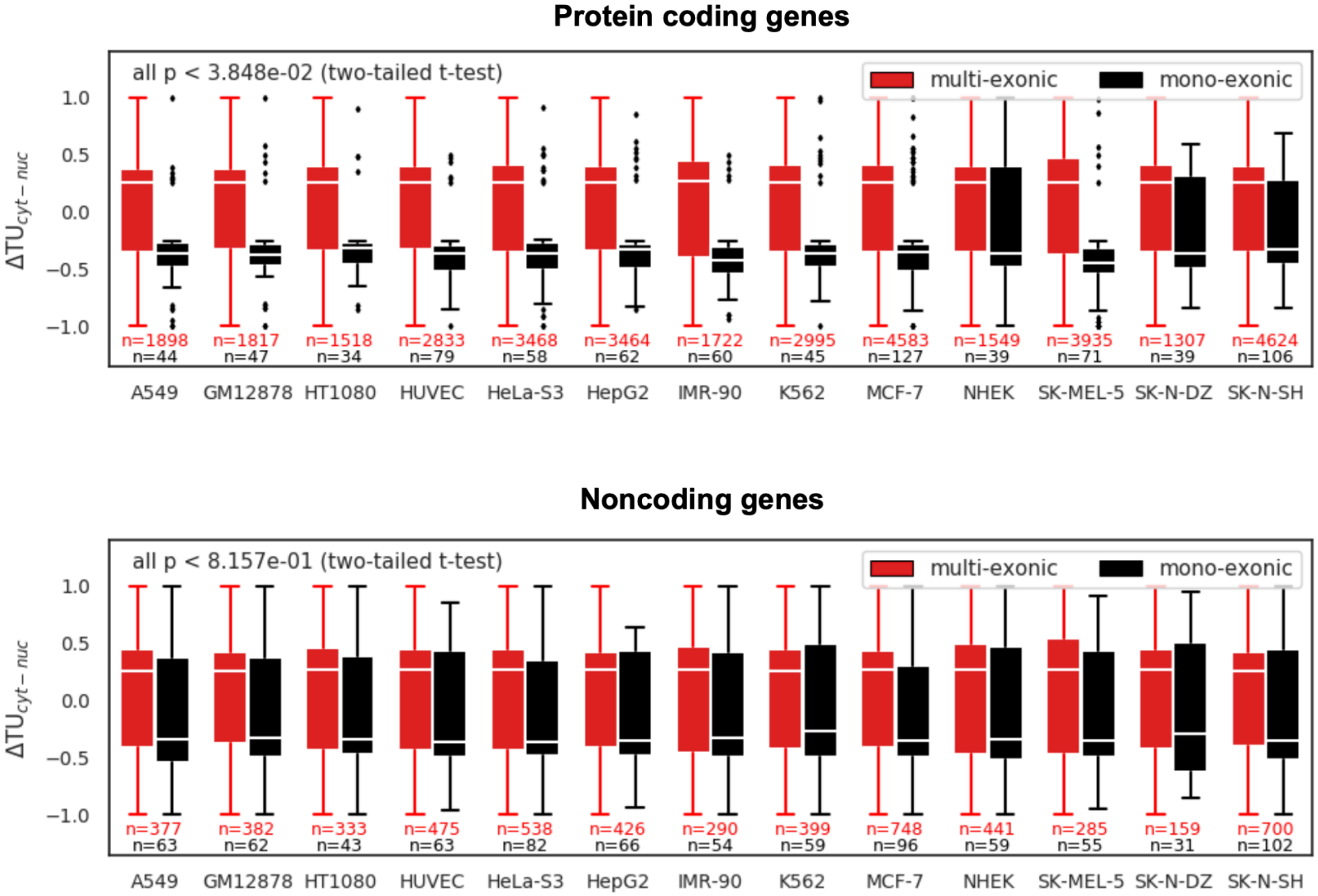
GO(Gene Ontology) enrichment analysis for 2766 switching genes in SK-N-SH was conducted by g:Profiler(PMID: 31066453) with following parameters:

- Hypergeometric test
- Benjamini–Hochberg correction
- FDR(False Discovery Rate) < 0.001

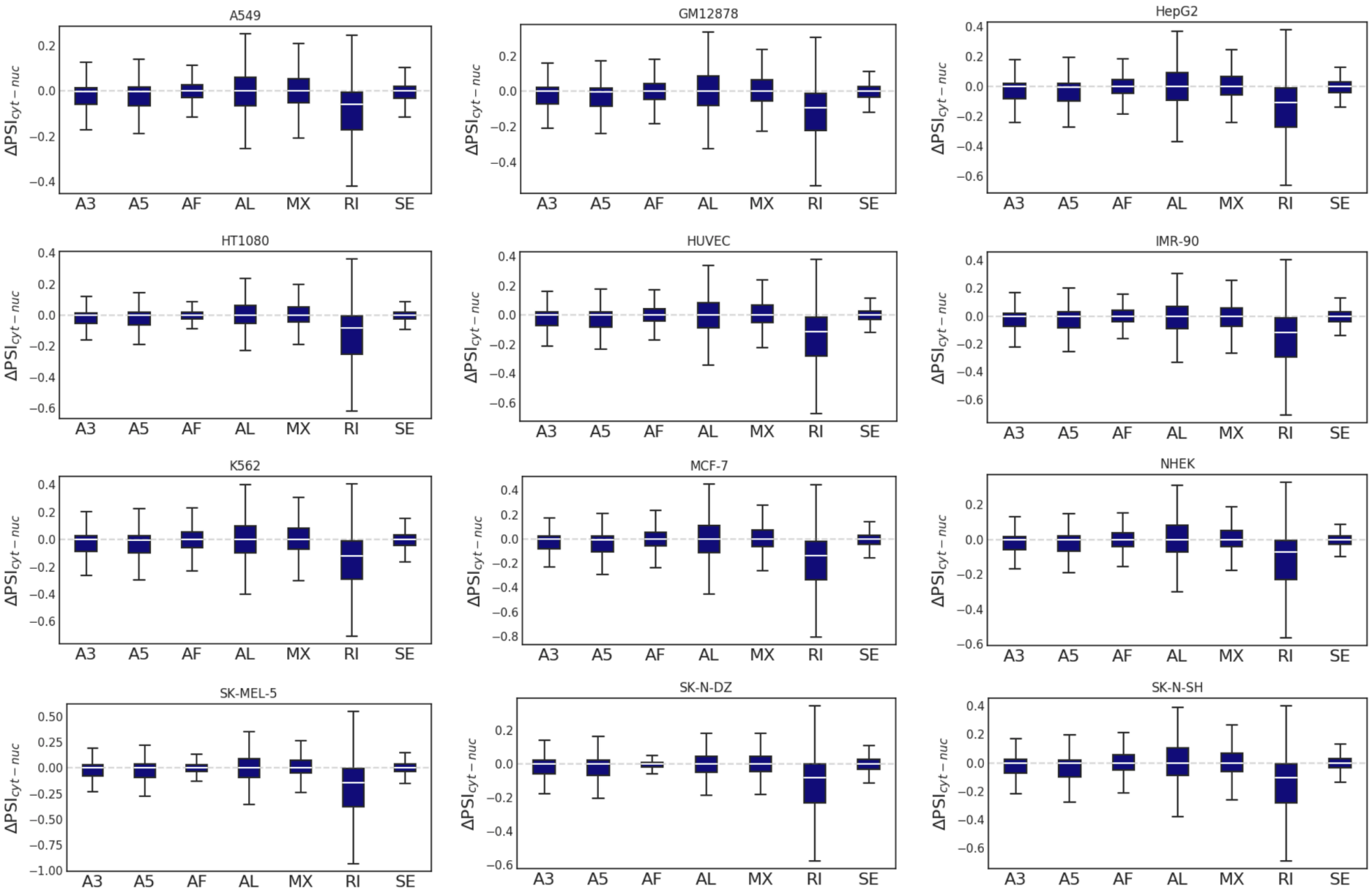
* MF, BP, CC represent Molecular Function, Biological Process, Cellular Component, respectively.



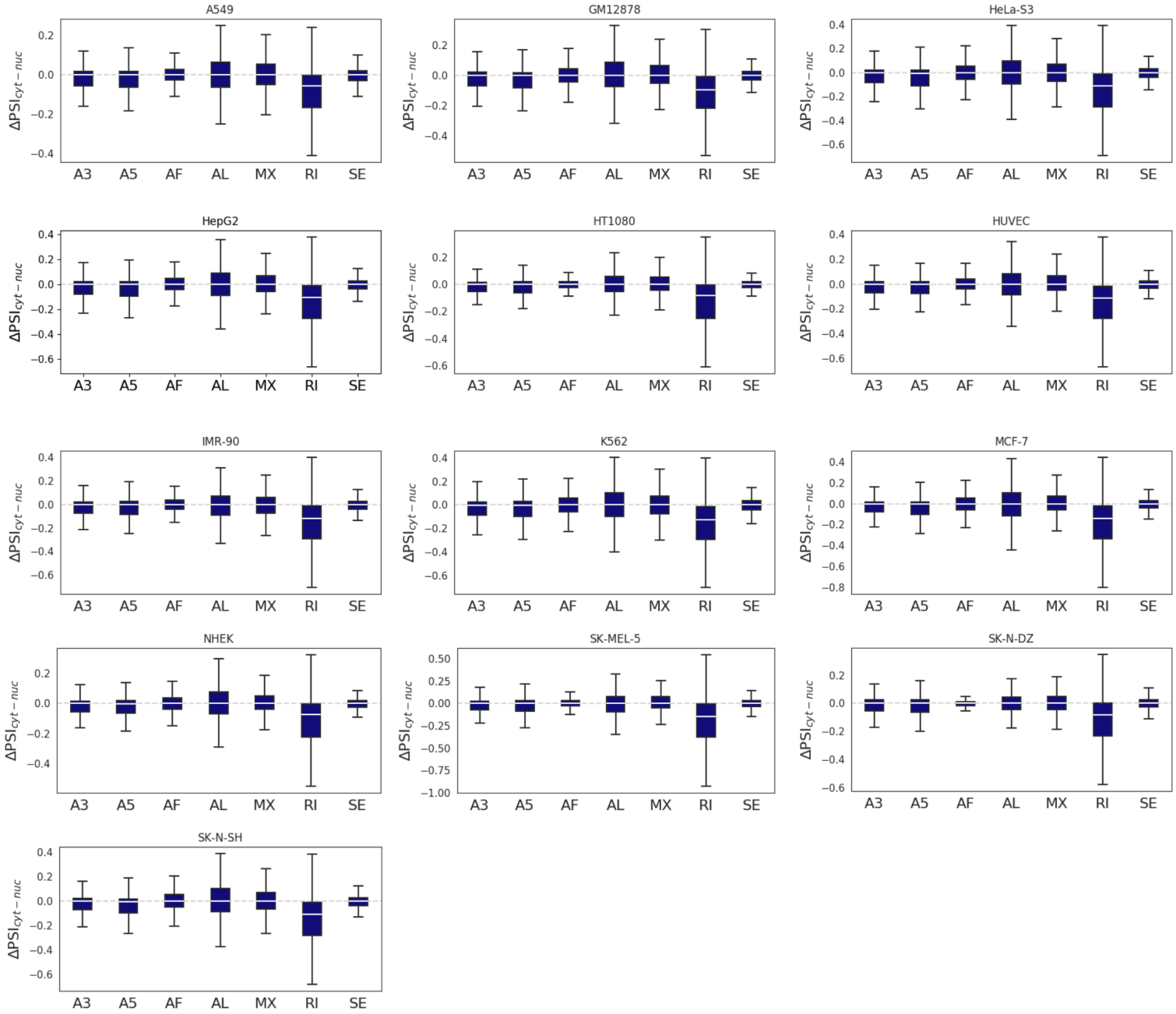
Supplementary Data 6. Comparison of ΔTU between protein-coding transcripts (green) and non-coding transcripts (black). Protein-coding transcripts possess higher ΔTU scores, which is consistent with our understanding that the transcript encoding protein is more prone to be located in the cytoplasm. Based on two-tailed t-test, we calculated the significant level of difference between ΔTU in protein-coding and non-coding transcripts in each cell line. The largest p value is 5.400e-24.



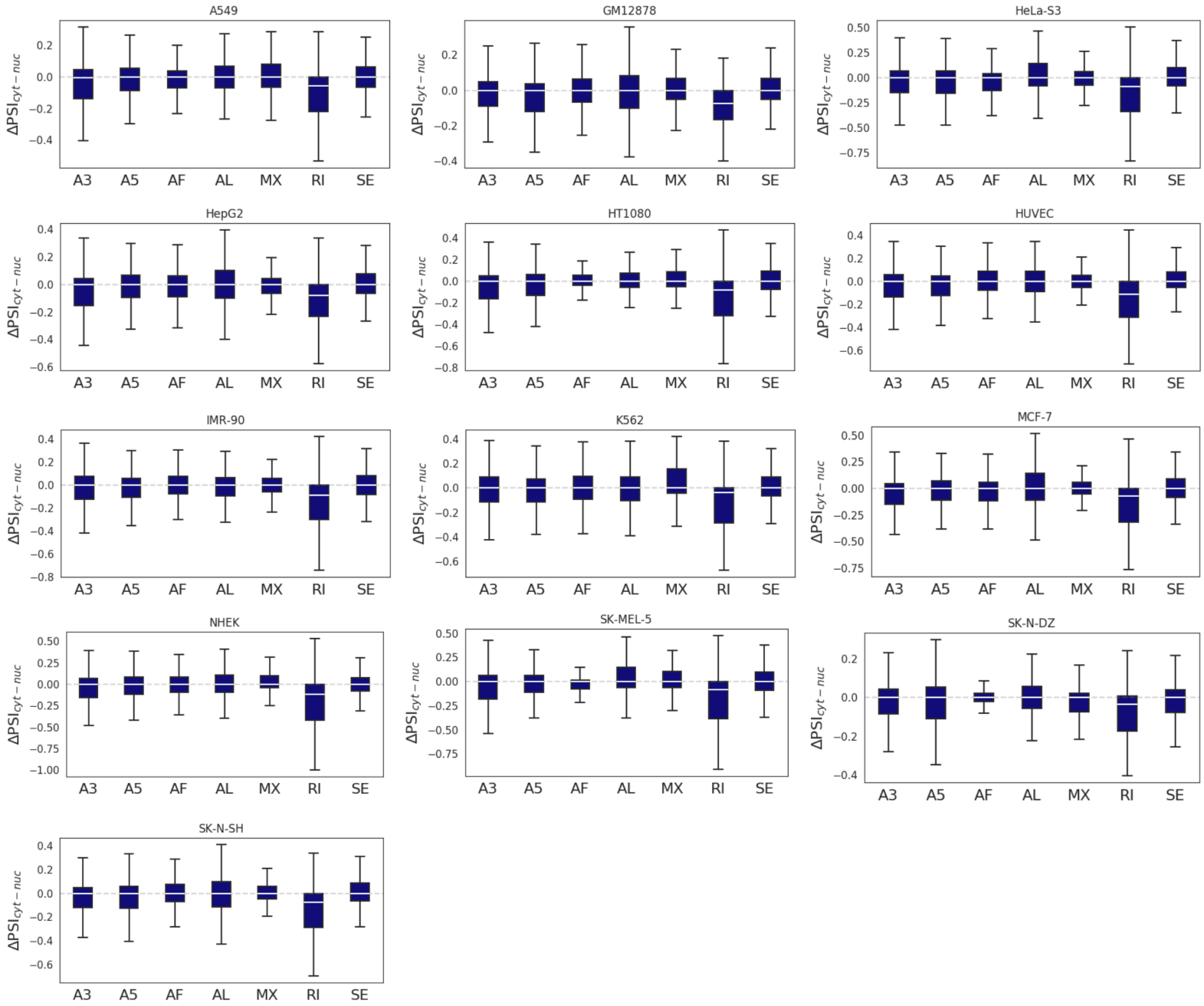
Supplementary Data 7. Comparison of ΔTU between mono-exonic (black) and multi-exonic (red) transcripts across 13 cell lines. Top and bottom panels for protein-coding and noncoding genes, respectively. ΔTU shows a significant positive correlation with splicing, indicating that splicing appears to be a dominant factor for RNA export from the nucleus. Based on two-tailed t-test, we calculated the significant level of difference between ΔTU in multi-exonic and mono-exonic transcripts in each cell line. The largest p value is 3.848e-02 and 8.157e-01 for protein-coding and non-coding genes, respectively.



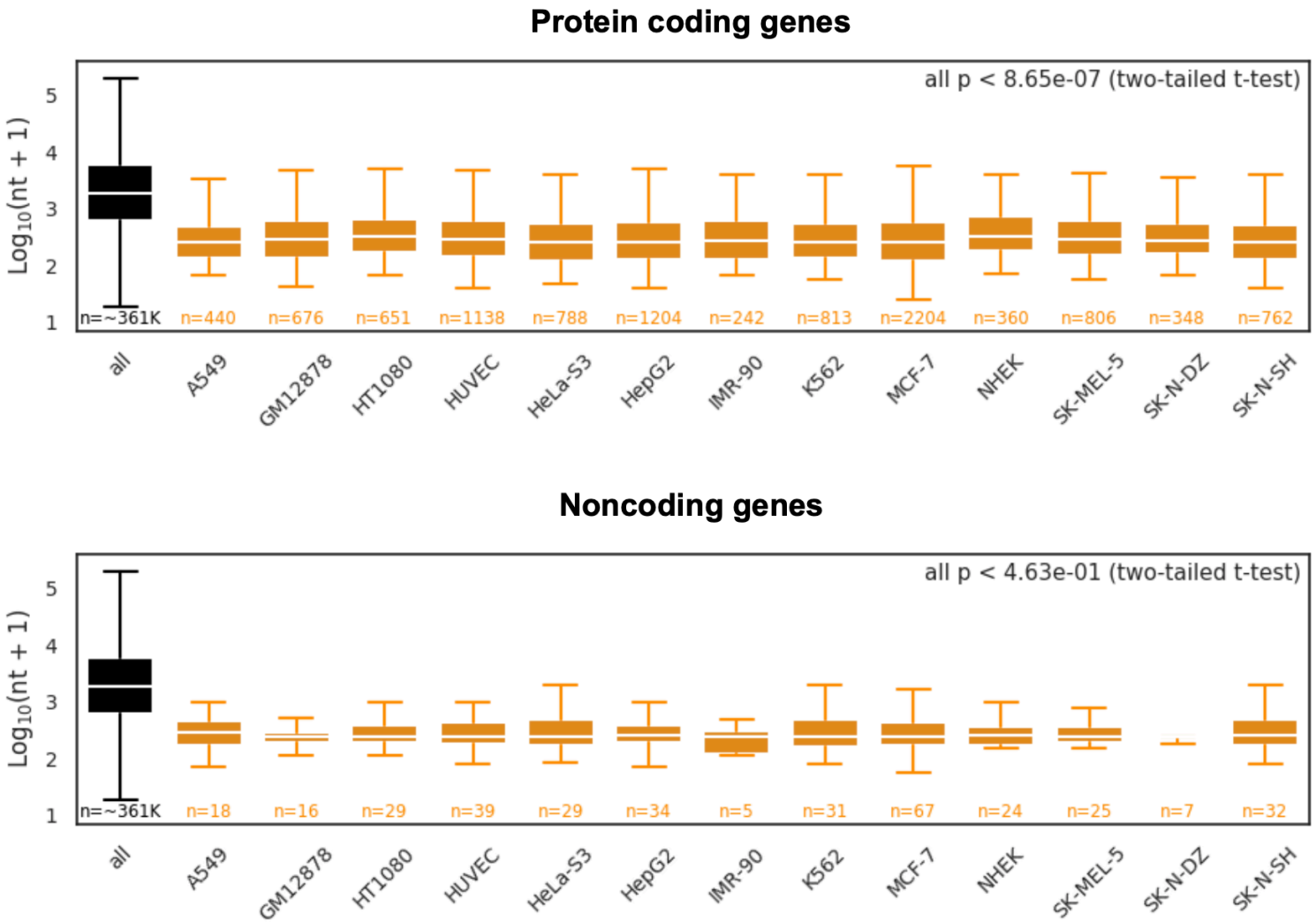
Supplementary Data 8. Comparison of alternative splicing patterns between cytoplasmic and nuclear transcripts. (all genes)



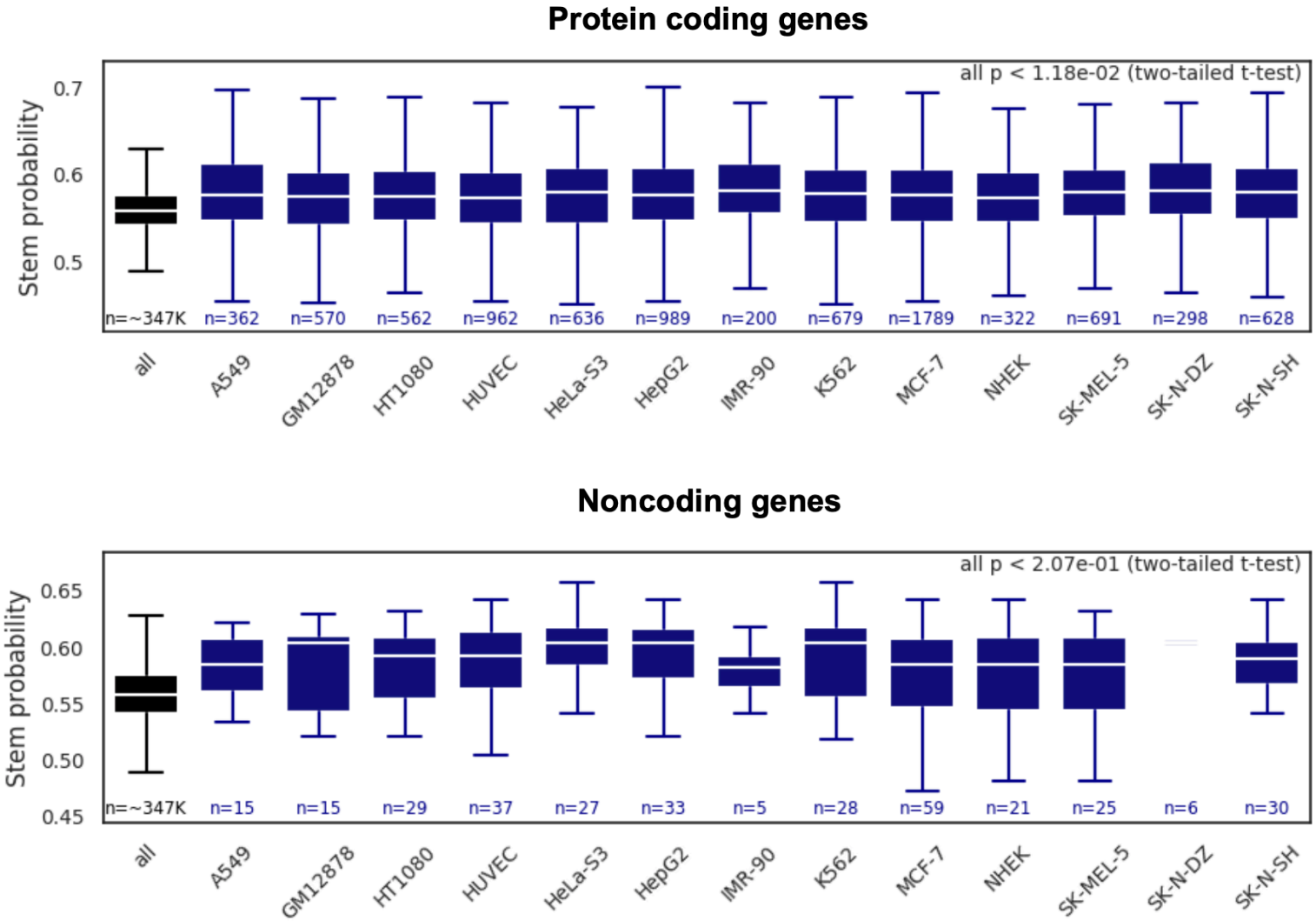
Supplementary Data 9. Comparison of alternative splicing patterns between cytoplasmic and nuclear transcripts. (protein-coding genes, continued)



Supplementary Data 9. Comparison of alternative splicing patterns between cytoplasmic and nuclear transcripts. (noncoding genes)



Supplementary Data 10. Comparison of length between all introns and nuclear RIs (ΔΨ<0 and p<0.05). Top and bottom panels for protein-coding and noncoding genes, respectively. P values: two-tailed t-test between all introns and nuclear RIs. The largest p value is 8.65e-07 and 4.63e-01 for protein-coding and non-coding genes, respectively.



Supplementary Data 11. Comparison of RNA secondary structure (stem probability) between all introns and nuclear RIs (ΔΨ<0 and p<0.05). Top and bottom panels for protein-coding and noncoding genes, respectively. P values: two-tailed t-test between all introns and nuclear RIs. The largest p value is 1.18e-02 and 2.07e-01 for protein-coding and non-coding genes, respectively.