Conserved long-range RNA structures associated with pre-mRNA processing of human protein-coding genes

Svetlana Kalmykova[1], Marina Kalinina[1], Stepan Denisov[1], Alexey Mironov[1], Dmitry Skvortsov[2], Roderic Guigo[3], and Dmitri Pervouchine[1]

[1]Skolkovo Institute of Science and Technology, 3 Nobel st, Moscow 143026, Russia

[2]Faculty of Chemistry, Moscow State University, Leninskiye Gory 1-73, 119234 Moscow, Russia

[3]Center for Genomic Regulation and UPF, Dr. Aiguader 88, Barcelona 08003, Spain

Abstract

The ability of nucleic acids to form double-stranded structures is essential for all living systems on Earth. While DNA employs it for genome replication, RNA molecules fold into complicated secondary and tertiary structures. Current knowledge on functional RNA structures in human protein-coding genes is focused on locally-occurring base pairs. However, chemical crosslinking and proximity ligation experiments have demonstrated that long-range RNA structures are highly abundant. Here, we present the most complete to-date catalog of conserved long-range RNA structures in the human transcriptome, which consists of 1.1 million pairs of conserved complementary regions (PCCRs). PCCRs tend to occur within introns proximally to splice sites, suppress intervening exons, circumscribe circular RNAs, and exert an obstructive effect on cryptic and inactive splice sites. The double-stranded structure of PCCRs is supported by a significant decrease of icSHAPE nucleotide accessibility, high abundance of A-to-I RNA editing sites, and frequent nearby occurrence of forked eCLIP peaks. Introns with PCCRs show a distinct splicing pattern in response to RNA Pol II slowdown suggesting that splicing is widely affected by co-transcriptional RNA folding. Additionally, transcript starts and ends are strongly enriched in regions between complementary parts of PCCRs, leading to an intriguing hypothesis that RNA folding coupled with splicing could mediate co-transcriptional suppression of premature cleavage and polyadenylation events. PCCR detection procedure is highly sensitive with respect to *bona fide* validated RNA structures at the expense of having a high false positive rate, which cannot be reduced without loss of sensitivity. The catalog of PCCRs is visualized through a UCSC Genome Browser track hub to facilitate further genome research on long-range RNA structures.

Introduction

Double-stranded structure is a key feature of nucleic acids that enables replicating the genomic information and underlies fundamental cellular processes [1, 2]. Many RNAs adopt functional secondary structures, and mRNAs are no exception although their main role is to encode proteins [3, 4, 5, 6]. In eukaryotes, RNA structure affects gene expression through modulating all steps of pre-mRNA processing including splicing [7], cleavage and polyadenylation [8], and RNA editing [9]. The loss of functional RNA structure has been increasingly reported as implicated in hereditary disease and cancer [10, 11, 12, 13].

Recent progress in high-throughput sequencing techniques enabled several experimental strategies to determine RNA structure *in vivo* [14, 15, 16]. Chemical RNA structure probing can reveal which bases are single- or double-stranded, but it cannot determine which nucleotides form base pairs [17, 18, 19, 20, 21]. In order to identify the interacting partners, RNA structure probing has to be combined with *de novo* RNA structure prediction [22], but due to a number of technical limitations such methods cannot account for base pairings that are far apart [23]. Photo-inducible RNA crosslinking and proximity ligation assays, however, revealed highly abundant double-stranded regions across the human transcriptome that are composed of long-range base pairings [14, 24, 25, 26].

To date, a number of functional long-range RNA structures in protein-coding genes have been characterized [27, 28]. The first reference appeared under the term "panhandle structure", which was coined in by virologists in the 1980's to refer to a complementary base pairing between the 5'-end and 3'-end of the RNA genome of several segmented negative-stranded viruses [29]. Since then, dozens of long-range RNA structures that modulate pre-mRNA splicing have been confirmed experimentally [27, 28]. The use of high-throughput assays for large-scale profiling of intronic pre-mRNA structure is still limited because the intronic signal is often reduced or missing, which is also a common problem in other methods [30]. Computational identification of long-range RNA structure therefore remains a great challenge in RNA biology.

Comparative genomics provides a powerful alternative to *de novo* RNA folding by detecting signatures of evolutionary covariation [31, 32]. In previous reports, we presented a methodology that implemented simultaneous alignment and RNA folding using *k*-mers [28, 33, 34]. However, in the vast majority of cases, sequence alignments have insufficient variation to detect covariations [34]. This motivated us to revisit this question, this time with a much simpler approach that finds pairs of conserved complementary regions (PCCR) in pre-aligned evolutionarily-conserved regions. The method named PREPH (prediction of panhandles) efficiently finds long nearly perfect complementary matches in pairs of conserved sequences. By construction, it has a high false positive rate (in some cases up to 50%), but on the other hand it is highly sensitive to *bona fide* validated RNA structures, and its false positive rate cannot be reduced without losing sensitivity.

In spite of admixture of false positives, we observed an overall association of PCCRs with pre-mRNA splicing, e.g., their occurrence within introns proximally to splice sites, suppression of exons, avoidance of branch points, obstructive effect on cryptic and inactive splice sites, etc. When juxtaposed with the

high-throughput RNA structure probing data, PCCRs reveal a significant decrease of icSHAPE nucleotide accessibility, high abundance of A-to-I RNA editing sites, and frequent nearby occurrence of forked eCLIP peaks, which strongly support double-stranded structure of PCCRs. The experimentally validated RNA structures did not have sufficient statistical evidence in RNA structure probing, mutational, or compensatory substitution analyses due to sparsity of these data. This motivated us to combine the catalog of PCCRs with additional experimental characterization to create a resource that could be of great interest to the bioinformatics community. We provide it as a reference dataset that can be visualized through a UCSC Genome Browser track hub [35].

Results

Conserved Complementary Regions

Conserved complementary regions (CCR) were identified in intronic regions of 20,345 annotated human protein-coding genes (Figure 1A). The nucleotides that belong to at least one annotated exon were excluded from the search because exonic sequences evolve under selective constraints on the aminoacid sequence. The remaining intronic regions were extended by 10 nts to within flanking exons to account for base pairings that overlap splice sites. This procedure resulted in 326,363 intronic regions (median length 1,178 nts) which comprise approximately 30% of the genome. These regions were intersected with the `phastConsElements` track of the genomic intervals that are highly conserved between 100 vertebrates [36]. This step resulted in 2,244,713 shorter fragments (median length 16 nts), which altogether cover approximately $5.7 \cdot 10^7$ nucleotides (5.3% of the previous set). In what follows, this set will be referred to as *conserved intronic regions*.

The PREPH pipeline (Figure 1B) was applied to all pairwise combinations of conserved intronic regions that are located not more than 10,000 nts apart from each other and belong to the same gene (see Methods for details). With the default parameters, it yielded 1,238,794 pairs of conserved complementary regions (PCCR), on average 92 PCCR per gene, with 95% of genes having not more than 361 PCCRs. Genes with exceptionally high PCCR counts were excluded from further analysis (Table S1). Additional exclusion of regions that overlap with the repeat regions, small non-coding genes residing in introns (tRNA, snoRNA, miRNA), transcription factors binding sites (see Supplementary Methods) resulted in the final set of 1,134,601 PCCRs, which were analyzed further.

The median predicted free energy of hybridization and the median length of CCRs were −17.2 kcal/mol and 13 nts, respectively, with the frequency distribution decaying towards longer and more stable structures (Figure 1C and S1A). PCCR were classified by their energies into the following four groups, which will be represented by a uniform color scheme throughout all figures: group I from −15 to −20 kcal/mol (green), group II from −20 to −25 kcal/mol (yellow), group III from −25 to −30 kcal/mol (orange), and group IV below −30 kcal/mol (red). The method showed a high sensitivity with respect to functional long-range

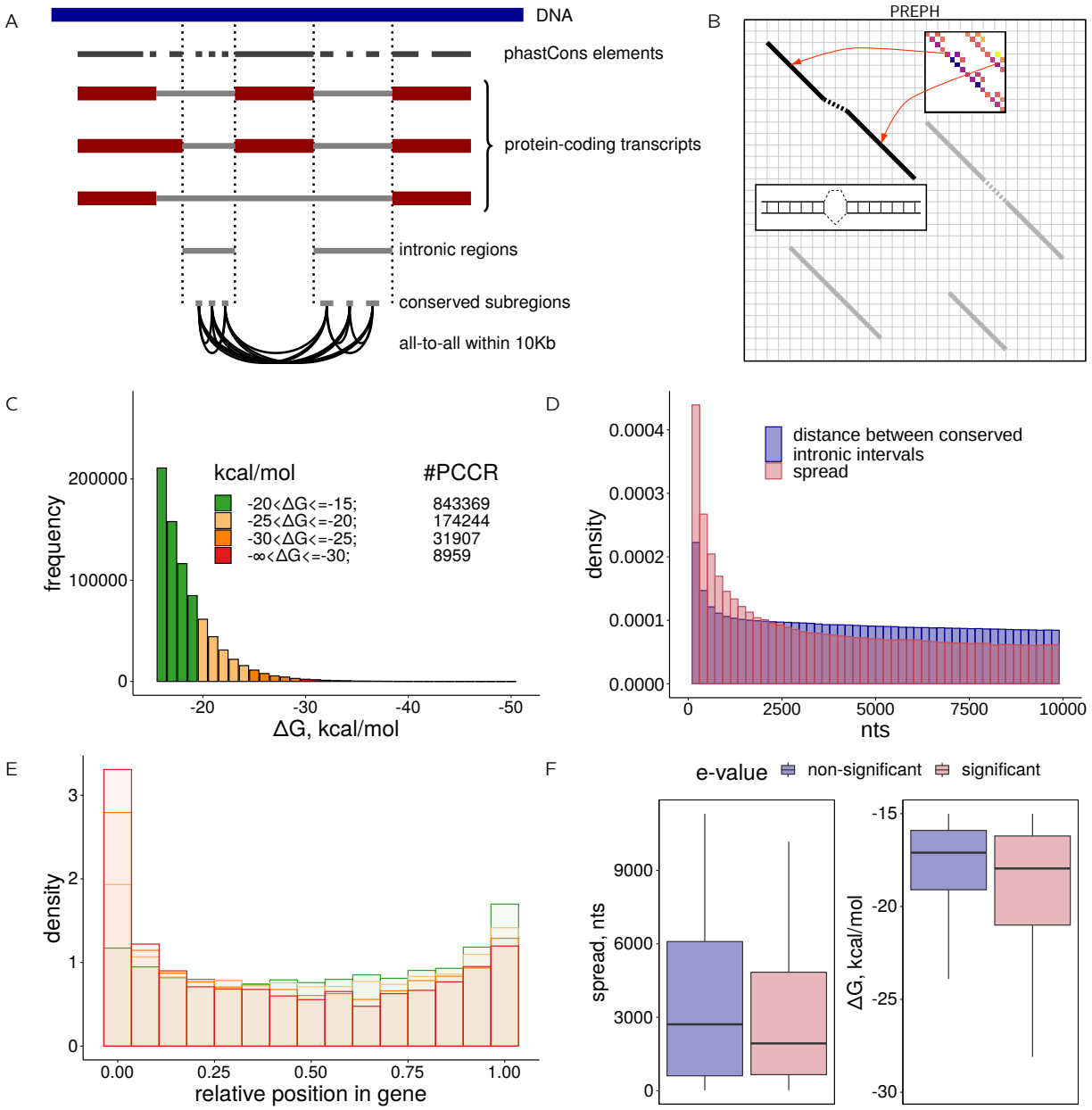


Figure 1: Pairs of Conserved Complementary Regions (PCCR). **(A)** PCCRs are identified in conserved intronic regions that are less than 10,000 nt apart from each other. **(B)** PCCRs are identified in a pair of input sequences by sparse dynamic programming implemented in PREPH. **(C)** The distribution of PCCR energy. The four energy groups are: group I from −15 to −20 kcal/mol (green), group II from −20 to −25 kcal/mol (yellow), group III from −25 to −30 kcal/mol (orange), and group IV below −30 kcal/mol (red). **(D)** The distribution of PCCR spread (i.e., distance between two CCRs in a pair). **(E)** The distribution of p, the relative position of PCCR in the gene. **(F)** The distribution of energy and spread of PCCRs with statistically significant compensatory substitutions (E-value < 0.05).

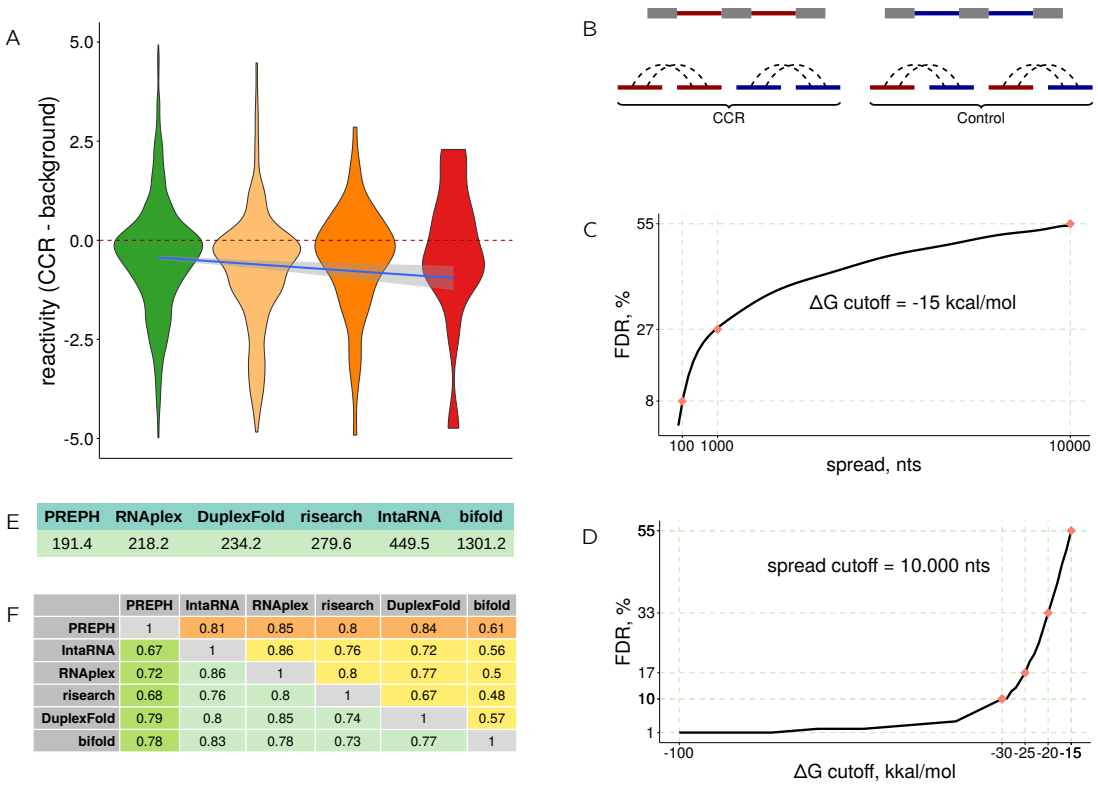
RNA structures listed in Table 1 from [23] (Figure S2). The structure responsible for splicing of the intron between exons 9 and 10 in *SF1* [33] was in group III (-25.1 kcal/mol); the structure associated with exon 46–52 skipping in *DST* [34] was in group II (−24.3 kcal/mol); a long-distance intronic interaction that regulates *PLP1/DM20* splicing [37] was in group I (−15.8 kcal/mol); the structure that was suggested to

regulate mutually exclusive exon choice in *DNM1* [38] was in group III (-25.9 kcal/mol). The RNA bridge in *ENAH* [39] also belongs to group I (-19.4 kcal/mol) and spreads over 1,700 nts, indicating that less stable RNA structures are not less functional or less interesting than the others.

The frequency distribution of the distances between two CCRs in a pair, here referred to as *spread*, peaks at short distances, however decaying to a non-zero baseline (Figure 1D). We found that it mirrors the frequency distribution of pairwise distances between conserved intronic regions that were passed to PREPH as an input. We next asked whether PCCRs were distributed uniformly along the gene, or they tend to accumulate in certain gene parts. To quantify the position of a PCCR within a gene, we introduced p, a measure of the relative position (see Methods), which changes from 0% for regions located in the very 5'-end of the gene to 100% for regions located in the very 3'-end. The metric p can be computed for a PCCR as a whole to represent its relative position, or for of each its CCR separately. The location of PCCR as a whole was not uniform, with two pronounced modes at the 5'-end and at the 3'-end (Figure 1E). This enrichment was also prominent in the distribution of separate CCRs (Figure S1B) and didn't depend on the density of conserved regions along the gene (Figure S1D). It could, in part, be due to higher chances of CCR to occur in more constrained regions, as we found that CCR have a higher average degree of evolutionary conservation, measured by the average phastCons score, than their adjacent conserved regions (MW test, $P < 2.2 * 10^{-16}$). PCCRs with statistically significant compensatory substitutions (see Mutations section) were on average more stable and less spread than other PCCRs (Figure 1F). The ontology annotation of genes with PCCRs was enriched with terms related to morphogenesis and development of the central nervous system when compared to the genes of the same length but without PCCRs (Figure S1C).

Benchmark

PCCRs presumably represent the regions in the pre-mRNA that have an increased capacity to base-pair. The propensity of individual nucleotides to base-pair was assessed at the transcriptome-wide level by measuring the nucleotide flexibility score with icSHAPE method [40]. We compared the average icSHAPE reactivity within CCR with that in a control set of intervals located nearby (Figure 2A). Indeed, the average icSHAPE reactivity of nucleotides within CCR was significantly lower as compared to a locally-matched control (Wilcoxon test, $P < 10^{-80}$), and the difference increased by the absolute value with increasing structure stability ($\hat{\beta} = 0.06 \pm 0.02$). Furthermore, we compared our predictions with the long-range RNA-RNA interactions that were assessed experimentally in PARIS [25] and LIGR-seq [26] by computing the precision and recall metrics (see Methods). In comparison to PARIS, the precision was 0.47% and the recall was 96.27%. For LIGR-seq, the precision and recall were 7.86% and 25.8%, respectively. Obviously, low precision figures arise from very sparse coverage of the transcriptome by PARIS and LIGR-seq. For both datasets, however, the precision was higher for more stable and less spread PCCR, while the recall was lower. We consider the agreement with RNA structure probing experiments as an indication that our predicted set at least partially corresponds to double-stranded regions. Below we estimate what proportion of PCCR could be



E

PREPH	RNAplex	DuplexFold	risearch	IntaRNA	bifold
191.4	218.2	234.2	279.6	449.5	1301.2

F

	PREPH	IntaRNA	RNAplex	risearch	DuplexFold	bifold
PREPH	1	0.81	0.85	0.8	0.84	0.61
IntaRNA	0.67	1	0.86	0.76	0.72	0.56
RNAplex	0.72	0.86	1	0.8	0.77	0.5
risearch	0.68	0.76	0.8	1	0.67	0.48
DuplexFold	0.79	0.8	0.85	0.74	1	0.57
bifold	0.78	0.83	0.78	0.73	0.77	1

Figure 2: Benchmark. **(A)** icSHAPE reactivity of nucleotides within CCR vs. background reactivity of nearby nucleotides in the four energy groups (see Figure 1C for the color code). **(B)** The rewiring procedure for the estimation of false positive rate (FDR). **(C)** FDR as a function of spread at the constant energy -15 kcal/mol. **(D)** FDR as a function of energy at the constant spread $10,000$ nt. Time efficacy **(E)** and accuracy **(F)** of PREPH (see Supplementary Methods).

false positives.

One way to estimate the rate of false positive PCCRs is to apply the same analysis that was used to identify the true PCCRs to a set of control sequences that should not base-pair. In previous works, it was described as the so-called "re-wiring" approach [34]. Here, we used the same strategy by applying PCCR to chimeric sequence sets that were sampled randomly from different genes (Figure 2B) controlling for their nucleotide composition and length, which confound this comparison. The false discovery rate (FDR), defined here as the number of false positive predictions as a fraction of the total number of predictions, depends on PCCR energy and spread (Figure 2C and 2D), decaying from 55% to 10% when the energy threshold increases from -15 to -30 kcal/mol, and also from 55% to 8% when the spread limit drops to 100 nts. This indicates that a half of group I predictions of PREPH alone are false positives, however these are very pessimistic estimates since the approach greatly overestimates the actual FDR figure [34]. It is based on the assumption that pre-mRNAs of different genes do not interact and CCR in different genes evolve independently. Sampling chimeric sequences from paralogous genes also contributes to the overestimation

of FDR. We therefore conclude that at least 50% of our predictions must be true positives.

Next, we compared the accuracy and runtime of PREPH to those of other programs such as IntaRNA2.0 [41], RIsearch2 [42], RNAplex [43], DuplexFold [44] and bifold [44]. As a benchmark for the time efficiency, we used a set of 1000 pairs of randomly chosen conserved intronic sequences from the human genome (see Methods). PREPH showed the quickest result compared to the other programs (191.4 sec) (Figure 2E). At that, the equilibrium free energies of the predictions by PREPH correlated reasonably well with those of RNAplex, IntaRNA, and Duplexfold (Figure S3). Next, we evaluated the accuracy of PREPH with respect to that of the other programs by comparing optimal free energy structures (see Methods). The number of base pairs that were common between PREPH and each of the other programs as a fraction of the number of base pairs predicted by PREPH alone was used as a measure of specificity (Figure 2F). PREPH showed the specificity above 80% with respect to all other programs except bifold, however the latter was not in agreement with all other programs. Thus, PREPH allows for computationally-efficient detection of PCCRs without significant loss of accuracy compared to other methods. The computation time of PREPH on the complete dataset of conserved intronic regions was 4 hrs (15 threads, 1200 MHz CPUs each).

Splicing

Previous studies reported that long-range base pairings are positioned non-randomly with respect to splicing signals [34, 33, 28]. Here, we elaborated on this result using the classification shown in Figure 3. If a PCCR overlaps an intron, it can be located either entirely within the intron (inside), or the intron can be located entirely within PCCR (outside), or the two intervals intersect (crossing). The tendency of RNA structure to prefer one of these categories is characterized by the number of PCCRs in them as compared to their counts in a certain control set (each PCCR may be counted more than once if it is located inside one intron and crosses another). Numerically, it is measured by the *enrichment* metric, defined as the fraction of the number of PCCRs found originally to the number of PCCRs in the control set.

Here, we use two control sets. In the first, referred to as *random shift*, we randomly move each PCCR within its gene while maintaining its spread. The resulting pseudo-PCCRs have exactly the same spread and belong to the same genes as the original PCCRs (Figure 3A). However, since not all positions along the gene are equally likely, PCCR frequencies are confounded by position within the gene. To account for this, we used another control procedure, referred to as *random gene*. For each gene and a PCCR in it, we create a pseudo-PCCR at the same relative position as was the original PCCR, but in a randomly chosen gene of the same length. The resulting pseudo-PCCRs have exactly the same spread and the same relative positions as the original PCCRs, but in a different gene (Figure 3A). Control procedures were repeated 40 times to estimate the significance of enrichment or depletion with respect to random sampling.

Comparison of PCCRs with the set of annotated human introns by random shifts revealed a strong enrichment of PCCRs inside introns, depletion of PCCRs outside introns, and almost no statistically discernible difference for crossing PCCRs (Figure 3B). The magnitude of the enrichment and depletion increased with

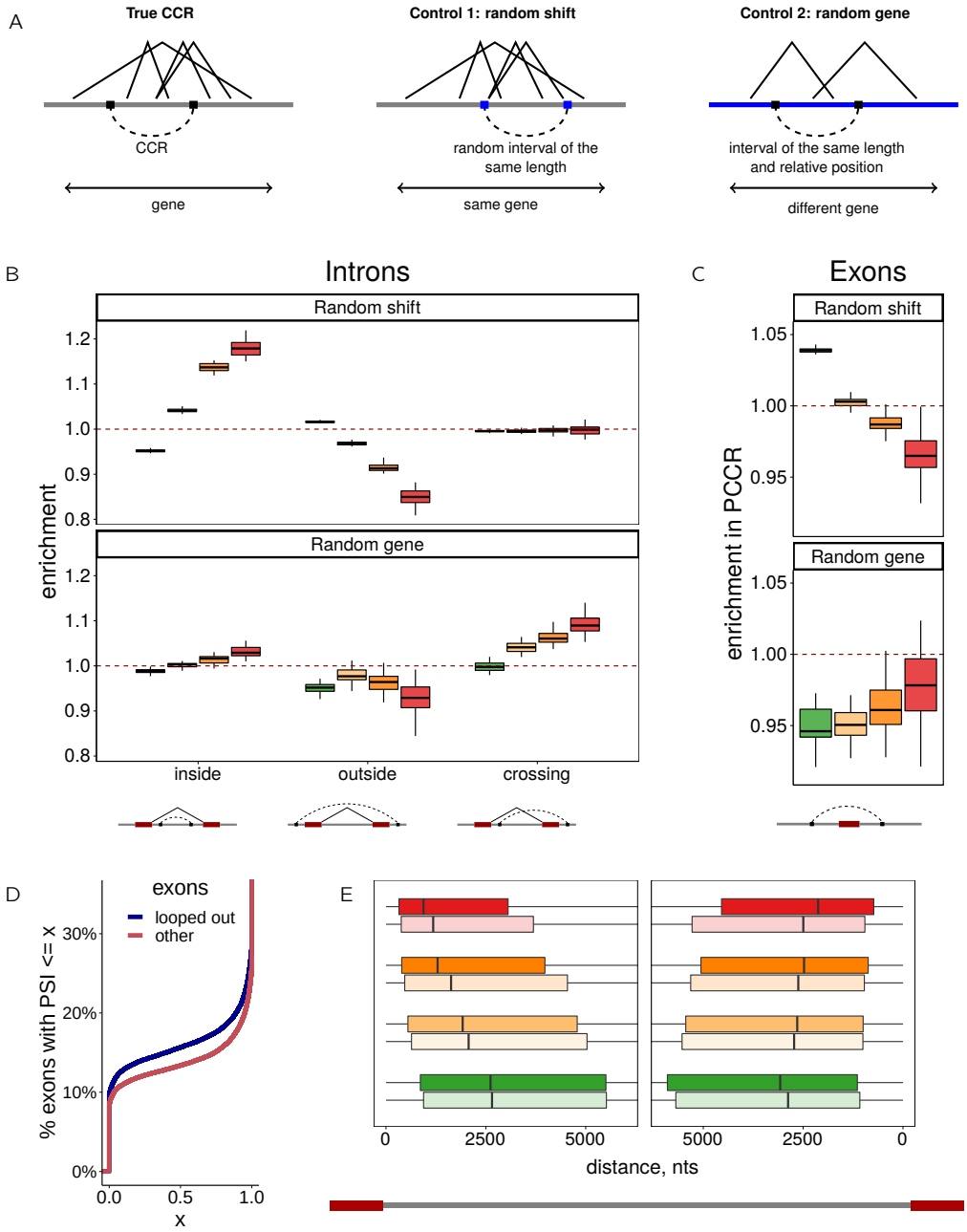


Figure 3: Splicing. **(A)** The two control procedures, random shift and random gene. **(B)** PCCRs are enriched inside and depleted outside introns. Boxplots correspond to random shift (top) and random gene controls (bottom). The color code is as in Figure 1C. **(C)** Exons looped-out by PCCRs are depleted. **(D)** The cumulative density of exon inclusion rates for exons inside PCCR (looped-out) vs. other exons. **(E)** PCCRs are located closer to intron ends than expected by chance.

increasing structure stability. To rule out the possibility that preferential PCCR positioning within introns is not a consequence of uneven distribution of longer introns along the gene, with 5'-introns being on average longer [33, 28], we repeated this analysis using random gene control, and the enrichment and depletion of PCCRs remained significant (Figure 3B). A similar procedure revealed strong depletion of PCCRs that

loop-out exons, which also becomes stronger as PCCR stability increases (Figure 3C). The average inclusion rate of exons that were looped out by PCCR tends to be lower than that of exons that are not surrounded by PCCR (Figure 3D). These results indicate that PCCRs generally tend to avoid placing exons in a loop, which would promote their exclusion.

The ends of PCCRs tend to be closer to exon boundaries than would be expected by chance from random shifts, and the effect is again stronger for more stable structures (Figure 3E). Contrary to what was reported earlier for *D. melanogaster*, there is no substantial depletion of CCR in polypyrimidine tracts (PPT), defined as nucleotides in positions $-50 \cdots -1$ upstream of an annotated exon, as compared to nucleotides in positions $-100 \cdots -51$ (Figure S4). This indicates a large regulatory potential for splicing since a large fraction of CCR that overlap PPT (43.2%) also block the acceptor splice site. In addition, there were on average 20% less overlaps of CCR with intronic branch points [45] than would be expected from random shift control, i.e., CCRs tend to not overlap intronic branch points.

It was suggested previously that one of the functions of RNA secondary structure is to suppress cryptic splice sites [46]. To test this, we identified intronic motifs with high similarity to donor and acceptor site consensus sequences. Further, we analysed a large compendium of RNA-seq samples from the Genotype Tissue Expression Project (GTEx) to identify splice sites that are used at low and high frequency [47]. CCRs overlapping highly expressed (active) splice sites were strongly depleted, while CCRs overlapping splice sites with low read support (inactive) were enriched (Figure 4A). The magnitude of this effect grows with increasing structure energy, however this trend could result from lower GC-content and, consequently, lower stability of stronger splice sites. Similarly, CCRs overlapping non-expressed sequences with high similarity to splice sites (cryptic) were also enriched, however to a lesser extent. We also found that the largest enrichment of CCRs overlapping candidate cryptic splice sites was among PCRR with small spread, in agreement with the previous findings that cryptic splice sites are suppressed by the local RNA secondary structure [48].

In order to investigate how PCCRs are related to circular RNAs (circRNA), we used circRNAs from the tissue-specific circular RNA database (TCSD) that were expressed in liver [49]. Both random shift and random gene control revealed a pattern opposite to that of linear introns, in which PCCRs were enriched outside circRNAs and depleted inside circRNAs (Figure 4B). This is in good agreement with the recently proposed feature of circular RNAs to be looped out by RNA secondary structure [50].

RNA editing

A widespread form of post-transcriptional RNA modification is the enzymatic conversion of adenosine nucleosides to inosine, a process called A-to-I editing, which is mediated by the ADAR family of enzymes [51, 52]. Since ADAR editing occurs in double-stranded RNA substrates [53, 54], we asked whether A-to-I editing sites are enriched among CCRs (Figure 4C). We observed a strong enrichment of ADAR edited sites in CCR compared to non-CCR parts of conserved intronic regions (OR $= 2.1 \pm 0.2$). While the odds ratio was nearly the same for PCCR with energy from -15 to -30 kcal/mol, it dramatically increased for PCCR with the

energy below −30 kcal/mol (2.1 ± 0.2 vs 17.7 ± 6.5), in agreement with previous observations that ADAR targets are long double-stranded RNAs [55]. The odds ratio was also significantly greater (1.5 ± 0.07 vs. 6.4 ± 1.3) for PCCR with higher evidence of compensatory mutations (E-value < 0.05), but it didn't depend significantly on the PCCR spread. These results were concordant with each other for the datasets of A-to-I editing sites from RADAR and REDIportal databases [56, 57] and additionally support double-stranded structure of PCCRs.

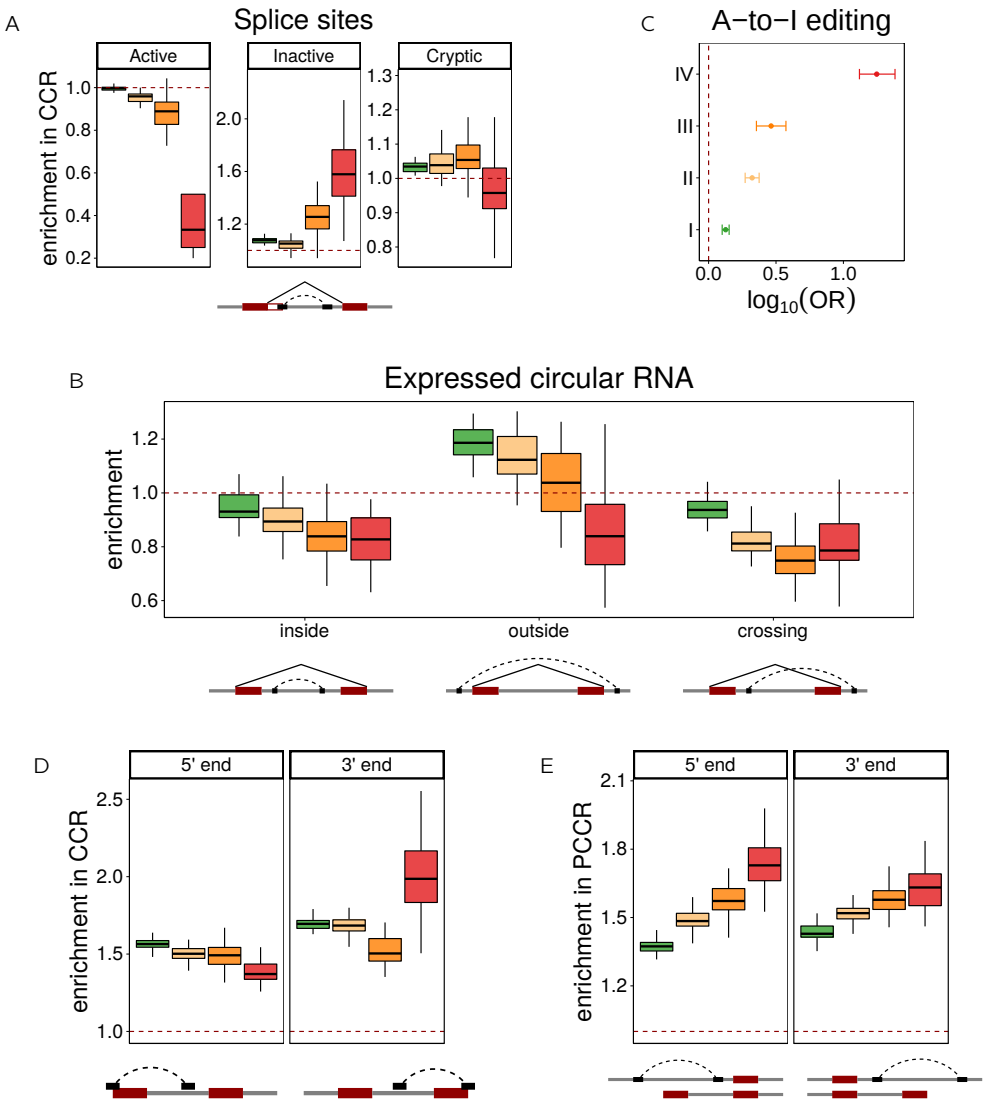


Figure 4: Splicing, 5'-end, and 3'-end processing. **(A)** Active splice sites are depleted, and inactive and cryptic splice sites are enriched within CCRs. **(B)** PCCRs are enriched outside circular (back-spliced) introns and depleted inside them, in contrast to linear introns. **(C)** A-to-I RNA editing sites are enriched within CCRs; OR denotes odds ratio. PCCRs are enriched with the annotated and expressed 5'-ends. **(D)** 5'- and 3'-ends of aberrant transcripts are enriched within CCRs. **(E)** 5'- and 3'-ends of aberrant transcripts are enriched inside PCCRs, i.e., in between two CCRs in a pair.

5'-end and 3'-end processing

Recent reports indicate that RNA structure is important for the 3'-end processing of human mRNAs [8], and that competing RNA base pairings could be involved in alternative splicing and polyadenylation in the 3'-variable region [58]. In this section we characterize the relationship between PCCRs, 5'-end, and 3'-end mRNA processing using the data from transcript annotation databases [59] and clusters of poly(A)-seq and CAGE-tags [60, 61].

First, we estimated the number of transcripts that start or end inside CCRs, including all incomplete and aberrant isoforms that are annotated in GENCODE (Figure 4E). Both 5'- and 3'-ends were enriched inside CCRs suggesting that double-stranded structures could be involved in the suppression of aberrant isoforms. Next, we asked whether the annotated 5'- and 3'-ends of transcripts were enriched within PCCRs, i.e. in the regions between CCRs. Indeed, both were found to be significantly enriched, and the magnitude of the enrichment increased with increasing PCCR stability (Figure 4D). The random gene control confirmed that the effect was not due to non-uniform distribution of PCCRs. To rule out the possibility that non-expressed isoforms contribute to the observed enrichment, we additionally examined the expressed CAGE-tags and poly(A)-seq clusters and found that they were also significantly enriched within PCCR (Figure S5). Since PCCRs are, in turn, enriched inside introns, this led us to think that there could be an RNA structure-mediated connection between splicing and pre-mRNA end processing.

Thus, we asked whether the annotated transcript ends, CAGE-tags, and poly(A)-seq clusters occur more frequently in introns that contain PCCRs. However, introns with PCCRs tend to be also longer than introns without PCCRs, which may affect the above frequencies. Hence, we subdivided all annotated introns into two groups, introns with PCCR (IWP) and introns without PCCR (IWO) and selected two samples from IWP and IWO, 40,061 introns each, with matching intron lengths (Figure S6). As a result, 12,777 (31.9%) of IWP contained at least one annotated 3'-end, while only 9,567 (23.9%) of IWO did so ($OR = 1.49 \pm 0.05$). Similarly, 12,093 IWP (30.2%) contained at least one annotated 5'-end, compared to 8,919 (22.3%) for IWO ($OR = 1.51 \pm 0.05$). The enrichment of both 5'- and 3'-ends in IWP raises an intriguing hypothesis about a possible link between long-range RNA structure, splicing, and end-processing (see Discussion).

Mutations

Mutations generally lower the stability of RNA secondary structure, and some of them are under evolutionary selection owing to their effects on the thermodynamic stability of pre-mRNA [62]. In agreement with this, the average phastCons score [36] of nucleotides within CCR were significantly greater than that in adjacent conserved regions (MW test, $P < 2.2 * 10^{-16}$), confirming that CCRs have a higher chance of occurrence in more constrained regions. In order to estimate the impact of human population polymorphisms, we analyzed SNPs from the 1000 Genomes project [63] and compared SNP density in CCR with that in the remaining part of conserved intronic regions. Germline SNPs were significantly underrepresented in CCR (0.0196 vs 0.0207, $P < 0.01$).

Next, for each SNP that occurs in a PCCR, we calculated the free energy change caused by the mutation and compared it to the free energy change that would have been observed if the same mutation occurred in a different nucleotide of the same CCR (Figure S7A and S7B). This comparison accounts for the confounding effect of mutational signatures. It turned out that actual SNPs destabilized the predicted RNA structures less than it would be expected by chance (Wilcoxon test, $P < 0.01$), suggesting that SNPs tend to minimize their impact on RNA structure stability when occurring in structured regions.

Individual compensatory mutations may lead to the loss of base pairings, but in combination they can restore it. To assess the number of compensatory substitutions, we found orthologs of human PCCRs in 100 vertebrate genomes and calculated the measure of covariation P for every species, defined as the proportion of complementary substitutions among complementary base pairs that had at least one substitution. The ratio of the observed P to the covariation expected from a random model was greater than one for all analyzed species and, in general, decreased with increasing phylogenetic distance (Figure S7C). This approach, however, does not take into account the structure of the phylogenetic tree. We therefore analyzed PCCRs using a method implemented in the R-scape program [64], which scores multiple independent occurrences of a complementary substitution on different branches of the phylogenetic tree. This allowed estimation of non-randomness of compensatory mutations as a product of the corresponding E-values of the constituent base pairs in each PCCR. Some of the known RNA structures were moderately significant (E-value = 0.28 for *ENAH*, E-value = 0.27 for *DNM1*, and E-value = 0.40 for *PLP1*), while others (e.g. *SF1*, *DST*) did not have sufficient number of substitutions to estimate the E-value. We did not pursue this analysis any further because only a small fraction of PCCRs had enough substitutions to estimate the E-value.

Compensatory mutations may also occur in the human population, but at a substantially lower frequency. To estimate the frequency of compensatory population polymorphisms, we identified PCCRs that contain SNPs with allele frequency of 1% and higher. Out of 82,994 base pairs in PCCRs that were affected by SNPs, only 40 showed sufficient support for compensatory mutations. After randomization, in which the pairs of complementary nucleotides were randomly switched, the respective average values were 83,101 and 0.67 (see Methods, Figure S7A). The density of compensatory mutations normalized to the number of base pairs with mutations were 0.048% for the actual base pairs and $(8.0 \pm 7.8) * 10^{-4}\%$ for the randomized set. From this we conclude that, despite compensatory mutations within PCCR are quite rare in the human population, they are significantly enriched (Poisson test, $P < 10^{-5}$).

RBP binding sites

Multiple lines of evidence indicate that binding of RNA-binding proteins (RBPs), which is crucial for co- and post-transcriptional RNA processing, depends on RNA secondary structure context [66, 67, 68]. To assess the association between long-range RNA structure represented by PCCRs and RBP binding, we computed eCLIP peak frequencies [69, 70] in the vicinity of CCRs and compared them to the background eCLIP peak frequencies in conserved intronic regions surrounding CCRs.

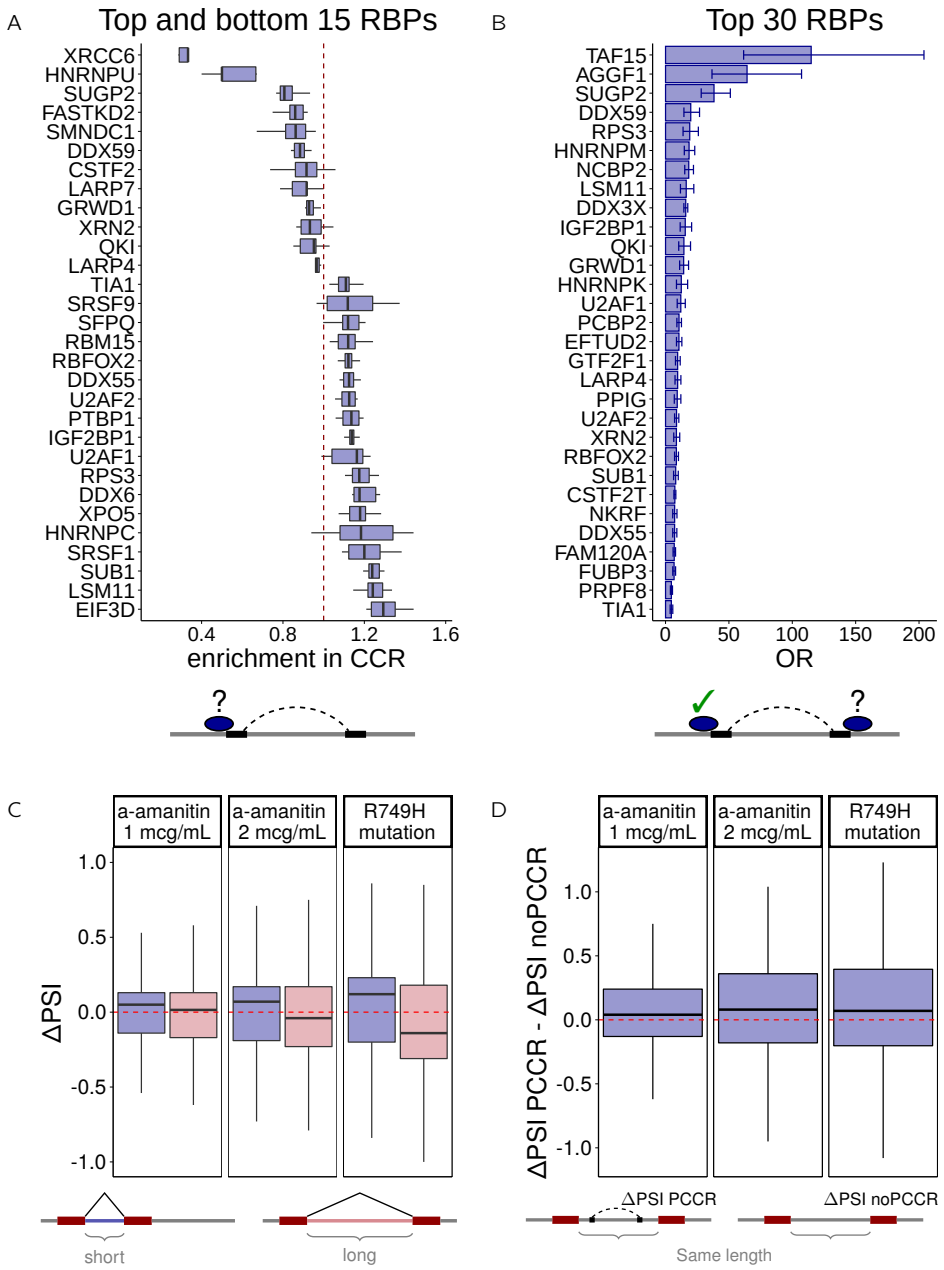


Figure 5: RNA-binding proteins (RBP). **(A)** Enrichment and depletion of eCLIP peaks overlapping CCRs. **(B)** The enrichment of forked eCLIP peaks (and depletion) expressed as the odds ratio (OR) of RBP binding near both CCR in a pair given that RBP binds near at least one CCR. **(C)** The change of inclusion rate ($\Delta\Psi$) of exons following short introns vs. exons following long introns in response to RNA Pol II slowdown with α-amanitin and in slow RNA Pol II mutants [65]. **(D)** The difference between the inclusion rate change of exons following introns with a PCCR ($\Delta\Psi$PCCR) and the inclusion rate change of exons following introns of the same length, but without PCCRs ($\Delta\Psi$noPCCR) in response to RNA Pol II slowdown.

Among factors that showed a substantial depletion of binding sites within CCR were *XRCC6* (a.k.a *Ku70*), a single-stranded DNA-dependent ATP-dependent helicase [71], and *QKI*, an ssRNA-binding protein

that plays a central role in myelinization [72]. On the other hand, binding sites of RBPs that are known to exert their function in conjunction with RNA structure, such as *RBFOX2* [39], and factors that favor increased structure over their motifs such as *SRSF9* and *SFPQ* [73], were significantly enriched (Figure 5A).

One of the features of eCLIP protocol is that in the presence of RNA secondary structure, an RBPs can crosslink with either RNA strand adjacent to the base paired region. Hence, we expect a higher chance of observing an eCLIP peak near the CCR given that the other CCR in the pair contains a nearby peak of the same RBP (it will be referred to as *forked* eCLIP peak). To estimate the magnitude of this association, we computed the respective odds ratio and found that the majority of RBPs (65 out of 74, P-value $< 10^{-12}$) indeed have a substantially higher likelihood of binding close to a CCR given that they bind the other CCR in a pair (Figure 5B). It can be regarded as independent evidence for double-stranded structure of PCCRs. Interestingly, the highest degree of this association was observed for *TAF15*, a TBP-associated factor 15, which is not a dsRNA binding protein itself, but interacts with FUS, which is capable of binding dsRNA [74]. This indicates, on one hand, that RBP binding is inseparable from the surrounding RNA structure and, on the other hand, that eCLIP peaks may not reflect the actual binding positions since they are affected by intramolecular RNA structure and interactions with other RBPs.

RNA Pol II elongation speed

The kinetic model of cotranscriptional splicing suggests that RNA Pol II elongation slowdown expands the "window of opportunity" for the recognition of weak splice sites, thereby increasing the rate of inclusion of upstream exons [65, 75]. Besides this direct impact on splice site recognition, slow RNA Pol II elongation may also affect the way the transcript folds, which is another important determinant of how the transcript will be processed by the splicing machinery [76]. To investigate the role of long-range RNA structure in co-transcriptional splicing, we performed RNA-seq experiments, in which we used α-amanitin to slow down the RNA Pol II elongation speed [77], and additionally analyzed publicly available data on the impact of RNA Pol II elongation speed on splicing [65].

The expected consequence of the RNA Pol II slowdown is that the inclusion rate of exons that follow short introns will increase, and the inclusion rate of exons that follow long introns will decrease. Indeed, this trend was observed both when RNA Pol II elongation speed was decreased by α-amanitin and in the slow RNA Pol II mutant R749H (Figure 5C) [65]. To check whether RNA Pol II slowdown differently affects introns with and without PCCRs, we matched each exon that follows an intron containing a PCCR with a randomly chosen exon that follows an intron of the same length, but without PCCRs (Figure S6). The difference in inclusion rates of these matched exons showed that exons that follow an intron with a PCCR tend to be more included than exons following an intron without PCCRs at both concentrations of the inhibitor and in R749H RNA Pol II mutant (Figure 5D). This can be considered as evidence for RNA Pol II slowdown to affect exon inclusion through pre-mRNA folding in addition to splice site recognition. Namely, slower RNA Pol II elongation speed may not only facilitate processing of upstream splice sites by the spliceosome, but

also allow sufficient time for folding intronic RNA structures, thus promoting exon inclusion.

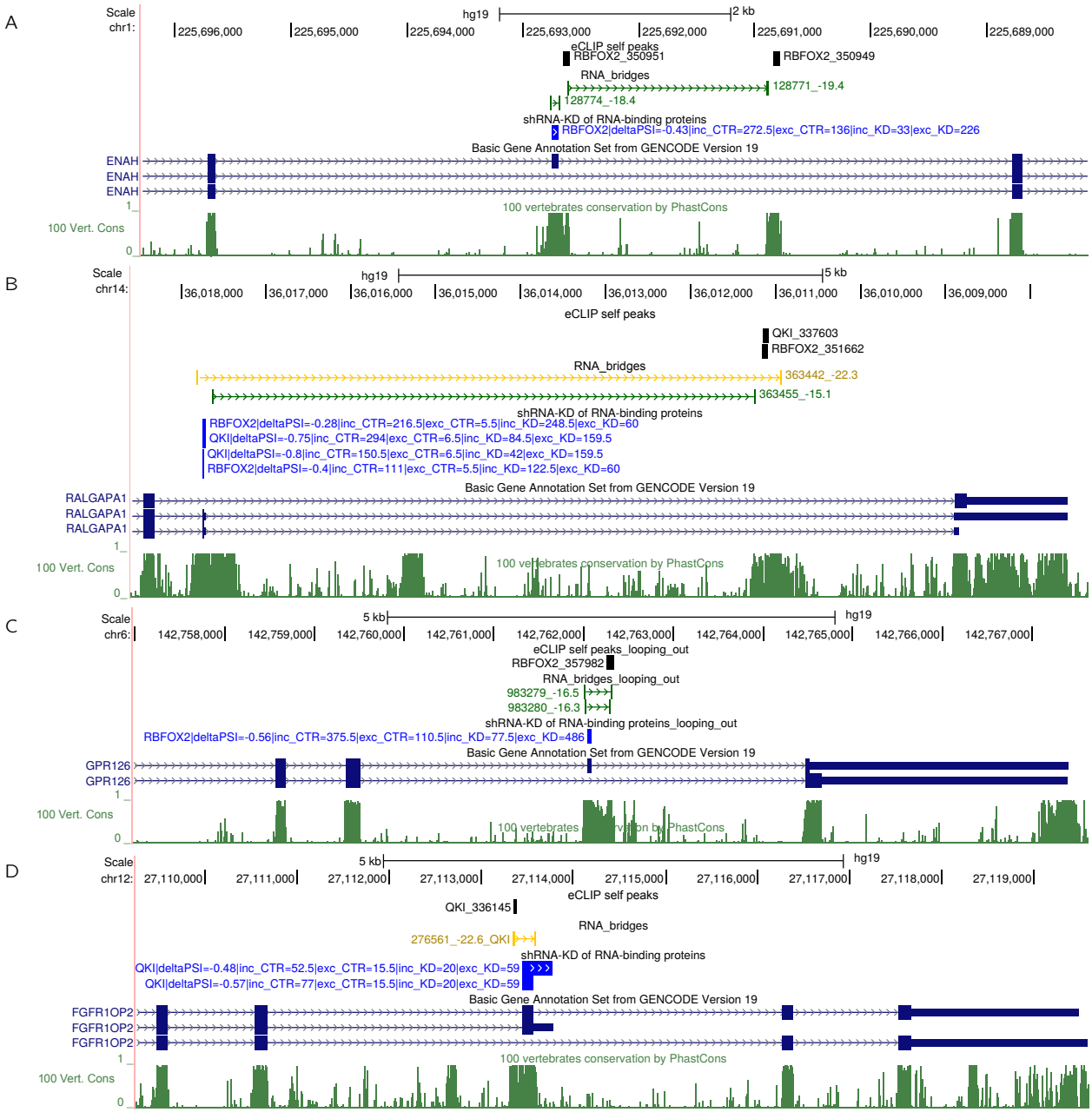


Figure 6: Case studies. **(A)** An RNA bridge in *ENAH* gene brings a distant RBFOX2 binding site into proximity of the regulated cassette exon [39]. The exon inclusion rate decreases under RBFOX2 depletion ($\Delta\Psi = -0.43$). **(B)** The predicted RNA bridge in *RALGAPA1* brings distant binding sites of RBFOX2 and QKI to the exon. The exon prominently responds to the depletion of these two factors ($\Delta\Psi = -0.28$ and $\Delta\Psi = -0.75$, respectively). **(C)** A cassette exon in *GPR126* is looped out by a PCCR overlapping an eCLIP peak of RBFOX2 and responds to RBFOX2 depletion ($\Delta\Psi = -0.56$). **(D)** An alternative terminal exon in FGFR1OP2 is looped out by a PCCR overlapping an eCLIP peak of *QKI* and significantly responds to *QKI* depletion ($\Delta\Psi = -0.48$). In all panels, Q-value < 0.01.

Case studies

In previous works, we presented a number of conserved long-range RNA structures with a potential impact on splicing based on sequence conservation patterns (*SF1*, *ZFX*,*HNRNPK*), overlaps of the double-stranded regions with splice sites or PPT (*SLC39A7*, *SRSF7*, *ZNF384*), and association with complex splicing patterns such as multiple exon skip (*DST*) [33, 34]. Here, we focus on the association of long-range RNA structures with RBP binding.

First, we identified potential RNA bridges, i.e., long-range RNA structures that bring an RBP binding site closer to the regulated exon, by searching for candidate binding sites of all RBPs profiled by eCLIP that were located within 50 nts from a CCR, on one hand, and exons located within 50 nts from its mate CCR, on the other hand. To increase specificity, we additionally required that exon inclusion rate significantly respond to shRNA-KD of the same factor using the data on RBP knockdowns produced by the ENCODE Consortium (see Methods) [70]. This procedure yielded a set of 340 candidate RNA bridges (Supplementary Data File 1), including the RNA bridge that controls the inclusion of exon 12 of *ENAH* gene (Figure 6A). A pair of CCR with the hybridization energy -19.4 kcal/mol is surrounded by *RBFOX2* peaks on both sides, likely reflecting the cross-linking of *RBFOX2* next to the double-stranded region, and the inclusion of exon 12 drops by 43% ($\Delta\Psi = -0.43$) upon *RBFOX2* knockdown. Remarkably, a shorter PCCR with the hybridization energy -18.4 kcal/mol also loops out the exon 12. As a novel example, we describe a candidate RNA bridge in the 3'-end of *RALGAPA1* gene, which encodes the major subunit of the RAL-GTPase activating protein . In this gene, a group of nested PCCRs approximates binding sites of *RBFOX2* and *QKI* to the penultimate poison exon (Figure 6B). The knockdown of each factor promotes exon skipping, suggesting that exon inclusion depends on binding of *RBFOX2* and *QKI* through an RNA bridge.

In a similar way, we identified candidate RNA structures that loop out exons by searching for PCCRs that surround an exon and contain an RBP binding site within one of its CCRs. Additionally, we required that the exon significantly responds to RBP knockdown. This procedure yielded a set of 1362 candidate RNA loop-outs (Supplementary Data File 2). Among them there were two nested RNA structures looping-out exon 24 of *GPR126*, human G protein-coupled receptor 126, in which one of the CCRs overlaps *RBFOX2* binding site, and the exon responds to *RBFOX2* knockdown (Figure 6C). Another example is the alternative 3'-end exon in *FGFR1OP2*, Fibroblast Growth Factor Receptor 1 Oncogene Partner 2, which becomes significantly less included upon *QKI* knockdown and, at the same time, is looped out by a PCCR that overlaps *QKI* binding site (Figure 6D).

To visualize these predictions, we created a track hub (https://github.com/kalmSveta/PCCR) for the UCSC Genome Browser [35]. The PCCR tracks are grouped by the energy (group I, II, III, and IV), spread, and E-value from R-scape. Additional data reported along with PCCR are the response of exons to RNA Pol II elongation slowdown and icSHAPE reactivity according to RNA probing data. The tables listing all PCCRs predicted for GRCh37 and GRCh38 genome assemblies are available as Supplementary Data Files 3 and 4, respectively. The predictions of RNA bridges and looping-out PCCRs are summarized in a standalone

track (https://github.com/kalmSveta/RNA-bridges) along with eCLIP data and exon responses to shRNA knockdowns.

Discussion

It has been increasingly acknowledged that RNA structure plays a critical role in the regulation of eukaryotic gene expression at all steps from transcription to translation, but very little attention has been paid to long-range RNA structure. From the thermodynamic standpoint, long-range base pairings contribute to the enthalpy of RNA folding as much as local base pairings do, and the corresponding energy figures exceed by an order of magnitude the typical folding energies of globular protein domains [78]. However, since RNA structure affects, and is itself strongly affected by RBP binding, reliable predictions of long-range RNA structure in full-length eukaryotic transcripts do not seem feasible at the current state of the art. Instead of the detailed structure, one could consider as a proxy for RNA structure the core of highly stable and evolutionarily conserved double-stranded regions, different combinations of which may represent one or several physiologically-relevant folds. Here, we perform for the first time a systematics study of such elements across the human transcriptome and present a catalog of PCCRs that is substantially larger than other similar lists [34].

The existence of associations between long-range RNA structure and splicing has been suggested in previous studies [7], including our earlier reports [34, 33, 28]. The extended catalog of PCCRs presented here demonstrates that the trends proposed for smaller sets also hold at the transcriptome-wide level, namely the preference of PCCRs to be positioned within introns proximally to splice sites [28], lower inclusion rate of looped-out exons [79, 80], circumscription of circRNAs [81], avoidance of intronic branch points [82, 83], and generally obstructive effect on splice sites that are implicated in double-stranded structure [84, 85, 86]. This analysis additionally demonstrates that PCCRs are concordant with the existing RNA structure probing and photo-inducible RNA crosslinking data. PCCR show a considerable depletion of SNVs and enrichment of compensatory substitutions. The observed correlations of PCCR with A-to-I RNA editing sites not only reconfirm the well-known biogenesis of ADAR-mediated pathway [87], but also additionally support the validity of the present catalog of long-range PCCRs.

High-throughput assays aimed at the determination of RNA elements responsible for RBP binding, both *in vivo* [70] and *in vitro* [73], have underscored the importance of RNA structural context for the assembly of RNA-protein complexes. Here we report that the structural context of RBP binding specificity must also include long-range base pairings, with preference for some RBPs and avoidance for the others. In eCLIP data, we observe traces of multi-molecular complexes with patterns that are specific to double-stranded regions spanning very large distances. This, on one hand, gives much better understanding of cross-linking and immunoprecipitation assays and, on the other hand, indicates that abundant base pairings may exist between distant RNA regions.

The components of RNA processing machinery operate in a strict coordination not only in space but also

in time. The kinetic profile of RNA Pol II elongation has a significant impact on alternative splicing [88]. Slow RNA Pol II elongation generally opens a window of opportunity for weak splice sites to be recognized, leading to higher inclusion of alternative exons, although in some cases the effect can be the opposite [89, 90]. Slow RNA Pol II elongation may also influence poly(A) site choice by enhancing the recognition of suboptimal polyadenylation signals [91]. Consistent with this, we see an increased inclusion of exons preceded by shorter introns. However, we also observe an additional component in this general trend that is associated with RNA structure, which suggests that attenuation mechanisms may play a role in eukaryotic mRNA processing [92, 93].

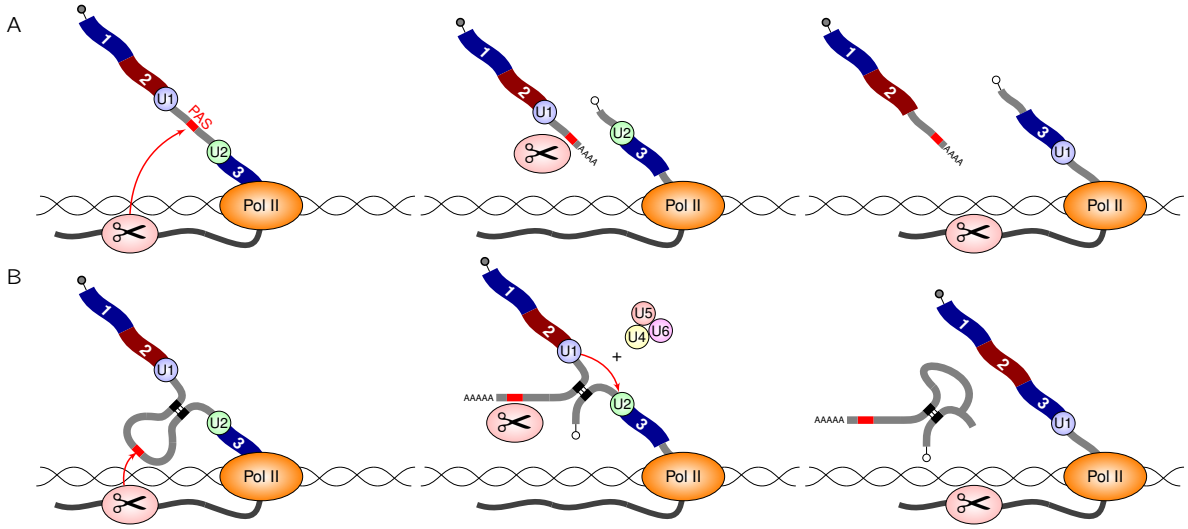


Figure 7: RNA structure can mediate coupling of splicing and cleavage/polyadenylation (a hypothesis). **(A)** In the absence of RNA structure, the first active PAS can be recognized and used for cleavage/polyadenylation. The elongation by RNA Pol II may continue. **(B)** In the presence of RNA structure, the same PAS can be recognized and used. However, if splicing occurs while the pieces of the cleaved transcript are held together by RNA structure, the mRNA will not be affected by cleavage/polyadenylation. This could be a general mechanism of suppression of premature transcription termination. The switch between (A) and (B) depends on the rates of individual RNA processing steps as well as on the rate of RNA Pol II elongation.

RNA structure is implicated in the recognition of polyadenylation signals (PAS) by cleavage and polyadenylation specificity factor (CPSF) and facilitates the 3'-end processing by juxtaposing PAS and cleavage sites that are otherwise too far apart [94, 8]. Functional RNA structures in 5'-UTRs are also implicated in the regulation of translation [11]. While these reports mostly concern local RNA structure, an intriguing finding of this work is the link between long-range RNA structure and pre-mRNA 3'-end processing, which is manifested by the enrichment of poly(A)-seq clusters within PCCRs (Figure 4E and 4D). It indicates that mechanisms other than sequestration or spatial convergence of PAS and cleavage sites may be involved, including co-transcriptional coupling between splicing and 3'-end processing [95, 96]. An example of RNA structure-mediated switch that controls both splicing and polyadenylation was described earlier in the *Nmnat* gene in *D. melanogaster* [28].

Human genes contain thousands of dormant intronic PASs that are suppressed, at least in part, by U1 small nuclear ribonucleoproteins in a process called telescripting [97]. While the exact mechanism of this suppression is not known, many intronic PASs were also found to be associated with *CstF64*, an essential pre-mRNA 3'-processing factor, *in vivo* [98]. This suggests that cleavage and polyadenylation machinery may actually operate in all introns constitutively, whereas the results of pre-mature intronic polyadenylation could be removed by co-transcriptional splicing. Here, the essential role of RNA structure could be to stabilize the transcript through intramolecular base pairings despite disruption of the linear sequence by cleavage and allowing sufficient time for co-transcriptional splicing to occur (Figure 7A). This would explain high enrichment of poly(A)-seq clusters inside PCCR. The same mechanism could work in a regulatory mode, where the outcome of alternative polyadenylation must depend on RNA structure and timing of co-transcriptional splicing (Figure 7B).

Exhaustive transcriptome annotations such as ENSEMBL contain transcript isoforms that show conspicuous association between 5'-end and 3'-end processing [99]. A typical example is an intron that contains a 3'-end of protein-coding transcript and a 5'-end of another, usually non-coding transcript (Figure S8). Here, we observe that it is 2.86 ± 0.10 times more likely to see a 5'-end in an intron that contains a 3'-end, and hence the enrichment of poly(A)-seq clusters within PCCRs also implies the enrichment of 5'-ends and CAGE clusters. These observations are, at least in part, physiological because U1 snRNP knockdown causes splicing inhibition and premature cleavage and polyadenylation in numerous introns near the start of the transcript [100], and there are multiple lines of evidence for the existence of PAS in 5'-UTR in plants [101, 102]. Therefore, the same RNA structure-mediated mechanism of coupling between cleavage, polyadenylation, and splicing may also exist in 5'-UTRs, however the link between these processes and RNA recapping is currently missing. As the existence of decapping and recapping pathways has already been established [103, 104], we expect that establishing connections between all steps of RNA processing will soon be a matter of intensive research.

Conclusion

Eukaryotic pre-mRNAs are structured macromolecules with base pairings that span long distances, and secondary structure plays a critical role in their processing. Here, we present the most complete to-date catalog of conserved complementary regions in human protein-coding genes and provide their extended characterization by other methods. While the overall false discovery rate could be high, different subsets of this catalog have a substantially lower proportion of false positives. We offer this catalog for common use and provide its convenient visualization through a UCSC Genome Browser track hub.

Acknowledgements

RNA-sequencing was funded by Skolkovo Institute of Science and Technology. The authors would like to thank Skoltech Genomics Core facility staff (Margarita Ezhova, Maria Logacheva) for library preparation and sequencing, and the `Cobrain` computer cluster for access to computational facilities. S.K., S.D. and D.P. were supported by Russian Foundation for Basic Research grant 18-29-13020-MK and by Skolkovo Institute of Science Technology Research Grant RF-0000000653. M.K. and D.S. were supported by Russian Foundation for Basic Research grant 18-29-13020-MK.

Author contributions

D.P. and S.K. designed and carried out the analysis. M.K. and D.S. performed the experiments. S.D. and A.M. analyzed the evolutionary part. All authors analyzed the data. D.P. and S.K. wrote the paper.

Methods

Genomes and transcript annotations

February 2009 (hg19, GRCh37) and December 2013 (hg38, GRCh38) assemblies of the human genome were downloaded from Genome Reference Consortium [105]. These assemblies were used with GENCODE transcript annotations v19 and v33, respectively [106]. Only genes labelled "protein coding" were analyzed. Transcript annotations were parsed by custom scripts to extract the coordinates of exons and introns. Throughout the paper, all results are reported for GRCh37 assembly unless the contrary is specified. The Genome Browser track hubs (see below) contain tracks for both GRCh37 and GRCh38 assemblies. The intronic regions were defined as the longest continuous segments within genes which don't overlap any annotated exons. The intronic regions were extended by 10 nts into both flanking exons. Next, the intronic regions were intersected with the phastCons-derived set of conserved regions (phastCons elements track for alignment of 99 vertebrates genomes to the human genome [107]).

PREPH

The PREPH utility uses a k-mer-based technique, which is similar to previously published IRBIS method [34], to identify all pairs of nearly-perfect complementary regions in a given pair of sequences. At the preparatory step, PREPH pre-computes a $4^k \times 4^k$ table containing helix hybridization energies for all pairs of k-mers (default $k = 5$) that are either perfectly Watson-Crick complementary, or contain a few GT base-pairs (by default at most two) using energy tables from Vienna RNA package [108]. The dynamic programming matrix is computed by local Smith-Waterman algorithm using a limited set of structural rules: initiating a k-nt-long helix, extending a helix by a stacking base pair, and adding to a helix a short internal loop or

bulge with up to m nucleotides in each strand (default $m = 2$), followed by another helix. To speed up backtracking, PREPH uses auxiliary matrices to store intermediate structures, and reports non-intersecting pairs of complementary regions passing the energy threshold (by default -15 kcal/mol). Here, two pairs of complementary regions, x_1 complementary to y_1 and x_2 complementary to y_2, are referred to as intersecting if x_1 has common nucleotides with x_2 and also y_1 has common nucleotides with y_2. That is, two pairs of complementary regions may intersect by only one, not both interacting strands. The reduced scoring scheme, optimized back-tracking, and indexing of the initial sequences by k-mers result in a great improvement of computation speed. The detailed description of PREPH is exempt to Supplementary Methods. PREPH software is available at github (https://github.com/kalmSveta/PrePH).

Relative position within the gene

The relative position of a genomic interval $[x, y]$ in the containing gene $[a, b]$, where x, y, a, and b are genomic coordinates on the plus strand, was calculated as $p = \frac{x-a}{(y-x)-(b-a)+1}$ for the genes on the positive strand. For the genes on the negative strand the value of $1 - p$ was used instead.

Gene Ontology (GO) enrichment analysis

Since genes with PCCRs tend to be longer than genes without PCCRs, we randomly matched each gene with PCCRs to a gene without PCCR of approximately the same length. The enrichment of GO terms [109, 110] between these two gene sets was calculated with clusterProfiler R package [111].

Benchmarking

The accuracy and runtime of PREPH were compared with IntaRNA2.0 [41], RIsearch2 [42], RNAplex [43], DuplexFold [44] and bifold [44]. PREPH was run with the following parameters: k-mer length is 5 nts, maximal distance between complementary regions is 10,000 nts, the minimal length of the aligned regions is 10 nts, the energy threshold is -15 kcal/mol, the maximal number of Wobble pairs in a k-mer is 2. The parameters for the other programs are listed below.

Speed comparison

To benchmark the time efficiency, we use a set of 1000 pairs of randomly chosen conserved intronic sequences from the human genome. The sequences were 50 to 500 nucleotides-long and contained nearly perfect sequence complementarity. All the programs were run with the energy threshold set to -15 kcal/mol. `IntaRNA2.0` *outOverlap* parameter was set to *B*, which allowed overlap for interacting subsequences for both target and query; *n* parameter was set to 100 to limit the maximal number of suboptimal structures; *qAcc* and *tAcc* were set to N to omit the computation of accessibility. `RIsearch2` seed length was set to 5, the length of flanking sequences considered for seed extension was set to 50. `RNAplex` *fast-folding* parameter

was set to *f2* to allow the structure to be computed based on the approximated model. `DuplexFold` and `bifold` maximum loop/bulge size was set to two. All other parameters were left at their default values. The computations were carried out on Intel R Core TM i5-8250U CPU with 1.60 GHz.

Optimal energy and structure comparison

For the comparison of MFE between different methods, we used simulated data with 1000 pairs of nearly perfect sequence complementarity, which were from 10 to 50 nts. All the programs were run with the energy threshold set to -10 kcal/mol. The other parameters were the same as before. Pearson correlation coefficients were computed between energies of the predicted optimal structures. To compare the predictions of PREPH with predictions of other programs at the level of individual base pairs, we computed the following metric

$$\text{Score} = \frac{|S_1 \cap S_2|}{|S_1|},$$

where S_1 is the set of base pairs predicted by PREPH, S_2 is the set of base pairs predicted by the other program, and $S_1 \cap S_2$ is the common set of base pairs ($|S|$ denotes cardinality of a set).

RNA probing data

The icSHAPE reactivity scores [112] were mapped from GRCh38 to GRCh37 human genome assembly by LiftOver [113]. The reactivity score of a CCR was calculated as the average reactivity of its base paired nucleotides, for which the icSHAPE reactivity score was available. The background reactivity was calculated as the average reactivity of the same number of nucleotides chosen at random outside the CCR, but within its conserved intronic region. Wilcoxon signed rank test was applied to test matched samples of reactivity score differences for departures from zero. The nucleotides with the reactivity score of 0.5 and below were regarded as base-paired and used to calculate the precision and recall (see below).

The coordinates of base paired regions from PARIS experiments [25] were mapped from GRCh38 to GRCh37 by Liftover [113]. Additionally, the coordinates of base paired regions from LIGR-seq data were used [26]. The genomic coordinates of the ligation site in both parts of the split-read were identified and extended by the flanks of 200 nts upstream of the ligation site (for the 3'-part of the split-read) or downstream (for the 5'-part of the split-read). In order to evaluate the precision and recall, we selected conserved intronic regions with at least one nucleotide overlapping the experimentally validated structures and confined our analysis to PCCRs in these regions. A CCR was classified as a true positive if it had at least one common nucleotide with an experimentally validated structure.

Cell culture, treatments and RNA purification

A549 cell line was maintained in DMEM/F-12 medium containing 10% fetal bovine serum, 50 U/ml penicillin, and 0.05 mg/ml streptomycin (all products from Thermo Fisher Scientific) at 37°C in 5% CO_2. For α-

amanitin (Sigma) treatments 1 and 2 μg/mL of α-amanitin was added to cells at 50-70% confluency. After 24h of treatment, cells were harvested, total RNA was isolated using PureLink RNA Mini Kit (Thermo Fisher Scientific). Poly(A)$^+$ mRNA was purified using Dynabeads Oligo(dT) 25 (Thermo Fisher Scientific) following the manufacturer's instructions.

Library preparation and RNA sequencing

Illumina cDNA libraries were constructed using NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England BioLabs) following the manufacturer's protocol with the only modification - the change of fragmentation time from 15 to 10 minutes. cDNA libraries were sequenced using the NextSeq500 (Illumina, San Diego, CA USA) instrument; 33–41 million raw reads were obtained for each sample with a 75 bp read length.

Splicing quantification

RNA-seq data of poly(A)$^+$ RNA from the HepG2 cell line were downloaded in BAM format from the ENCODE Consortium website [114]. Short-hairpin shRNA knock-down (shRNA-KD) of 250 RBPs followed by RNA-seq data [70] were downloaded in BAM format from ENCODE data repository [115, 114] (Table S2). Poly(A)$^+$ RNA from wild-type *Amr* and *Rpb1* C4/R749H mutant HEK293 cells treated with α-amanitin for 42 h were downloaded from the Gene Expression Omnibus (GSE63375) [65]. RNA-seq data of poly(A)$^+$ RNA data from the A549 cell line treated with α-amanitin were obtained as explained below and mapped to the GRCh37 human genome assembly using STAR aligner [116]. The coordinates of circRNAs expressed in liver tissue and their associated SRPTM metrics (the number of circular reads per number of mapped reads per read length) were obtained from TCSD database [49]. The genomic coordinates of adenine branch point nucleotides were selected from the validated set of branchpoints expressed in K562 cells [45].

RNA-seq experiments were processed by IPSA pipeline to obtain split read counts supporting splice junctions [117]. Split read counts were filtered by the entropy content of the offset distribution, annotation status and canonical GT/AG dinucleotides at splice sites, and pooled between bioreplicates. The exon inclusion rate (Ψ, PSI, or Percent-Spliced-In) was calculated according to the equation

$$\Psi = \frac{inc}{inc + 2 * exc},$$

where inc is the number of reads supporting exon inclusion and exc is the number of reads supporting exon exclusion. Ψ values with the denominator below 10 were considered unreliable and discarded. Differential exon inclusion between a pair of conditions (shRNA-KD vs. non-specific control and α-amanitin vs. untreated control) was assessed as described previously [118].

Cryptic and actively-expressed splices sites in the human transcriptome were identified using genomic alignments of RNA-seq samples from the GTEx Consortium [47]. Splice sites with the canonical GT/AG dinucleotides were called from split read alignments and ranked by the total number of supporting split

reads pooled across all 8,551 samples. The top 2% (respectively, bottom 2%) of splice sites among those supported by at least three split reads were referred to as active (respectively, inactive). In order to identify cryptic splice sites, we applied the same strategy as [119] by scanning the intron sequences for any sites that have a MaxEntScan score >500 for donor sites and >650 for acceptor sites [120] and excluding splice sites that were detected in GTEx. MaxEntScan score thresholds were chosen to have a comparable number of splice sites as in active and inactive sets above.

5'-end and 3'-end RNA processing

The genomic coordinates of human poly(A) sites profiled by high-throughput sequencing of 3'-ends of polyadenylated transcripts (poly(A)-seq) that were supported by 20 or more reads and intersected with the annotated transcript ends in GENCODE database were used [60]. Similarly, clusters of human CAGE tags expressed in the HepG2 cell line that were supported by RPKM of at least 10 and intersected with the annotated transcript starts in GENCODE database were used [121].

RNA editing

Adenosine-to-inosine (A-to-I) RNA editing sites were obtained from RADAR [56] and REDIportal [57] databases. To compare the density of RNA editing sites in CCR and that in the adjacent conserved regions, we computed the number of adenosine residues that are RNA editing sites within CCR and compared it to the respective figures for conserved intronic regions of the same length outside of CCR. Odds ratio was calculated for the contingency table of adenosine residues that are/are not RNA editing sites within/outside CCR.

eCLIP

Enhanced cross-linking and immunoprecipitation (eCLIP) peaks for 74 RBPs assayed in HepG2 human cell line were downloaded from the NCODE data repository in bed format [115, 114] (Table S3). The peaks were filtered by the conditions logFC ≥ 3 and P-value < 0.001. Since the agreement between two replicates was moderate, we use the pooled set of eCLIP peaks. To quantify the association between RBP binding and individual CCR, we calculated for each RBP the number of CCRs that intersected with its eCLIP peaks by at least 50% of the CCR length. This number was compared to the respective number of intersections obtained in the random shift control. To quantify the association of RBP binding with both CCRs in a PCCR, we constructed a contingency table for the number of PCCRs that had/didn't have eCLIP peaks in left/right CCR for each RBP, and computed the respective odds ratio (OR). The 95% confidence interval for the odds ratio was calculated from Fisher test P-value.

Population polymorphisms

For the evaluation of mutation enrichment in CCR, we used genotyping data from phase 3 of the 1000 Genomes Project for 2504 individuals [63] and computed the density of SNPs (SNVs present in more than 0.1% of individuals) in stacked nucleotide pairs of CCR and a similar quantity in conserved regions outside CCRs (regions of each type were merged with `bedtools merge` [122] before calculating the densities).

To assess the impact of SNPs on RNA structure stability, we calculated PCCR energy for the mutated sequence using energy parameters [108] and compared it to the respective energy of the structure with SNPs of the same substitution type as observed originally, but introduced at a different position. We selected PCCRs with the free energy change greater than 2 kcal/mol by absolute value and compared the two energy sets using Wilcoxon signed-rank test.

To evaluate the enrichment of compensatory mutations in PCCRs, we selected SNVs that occur in more than 1% of donors in the 1000 Genome Project and intersected their list with the list of base pairs in PCCRs. Among them we estimated the number of base pairs, in which a compensatory mutation occurred in more than 1% of donors. To estimate the expected number of base pairs with compensatory mutations, we first subdivided base pairs into groups composed of the same base pair types (AT, TA, CG, GC, TG, GT) located on the same chromosome with the same number of SNP donors. Then we then randomly interchanged ("re-wired") base pairs within each group, e.g. A_1T_1 and A_2T_2 were replaced by A_1T_2 and A_2T_1 and applied the same procedure again (Figure S7A). This randomization procedure was repeated 1000 times.

Phylogenetic analysis

To assess the degree of evolutionary conservation of a CCR, we computed the difference between the average PhastCons conservation score of all its nucleotides and the average PhastCons conservation score of the same number of nucleotides in its flanking regions within the same `phastConsElements` interval.

We assessed the number of complementary substitutions in PCCRs that occurred between human and a vertebrate species by identifying orthologs of human PCCRs in whole genome alignments of the human genome with the genomes of 99 vertebrates [123]. For each species, we selected regions that align to both CCR in a pair without gaps and computed the measure of covariation P defined as the number of base pairs within all PCCRs with compensatory substitutions as a fraction of the number of base pairs within all PCCRs that had a substitution in at least one of the two complementary nucleotides. The expected proportion of compensatory substitutions was computed based on the assumption of independence.

To test whether complementary substitutions support the presence of a conserved RNA secondary structure in each individual PCCR, we extracted non-gapped orthologs of human PCCRs from multiple sequence alignments and applied R-scape tool [64] to stem-loop structures obtained from PCCR sequences via linking through an artificial linker using a pre-computed phylogenetic tree of vertebrates from [123]. The E-values reported by R-scape were multiplied for each base pair in a PCCR, and the obtained products were adjusted

with Benjamini-Hochberg correction.

Statistical analysis

The data were analyzed and visualized using R statistics software version 3.4.1 and ggplot2 package. Non-parametric tests were performed by built-in R functions using normal approximation with continuity correction. Error bars in all figures and the numbers after the $\pm$ sign represent 95% confidence intervals. One-sided P-values are reported throughout the paper.

References

[1] Breaker RR. Riboswitches and the RNA world. Cold Spring Harb Perspect Biol. 2012 Feb;4(2).

[2] Bowman JC, Hud NV, Williams LD. The ribosome challenge to the RNA world. J Mol Evol. 2015 Apr;80(3-4):143–161.

[3] Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet. 2016 Jan;17(1):47–62.

[4] Marchese FP, Raimondi I, Huarte M. The multidimensional mechanisms of long noncoding RNA function. Genome Biol. 2017 10;18(1):206.

[5] Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature. 2012 Feb;482(7385):339–346.

[6] Silverman IM, Li F, Gregory BD. Genomic era analyses of RNA secondary structure and RNA-binding proteins reveal their significance to post-transcriptional regulation in plants. Plant Sci. 2013 May;205-206:55–62.

[7] Baralle FE, Singh RN, Stamm S. RNA structure and splicing regulation. Biochim Biophys Acta Gene Regul Mech. 2019;1862(11-12):194448.

[8] Wu X, Bartel DP. Widespread Influence of 3'-End Structures on Mammalian mRNA Processing and Stability. Cell. 2017 May;169(5):905–917.

[9] Rieder LE, Reenan RA. The intricate relationship between RNA structure, editing, and splicing. Semin Cell Dev Biol. 2012 May;23(3):281–288.

[10] Garcia-Lopez A, Tessaro F, Jonker HRA, Wacker A, Richter C, Comte A, et al. Targeting RNA structure in SMN2 reverses spinal muscular atrophy molecular phenotypes. Nat Commun. 2018 05;9(1):2032.

[11] Leppek K, Das R, Barna M. Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat Rev Mol Cell Biol. 2018 03;19(3):158–174.

[12] Bernat V, Disney MD. RNA Structures as Mediators of Neurological Diseases and as Drug Targets. Neuron. 2015 Jul;87(1):28–46.

[13] Singh NN, Singh RN. How RNA structure dictates the usage of a critical exon of spinal muscular atrophy gene. Biochim Biophys Acta Gene Regul Mech. 2019;1862(11-12):194403.

[14] Ramani V, Qiu R, Shendure J. High-throughput determination of RNA structure by proximity ligation. Nat Biotechnol. 2015 Sep;33(9):980–984.

[15] Mortimer SA, Kidwell MA, Doudna JA. Insights into RNA structure and function from genome-wide studies. Nat Rev Genet. 2014 Jul;15(7):469–479.

[16] Graveley BR. RNA Matchmaking: Finding Cellular Pairing Partners. Mol Cell. 2016 07;63(2):186–189.

[17] Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature. 2014 Jan;505(7485):696–700.

[18] Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature. 2014 Jan;505(7485):701–705.

[19] Spitale RC, Flynn RA, Zhang QC, Crisalli P, Lee B, Jung JW, et al. Structural imprints in vivo decode RNA regulatory mechanisms. Nature. 2015 Mar;519(7544):486–490.

[20] Flynn RA, Zhang QC, Spitale RC, Lee B, Mumbach MR, Chang HY. Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE. Nat Protoc. 2016 Feb;11(2):273–290.

[21] Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, et al. Landscape and variation of RNA secondary structure across the human transcriptome. Nature. 2014 Jan;505(7485):706–709.

[22] Watters KE, Lucks JB. Mapping RNA Structure In Vitro with SHAPE Chemistry and Next-Generation Sequencing (SHAPE-Seq). Methods Mol Biol. 2016;1490:135–162.

[23] Pervouchine DD. Towards Long-Range RNA Structure Prediction in Eukaryotic Genes. Genes (Basel). 2018 Jun;9(6).

[24] Aw JG, Shen Y, Wilm A, Sun M, Lim XN, Boon KL, et al. In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. Mol Cell. 2016 05;62(4):603–617.

[25] Lu Z, Zhang QC, Lee B, Flynn RA, Smith MA, Robinson JT, et al. RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. Cell. 2016 May;165(5):1267–1279.

[26] Sharma E, Sterne-Weiler T, O'Hanlon D, Blencowe BJ. Global Mapping of Human RNA-RNA Interactions. Mol Cell. 2016 05;62(4):618–626.

[27] Jin Y, Dong H, Shi Y, Bian L. Mutually exclusive alternative splicing of pre-mRNAs. Wiley Interdiscip Rev RNA. 2018 05;9(3):e1468.

[28] Raker VA, Mironov AA, Gelfand MS, Pervouchine DD. Modulation of alternative splicing by long-range RNA structures in Drosophila. Nucleic Acids Res. 2009 Aug;37(14):4533–4544.

[29] Hsu MT, Parvin JD, Gupta S, Krystal M, Palese P. Genomic RNAs of influenza viruses are held in a circular conformation in virions and in infected cells by a terminal panhandle. Proc Natl Acad Sci USA. 1987 Nov;84(22):8140–8144.

[30] Van Nostrand EL, Huelga SC, Yeo GW. Experimental and Computational Considerations in the Study of RNA-Binding Protein-RNA Interactions. Adv Exp Med Biol. 2016;907:1–28.

[31] Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, et al. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol. 2006 Apr;2(4):e33.

[32] Rivas E, Clements J, Eddy SR. Estimating the power of sequence covariation for detecting conserved RNA structure. Bioinformatics. 2020 Feb;.

[33] Pervouchine DD, Khrameeva EE, Pichugina MY, Nikolaienko OV, Gelfand MS, Rubtsov PM, et al. Evidence for widespread association of mammalian splicing and conserved long-range RNA structures. RNA. 2012 Jan;18(1):1–15.

[34] Pervouchine DD. IRBIS: a systematic search for conserved complementarity. RNA. 2014 Oct;20(10):1519–1531.

[35] Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, et al. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics. 2014 Apr;30(7):1003–1005.

[36] Margulies EH, Blanchette M, Haussler D, Green ED. Identification and characterization of multi-species conserved sequences. Genome Res. 2003 Dec;13(12):2507–2518.

[37] Taube JR, Sperle K, Banser L, Seeman P, Cavan BC, Garbern JY, et al. PMD patient mutations reveal a long-distance intronic interaction that regulates PLP1/DM20 alternative splicing. Hum Mol Genet. 2014 Oct;23(20):5464–5478.

[38] Suyama M. Mechanistic insights into mutually exclusive splicing in dynamin 1. Bioinformatics. 2013 Sep;29(17):2084–2087.

[39] Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, et al. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat Struct Mol Biol. 2013 Dec;20(12):1434–1442.

[40] Li P, Shi R, Zhang QC. icSHAPE-pipe: A comprehensive toolkit for icSHAPE data analysis and evaluation. Methods. 2019 Oct;.

[41] Mann M, Wright PR, Backofen R. IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. Nucleic Acids Res. 2017 07;45(W1):W435–W439.

[42] Alkan F, Wenzel A, Palasca O, Kerpedjiev P, Rudebeck AF, Stadler PF, et al. RIsearch2: suffix array-based large-scale prediction of RNA-RNA interactions and siRNA off-targets. Nucleic Acids Res. 2017 05;45(8):e60.

[43] Tafer H, Hofacker IL. RNAplex: a fast tool for RNA-RNA interaction search. Bioinformatics. 2008 Nov;24(22):2657–2663.

[44] Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics. 2010 Mar;11:129.

[45] Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, et al. Genome-wide discovery of human splicing branchpoints. Genome Res. 2015 Feb;25(2):290–303.

[46] Buratti E, Baralle FE. Influence of RNA secondary structure on the pre-mRNA splicing process. Mol Cell Biol. 2004 Dec;24(24):10505–10514.

[47] Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015 May;348(6235):660–665.

[48] Deshler JO, Rossi JJ. Unexpected point mutations activate cryptic 3' splice sites by perturbing a natural secondary structure within a yeast intron. Genes Dev. 1991 Jul;5(7):1252–1263.

[49] Xia S, Feng J, Lei L, Hu J, Xia L, Wang J, et al. Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes. Brief Bioinformatics. 2017 Nov;18(6):984–992.

[50] Ottesen EW, Luo D, Seo J, Singh NN, Singh RN. Human Survival Motor Neuron genes generate a vast repertoire of circular RNAs. Nucleic Acids Res. 2019 04;47(6):2884–2905.

[51] Walkley CR, Li JB. Rewriting the transcriptome: adenosine-to-inosine RNA editing by ADARs. Genome Biol. 2017 10;18(1):205.

[52] Bass BL. RNA editing by adenosine deaminases that act on RNA. Annu Rev Biochem. 2002;71:817–846.

[53] Hogg M, Paro S, Keegan LP, O'Connell MA. RNA editing by mammalian ADARs. Adv Genet. 2011;73:87–120.

[54] Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem. 2010;79:321–349.

[55] Liddicoat BJ, Piskol R, Chalk AM, Ramaswami G, Higuchi M, Hartner JC, et al. RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. Science. 2015 Sep;349(6252):1115–1120.

[56] Ramaswami G, Li JB. RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. 2014 Jan;42(Database issue):D109–113.

[57] Picardi E, D'Erchia AM, Lo Giudice C, Pesole G. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. Nucleic Acids Res. 2017 01;45(D1):D750–D757.

[58] Pan H, Shi Y, Chen S, Yang Y, Yue Y, Zhan L, et al. Competing RNA pairings in complex alternative splicing of a 3' variable region. RNA. 2018 11;24(11):1466–1480.

[59] Wright JC, Mudge J, Weisser H, Barzine MP, Gonzalez JM, Brazma A, et al. Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. Nat Commun. 2016 06;7:11778.

[60] Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, et al. A quantitative atlas of polyadenylation in five mammals. Genome Res. 2012 Jun;22(6):1173–1183.

[61] Fejes-Toth K, Sotirova V, Sachidanandam R, Assaf G, Hannon GJ, Kapranov P, et al. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. Nature. 2009 Feb;457(7232):1028–1032.

[62] Chamary JV, Hurst LD. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. Genome Biol. 2005;6(9):R75.

[63] Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.

[64] Rivas E, Clements J, Eddy SR. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. Nat Methods. 2017 01;14(1):45–48.

[65] Fong N, Kim H, Zhou Y, Ji X, Qiu J, Saldi T, et al. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. Genes Dev. 2014 Dec;28(23):2663–2676.

[66] Taliaferro JM, Lambert NJ, Sudmant PH, Dominguez D, Merkin JJ, Alexis MS, et al. RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation. Mol Cell. 2016 10;64(2):294–306.

[67] Kazan H, Ray D, Chan ET, Hughes TR, Morris Q. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLoS Comput Biol. 2010 Jul;6:e1000832.

[68] Li X, Quon G, Lipshitz HD, Morris Q. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA. 2010 Jun;16(6):1096–1107.

[69] Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods. 2016 06;13(6):508–514.

[70] Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, et al. A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. bioRxiv. 2018;Available from: `https://www.biorxiv.org/content/early/2018/10/05/179648`.

[71] Willis DM, Loewy AP, Charlton-Kachigian N, Shao JS, Ornitz DM, Towler DA. Regulation of osteocalcin gene expression by a novel Ku antigen transcription factor complex. J Biol Chem. 2002 Oct;277(40):37280–37291.

[72] Aberg K, Saetre P, Jareborg N, Jazin E. Human QKI, a potential regulator of mRNA expression of human oligodendrocyte-related genes involved in schizophrenia. Proc Natl Acad Sci USA. 2006 May;103(19):7482–7487.

[73] Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, et al. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. Mol Cell. 2018 06;70(5):854–867.

[74] Wang X, Schwartz JC, Cech TR. Nucleic acid-binding specificity of human FUS protein. Nucleic Acids Res. 2015 Sep;43(15):7535–7543.

[75] Schor IE, Gómez Acuña LI, Kornblihtt AR. Coupling between transcription and alternative splicing. Cancer Treat Res. 2013;158:1–24.

[76] Saldi T, Cortazar MA, Sheridan RM, Bentley DL. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. J Mol Biol. 2016 06;428(12):2623–2635.

[77] Rudd MD, Luse DS. Amanitin greatly reduces the rate of transcription by RNA polymerase II ternary complexes but fails to inhibit some transcript cleavage modes. J Biol Chem. 1996 Aug;271(35):21549–21558.

[78] Pace CN. Conformational stability of globular proteins. Trends Biochem Sci. 1990 Jan;15(1):14–17.

[79] Warf MB, Berglund JA. Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem Sci. 2010 Mar;35(3):169–178.

[80] Solnick D, Lee SI. Amount of RNA secondary structure required to induce an alternative splice. Mol Cell Biol. 1987 Sep;7(9):3194–3198.

[81] Welden JR, Stamm S. Pre-mRNA structures forming circular RNAs. Biochim Biophys Acta Gene Regul Mech. 2019;1862(11-12):194410.

[82] Goguel V, Wang Y, Rosbash M. Short artificial hairpins sequester splicing signals and inhibit yeast pre-mRNA splicing. Mol Cell Biol. 1993 Nov;13(11):6841–6848.

[83] Jacquenet S, Ropers D, Bilodeau PS, Damier L, Mougin A, Stoltzfus CM, et al. Conserved stem-loop structures in the HIV-1 RNA region containing the A3 3' splice site and its cis-regulatory element: possible involvement in RNA splicing. Nucleic Acids Res. 2001 Jan;29(2):464–478.

[84] Estes PA, Cooke NE, Liebhaber SA. A native RNA secondary structure controls alternative splice-site selection and generates two human growth hormone isoforms. J Biol Chem. 1992 Jul;267(21):14902–14908.

[85] Singh NN, Singh RN, Androphy EJ. Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. Nucleic Acids Res. 2007;35(2):371–389.

[86] McManus CJ, Graveley BR. RNA structure and the mechanisms of alternative splicing. Curr Opin Genet Dev. 2011 Aug;21(4):373–379.

[87] Ramaswami G, Deng P, Zhang R, Anna Carbone M, Mackay TFC, Billy Li J. Genetic mapping uncovers cis-regulatory landscape of RNA editing. Nat Commun. 2015 Sep;6:8194.

[88] de la Mata M, Muñoz MJ, All'o M, Fededa JP, Schor IE, Kornblihtt AR. RNA Polymerase II Elongation at the Crossroads of Transcription and Alternative Splicing. Genet Res Int. 2011;2011:309865.

[89] de la Mata M, Alonso CR, Kadener S, Fededa JP, Blaustein M, Pelisch F, et al. A slow RNA polymerase II affects alternative splicing in vivo. Mol Cell. 2003 Aug;12(2):525–532.

[90] Dujardin G, Lafaille C, de la Mata M, Marasco LE, Muñoz MJ, Le Jossic-Corcos C, et al. How slow RNA polymerase II elongation favors alternative exon skipping. Mol Cell. 2014 May;54(4):683–690.

[91] Pinto PA, Henriques T, Freitas MO, Martins T, Domingues RG, Wyrzykowska PS, et al. RNA polymerase II kinetics in polo polyadenylation signal selection. EMBO J. 2011 May;30(12):2431–2444.

[92] Wright S. Regulation of eukaryotic gene expression by transcriptional attenuation. Mol Biol Cell. 1993 Jul;4(7):661–668.

[93] Naville M, Gautheret D. Transcription attenuation in bacteria: theme and variations. Brief Funct Genomic Proteomic. 2009 Nov;8(6):482–492.

[94] Graveley BR, Fleming ES, Gilmartin GM. RNA structure is a critical determinant of poly(A) site recognition by cleavage and polyadenylation specificity factor. Mol Cell Biol. 1996 Sep;16(9):4942–4951.

[95] Bentley DL. Coupling mRNA processing with transcription in time and space. Nat Rev Genet. 2014 03;15(3):163–175.

[96] Movassat M, Crabb TL, Busch A, Yao C, Reynolds DJ, Shi Y, et al. Coupling between alternative polyadenylation and alternative splicing is limited to terminal introns. RNA Biol. 2016 07;13(7):646–655.

[97] Oh JM, Di C, Venters CC, Guo J, Arai C, So BR, et al. U1 snRNP telescripting regulates a size-function-stratified human genome. Nat Struct Mol Biol. 2017 Nov;24(11):993–999.

[98] Yao C, Biesinger J, Wan J, Weng L, Xing Y, Xie X, et al. Transcriptome-wide analyses of CstF64-RNA interactions in global regulation of mRNA alternative polyadenylation. Proc Natl Acad Sci USA. 2012 Nov;109(46):18773–18778.

[99] Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, et al. Ensembl 2019. Nucleic Acids Res. 2019 Jan;47(D1):D745–D751.

[100] Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, et al. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature. 2010 Dec;468(7324):664–668.

[101] Zhu Y, Vaughn JC. Experimental Verification and Evolutionary Origin of 5'-UTR Polyadenylation Sites in Arabidopsis thaliana. Front Plant Sci. 2018;9:969.

[102] de Lorenzo L, Sorenson R, Bailey-Serres J, Hunt AG. Noncanonical Alternative Polyadenylation Contributes to Gene Regulation in Response to Hypoxia. Plant Cell. 2017 Jun;29(6):1262–1277.

[103] Ignatochkina AV, Takagi Y, Liu Y, Nagata K, Ho CK. The messenger RNA decapping and recapping pathway in Trypanosoma. Proc Natl Acad Sci USA. 2015 Jun;112(22):6967–6972.

[104] Del Valle Morales D, Trotman JB, Bundschuh R, Schoenberg DR. Inhibition of cytoplasmic cap methylation identifies 5' TOP mRNAs as recapping targets and reveals recapping sites downstream of native 5' ends. Nucleic Acids Res. 2020 Apr;48(7):3806–3815.

[105] Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. PLoS Biol. 2011 Jul;9(7):e1001091.

[106] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012 Sep;22(9):1760–1774.

[107] Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005 Aug;15(8):1034–1050.

[108] Lorenz R, Bernhart SH, Hoener Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011 Nov;6:26.

[109] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25–29.

[110] authors listed N. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019 Jan;47(D1):D330–D338.

[111] Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012 May;16(5):284–287.

[112] Sun L, Fazal FM, Li P, Broughton JP, Lee B, Tang L, et al. RNA structure maps across mammalian cellular compartments. Nat Struct Mol Biol. 2019 04;26(4):322–330.

[113] Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 2006 Jan;34(Database issue):D590–598.

[114] Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, et al. ENCODE data at the ENCODE portal. Nucleic Acids Res. 2016 Jan;44(D1):D726–732.

[115] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep;489(7414):57–74.

[116] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan;29(1):15–21.

[117] Pervouchine DD, Knowles DG, Guigó R. Intron-centric estimation of alternative splicing from RNA-seq data. Bioinformatics. 2013 Jan;29(2):273–274.

[118] Pervouchine D, Popov Y, Berry A, Borsari B, Frankish A, Guigó R. Integrative transcriptomic analysis suggests new autoregulatory splicing events coupled with nonsense-mediated mRNA decay. Nucleic Acids Res. 2019 06;47(10):5293–5306.

[119] Bretschneider H, Gandhi S, Deshwar AG, Zuberi K, Frey BJ. COSSMO: predicting competitive alternative splice site selection using deep learning. Bioinformatics. 2018 07;34(13):i429–i437.

[120] Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol. 2004;11(2-3):377–394.

[121] Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012 Sep;489(7414):101–108.

[122] Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014 Sep;47:1–34.

[123] Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 2004 Apr;14(4):708–715.