RNA BIOLOGY
2023, VOL. 20, NO. 1, 563–572
https://doi.org/10.1080/15476286.2023.2242649

Taylor & Francis
Taylor & Francis Group

BRIEF COMMUNICATION

OPEN ACCESS Check for updates

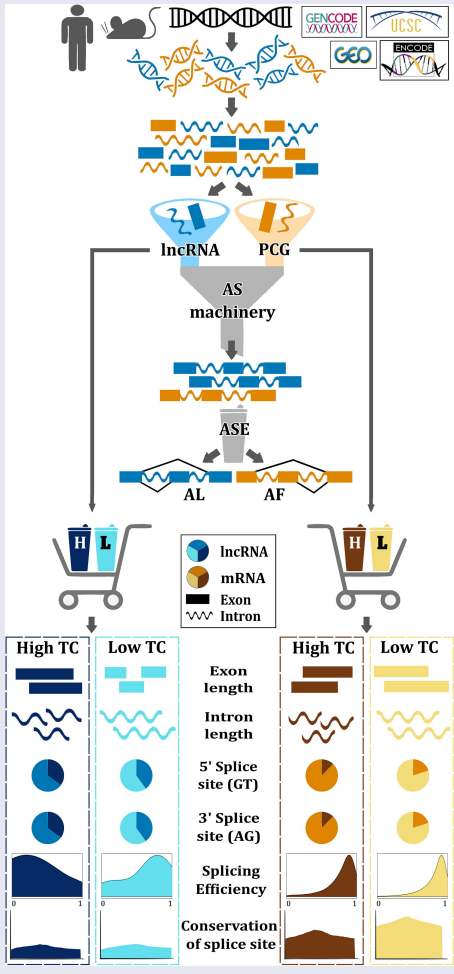
Inefficient splicing of long non-coding RNAs is associated with higher transcript complexity in human and mouse

Koushiki Basu, Anubha Dey, and Manjari Kiran

Department of Systems and Computational Biology, School of Life Sciences, University of Hyderabad, Hyderabad, India

ABSTRACT

Recent reports show that long non-coding RNAs (lncRNAs) have inefficient splicing and fewer alternative splice variants than mRNAs. Here, we have explored the efficiency of lncRNAs and mRNAs in producing various splice variants, given the number of exons in humans and mice. Intriguingly, lncRNAs produce more splice variants per exon, referred to as Transcript Complexity, than mRNAs. Most lncRNA splice variants are the product of the alternative last exon and exon skipping. LncRNAs and mRNAs with higher transcript complexity have shorter intron lengths. Longer exon length and GC/AG at 5′/3′ splice sites are associated with higher transcript complexity in lncRNAs. Lastly, our results indicate that inefficient splicing of lncRNAs may facilitate multiple introns splicing and, thus, more spliced products per exon.



ARTICLE HISTORY
Revised 23 July 2023
Accepted 26 July 2023

KEYWORDS
Long non-coding RNA; messenger RNA; transcript complexity; alternative splicing; splice variant

CONTACT Manjari Kiran manjari.hcu@uohyd.ac.in Department of Systems and Computational Biology, School of Life Sciences, University of Hyderabad, Hyderabad, Telangana 500 046, India

Supplemental data for this article can be accessed online at https://doi.org/10.1080/15476286.2023.2242649.

Introduction

Over the years, lncRNAs gained attention and from being considered junk became an essential regulatory layer of the transcriptome. lncRNAs are significant players in transcriptional regulation, guiding molecules, alternative splicing, protein translation, chromatin modifications, and spatial conformation [1,2]. Apart from regulatory functions, lncRNAs are the cornerstone of innumerable disorders of complex aetiology, such as cancer, diabetes, autoimmune disorders, and many more [3,4]. The deregulation of lncRNAs is associated with the development and progression of various cancer types, making them suitable as biomarkers for cancer diagnosis and prognosis [5,6].

Although most lncRNAs and mRNAs are transcribed by RNA Polymerase II and undergo the same RNA processing steps, i.e. capping, splicing, and polyadenylation, lncRNAs show lower expression levels and higher tissue-specificity compared to mRNAs [7,8]. Alternative splicing increases the transcriptome diversity of both lncRNAs and mRNAs. Based on some initial investigations, lncRNAs are inefficient in splicing compared to mRNAs and have a lower number of spliced products (transcripts) [9–12]. Recently, splicing efficiency has been measured by the number of RNA-seq reads mapped to spliced versus unspliced transcripts [10]. As reported in several studies, different organizational and structural features such as average intron length, 5' and 3' splice site dinucleotides, splice site strength and evolutionary conservation, higher thymidine content in the polypyrimidine tract, and consensus sequence of branch points may explain the lower number of spliced variants of a gene [10,12–14]. Also, high epigenetic regulation (e.g. H3K9me3 histone modification), lower interaction with splicing factors (e.g. U2AF65 binding), fewer SR proteins binding sites, and absence of RNA polymerase II phosphorylation over 5′ splice site reduce the efficiency of alternative splicing [10,12,14–16].

In this study, we explored the 'Transcript Complexity' (TC) mRNAs and lncRNAs considering the number of transcripts (splice variants) per exon. Also, we investigated how different structural, organizational, and lncRNA-specific features contribute to the higher TC of lncRNAs. We have used gene annotations from the GENCODE project to calculate TC [17]. The results are compared for transcripts with different Transcript Support Levels (TSLs) to confirm that the findings are not because of poorly supported transcript models. Strikingly, all the analyses revealed that the TC of lncRNAs is significantly higher than mRNAs. To further validate our observation, the study was repeated using annotation data from other databases and removing the targets of nonsense-mediated mRNA decay (NMD). We have compared different factors involved in TC and how they are distinct in human and mouse lncRNAs and mRNAs. The analysis of splicing features revealed a significant distinction between alternative splicing events in lncRNAs and mRNAs. We observed that lncRNAs with high TC are associated with shorter intron length, longer exon length, GT > GC at 5' splice site, low splice site strength, and splicing inefficiency. Finally, using linear regression, we found that exon length and 5' splice site dinucleotide are the determinant features of TC for lncRNAs.

Materials and methods

Data collection

The list of protein-coding mRNAs and lncRNAs is collected from the GENCODE. For humans, data from the release v38 annotated on the genome sequence GRCh38 (gencode.v38.basic.annotation.gtf.gz) and for mouse release M28 annotated on the genome sequence GRCm39 (genocode.vM28.basic.annotation.gtf.gz) was used. Protein-coding mRNAs and lncRNAs were selected from the basic annotation file when the gene, transcript, and exon were indicated as 'protein_coding' and 'lncRNA', respectively. Table S1 shows the total number of genes, transcripts, and exons extracted from the annotation file and Transcript Per Gene (TPG) and Exon Per Gene (EPG) values were calculated for both species. TPG is calculated by dividing the total number of transcripts by the total number of genes. Similarly, EPG is calculated by dividing the total number of exons by the total number of genes. Also, mRNAs and lncRNAs are categorized based on the Transcript Support Level data for some analyses. Transcript Support Levels indicate how well supported a transcript model is based on mRNA and EST alignments from UCSC and Ensembl [18]. The total number of genes with different TSLs is included in Table S3.

$$TPG = \frac{Total\ number\ of\ Transcripts}{Total\ number\ of\ Genes}$$

$$EPG = \frac{Total\ number\ of\ Exons}{Total\ number\ of\ Genes}$$

The validation of the results obtained by using gene annotation from GENCODE release v38 for humans was obtained by repeating the analysis with annotation data from three different databases: the Refseq database at NCBI (GRCh38_latest_genomic.gtf) [19], the Ensembl database (Homo_sapiens.GRCh38.109.gtf) [20] and the T2T Consortia data from GENCODE (gencode.v43.basic.annotation.txt) [21] (Table S11).

Transcript complexity (TC) calculation

As mentioned, a gene's TC is calculated as the number of transcripts per exon. For the calculation of TC for each gene, the number of transcripts and exons are extracted from basic annotation datasets (v38 for humans and vM28 for mice).

$$TC = \frac{Number\ of\ Transcripts}{Number\ of\ Exons}$$

Nonsense-mediated mRNA decay data

The enhanced UV crosslinking and immunoprecipitation followed by high-throughput sequencing data for NMD analysis is obtained from the eCLIP experiment on HepG2 against UPF1 (GSE177267) [22]. The data was annotated using annotatePeaks.pl command from the HOMER package [23] and 'protein_coding' genes were filtered out as the targets of the NMD pathway. These protein-coding mRNAs were then removed from the basic annotation data for humans resulting in 18,003 mRNAs. The TC of lncRNAs and mRNAs (– NMD targets) was compared to show that the unstable transcripts did not affect the result.

Alternative splicing analysis

The splicing event for each gene was characterized from the GTF files of mRNA and lncRNA annotations from the GENCODE database and the SUPPA tool [24]. The SUPPA tool assigns alternative splicing events involving GC-AG and GT-AG introns. The alternative splicing events (ASE) are classified into the following types using the SUPPA tool: skipping exons (SE), alternative 5'/3' splice site (A5/A3 or SS if both are considered together), mutually exclusive exons (MX), retained intron (RI) and alternative first/last exon (AF/AL or FL if both are considered together).

Exon and intron sequence analysis

Exon and intron sequences were retrieved using the Table Browser tool from UCSC using human GRCh38 and mouse GRCm39 sequences [25]. All single-exon genes are excluded from intron analysis as they are not subjected to splicing. For each transcript, the total intron length and exon length are calculated. From each intron sequence, 5' splice site and 3' splice site dinucleotide were extracted. The splice site strength of humans was computed using the MaxEntScan web tool, which generalizes most prior probabilistic models of sequence motifs and is established on the 'Maximum Entropy Principle [26]. MaxEntScan web tool considers adjacent and non-adjacent dependencies between positions and predicts the strength of the splicing sequences.

MaxEntScan:score5ss takes the FASTA file consisting of sequence motifs of 9 nucleotides (3 bases in exon and 6 bases in intron) as input and scores the 5' splice site. Similarly, MaxEntScan:score3ss was used to score the 3" splice site, and each sequence motif in the FASTA file was 23 nucleotides long (20 bases in intron and 3 bases in exon). The strength of the 5'and 3" splice site was computed using all the provided scoring models (Maximum Entropy Model (MAXENT), Maximum Dependence Decomposition Model (MDD), First-order Markov Model (MM), and Weight Matrix Model (WMM)). The FASTA file of the sequence motifs used as input for the MaxEntScan tool is obtained using the "GencoDymo" R package version 0.2.1 [27].

Conservation analysis

The conservation of each intron's 5' splice sites and 3' splice sites were evaluated using phastCons data from UCSC [28]. PhastCons score for multiple alignments of 99 vertebrate genomes to the human genome (hg38.phastCons100way.bw) and 34 genomes to the mouse genome (mm39.phastCons35way.bw) was downloaded. For each intron sequence, the position of the 5' and 3' splice sites were extracted from the intron sequence data (from UCSC), and the phastCons conservation score was assigned for each position. Further, the score was assigned for 10 nucleotides upstream and downstream for each 5'and 3' splice site.

Splicing efficiency analysis

The high-throughput sequencing data for expression profiling of human coding and non-coding genes were retrieved from the GEO database for splicing efficiency analysis [29,30]. The BAM files of the HEK293 cell line labelled with 4-thiouridine for 15 min (GSM2257731), 30 min (GSM2257734), 45 min (GSM2257737), and 60 min (GSM2257740) were considered for the analysis. For each alignment file, the splicing efficiency of the individual intron was quantified using the SPLICE-q tool [31]. SPLICE-q takes the BAM file with RNA-seq reads and the GTF file of the reference genome as input and annotates the introns and the splice junctions of the data. It then selects the split and unsplit junction reads and computes splicing efficiency based on the reads' coverage and concise idiosyncratic gapped alignment report (CIGAR). The formula used by SPLICE-q to calculate splicing efficiency is-

$$SE_i = \frac{\sum_{j\in\{5',3'\}} S_i^j}{\sum_{j\in\{5',3'\}} (S_i^j + N_i^j)}; 0 < SE_i < 1$$

where, i is intron and S and N represent the split and unsplit reads for the 5' and 3' splice sites.

Statistical analysis

R version 4.2.1 was utilized for data analysis and statistics. The Wilcoxon rank-sum test was used to compare the median of TC for lncRNAs and mRNAs. The same test was also used to compare the median of the number of exons, intron length, exon length, and 5' and 3' splice site strength of lncRNAs and mRNAs with low and high TC (Table S1, S4, S6, S9, S11). The Kolmogorov–Smirnov test was performed to test the difference in cumulative distribution function of the splice site strength of lncRNAs and mRNAs with low and high TC (Table S1, S9, S10, S11). The correlation between the number of transcripts and the number of exons; TC and the number

of exons; TC and splicing efficiency of lncRNA and mRNA was calculated using Spearman's rank correlation test (Table S1, S2, S10).

Linear model for the determinant feature of TC

The data for the linear model includes all the features of TC: exon length, intron length, 5' and 3' splice site dinucleotide, conservation score, and strength of 5' and 3' splice site. Since the PhastCons score data from UCSC [25] is unavailable for most genes, they are excluded from the final dataset. Each feature and different combinations of features were tested against TC using the caret R package [32], and linear regression with the base R function lm was used to determine R^2 values. The dataset was divided into 75%–25% for training and testing. The sign of correlation was determined with the cor function. We performed 1000 permutations of splitting the data into training and testing datasets to obtain the distribution of R^2 for each feature and model.

Results and discussions

Positive correlation between the transcript complexity and number of exons

Similar to mRNAs, lncRNAs are also composed of exons and introns. As previously reported, we also found that the TPG for mRNAs is 1.5 times more than lncRNAs in humans and mice. Similarly, the calculated EPG for mRNA is much higher compared to lncRNA in both species (Table S1). Since most genes have 1 to 4 transcripts, the reports till date are mostly on genes with 1–4 transcripts. However, some genes may produce more transcripts, given the number of exons. We measured the splice variant efficiency by counting the number of transcripts reported for a gene for the number of exons. We referred to this as a measure of the gene's TC. We first compared the TC of mRNA and lncRNA genes based on the number of transcripts per exon. As expected, there exists a positive correlation between the number of transcripts and the number of exons for both lncRNAs and mRNAs in humans ($\rho_{lncRNA} = 0.57$, $\rho_{mRNA} = 0.62$) and mice ($\rho_{lncRNA} = 0.52$, $\rho_{mRNA} = 0.47$) (Spearman correlation test, p-value <2.2e-16) (Figure S1 and Table S2). It was observed that with an increasing number of exons, the number of transcripts increases exponentially for lncRNAs and mRNAs (Figure S1B). The results remain the same for genes with different Transcript Support Levels and with different annotation datasets (Table S2, S11, Figure S4A).

The transcript complexity of lncRnas is higher than protein-coding genes

The distribution of TC shows a slight bimodal distribution for both lncRNAs and mRNAs. The median TC of lncRNAs is higher than mRNAs in both humans ($median_{lncRNA} = 0.5$, $median_{mRNA} = 0.25$) and mice ($median_{lncRNA} = 0.33$, $median_{mRNA} = 0.21$) (Wilcoxon rank-sum test, P-value <2.2e-16) (Figure 1A and Table S2, S4). The results remain the same for genes with different Transcript Support Levels and with different annotation datasets (Table S2, S4, S11, and Figure S2, S4B). The median TC shows that the transcript complexity of lncRNAs is higher than the mRNAs even after removing NMD-targeted mRNAs (Figure S3).

Alternative splicing events in long non-coding RNAs and protein-coding genes

As alternative splicing events (ASE) play an important role in increasing transcriptomic diversity from a single gene, we next sought to test ASE contributing to higher TC for lncRNAs than mRNAs. It is observed that for both humans and mice lncRNAs, most of the transcripts are generated by Alternative Last Exons (AL), whereas for mRNAs, it is Alternative First Exons (AF) (Figure 1B). After Alternative First Exons, mRNAs' most prevalent alternative splicing events are either AL for humans or Skipping Exons (SE) for mice. In contrast, the second prevalent alternative splicing event for lncRNAs is SE, followed by AF for humans and mice (Figure 1B). Although, AF is the most prevalent ASE for mRNA in different annotation dataset, SE is the most prevalent for lncRNA (Figure S4C). ASE does not affect the EPG values in humans and mice as there is no difference in median of the number of exons and the calculated EPG values for different ASE (Table S5). The lncRNAs and mRNAs are further divided based on median TC into two groups (please see *Methods and Materials*). It is observed that for mRNAs with either low or high TC, most of the transcripts are generated by Alternative First Exons (AF) in both humans and mice. On the contrary, lncRNAs with high TC generate most of the transcripts through Alternative Last Exons (AL), and the most prevalent ASE for lncRNAs with low TC is Skipping Exons (SE) in both species (Figure S5B). Thus, AL is the most common type of ASE in lncRNAs and AF in mRNAs with high TC.

Shorter intron lengths and longer exon lengths are associated with higher transcript complexity

As reported earlier, average exon and intron length may explain the difference in the number of splice variants and splicing efficiency in lncRNAs and mRNAs [12]. Thus, exon and intron length may also contribute to the TC of the gene. It was observed that mRNAs have higher median exon and intron lengths than lncRNAs for humans and mice (Table S6). To better understand features contributing to higher TC, we divided lncRNAs and mRNAs into two groups (low and high TC) based on the median TC of all the genes (0.33 for humans and 0.25 for mice). It is observed that genes (lncRNAs/mRNAs) with higher TC have shorter intron lengths than genes with lower TC in humans and mice (Wilcoxon rank-sum test, $p < 2.2e\text{-}16$). However, lncRNAs with higher TC have longer exons than lncRNAs with lower TC in both species (Wilcoxon rank-sum test, $p < 2.2e\text{-}16$) (Table S6). Thus, there exists

a strong association between intron length and TC, irrespective of the coding potential of the gene. However, longer exon length is associated with lncRNAs with high TC, whereas an opposite trend is observed in the case of mRNAs (Figure 2A).

GT > GC splice site and poor splice site strength are associated with higher transcript complexity

Several splice site-specific features, such as dinucleotide at 5' and 3' splice sites and splice site strength, can affect the alternative inclusion or exclusion of particular exons, thereby affecting TC. The introns of both lncRNAs and mRNAs are divided into groups based on low and high TC, and the percentage of introns in each group is calculated for both humans and mice. It is perceived that most introns have GT and AG as the 5'and 3' splice sites, respectively, for humans and mice. We found a high fraction of introns with GT at the 5' splice site and AG at 3' for low TC genes (lncRNAs/mRNAs) compared to high TC genes in humans and mice. GC at the 5' splice site is another predominant dinucleotide for lncRNAs introns for both low and high TC (Figure 2A and Table S7, S8). Reports suggest that the GC splice site is an intrinsically weak donor site and is present in only 1% of human introns [33,34]. Efficient selection of GC over GT at the 5' splice site is associated with some strong exonic splicing enhancers and weak splicing silencers [34].

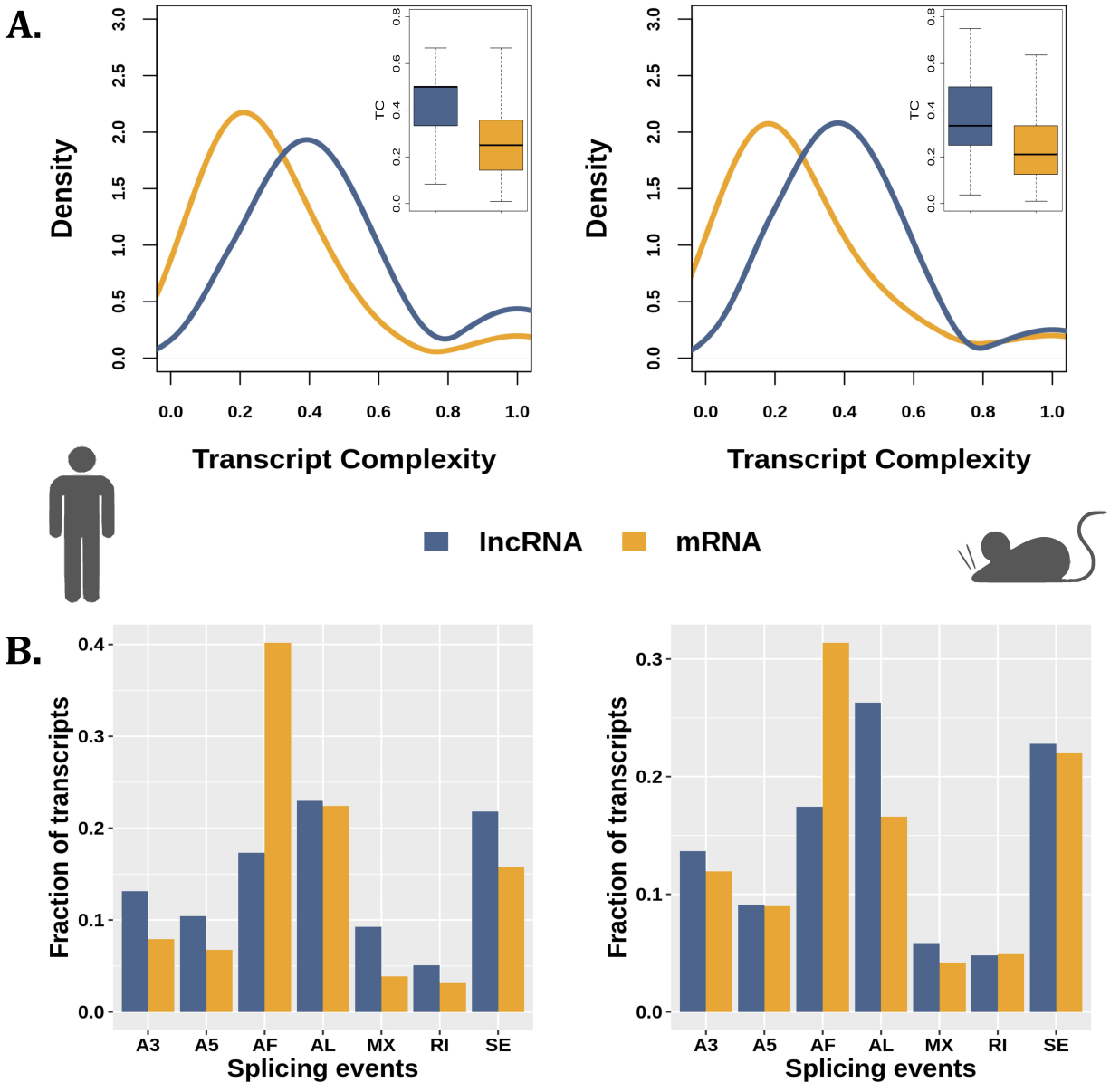


Figure 1. (A) Distribution of Transcript Complexity (TC) for lncRnas and mRnas in humans and mice and (B) Distribution of the fraction of transcripts generated by different alternative splicing events for lncRnas and mRnas in humans and mice.

We next compared the strength of regions near 5' and 3' splice sites. MaxEntScan tool was used to assign the computationally predicted strength score to 5' and 3' splice sites for humans. For the donor splice site, MAXENT, MDD, MM, and WMM scoring models were used. It is observed that for each choice of scoring model, lncRNA with high TC have the lowest splice site strength, followed by lncRNA with low TC and mRNAs (Wilcoxon rank-sum test <9.782e-05 and Kolmogorov–Smirnov test < 1.166e-10) (Figure 2B and Table S9). The result corroborates with previously reported inefficiency of splicing in lncRNAs compared to mRNAs [11]. Similarly, for the acceptor splice site, MAXENT, MM, and WMM were used, and a similar trend of splice site strength was observed (Wilcoxon rank-sum test <1.589e-05 and Kolmogorov–Smirnov test <2.642e-14) (Figure 2B and Table S9).

In contrast, there is almost no difference between the strength of mRNAs with low and high TC near 5' splice sites for MM and WMM scoring models and 3' splice sites for MAXENT scoring models. The low strength of the lncRNA splice site regions can affect the splicing in two possible ways i) by affecting the splicing silencer or enhancer and ii) by selecting a non-canonical splicing factor and thus promoting different inclusion and exclusion events. The reduced splicing efficiency of lncRNAs has been associated with low 5' splice site strength [14]. The inefficiency in lncRNAs splicing may allow different non-canonical splicing events, thereby, more TC.

Low multi-vertebrate conservation of splice sites in introns of mRNA with high transcript complexity

Previous reports indicate that introns having a higher conservation score of splice sites have a low frequency of alternative splicing [35]. Thus, the conservation score of splice sites in introns should be inversely associated with the TC of a gene. We noticed that for both 5' and 3' splice sites, lncRNAs having low and high TC have low conservation scores and are almost constant across upstream and downstream of the splice site position in both species. This observation can be because of unavailability of the PhastCons score downloaded from UCSC for most of the genes. For mRNAs, it is observed that genes with high TC have low conservation scores than genes with low TC at both 5' and 3' splice sites in humans and mice (Figure 3A

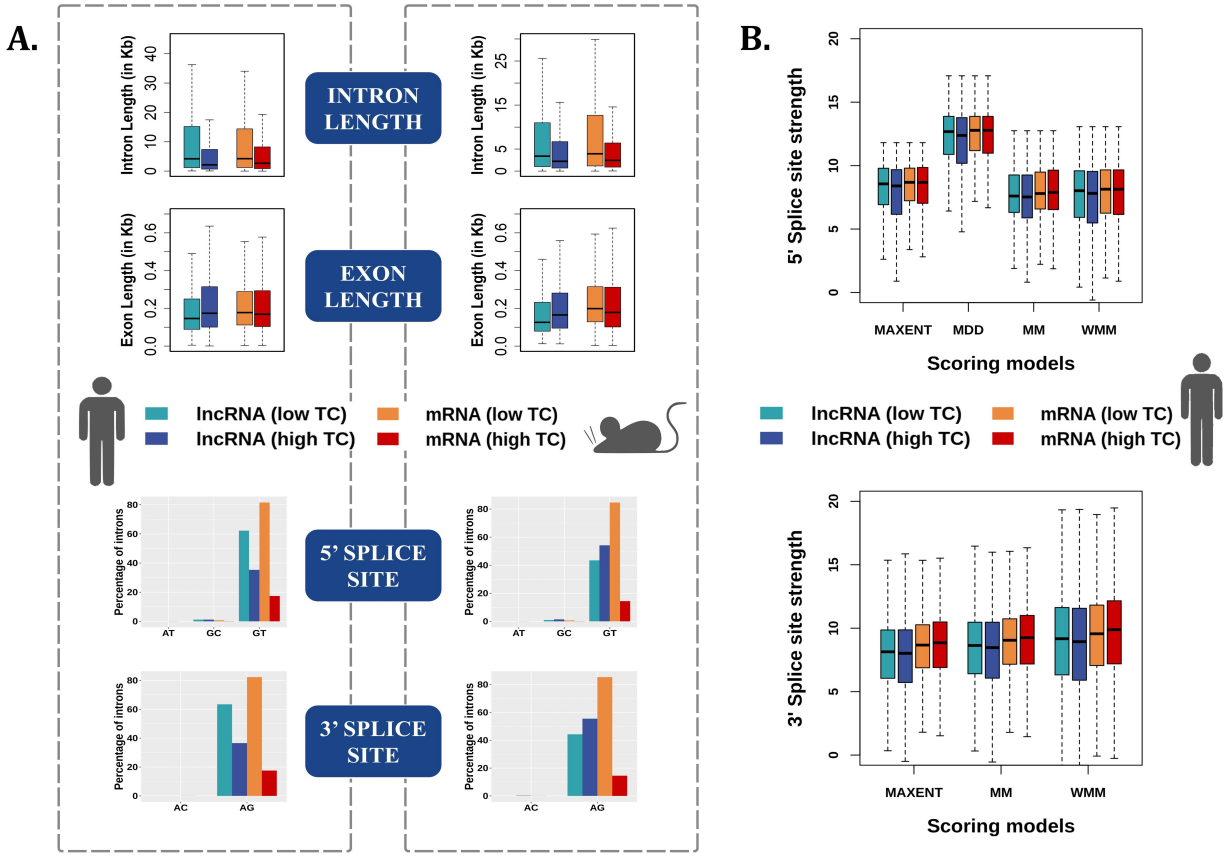


Figure 2. (A) Distribution of intron length, exon length, 5' splice sites, and 3' splice sites in lncRnas and mRnas with low and high TC in both human and mouse (B) 5' splice site strength (top) and 3' splice site strength (bottom) of lncRnas and mRnas with low and high TC for human.

and Figure S6). Taking a clue from this, low conservation or strength of splice sites for any gene may contribute to the flexibility of exon to be spliced in or out and thus may contribute to more TC.

Low splicing efficiency of long non-coding RNAs is associated with high TC

Since, the splicing inefficiency of long non-coding RNAs has been linked with low 5' splice site strength, we next examined if the high TC of lncRNAs is due to its splicing inefficiency. We used the nascent RNA dataset available for the HEK293 cell line labelled with 4-thiouridine for 15, 30, 45, and 60 min (please see *Methods and Materials*). Surprisingly, we found a negative correlation between the TC of lncRNAs and their splicing efficiency (fraction of reads mapped on spliced vs. unspliced forms) (for 15 min: -0.592_{lncRNA} vs -0.07_{mRNA}; for 30 min: -0.635_{lncRNA} vs -0.14_{mRNA}; for 45 min: -0.581_{lncRNA} vs -0.132_{mRNA}; for 60 min: -0.695_{lncRNA} vs -0.097_{mRNA}; P value < 2.2e-16). Overall, the distribution of splicing efficiency of lncRNAs is lower than mRNAs for all time points (P value < 2.2e-16). Among the two categories in high and low TC in lncRNAs and mRNAs, lncRNAs with high TC have the least splicing efficiency, and mRNAs with low TC have the highest splicing efficiency (Figure 3B, Figure S7). The splicing efficiency of lncRNAs is lower than mRNA genes with matched numbers of exons (Figure S7). Splicing efficiency shows a reverse pattern to TC (Figure S1B, Figure S7, and Table S10), suggesting a strong association of the splicing efficiency of a gene with its TC.

Exon length and 5' splice site predict the transcript complexity for lncRnas

To find the determinant feature of TC, we first evaluated the individual features: exon length, intron length, 5' and 3' splice site, 5' and 3' their strength and coding potential. As expected, exon length and coding potential were the most determinant features. We then 13 linear models for lncRNAs and mRNAs separately. Exon length, intron length, 5' and 3' splice site and their strength, and combinations of these parameters are used to develop models for lncRNAs and mRNAs. The first six models consider individual features. For both datasets, exon length is the most important determining feature (Figure 4). The second most important feature is the 5' splice site for lncRNA and intron length for mRNA. Further, the linear model is built and evaluated using a combination of two important features. Interestingly, the linear models having exon length as one of the features out of the two show the best R^2 values. There is no significant increase in the R^2 values by adding additional features, as shown in the last four models. Thus, exon length and 5' splice site are determining features for both lncRNA and mRNA datasets. Longer exon length and GC at 5' splice site are associated with higher TC in lncRNAs, whereas shorter exon length and GT at a 5' splice site are associated with higher TC in mRNAs. Shorter exon length combined with shorter intron length can also be associated with higher TC in mRNAs.

LINC01934 is a long non-coding RNA with 8 transcripts spliced product of 36 exons. The transcript complexity is below the median TC of the genome; therefore, an example of low TC lncRNAs has 136 nucleotides median exon length, and GT at 5'splice site. LAMTOR5 Antisense RNA 1, which acts as sponge lncRNA and regulates cancer progression, is an example of high TC lncRNA [36,37]. It has 17 transcripts as splice products of 36 exons with 178 nucleotides median exon length and prevalence of both GT and GC at 5' splice site. ANKRD44 (Ankyrin Repeat Domain 44) is a protein-coding gene associated with diseases, such as epilepsy and cancer, etc., is a mRNA with low TC [38,39]. This gene has only 5 transcripts splice products from 35 smaller exons (88 nucleotides, median exon length) and comparatively longer introns (2875 nucleotides, median intron length).

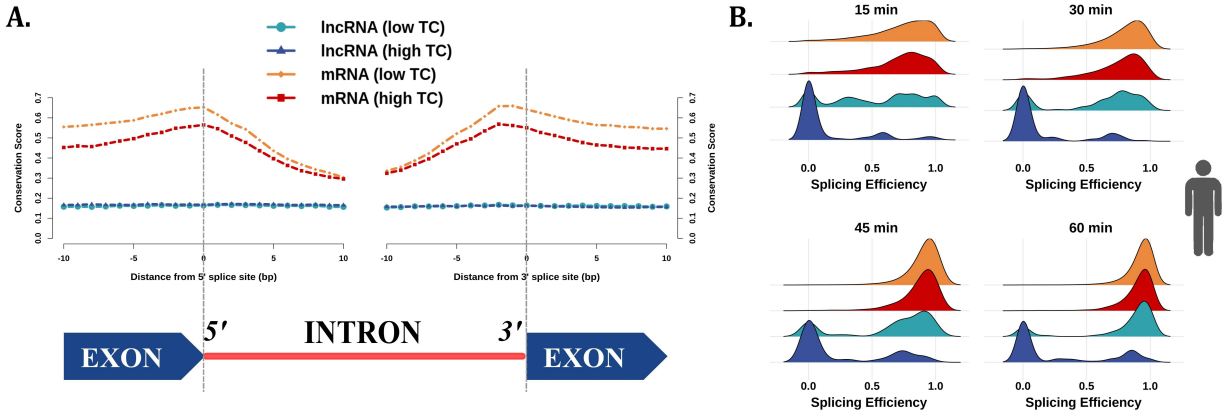


Figure 3. (A) Distribution of PhastCons conservation score of 5'and 3' splice sites and ±10 base pairs in lncRnas and mRnas with low and high TC for human (B) Splicing efficiency of low and high TC lncRnas and mRnas in human [GEO Dataset: GSE84722; *Homo sapiens* HEK293 cell labelled with 4-thiouridine for 15 minutes (Sample: GSM2257731), 30 minutes (Sample: GSM2257734), 45 minutes (Sample: GSM2257737) and 60 minutes (Sample: GSM2257740)].

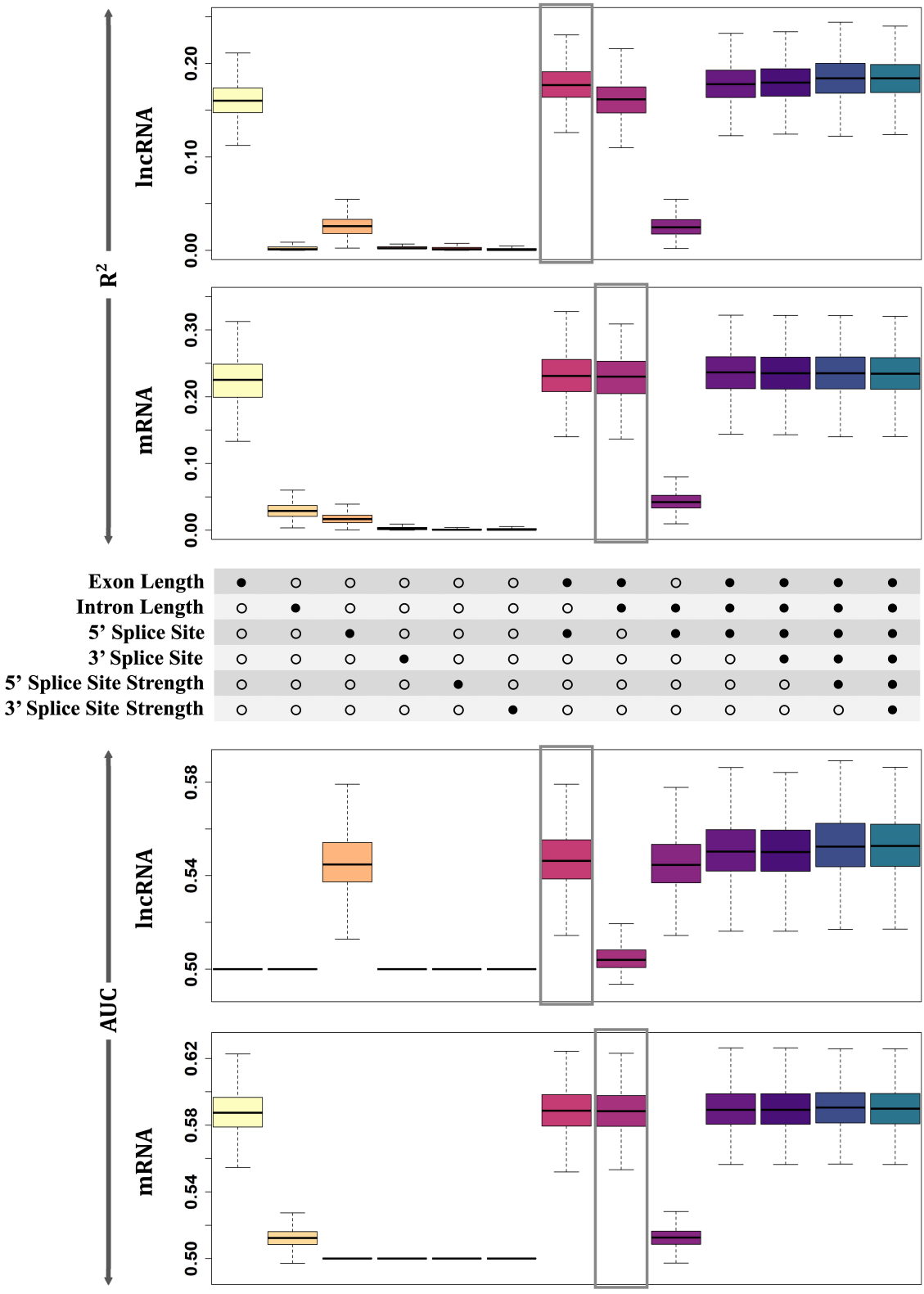


Figure 4. Depicts the R^2 for each model taking different features of TC calculated based on 1000 permutations of splitting data into training and testing sets for lncRnas and mRnas of human.

An example from protein-coding gene with high TC is NDUFS2 (NADH: Ubiquinone Oxidoreductase Core Subunit S2) which is involved in respiratory electron transport and mitochondrial complex I deficiency disease has 14 transcripts spliced from 35 exons and smaller intron length (261 nucleotides, median intron length) [40] (Table S12). These examples strengthen the broad generalizations made in this study.

Conclusion

In the present study, we measured the TC of a gene considering the number of exons. Most previous studies have compared the splicing efficiency of genes by counting the number of transcripts or calculating reads mapped to spliced to unspliced form. Here, we compared the splice variant efficiency of a gene based on the number of transcripts per exon and referred to it as TC. Previous studies focused on the splicing efficiency of the transcript and have reported strong dependences of splicing kinetics on the nature and position of 5' splice site flanking sequences [14]. The low strength of the 5' splice site and intron length has been associated with inefficient lncRNA splicing. Our findings suggest that lncRNAs have higher TC or multiple splice variants per exon than protein-coding mRNAs.

It is also observed that most of the transcripts (splice variants) of lncRNAs are generated by Alternative Last Exons (AF). In contrast, transcripts of mRNAs are the product of Alternative First Exon (AL) in humans and mice. We showed that lncRNAs with high TC are associated with shorter intron lengths and longer exon lengths in humans and mice. As reported previously, we also found GT and AG as the major 5' and 3' splice sites, respectively. Our result indicates that GC at the 5' splice site is another predominant dinucleotide in lncRNAs with high TC and lower splice site strength than mRNAs. Multi-vertebrate conservation analyses revealed that lncRNAs with low and high TC have low conservation scores at both 5' and 3' splice sites and are almost constant upstream and downstream of the position. In contrast, mRNAs with low TC have higher conservation scores than genes with high TC at 5' and 3' splice sites in humans and mice. The splicing efficiency analyses on different datasets show that lncRNAs with high TC undergo inefficient splicing.

This study opens a pandora of questions needed to be explored further. Is higher TC of lncRNAs dependent on the sequence length between the 3' splice site and the branch point or shorter polypyrimidine tract (PPT), resulting in inefficient splicing and more TC? Are these transcripts enriched by specific SRSF binding motifs, splicing silencers, and enhancers? Is there binding of differentially expressed RNA binding proteins which inhibit the splicing of groups of lncRNAs resulting in their nuclear retention? Is TC linked to the sub-cellular localization of the transcripts? How does the binding affinity of U1 to 5'ss upstream or the binding of cofactors on the branching point affect the TC? Does GT > GC affect the splicing kinetics and, thus, the TC? What is the role of 5' splice site sequence strength in deciding alternative splicing?

Although there are many unanswered questions, this study is the first to report a comparison of TC of lncRNAs and mRNAs in both human and mouse transcriptomes and relate it to splicing efficiency and associated sequence features. This study also reports that the exon length and 5' splice site are the determinant features of Transcript Complexity in humans. Further studies on finding mechanisms and linking splicing efficiency to TC would help to understand alternative splicing in lncRNAs.

Acknowledgments

We thank the members of MKlab for the helpful discussions.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

KB is a registered Integrated Systems Biology master's student and AD is a registered PhD student at the University of Hyderabad. A part of the work is funded by a Start-up Research Grant (SRG/2020/002146) awarded to MK from the Science and Engineering Research Board, Department of Science and Technology (SERB, DST), Government of India. AD and MK also acknowledge funding support from UoH-IoE Grant (UoH-IoE-RC2-21-012).

ORCID

Manjari Kiran http://orcid.org/0000-0003-0153-7072

Data availability statement

All the processed data and codes are available on https://github.com/Koushiki26/Transcript_Complexity.

References

[1] Mattick JS. The state of long non-coding RNA biology. Noncoding RNA. 2018;4(3):17. doi: 10.3390/ncrna4030017
[2] Jandura A, Krause HM. The new RNA world: growing evidence for long noncoding RNA functionality. Trends Genet. 2017;33(10):665–676. doi: 10.1016/j.tig.2017.08.002
[3] Wang J, Wei F, Zhou H. Advances of lncRNA in autoimmune diseases. Front Lab Med. 2018;2(2):79–82. doi: 10.1016/j.flm.2018.07.004
[4] DiStefano JK. The emerging role of long noncoding RNAs in human disease. Methods Mol Bio. 2018;1706:91–110. doi: 10.1007/978-1-4939-7471-9_6
[5] Bolha L, Ravnik-Glavač M, Glavač D. Long noncoding RNAs as biomarkers in cancer. Dis Markers. 2017;2017:1–14. doi: 10.1155/2017/7243968
[6] Kiran M, Chatrath A, Tang X, et al. A prognostic signature for lower grade gliomas based on expression of long non-coding RNAs. Mol Neurobiol. 2019;56(7):4786–4798. doi: 10.1007/s12035-018-1416-y
[7] Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. Nat Rev Genet. 2009;10(3):155–159. Internet. doi: 10.1038/nrg2521
[8] Cabili MN, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev Internet. 2011;25(18):1915–1927. doi: 10.1101/gad.17446611
[9] Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 2012;22(9):1775–1789. doi: 10.1101/gr.132159.111
[10] Melé M, Mattioli K, Mallard W, et al. Chromatin environment, transcriptional regulation, and splicing distinguish lincRnas and mRnas. Genome Res Internet. 2017;27(1):27–37. doi: 10.1101/gr.214205.116
[11] Tilgner H, Knowles DG, Johnson R, et al. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for

lncRnas. Genome Res. 2012;22(9):1616–1625. doi: 10.1101/gr.134445.111
[12] Khan MR, Wellinger RJ, Laurent B. Exploring the alternative splicing of long noncoding RNAs. Trends Genet. 2021;37(8):695–698. Internet. doi: 10.1016/j.tig.2021.03.010
[13] Wachutka L, Caizzi L, Gagneur J, et al. Global donor and acceptor splicing site kinetics in human cells. Elife. 2019;8. doi: 10.7554/eLife.45056
[14] Krchňáková Z, Thakur PK, Krausová M, et al. Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5′ splice-site sequences due to weak interactions with SR proteins. Nucleic Acids Res. 2019;47(2):911–928. doi: 10.1093/nar/gky1147
[15] Gonzalez I, Munita R, Agirre E, et al. A lncRNA regulates alternative splicing via establishment of a splicing-specific chromatin signature. Nat Struct Mol Biol Internet. 2015;22(5):370–376. doi: 10.1038/nsmb.3005
[16] Ramanouskaya TV, Grinev VV. The determinants of alternative RNA splicing in human cells. Mol Genet Genomic. 2017;292(6):1175–1195. doi: 10.1007/s00438-017-1350-0
[17] Frankish A, Diekhans M, Ferreira A-M, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):D766–D773. doi: 10.1093/nar/gky955
[18] Yates A, Akanni W, Amode MR, et al. Ensembl 2016. Nucleic Acids Res. 2016;44(D1):D710–D716. doi: 10.1093/nar/gkv1157
[19] O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–D745. doi: 10.1093/nar/gkv1189
[20] Cunningham F, Allen JE, Allen J, et al. Ensembl 2022. Nucleic Acids Res. 2022;50(D1):D988–D995. doi: 10.1093/nar/gkab1049
[21] Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. Science (1979). 2022;376:44–53.
[22] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74. doi: 10.1038/nature11247
[23] Heinz S, Benner C, Spann N, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576–589. doi: 10.1016/j.molcel.2010.05.004
[24] Trincado JL, Entizne JC, Hysenaj G, et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 2018;19(1):40. doi: 10.1186/s13059-018-1417-1
[25] Karolchik D. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32(90001):493D–496. doi: 10.1093/nar/gkh103
[26] Yeo G, Burge CB Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Proceedings of the seventh annual international conference on Computational molecular biology - RECOMB '03. New York, New York, USA: ACM Press; 2003. p. 322–331.
[27] Abou Alezz M, Celli L, Belotti G, et al. GC-AG introns features in long non-coding and protein-coding genes suggest their role in gene expression regulation. Front Genet. 2020;11:11. doi: 10.3389/fgene.2020.00488
[28] Siepel A, Haussler D Combining phylogenetic and hidden Markov models in biosequence analysis. Proceedings of the seventh annual international conference on Computational molecular biology - RECOMB '03. New York, New York, USA: ACM Press; 2003. p. 277–286.
[29] Mukherjee N, Calviello L, Hirsekorn A, et al. Integrative classification of human coding and noncoding genes through RNA metabolism profiles. Nat Struct Mol Biol. 2017;24(1):86–96. doi: 10.1038/nsmb.3325
[30] Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2012;41(D1):D991–D995. doi: 10.1093/nar/gks1193
[31] de Melo Costa VR, Pfeuffer J, Louloupi A, et al. SPLICE-q: a Python tool for genome-wide quantification of splicing efficiency. BMC Bioinf. 2021;22:368. doi: 10.1186/s12859-021-04282-6
[32] Kuhn M. Building predictive models in *R* using the **caret** package. J Stat Softw. 2008;28(5):28. doi: 10.18637/jss.v028.i05
[33] Thanaraj TA. Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. Nucleic Acids Res. 2001;29(12):2581–2593. doi: 10.1093/nar/29.12.2581
[34] Kralovicova J, Hwang G, Asplund AC, et al. Compensatory signals associated with the activation of human GC 5′ splice sites. Nucleic Acids Res. 2011;39(16):7077–7091. doi: 10.1093/nar/gkr306
[35] Baek D, Green P. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. Proceedings of the National Academy of Sciences. 2005;102(36):12813–12818. doi: 10.1073/pnas.0506139102
[36] Zaniani NR, Oroujalian A, Valipour A, et al. LAMTOR5 expression level is a biomarker for colorectal cancer and lncRNA LAMTOR5-AS1 predicting miRNA sponging effect. Mol Biol Rep. 2021;48(8):6093–6101. doi: 10.1007/s11033-021-06623-3
[37] Pu Y, Tan Y, Zang C, et al. LAMTOR5-AS1 regulates chemotherapy-induced oxidative stress by controlling the expression level and transcriptional activity of NRF2 in osteosarcoma cells. Cell Death Dis. 2021;12(12):1125. doi: 10.1038/s41419-021-04413-0
[38] Kurki MI, Gaál EI, Kettunen J, et al. High risk population isolate reveals low frequency variants predisposing to intracranial aneurysms. PLoS Genet. 2014;10(1):e1004134. doi: 10.1371/journal.pgen.1004134
[39] La Ferla M, Lessi F, Aretini P, et al. ANKRD44 gene silencing: a putative role in trastuzumab resistance in Her2-Like breast cancer. Front Oncol. 2019;9. doi: 10.3389/fonc.2019.00547
[40] Dunham-Snary KJ, Wu D, Potus F, et al. Ndufs2, a core subunit of mitochondrial complex I, is essential for acute oxygen-sensing and hypoxic pulmonary vasoconstriction. Circ Res. 2019;124(12):1727–1746. doi: 10.1161/CIRCRESAHA.118.314284