Human LUC7 proteins impact splicing of two major subclasses of 5' splice sites

7 December, 2022

Connor J. Kenny[1], Michael P. McGurk[1] and Christopher B. Burge[1]*

[1]Department of Biology
Massachusetts Institute of Technology
Cambridge, MA 02142

*Address correspondence to: cburge@mit.edu

ABSTRACT

Human LUC7 family proteins associate with the U1 small nuclear ribonucleoprotein (snRNP) complex. Mutation or deletion of *LUC7L2* is associated with myeloid neoplasms, and depletion of *LUC7L2* alters cellular metabolism. Here, we describe distinctive 5' splice site (5'SS) features of exons impacted by each of the three human LUC7s. We find that *LUC7L2* and *LUC7L* enhance splicing of “right-handed” 5'SS with stronger consensus matching on the intron side of the near-invariant /GU, while *LUC7L3* preferentially enhances splicing of “left-handed” 5'SS with stronger consensus matching on the exon side of the splice junction. Specificity for right- or left-handed 5'SS is conferred by the distinct structured N-terminal domains of LUC7L2 and LUC7L3. Evolutionary analysis shows that divergence of *LUC7L3* and *LUC7L2* subfamilies occurred prior to the divergence of plants from animals/fungi, and suggests that loss of the *LUC7L3* subfamily from the fungal lineage contributed to the predominance of right-handed 5'SS in fungi.

INTRODUCTION

In eukaryotes, protein-coding regions of the genome are often interrupted by non-coding introns, which are recognized by ribonucleoprotein complexes and excised by the spliceosome. Conserved motifs in the pre-mRNA transcript mark intron boundaries, of which 5' splice site (5'SS) sequences at the 5' ends of introns are critically important. In the first step of pre-mRNA splicing, U1 snRNP is recruited to the exon-intron boundary through RNA:RNA base pairing between the 5' end of U1 snRNA and the 5'SS (Krämer et al., 1984). U1 snRNP binding is supplemented by recognition of the downstream 3' splice site and branch point sequence by U2AF and U2 snRNP, leading to E complex formation and subsequent recruitment of a pre-assembled tri-snRNP (pre-B complex). Upon binding of the tri-snRNP, the U1 snRNA:5'SS duplex is unwound on the intron side by RNA helicase hDDX23/yPrp28, allowing a portion of the 5'SS to base-pair to U6 snRNA and permitting the release of U1 snRNP from the spliceosome (Staley and Guthrie, 1999). This catalytic spliceosome then undergoes a series of complex rearrangements to catalyze two transesterification reactions, resulting in ligation of the two exons and excision of the intervening intron as an RNA lariat.

The core pre-mRNA splicing process and its individual protein complexes are highly conserved across eukaryotes, from *Saccharomyces cerevisae* to plants to humans. However, the splice site motifs of *S. cerevisae* and mammalian introns are surprisingly different. In mammals, 5'SS at the exon-intron boundary are moderately conserved in the last three bases of the exon and the first six bases of the intron (Figure 1A). In contrast, yeast 5'SS lack a clear consensus in the upstream exon, but the intronic bases of 5'SS are nearly invariant. Although many individual human 5'SS are highly conserved across species, genome-wide analyses in humans revealed that the spectrum of possible sequences that can serve as a 5'SS is quite broad and spans thousands of different 9-mer sequences (Roca et al., 2012). Furthermore, early analyses of human 5'SS composition identified two clusters of U2-type 5'SS, with 5'SS consensus bases clustered either on the exon side or the intron side (Yeo and Burge, 2004; Carmel et al., 2004, Roca et al., 2008, Figure S1A). However, the diversity of human 5'SS complicates untangling the principles of 5'SS selection, and potential functional properties of these two 5'SS clusters have not been explored.

The U1 snRNP complexes of human and *S. cerevisae* have distinct compositions. Yeast U1 snRNP has 17 stably associated polypeptides, while in human U1 snRNP, homologs of these 17 polypeptides divide into two categories: 10 stable "core" proteins and 5 weakly associated "auxiliary" proteins (Li et al., 2017). (Yeast U1 protein Snu56 has no known human homolog and the yPrp39/yPrp42 heterodimer is replaced by a PRP39 homodimer in human.) Of the known human U1 snRNP auxiliary proteins, the LUC7 family of RNA-binding proteins (RBPs) are particularly intriguing. In *S. cerevisae*, biochemical (Fortes et al., 1999; Puig et al., 2007), genetic (Agarwal et al., 2016) and structural data (Bai et al., 2018, Plaschka et al., 2018, Li et al., 2019) indicate that yLUC7p directly stabilizes the U1 snRNA:5'SS duplex via extensive interactions with the U1 snRNA phosphodiester backbone. While yeast possess a single LUC7 protein, yLuc7p, mammals possess three: the closely-related LUC7L and LUC7L2, and the more divergent LUC7L3, all of which possess a highly conserved domain structure. The strong similarity of these proteins suggests that human LUC7 proteins also interact with and stabilize the human U1 snRNA:5'SS duplex, raising the question of whether the three paralogs possess redundant or distinct functions. One possibility is that human U1 auxiliary proteins permit more flexible recognition of diverse 5'SS in humans.

Indeed, recent work characterizing the human LUC7 family found that all three LUC7 proteins crosslink with the first 11 of nucleotides U1 snRNA (Daniels et al., 2021, Jourdain et al., 2021), including the positions that base-pair to 5'SS. In addition, LUC7L2 and LUC7L3 crosslink widely near 5'SS (and also 3'SS) in pre-mRNAs, and depletion of each factor impacts the splicing of several hundred to thousands of exons, LUC7L crosslinks more broadly in introns. Curiously, knockdown (KD) of the different LUC7 paralogs impacted distinct sets of exons (Daniels et al., 2021), suggesting that the three proteins play distinct roles in splicing. Furthermore, depletion of *LUC7L2*, but not *LUC7L* or *LUC7L3*, altered cellular metabolism by promoting oxidative phosphorylation at the expense of glycolysis, supporting functional distinction between the paralogs (Jourdain et al., 2021). These observations suggest that human LUC7 paralogs likely interact with U1 snRNP independently and may contribute to recognition of specific 5'SS.

Here, we investigated the distinct splicing roles of the human LUC7 family. We show that different human LUC7 paralogs impact the splicing of thousands of exons in a manner critically dependent on the relative degree of consensus matching at positions upstream

versus downstream of the nearly-invariant 5'SS /GU (where '/' denotes the splice junction). The ¼ of human 5'SS with consensus matching primarily on the exon side are often dependent on levels of *LUC7L3* for efficient splicing, while ¼ with consensus matching primarily on the intron side are often dependent on *LUC7L2* and *LUC7L*. Together, our analyses identify two distinct subclasses of "left-handed" and "right-handed" 5'SS that appear to differ in their regulation, and suggest that the presence of multiple LUC7s contributes to the maintenance of 5'SS diversity in animals and plants.

Results

LUC7 family members impact distinct subclasses of 5'SS

All human LUC7 family members possess two N-terminal ZNFs, but the LUC7L3-ZNF2 is highly divergent compared to its paralogs (Figure S1B). In yeast pre-B complex, yLuc7p-ZNF2 forms salt bridges with the phosphodiester backbone of U1 snRNA while duplexed to the 5'SS (Bai et al., 2018) (Figure 1B). Since human LUC7 paralogs also interact with human 5'SS (Puig et al., 2007, Jourdain et al., 2021, Daniels et al., 2021), we decided to reanalyze publicly available RNA-seq data from knockout (KO) or knockdown (KD) of these genes in human cells, including *LUC7L2* KO in K562 erythroleukemia and HeLa cells, and *LUC7L*, *LUC7L2*, *LUC7L3* KD in K562 cells. Analysis of 5'SS strength (Yeo and Burge, 2004) indicated that exons sensitive to depletion of LUC7 family members had either no significant difference in mean 5'SS strength relative to unchanged exons or had differences that were significant, but extremely small in magnitude, ranging from 0 to 0.25 bits (Figure S1C). These observations suggest 5'SS strength is not a strong predictor of responsiveness to levels of Luc7 paralogs.

Despite having similar strength, the 5'SS of exons impacted by depletion of *LUC7L2* and *LUC7L3* had strikingly different compositions compared to unchanged exons (Figure 1C). Specifically, KO of *LUC7L2* led to increased skipping of exons with consensus matching at positions +4 and +5 on the intron side of the 5'SS motif, and increased inclusion of exons with consensus matching at positions –2 and –1 in the exon. In contrast, KD of *LUC7L3* led to the opposite pattern, with increased skipping of exons with consensus bases at –2 and –1, and inclusion of exons with consensus matching at positions +4 and +5. No differences were observed in the logos for the upstream or downstream 3'SS, or the upstream 5'SS relative to affected exons (Figure S1D), supporting that the composition of the 5'SS of the impacted exon is critical for LUC7 regulation.

5'SS Balance predicts sensitivity to *LUC7L2* versus *LUC7L3*

The 5'SS logos (Figure 1C) suggested that the extent of consensus matching on each side of the /GU was an important determinant of sensitivity to *LUC7L2* versus *LUC7L3* depletion. To test this hypothesis, we devised a simple measure called the 5'SS Balance score, defined as the number of consensus matches on the intron side at positions +3, +4 and +5 minus the

number of consensus matches on the exon side at positions –3, –2 and –1 (Figure 1D), counting C or A as consensus at –3 and G or A as consensus at +3, and ignoring the weakly conserved +6 position for the moment. The 5'SS Balance therefore has a range of –3 to +3. Negative scores are assigned to “left-handed” (LH) 5'SS which have more consensus matching on the exon side; positive scores are assigned to “right-handed” (RH) 5'SS with more consensus matching on the intron side 3' of the /GU; and a “balanced” score of 0 is assigned to exons with similar consensus matching on both sides of the /GU.

Comparing the mean 5’SS Balance score of perturbed exons showed large differences, supporting the idea that *LUC7L2* promotes the splicing of exons with RH 5’SS and represses exons with LH 5’SS, and that *LUC7L3* promotes splicing of exons with LH 5’SS and represses splicing of exons with RH 5'SS (Figure 1E). Similar effects were observed for *LUC7L2* KD in K562 cells and *LUC7L2* KO in HeLa. Depletion of *LUC7L* gave results similar in direction to depletion of *LUC7L2*, but more modest in magnitude. These observations suggest the existence of subclasses of LH 5’SS promoted by *LUC7L3* and RH 5’SS promoted by *LUC7L2* and *LUC7L*.

To assess the impacts of LUC7 proteins on individual 5’SS sequences, we used a statistic calculating the enrichment of a particular 5'SS 9mer sequence (spanning positions –3 to +6) in exons significantly included or skipped in a particular RNA-seq dataset. This measure, which we call “5'SS Enrichment”, is a Z-score relative to a Dirichlet-multinomial distribution based on the number of occurrences of the 9mer as a 5'SS in exons included, unchanged or skipped in an RNA-seq dataset: positive values indicate enrichment in included exons, negative values enrichment in skipped exons, while values near zero indicate absence of bias for inclusion/skipping (or presence in unchanged exons only). As a test of this statistic, we applied it to a recently published RNA-seq data in which mutant U1 snRNA (g.3a>c) was overexpressed (Shuai et al., 2019; Figure S2A). This analysis recovered very strong and significant enrichment of 5'SS with +6G (complementary to U1 snRNA position 3) among exons included following overexpression of the mutant snRNA (Figure S2B-D), supporting its utility. Application of 5'SS Enrichment to LUC7 family depletions uncovered highly divergent sets of splice sites enriched at exons sensitive to depletion of *LUC7L2* or *LUC7L3*. This analysis showed, in all analyzed datasets, that individual 5'SS with strong Enrichment in the

LUC7L2 or *LUC7L3* depletions were predominantly those with low or high 5'SS Balance scores, respectively (Figure 2A,B, Figure S2E).

In order to provide a more empirical basis for our classifications, we developed a LUC7 score based on LUC7 perturbation data. Hierarchical clustering of 5'SS Enrichment scores across available *LUC7L2* or *LUC7L3* datasets yielded clear groups of LUC7L3-promoted/LUC7L2-repressed with LH motifs, and of LUC7L2-promoted/LUC7L3-repressed 5'SS with RH motifs (Figure 2C). Approximately one quarter of all human protein-coding exons have 5'SS in the LH set, and a similar number in the RH set (Figure 2D); the high frequency of both subclasses among protein-coding exons suggested that both subclasses are generally capable of efficient splicing. Constructing position-specific frequency weight matrices (PWMs) from the 5'SS in each cluster, we define the LUC7 score (or "empirical Balance score") of any 5'SS 9-mer as the $\log_2$ ratio of the *LUC7L2* PWM probability over that of the *LUC7L3* PWM. This score is strongly associated with the direction and magnitude of splicing change in response to perturbation of L2 or L3 (Figure 2E, Figure S3A). The LUC7 score was strongly correlated with the 5'SS Balance score (r = 0.86, $p < 2.2e-16$), as expected, but captures additional sequence features such as the preference for U at position +6 in L2-promoted but not L3-promoted 5'SS, and weights the matches to consensus differently at different positions, unlike the Balance score. Notably, the or LUC7 score is completely uncorrelated with splice site strength (Figure 2F, r = 0.007, p = 0.65), indicating that it captures an orthogonal 5'SS property (Figure S3B,C).

Together, these results indicate LH and RH splice sites are widespread and conserved, suggesting that they can be efficiently recognized, yet distinct in their dependence on different LUC7 family members.

LUC7 family members cross-regulate each other's splicing via RH 5'SS

Many splicing regulatory factors (SRFs) negatively auto-regulate the expression of their own gene at the level of splicing, and often negatively cross-regulate the expression of close paralogs through regulation of the same exons (Spellman et al., 2007; Leclair et al., 2020). This regulatory arrangement, which helps maintain a narrow range of activity for particular SRFs and SRF families, occurs in the LUC7 family as well (Daniels et al., 2021): specifically, *LUC7L2* represses the expression of *LUC7L* by promoting the inclusion of a protein-truncating

cassette (PTC) exon in *LUC7L* (Jourdain et al., 2021). Although PTC exons are not annotated in the other LUC7 paralogs, we noted the strongly conserved alternative second exon (AE2) in *LUC7L2* that replaces the coding sequence of a highly conserved N-terminal alpha helix that interacts with SM-ring component SmE in U1 snRNP. Sequence alignment of the LUC7L2 AE2 and LUC7L PTC exons revealed that they are 82% identical at the nucleotide level, implying a common evolutionary origin; no comparable exon was found in the *LUC7L3* gene.

Consistent with previous studies, we found that overexpression of either *LUC7L* or *LUC7L2* promoted inclusion of the *LUC7L* PTC exon in a minigene assay (Figure 3A), suggesting that both paralogs repress *LUC7L* expression (Figure 3B,C, left). We further found the *LUC7L2* AE2 is regulated similarly to the *LUC7L* PTC (Figure 3B,C, right), supporting a pattern of negative auto- and cross-regulation analogous to that observed for *PTBP1*/*PTBP2* and other closely paralogous SRF genes, consistent with the idea that LUC7L and LUC7L2 proteins are at least partially redundant in function. In contrast, we found that *LUC7L3* overexpression represses the inclusion of the *LUC7L* and *LUC7L2* regulatory exons, promoting the canonical mRNA isoforms of these genes. This positive regulation by *LUC7L3* may serve to maintain a specific ratio of L2 + L to L3 proteins, perhaps to ensure that both RH and LH 5'SS are spliced efficiently (Figure 3D).

Notably, the 5'SS sequence of the homologous LUC7L PE and LUC7L2 AE2 are average strength and strongly right-handed in character (MaxEnt score = 8.7 bits, 5’SS Balance Score = +3 (RH), LUC7 Score = +8.8), and identical in sequence, suggesting evolutionary constraint. To determine whether LUC7 regulatory exons are explicitly dependent on their 5’SS sequence, we swapped the RH 5'SS of LUC7L2 AE2 with a LH 5’SS of comparable strength (MaxEnt score = 8.1 bits, 5'SS Balance Score = -2, LUC7 Score = –9.8). We found that the *LUC7L* and *LUC7L2* exons with this new 5'SS were spliced similarly to the wildtype (WT) under control conditions, but that this change abolished splicing regulation by *LUC7L2* (Figure 3E). Together, these observations indicate that the “righthandedness” of the regulatory exons in the *LUC7L* and *LUC7L2* genes is necessary for their regulation by *LUC7L2*; this feature of the 5'SS may also enable opposing regulation by *LUC7L3*.

Right-Handed 5'SS specificity is mediated by LUC7 structured N-terminal domains

Previous work demonstrated that ZNF2 of yLUC7p is indispensible for cell viability and makes position-specific contacts with U1 snRNA while duplexed to the 5'SS. Comparison of LUC7 paralogs and yLUC7p via multiple sequence alignment suggested that yLuc7p is more similar overall to LUC7L and LUC7L2, particularly in their ZNF2 domains (Figure S1C). In contrast, LUC7L3 is divergent in ZNF2, potentially contributing to functional differences relative to LUC7L/LUC7L2. Therefore, we hypothesized the divergent structured domains may mediate the specificity of LUC7 paralogs for distinct 5'SS subclasses. To test this hypothesis, we generated chimeric proteins by fusing the structured N-terminal region of LUC7L2 or LUC7L3 with the intrinsically disordered C-terminal region (IDR) of the other paralog (Figure 4A) and assessed the impacts on exons with different types of 5'SS. Using a small library of minigene constructs with different RH and LH 5'SS exons of interest, we observed that overexpression of WT *LUC7L* or *LUC7L2* invariably promoted the inclusion of exons with RH 5'SS and repressed inclusion of exons with LH 5'SS (Figure 4B), consistent with the pattern established earlier from depletion/RNA-seq data. On the other hand, overexpression of WT *LUC7L3* repressed inclusion of RH 5'SS and promoted inclusion of exons with LH 5'SS.

Considering the chimeric proteins, we found that LUC7L2 N-terminus fused to LUC7L3 C-terminus (LUC7L2::LUC7L3) matched the pattern of *LUC7L2* (or *LUC7L*) overexpression on both LH and RH 5'SS, though somewhat weaker in magnitude (Figure 4C). On the other hand, the reverse chimera, LUC7L3::LUC7L2, repressed RH 5'SS inclusion, often somewhat more potently than WT LUC7L3, and activated one of the two LH 5'SS studied. These observations indicate that the structured N-terminal domains of LUC7L2 and LUC7L3 are sufficient to confer the general pattern of repression/activation of LH versus RH 5'SS, and suggest that the C-terminal regions of these paralogs are similar but not identical in function.

Loss of *LUC7L3*-related genes is associated with altered 5'SS composition

A major difference between the human and *S. cerevisiae* 5'SS motifs is that the compositionally biased exonic positions at –1, –2, –3 in human have essentially no bias in yeast (Figure 1A). To understand whether different members of the Luc7 family may have shaped the evolution of 5'SS sequences, we assessed 5'SS composition and the distribution of Luc7 genes across 33 diverse eukaryotes, including members of each major eukaryotic

kingdom. We first compiled annotated /GT 5'SS from each genome and summarized the positional information content in each lineage (Figure 5A). This analysis revealed distinct patterns between clades with strong similarity between organisms within each kingdom. All studied fungi have strongly biased composition on the intron side, with little or no bias at any exonic position, implying the predominance of RH 5'SS in this kingdom (Kupfer et al., 2004). All animals considered have moderate biases at both exonic and intronic positions, consistent with the presence of both LH and RH 5'SS (Lim and Burge, 2001). Plants have slightly stronger bias on the exon side than animals, but otherwise resemble animals overall (Szczesniak et al., 2013), suggesting presence of both LH and RH subclasses 5'SS.

Next, we used multiple sequence alignment of Luc7 proteins to understand the evolutionary relationships between Luc7 family members (Figure 5B). This analysis indicated that the Luc7 family is extremely ancient, predating the divergence of plants, animals and fungi, supported by the presence of at least one Luc7 protein in every eukaryote studied. Notably, the N-terminal alpha helix, which interacts with SM-ring component SmE, is highly conserved across eukaryotes, supporting conservation of this interaction across eukaryotes.

The resulting dendrogram of the Luc7 family implies that human *LUC7L2* and *LUC7L* derive from a duplication event occurring in the last common ancestor (LCA) of jawed vertebrates approximately 400 million years ago (Figure 5C). Both the *LUC7L2* and *LUC7L*3 subfamilies are extremely ancient, with this split arising prior to the divergence of plants from animals/fungi. While animals and plants retained both LUC7L2-like and LUC7L3-like proteins, the *LUC7L3* subfamily appears to have been lost from the LCA of modern fungi after the divergence from animals. Strikingly, this distribution of subfamilies matches the distribution of 5'SS motifs, with presence of substantial bias at exonic positions and LH 5'SS observed only in lineages that possess a *LUC7L3* subfamily gene. Together, these results provide evidence that 5'SS sequences in eukaryotic lineages have coevolved with Luc7 family proteins.

Discussion

Across eukaryotes, the spliceosomal snRNPs and the snRNA sequences that base pair with the pre-mRNA are extremely well conserved, underscoring the ancient origins of the spliceosome (Schwartz et al. 2008). Yet despite this deep conservation, the splicing signals between different eukaryotic lineages are comparatively divergent. The contrast between the highly rigid 5'SS of yeast and the much more variable 5'SS in humans has long been noted. Various models have been offered to explain this difference, including the existence of more extensive proofreading and quality control mechanisms in mammalian splicing (Sales-Lee et al, 2021). Our data offer an alternative perspective, indicating that different LUC7 family proteins promote the splicing of distinct subclasses of LH and RH 5'SS in mammals, likely contributing to the maintenance of 5'SS diversity in mammalian genes. Loss of the *LUC7L3* subfamily in the LCA of fungi may have narrowed the range of sequences efficiently recognized as 5'SS in this lineage, simultaneously eliminating splicing-related constraints on the last three bases of exons. In contrast, maintenance of both Luc7 subfamilies in animals and plants likely enables more flexible recognition of 5'SS in these lineages.

In human cells, our findings support that different 5'SS subclasses depend on different LUC7 family members for efficient splicing. The idea that this dependence results from direct interactions is supported by several lines of evidence. First, structural data indicates that yLuc7p contacts U1 snRNA backbone positions that base-pair with the 5'SS (Bai et al., 2018), and the protein residues involved are highly conserved in mammalian LUC7 proteins, particularly in LUC7L2 and LUC7L. Biochemical data indicates that mammalian LUC7 proteins associate with U1 snRNP and cross-linking data shows both interaction with the 5' end of U1 snRNA and strong enrichment near 5'SS for both LUC7L2 and LUC7L3 (Puig et al., 2007; Daniels et al., 2021; Jourdain et al., 2021). To these data, we add the observations that exons sensitive to LUC7 family regulation have distinct composition, that the structured N-terminal region mediates LUC7 family specificity, and that specific 5'SS composition is required for *LUC7L2* regulation of splicing. Together, these observations support the model that LUC7L2 and LUC7L3 tune the splicing of distinct subclasses of 5'SS via interactions with U1 snRNP that promote recognition of RH and LH 5'SS, respectively.

Several factors add complexity to this picture. First, in both our RNA-seq analysis and minigene experiments we observed not only positive effects of *LUC7L2* on RH 5'SS, but also widespread negative effects on LH 5'SS, and vice versa for *LUC7L3*. For example, *LUC7L2* KO resulted in inclusion of exons with LH 5'SS, while overexpression repressed LH exon inclusion. The mechanism underlying such negative regulation is less clear. While it is possible that interactions with the U1 snRNA backbone that stabilize the base-paired conformation at certain positions also destabilize non-base paired conformations, evidence for such a mechanism is lacking. A simpler alternative is that human LUC7 family members simply compete with one another for access to the same site on U1 snRNP, so that over-expression of LUC7L3, for example, inhibits RH 5'SS by reducing the pool of available U1 snRNP for interaction with RH-5'SS-promoting LUC7L2 and LUC7L. LUC7 proteins may interact directly with each other (Daniels et al. 2021) (Jourdain et al., 2021), opening up additional potential modes of regulation. It is also possible that the impact on 5'SS selection occurs after initial recognition of the 5'SS by U1 snRNP. After U1 snRNP is released, U6 snRNA interacts with sequences on the intron side of the /GU, and U5 snRNA interacts with sequences on the exon side to position the catalytic center of the spliceosome at the exon-intron boundary (Lesser and Guthrie, 1993; Newman and Norman, 1992, Sontheimer and Stetiz, 1992). Therefore, it's possible to imagine a model in which LUC7L2 promotes a conformation of the pre-catalytic spliceosome that favors subsequent base-pairing interactions between the 5'SS and U6 snRNA, and LUC7L3 promotes an alternate confirmation in which U5 snRNA interactions are more important, and it is at this later step that the preference for RH or LH 5'SS occurs. Further mechanistic and structural studies will be needed to understand precisely how human LUC7 family proteins influence pre-mRNA splicing.

Analyses of the RNA-seq data also showed that not every exon with LH or RH 5'SS is sensitive to perturbations of LUC7s. Some exons may have sufficiently strong 3'SS, or redundant regulation by exonic splicing enhancers or other splicing regulatory elements, such that enhancement by the appropriate LUC7 family protein is not required for efficient splicing. It is also possible that different LUC7 paralogs may differentially interact with other splicing machinery or SRFs, or even require a specific exon context to mediate splicing control. For example, a massively parallel reporter assay examining 5'SS sequence requirements in three exon contexts revealed that *IKBKAP* (*ELP1*) exon 20 could only splice with a small cohort of

5'SS that resemble the RH subclass (Wong et al., 2018). Interestingly, this particularly restrictive context could be relaxed by increasing the strength of the upstream 3'SS. Our experiments with chimeric LUC7 proteins indicate that the C-terminal mixed-charge regions of LUC7L2 and LUC7L3 have similar, but not identical activities, which might contribute to any contextual requirements. The C-terminal region of LUC7L3 appears necessary to promote the inclusion of LH 5'SS; this could occur through interaction with distinct SRFs (Daniels et al. 2021) or by other means such as effects on subnuclear localization.

The strong similarity between the N-termini LUC7L and LUC7L2 and our observations that LUC7 structured N-terminal regions determine 5'SS subclass specificity suggest that LUC7L and LUC7L2 proteins may be partially redundant in function. This idea is supported by similar effects of overexpressing *LUC7L* and *LUC7L2* in our minigene assays, as well as similarities in 5'SS motifs impacted by depletion of these factors. Our observations of positive regulation of the *LUC7L2* canonical isoform by *LUC7L3* and negative regulation by *LUC7L* and *LUC7L2* extend previously observed regulatory relationships between LUC7 proteins (Daniels et al. 2021) (Jourdain et al., 2021). The frequent mutation or loss of *LUC7L2* in myeloid neoplasms (Singh et al., 2013) might therefore present therapeutic opportunities. For example, the similarity in activity between LUC7L and LUC7L2, the negative cross-regulation of these factors and synthetic lethality between *LUC7L* and *LUC7L2* (Jourdain et al., 2021) support that *LUC7L2*-deficient tumors might exhibit preferential sensitivity to inhibition of *LUC7L*.

In plants, the Luc7 family has been implicated in development and abiotic stress responses (de Francisco Amorim et al., 2018) and in yeast, yLuc7p localization changes in response to hypoxia and reoxygenation (Dastidar et al., 2012). Previous analyses of the metabolic effects of *LUC7L2* depletion in human cells raise the possibility that variation in LUC7 family expression across cell types or development (Jourdain et al., 2021) might contribute to metabolic changes in mammals.

Acknowledgements

We would like to thank all members of the Burge lab, as well as Eliezer Calo, Nima Jaberi-Lashkari, Alexis Jourdain, Vamsi Mootha, Byron Lee, Phillip A. Sharp, Seychelle Vos and Rui Zhao for their helpful discussions and feedback on the manuscript.

Author Contributions

C.J.K worked under the supervision of C.B.B to design, perform and analyze all experiments and computational analyses. M.P.M. assisted with the development of statistical approaches. C.J.K. and C.B.B. wrote the manuscript with input from all authors. All authors contributed to edits/revisions of the manuscript.

Methods Details

Cloning

Human *LUC7L*, *LUC7L2* and *LUC7L3* ORFs were amplified from human cDNA and cloned into pcDNA3.1(+)IRES GFP (Addgene #: 51406). Domain swap constructs were synthesized as gblocks from IDT and cloned into pcDNA3.1(+)IRES GFP. Exons and flanking intronic regions used for pSpliceExpress minigenes were PCR amplified from human male genomic DNA using primers with attB overhangs and subsequently recombined into pSpliceExpress using BP Clonase II (Thermo Fisher, cat no. 11789020)

Cell Culture & Chemical Transfection

HEK293T RMCE cell lines were cultured in Advanced DMEM supplemented with 5% FBS, 25 mM HEPES and Glutamax. For each experiment, cells were plated 24 hours in advance in 24-well plates. The following day, cells were transfected with 500 ng of 95:5 w/v of a cDNA overexpression vector and minigene reporter respectively using Lipofectamine LTX (Thermo Fisher, cat. no. 15338100)

Minigene Reporter Assay

RNA was extracted 24 hours after transfection using Qiagen RNeasy Mini kit (cat. no. 74104) according to manufacturer's instructions with the optional on-column DNAse digestion (cat. No 79254). RNA was eluted in nuclease free water and quantified using Nanodrop. For each RNA sample, we used 125 ng of RNA input into a 12.5 uL LunaScript Multiplex One Step Master Mix for RT-PCR (cat. no. E1555S). After PCR cycling, we mixed PCR samples with NEB 6X loading dye and loaded 5 uL of PCR products on a 3% agarose gel infused with ethidium bromide, which was run for 40 min at 150V. All presented minigene data was performed at least in triplicate. Images were acquired using Azure Biosystems c600 with UV imager.

Percent Spliced In Calculation

Agarose gel images were manually quantified using chromatograms in ImageJ. Percent Spliced In values were calculated by taking the signal intensity of the larger band and dividing it by the sum of the signal intensity of the included product and the skipped product.

Analysis of LUC7 paralog RNAseq data

We downloaded RNAseq data for *LUC7L2* KO in HeLa and K562 erythroleukemia cells (GSE157917) and *LUC7L*, *LUC7L2*, *LUC7L3* KD in K562 cells (E-MTAB-9709). We used rMATS 4.1.1 (Shen et al., 2014) with default settings, setting the minimum delta PSI ("Δψ") for significance to 0.0001.

In any given KD or KO experiment, only a subset of exons likely change in splicing, with most other exons truly having Δψ of zero. The Δψ estimates reported by rMATS, however, are direct summaries of the read counts and are affected by sampling variation in the read counts which may obscure biological signals in data, especially for exons with low read counts. Commonly, shrinkage estimates are used in differential expression analyses to account for this issue. Shrinkage considers the set of all effect sizes to constrain the noise in estimates from low read count events. Doing so, however, requires parameter estimates and associated uncertainty, which are not directly reported by rMATS.

For each exon, we reconstruct the effect size in log-odds scale δ from the rMATS read counts and approximate a standard deviation σ describing the uncertainty in δ using the rMATS p-value. We pass these effect sizes and standard deviations to ashr using the 'halfnormal' option to assume that the proportions of up- and down-regulated exons may be unequal (Stephens, 2017). Using these shrinkage estimates we reconstruct an estimate of $\Delta\psi^*$ (see Supplemental Methods).

5'SS Enrichment Calculation

To calculate a single statistic describing the likelihood of a given 5'SS to be significantly included vs significantly skipped we used a Dirichlet-multinomial model to calculate the log-odds of whether a given 5'SS was more likely to be involved in a significantly included event vs significantly skipped event. First we combined the counts of 5'SS from significant included and skipped events (FDR < 0.2) and their respective background sets, which were matched for PSI and expression, into a count matrix. Then we converted this count matrix into a frequency matrix, which we then used to simulate 2500 draws from the posterior distribution. For each

draw we calculated the log-odds of a given splice donor site being enriched in included vs skipped set. The distribution of log-odds generated from the Dirichlet-multinomial model was used to calculate the mean/maximum likelihood estimate (MLE) and the standard error, which were used to calculate a Z-statistic for every 5'SS 9-mer. Within this framework, positive Z-scores were 5'SS more likely to be included than skipped and negative Z-scores are 5'SS more likely to be skipped than included. For hierarchical clustering, we used 5'SS enrichment Z-scores to cluster 5'SS with ward.D2 linkage by their activity across each LUC7 paralog RNA-seq dataset.

Free energy predictions of U1, U5, U6 and 5'SS sequences

We used ViennaRNA 2.5 (Lorenz et al., 2011) to model free energy predictions between all 5'SS and the 5'end of U1 snRNA (ATACTTACCUG), U6 snRNA (ATACAGAGA) and U5 snRNA loop 1 (GCCUUUUAC) using RNAcofold with default parameters.

Phylogenetic and cross-kingdom splice site analysis

For our phylogenetic analysis, we manually chose 13 animals, 9 fungi and 11 plants from Esembl that represented diverse species in the each of the eukaryotic kingdoms. For each of these species, we used BLAST from UniProt to identify putative Luc7 paralogs and performed multiple sequence alignment using *MAFFT* (Katoh & Standley, 2013). We then created a maximum likelihood tree with RAxML (Stamakis, 2006) and annotated the final protein tree using ggtree (Yu et al., 2016). For each of the species used in our phylogenetic analysis, we downloaded their genomes and associated gene annotation files from Ensembl and extracted their splice site sequences using bedtools 2.29.2. For each species, we calculated the information content per position of the 5'SS.

Figure Legends

Figure 1. Human LUC7 proteins impact 5' splice sites with distinct composition.

A) Sequence logos of annotated 5'SS in yeast (top) and humans (bottom). **B)** Cartoon representation of yLuc7p contacting U1 snRNA:5'SS duplex. Red asterisks indicate reported locations on duplex where yLuc7p interacts directly with U1 snRNA (Bai et al., 2018). **C)** Representative sequence logos of 5'SS for all cassette exons (left) and exons with significantly increased or decreased inclusion upon depletion of *LUC7L2* or *LUC7L3*. **D)** Schematic of 5'SS Balance score, which is defined as the number of consensus bases at positions +3/+4/+5 minus the number of consensus bases at –3/–2/–1 of a 5'SS. **E)** Mean Balance score for included and skipped exons and their respective background sets. Error bars represent standard deviation. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$, Wilcoxon rank sum test relative to respective background set of unchanged exons. (See also Figure S1.)

Figure 2. Enrichment of individual 5' splice site motifs across LUC7 depletions defines two distinct clusters.

A) Volcano plots representing 5'SS enrichment Z scores for individual 5'SS 9-mer and the number of exons with significantly altered PSI upon LUC7 paralog depletion (above). Swarm and box plots showing distribution of 5'SS enrichment Z scores for each 5'SS Balance score (below). **B)** Sequence logos derived from individual 5'SS with significant Enrichment ($Z \geq 1.96$ for included 5'SS, $Z \leq -1.96$ for skipped 5'SS). **C)** Heat map of 5'SS enrichment Z-scores, clustered using Euclidean distance and ward.D2 linkage. **D)** Fraction of CCDS protein coding exons that belong to either the LH (red) or RH (blue) 5'SS clusters shown in (C). **E)** Correlation plot of LUC7 score versus mean Delta PSI for statistically significant exons for different 5'SS sequences (each dot is a different 5'SS). **F)** Heat map showing Pearson correlation between new measures – 5'SS Balance and LUC7 Score – versus standard 5'SS measures. (See also Figure S2.)

Figure 3. Regulation of *LUC7L2* AE2 by LUC7 proteins depends on a conserved right-handed 5'SS motif.

A) Schematic of pSpliceExpress minigene construct used, in which an internal exon of interest with flanking splice sites is inserted into a minigene expressing rat insulin exons 2 and 3 and

intervening intron. **B)** Bar plots of cassette exon's percent spliced in (by RT-PCR with primers in flanking exons) across all experiments; bars are color-coded to match LUC7 paralog colors in (D). Error bars represent standard error. Brackets indicate significant comparisons. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.00001$, Bonferroni-corrected *t*-test. **C)** Representative gel images from LUC7-related minigenes. Percent spliced in values shown at bottom of each lane. **D)** Proposed regulatory relationships between human LUC7 family members. **E)** Bar plot of PSI for *LUC7L2* AE2 minigene with mutagenized 5'SS (left). Representative RT-PCR gel images shown at right.

Figure 4. LUC7 specificity for subclasses of 5'SS depends on the structured N-terminal region.

A) Protein domain structure of human LUC7 proteins (canonical isoform) and experimentally investigated chimeric proteins. **B)** Bar plot summarizing qRT-PCR analysis of change ("delta") in percent spliced in from transfected minigenes containing different internal exons of introns following overexpression of LUC7 WT or chimeric cDNA (shown at right). Mean is plotted with error bars representing standard error, and bars are color-coded by the Balance score of the internal exon's 5'SS. **C)** Pearson correlation heat map of delta percent spliced values for each LUC7 WT or chimeric cDNA overexpressed in B.

Figure 5. Phylogenetic analysis suggests coevolution of Luc7 proteins and 5'SS motifs.

A) Position-specific information content (calculated as in Irimia et al., 2019) of 5'SS motifs across 33 eukaryotes, color-coded by clade, with clade average value shown with thick line. **B)** Maximum likelihood phylogenetic tree built from multiple sequence alignment of Luc7 proteins from 33 animal, fungal and plant species, adding *Trichomonas vaginalis* Luc7 protein as an outgroup. Two main clusters are shaded to indicate Luc7 subfamilies; individual proteins represented by symbols indicating clade of origin. **C)** Representation of presence/absence and likely duplication/loss events for Luc7 subfamilies, overlaid on an organismal phylogenetic tree.

Supplementary Figure Legends

Figure S1. 5' Splice site motifs, strength and LUC7 paralogs. (Related to Figure 1.)

A) Hierarchical clustering of Hamming distance between 5,000 randomly sampled /GT 5'SS sequence (top) using ward.D2 linkage (as done in Yeo & Burge, 2004). **B)** General domain structure of human LUC7 proteins (top) with multiple sequence alignment of N-terminal alpha helix that interacts with SM ring subunit SmE and ZNF2 (bottom). **C)** Mean 5'SS strength and standard deviation of LUC7-regulated exons and their respective background sets of unchanged exons matched for starting expression and PSI value. *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$, Wilcoxon rank sum test relative to respective background set. **D)** Summary of splice site sequence logos at upstream and downstream exons of LUC7-regulated cassette exons.

Figure S2. 5'SS Enrichment application example and LUC7 depletion analysis. (Related to Figure 2.)

A) Simplified schematic of 5' end of WT (top) and mutant (bottom) U1 snRNA 5' contacting the 5' splice site. **B)** Raw 5'SS enrichment (probability of a 5'SS enriched in significantly included exons relative to significantly skipped exons), grouped by +6 position. **C)** 5'SS enrichment Z-scores of significantly deregulated cassette exons from overexpression of mutant U1 snRNA, colored by base at position +6. **D)** Sequence logos of 5'SS from all significant cassette exons (left) and from individual 5'SS with significant Enrichment values (right, $Z \geq 1.96$ for included 5'SS, $Z \leq -1.96$ for skipped 5'SS). **E)** Volcano plots similar to Figure 2A for other LUC7 paralog KDs.

Figure S3. Relationship between LUC7 score and 5'SS features. (Related to Figure 2.)

A) Correlation plots similar to Figure 2E for LUC7 PWM score and the mean change in percent spliced in for statistically significant exons. **B)** Frequency of 5'SS related to LUC7 PWM score (top), eCDF representing relationship of LUC7 PWM Score and 5'SS Balance score (bottom). **C)** Scatter plot representing the relationship of LUC7 PWM score and 5'SS strength.

References

Agarwal, R., Schwer, B., Shuman, S., 2016. Structure–function analysis and genetic interactions of the Luc7 subunit of the *Saccharomyces cerevisiae* U1 snRNP. RNA 22, 1302–1310. https://doi.org/10.1261/rna.056911.116

Bai, R., Wan, R., Yan, C., Lei, J., Shi, Y., 2018. Structures of the fully assembled *Saccharomyces cerevisiae* spliceosome before activation. Science 360, 1423–1429. https://doi.org/10.1126/science.aau0325

Carmel, I., 2004. Comparative analysis detects dependencies among the 5' splice-site positions. RNA 10, 828–840. https://doi.org/10.1261/rna.5196404

Charenton, C., Wilkinson, M.E., Nagai, K., 2019. Mechanism of 5' splice site transfer for human spliceosome activation. Science 364, 362–367. https://doi.org/10.1126/science.aax3289

Crispino, J.D., Sharp, P.A., n.d. A U6 snRNA:pre-mRNA interaction can be rate-limiting for Ul-independent sphclng 11.

Daniels, N.J., Hershberger, C.E., Gu, X., Schueger, C., DiPasquale, W.M., Brick, J., Saunthararajah, Y., Maciejewski, J.P., Padgett, R.A., 2021. Functional analyses of human LUC7-like proteins involved in splicing regulation and myeloid neoplasms. Cell Reports 35, 108989. https://doi.org/10.1016/j.celrep.2021.108989

Dastidar, R.G., Hooda, J., Shah, A., Cao, T.M., Henke, R.M., Zhang, L., 2012. The nuclear localization of SWI/SNF proteins is subjected to oxygen regulation. Cell Biosci 2, 30. https://doi.org/10.1186/2045-3701-2-30

de Francisco Amorim, M., Willing, E.-M., Szabo, E.X., Francisco-Mangilet, A.G., Droste-Borel, I., Maček, B., Schneeberger, K., Laubinger, S., 2018. The U1 snRNP Subunit LUC7 Modulates Plant Development and Stress Responses via Regulation of Alternative Splicing. Plant Cell 30, 2838–2854. https://doi.org/10.1105/tpc.18.00244

Fortes, P., Bilbao-Cortes, D., Fornerod, M., Rigaut, G., Raymond, W., Seraphin, B., Mattaj, I.W., 1999. Luc7p, a novel yeast U1 snRNP protein with a role in 5' splice site recognition. Genes & Development 13, 2425–2438. https://doi.org/10.1101/gad.13.18.2425

Jourdain, A.A., Begg, B.E., Mick, E., Shah, H., Calvo, S.E., Skinner, O.S., Sharma, R., Blue, S.M., Yeo, G.W., Burge, C.B., Mootha, V.K., 2021. Loss of LUC7L2 and U1 snRNP

subunits shifts energy metabolism from glycolysis to OXPHOS. Molecular Cell 81, 1905-1919.e12. https://doi.org/10.1016/j.molcel.2021.02.033

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30, 772–780. https://doi.org/10.1093/molbev/mst010

Krämer, A., Keller, W., Appel, B., Lührmann, R., 1984. The 5′ terminus of the RNA moiety of U1 small nuclear ribonucleoprotein particles is required for the splicing of messenger RNA precursors. Cell 38, 299–307. https://doi.org/10.1016/0092-8674(84)90551-8

Kupfer, D.M., Drabenstot, S.D., Buchanan, K.L., Lai, H., Zhu, H., Dyer, D.W., Roe, B.A., Murphy, J.W., 2004. Introns and Splicing Elements of Five Diverse Fungi. Eukaryot Cell 3, 1088–1100. https://doi.org/10.1128/EC.3.5.1088-1100.2004

Leclair, N.K., Brugiolo, M., Urbanski, L., Lawson, S.C., Thakar, K., Yurieva, M., George, J., Hinson, J.T., Cheng, A., Graveley, B.R., Anczuków, O., 2020. Poison Exon Splicing Regulates a Coordinated Network of SR Protein Expression during Differentiation and Tumorigenesis. Molecular Cell 80, 648-665.e9. https://doi.org/10.1016/j.molcel.2020.10.019

Lesser, C.F., Guthrie, C., 1993. Mutations in U6 snRNA that Alter Splice Site Specificity: Implications for the Active Site. Science 262, 1982–1988. https://doi.org/10.1126/science.8266093

Li, X., Liu, S., Zhang, L., Issaian, A., Hill, R.C., Espinosa, S., Shi, S., Cui, Y., Kappel, K., Das, R., Hansen, K.C., Zhou, Z.H., Zhao, R., 2019. A unified mechanism for intron and exon definition and back-splicing. Nature 573, 375–380. https://doi.org/10.1038/s41586-019-1523-6

Lim, L.P., Burge, C.B., 2001. A computational analysis of sequence features involved in recognition of short introns. Proc. Natl. Acad. Sci. U.S.A. 98, 11193–11198. https://doi.org/10.1073/pnas.201407298

Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., Hofacker, I.L., 2011. ViennaRNA Package 2.0. Algorithms Mol Biol 6, 26. https://doi.org/10.1186/1748-7188-6-26

Newman, A.J., Norman, C., 1992. U5 snRNA interacts with exon sequences at 5′ and 3′ splice sites. Cell 68, 743–754. https://doi.org/10.1016/0092-8674(92)90149-7

Plaschka, C., Lin, P.-C., Charenton, C., Nagai, K., 2018. Prespliceosome structure provides insights into spliceosome assembly and regulation. Nature 559, 419–422. https://doi.org/10.1038/s41586-018-0323-8

Puig, O., Bragado-Nilsson, E., Koski, T., Seraphin, B., 2007. The U1 snRNP-associated factor Luc7p affects 5' splice site selection in yeast and human. Nucleic Acids Research 35, 5874–5885. https://doi.org/10.1093/nar/gkm505

Roca, X., Akerman, M., Gaus, H., Berdeja, A., Bennett, C.F., Krainer, A.R., 2012. Widespread recognition of 5' splice sites by noncanonical base-pairing to U1 snRNA involving bulged nucleotides. Genes & Development 26, 1098–1109. https://doi.org/10.1101/gad.190173.112

Roca, X., Krainer, A.R., Eperon, I.C., 2013. Pick one, but be quick: 5′ splice sites and the problems of too many choices. Genes Dev. 27, 129–144. https://doi.org/10.1101/gad.209759.112

Roca, X., Olson, A.J., Rao, A.R., Enerly, E., Kristensen, V.N., Borresen-Dale, A.-L., Andresen, B.S., Krainer, A.R., Sachidanandam, R., 2007. Features of 5'-splice-site efficiency derived from disease-causing mutations and comparative genomics. Genome Research 18, 77–87. https://doi.org/10.1101/gr.6859308

Sales-Lee, J., Perry, D.S., Bowser, B.A., Diedrich, J.K., Rao, B., Beusch, I., Yates, J.R., Roy, S.W., Madhani, H.D., 2021. Coupling of spliceosome complexity to intron diversity. Current Biology 31, 4898-4910.e4. https://doi.org/10.1016/j.cub.2021.09.004

Schwartz, S., Silva, J., Burstein, D., Pupko, T., Eyras, E., Ast, G., 2008. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. 18, 88–103. https://doi.org/10.1101/gr.6818908

Shen, S., Park, J.W., Lu, Z., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., Xing, Y., 2014. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. U.S.A. 111. https://doi.org/10.1073/pnas.1419161111

Shuai, S., Suzuki, H., Diaz-Navarro, A., Nadeu, F., Kumar, S.A., Gutierrez-Fernandez, A., Delgado, J., Pinyol, M., López-Otín, C., Puente, X.S., Taylor, M.D., Campo, E., Stein, L.D., 2019. The U1 spliceosomal RNA is recurrently mutated in multiple cancers. Nature 574, 712–716. https://doi.org/10.1038/s41586-019-1651-z

Singh, H., Lane, A.A., Correll, M., Przychodzen, B., Sykes, D.B., Stone, R.M., Ballen, K.K., Amrein, P.C., Maciejewski, J., Attar, E.C., 2013. Putative RNA-splicing gene LUC7L2 on 7q34 represents a candidate gene in pathogenesis of myeloid malignancies. Blood Cancer Journal 3, e117–e117. https://doi.org/10.1038/bcj.2013.16

Sontheimer, E.J., Steitz, J.A., 1993. The U5 and U6 Small Nuclear RNAs as Active Site Components of the Spliceosome. Science 262, 1989–1996. https://doi.org/10.1126/science.8266094

Spellman, R., Llorian, M., Smith, C.W.J., 2007. Crossregulation and Functional Redundancy between the Splicing Regulator PTB and Its Paralogs nPTB and ROD1. Molecular Cell 27, 420–434. https://doi.org/10.1016/j.molcel.2007.06.016

Staley, J.P., Guthrie, C., 1999. An RNA Switch at the 5′ Splice Site Requires ATP and the DEAD Box Protein Prp28p. Molecular Cell 3, 55–64. https://doi.org/10.1016/S1097-2765(00)80174-4

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690. https://doi.org/10.1093/bioinformatics/btl446

Stephens, M., 2016. False discovery rates: a new deal. Biostat kxw041. https://doi.org/10.1093/biostatistics/kxw041

Szcześniak, M.W., Kabza, M., Pokrzywa, R., Gudyś, A., Makałowska, I., 2013. ERISdb: A Database of Plant Splice Sites and Splicing Signals. Plant and Cell Physiology 54, e10–e10. https://doi.org/10.1093/pcp/pct001

Wong, M.S., Kinney, J.B., Krainer, A.R., 2018. Quantitative Activity Profile and Context Dependence of All Human 5′ Splice Sites. Molecular Cell 71, 1012-1026.e3. https://doi.org/10.1016/j.molcel.2018.07.033

Yeo, G., Burge, C.B., n.d. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals 18.

Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T., 2017. GGTREE : an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8, 28–36. https://doi.org/10.1111/2041-210X.12628

Figure 1

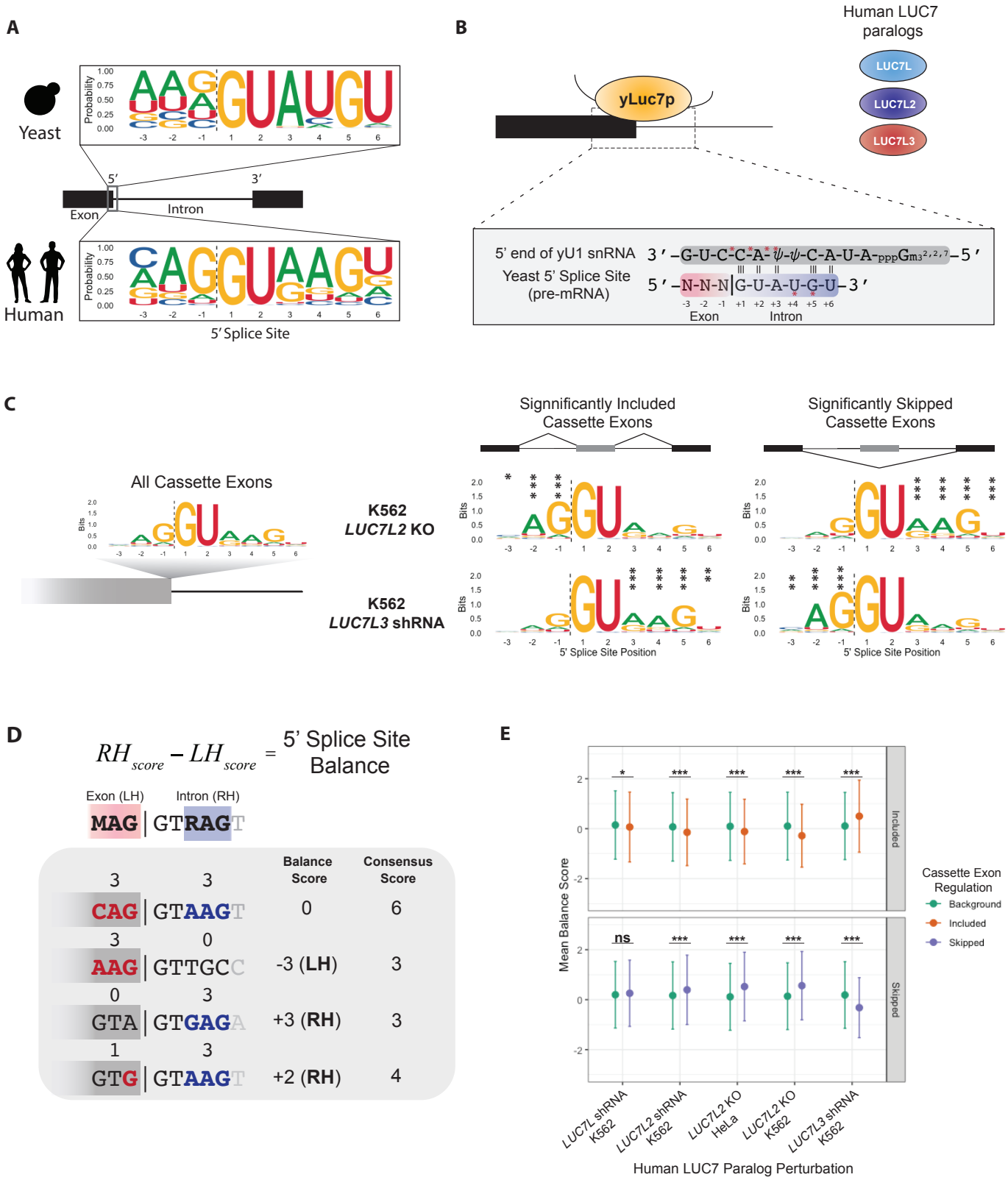

A
Yeast
Probability
5′
3′
Exon
Intron
Human
5′ Splice Site
B
Human LUC7 paralogs
LUC7L
LUC7L2
LUC7L3
yLuc7p
5′ end of yU1 snRNA
Yeast 5′ Splice Site (pre-mRNA)
Exon
Intron
C
All Cassette Exons
Signnificantly Included Cassette Exons
Significantly Skipped Cassette Exons
K562 LUC7L2 KO
K562 LUC7L3 shRNA
Bits
5′ Splice Site Position
D
RH score – LH score = 5′ Splice Site Balance
Exon (LH)
Intron (RH)
MAG | GTRAGT
Balance Score
Consensus Score
CAG | GTAAGT 0 6
AAG | GTTGCC -3 (LH) 3
GTA | GTGAGA +3 (RH) 3
GTG | GTAAGT +2 (RH) 4
E
Mean Balance Score
Included
Skipped
ns
Cassette Exon Regulation
Background
Included
Skipped
LUC7L shRNA K562
LUC7L2 shRNA K562
LUC7L2 KO HeLa
LUC7L2 KO K562
LUC7L3 shRNA K562
Human LUC7 Paralog Perturbation

Figure 2

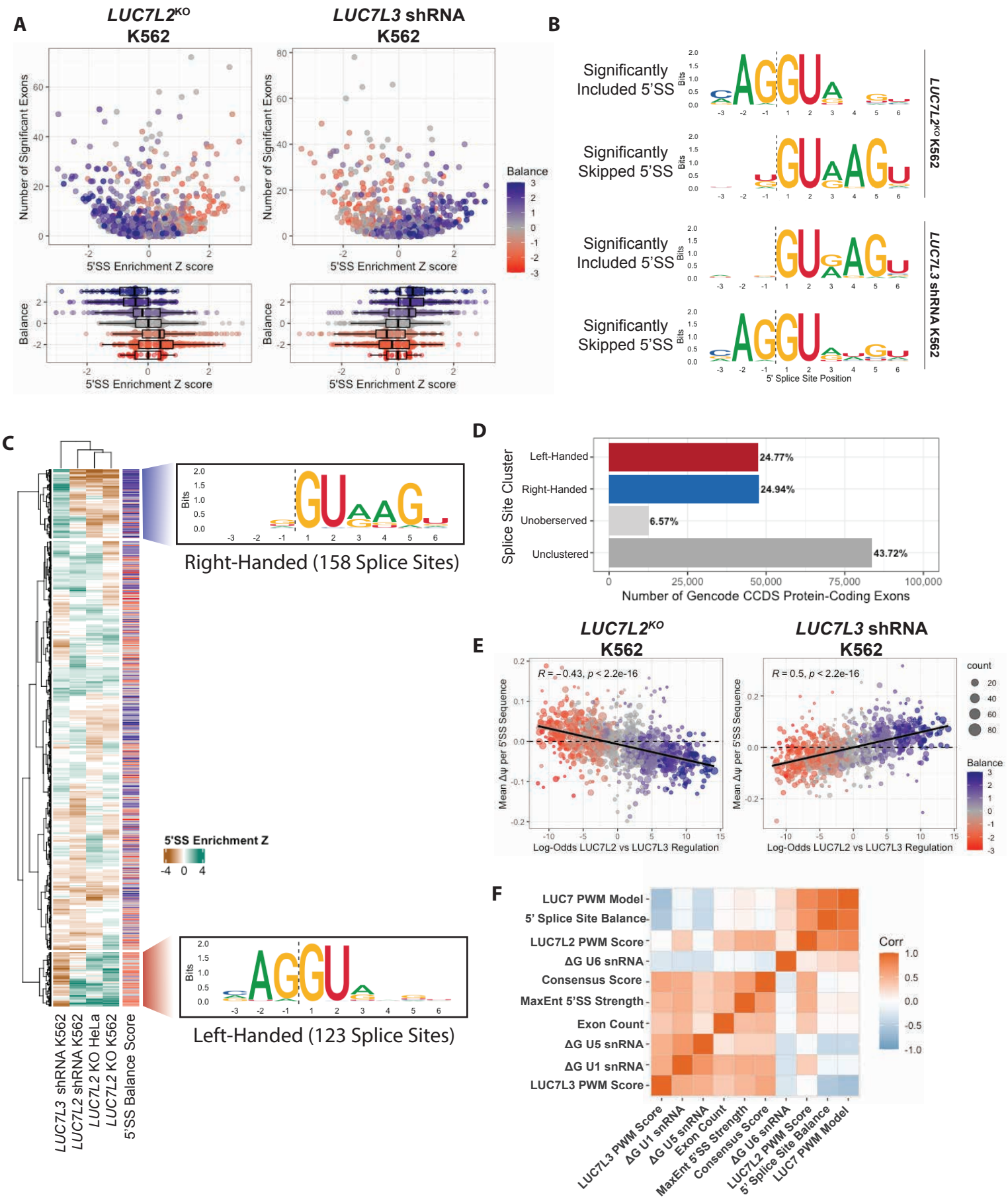

A
LUC7L2KO K562
LUC7L3 shRNA K562
Number of Significant Exons
5'SS Enrichment Z score
Balance
B
Significantly Included 5'SS
Significantly Skipped 5'SS
Significantly Included 5'SS
Significantly Skipped 5'SS
LUC7L2KO K562
LUC7L3 shRNA K562
Bits
5' Splice Site Position
C
Right-Handed (158 Splice Sites)
Left-Handed (123 Splice Sites)
5'SS Enrichment Z
LUC7L3 shRNA K562
LUC7L2 shRNA K562
LUC7L2 KO HeLa
LUC7L2 KO K562
5'SS Balance Score
D
Splice Site Cluster
Left-Handed
Right-Handed
Unoberserved
Unclustered
24.77%
24.94%
6.57%
43.72%
Number of Gencode CCDS Protein-Coding Exons
E
LUC7L2KO K562
LUC7L3 shRNA K562
R = -0.43, p < 2.2e-16
R = 0.5, p < 2.2e-16
Mean Δψ per 5'SS Sequence
Log-Odds LUC7L2 vs LUC7L3 Regulation
count
Balance
F
LUC7 PWM Model
5' Splice Site Balance
LUC7L2 PWM Score
ΔG U6 snRNA
Consensus Score
MaxEnt 5'SS Strength
Exon Count
ΔG U5 snRNA
ΔG U1 snRNA
LUC7L3 PWM Score
Corr

Figure 3

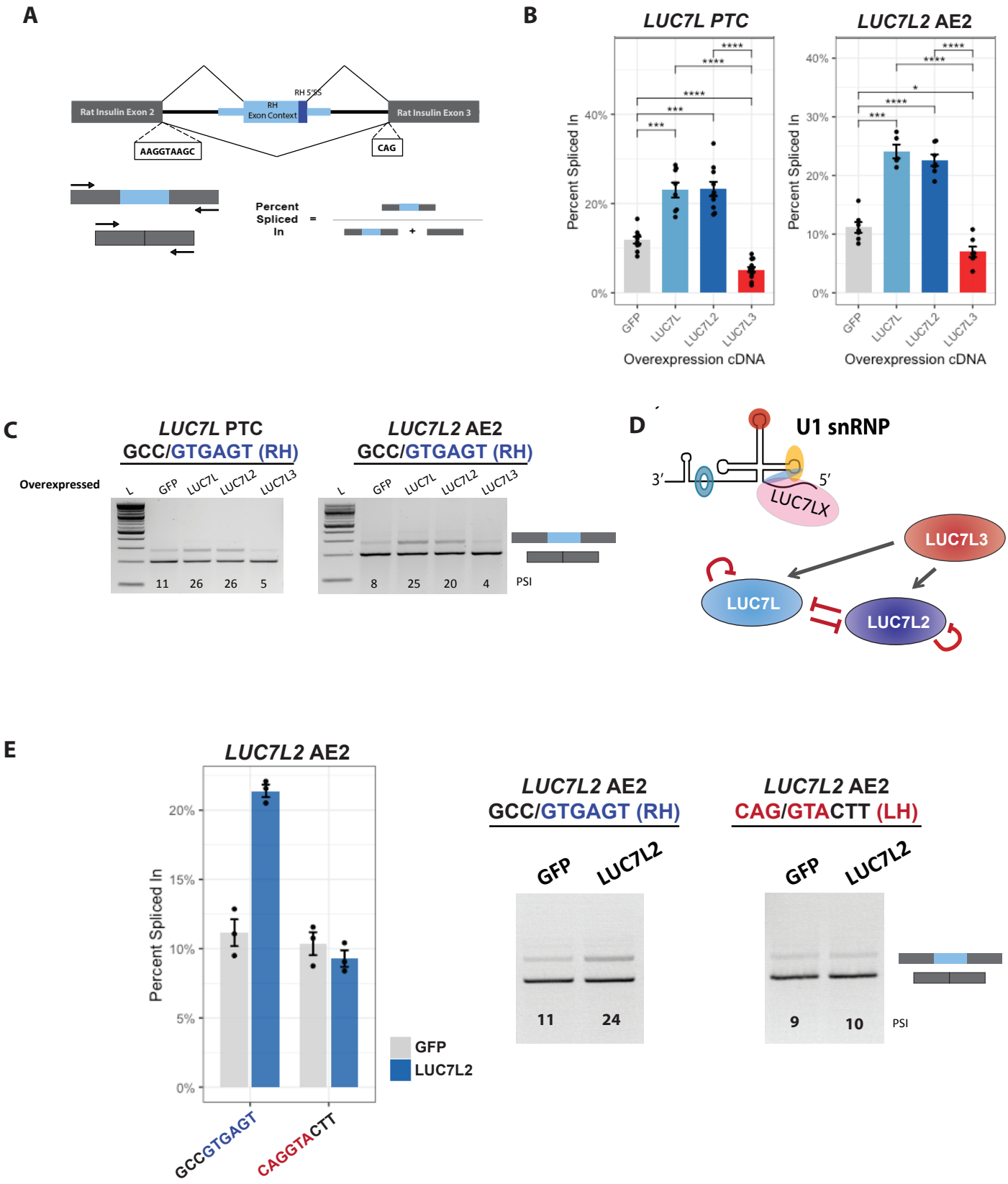

A
Rat Insulin Exon 2
RH 5'SS
RH Exon Context
Rat Insulin Exon 3
AAGGTAAGC
CAG
Percent Spliced In
B
LUC7L PTC
LUC7L2 AE2
Percent Spliced In
Overexpression cDNA
GFP
LUC7L
LUC7L2
LUC7L3
C
LUC7L PTC
GCC/GTGAGT (RH)
LUC7L2 AE2
GCC/GTGAGT (RH)
Overexpressed
L
11 26 26 5
8 25 20 4
PSI
D
U1 snRNP
3'
5'
LUC7LX
LUC7L3
LUC7L
LUC7L2
E
LUC7L2 AE2
Percent Spliced In
GFP
LUC7L2
GCCGTGAGT
CAGGTACTT
LUC7L2 AE2
GCC/GTGAGT (RH)
LUC7L2 AE2
CAG/GTACTT (LH)
11 24
9 10
PSI

Figure 4

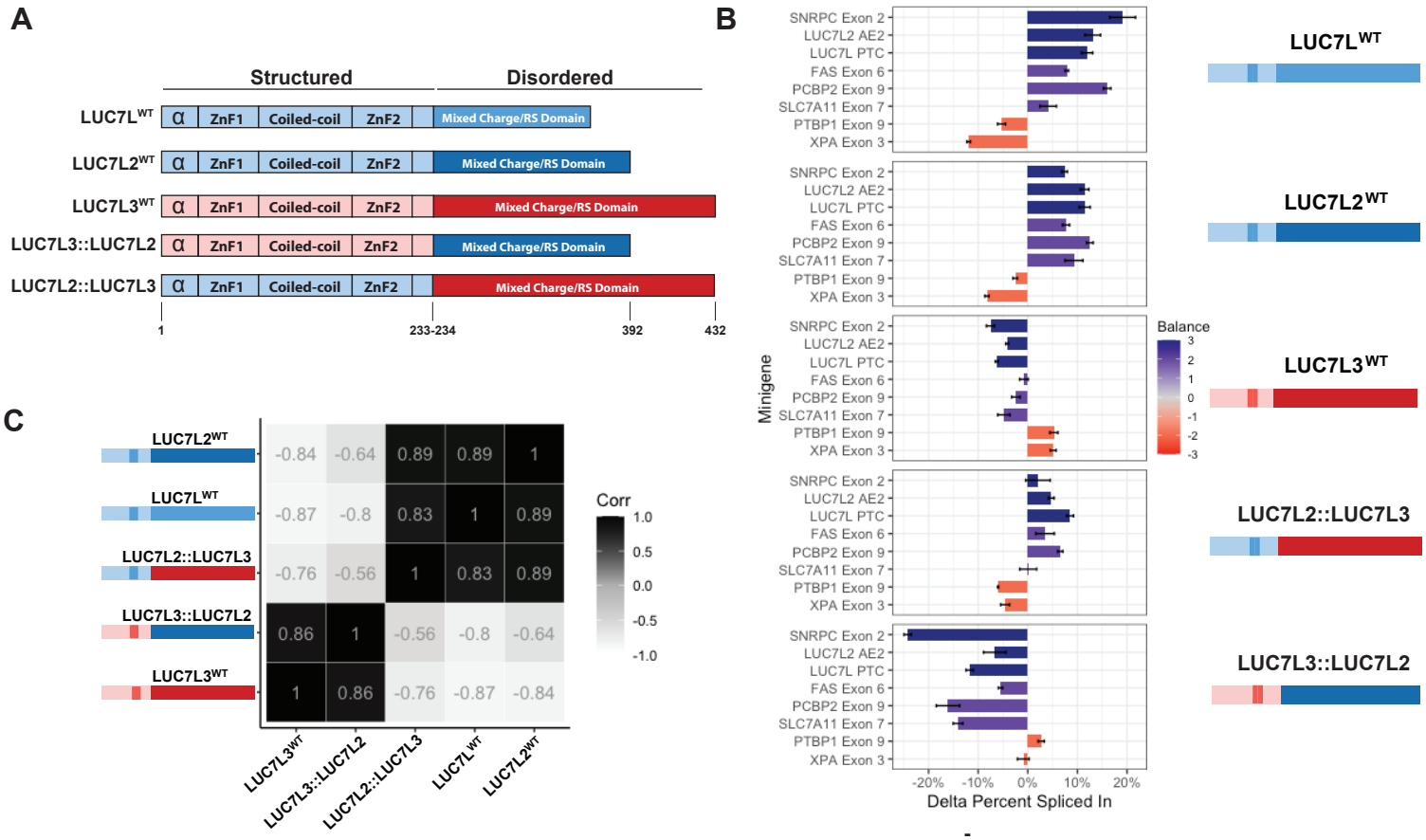

A
Structured
Disordered
LUC7L^WT
LUC7L2^WT
LUC7L3^WT
LUC7L3::LUC7L2
LUC7L2::LUC7L3
α
ZnF1
Coiled-coil
ZnF2
Mixed Charge/RS Domain
1
233-234
392
432
B
SNRPC Exon 2
LUC7L2 AE2
LUC7L PTC
FAS Exon 6
PCBP2 Exon 9
SLC7A11 Exon 7
PTBP1 Exon 9
XPA Exon 3
Minigene
Balance
Delta Percent Spliced In
-20%
-10%
0%
10%
20%
C
Corr
1.0
0.5
0.0
-0.5
-1.0
-0.84 -0.64 0.89 0.89 1
-0.87 -0.8 0.83 1 0.89
-0.76 -0.56 1 0.83 0.89
0.86 1 -0.56 -0.8 -0.64
1 0.86 -0.76 -0.87 -0.84

Figure 5

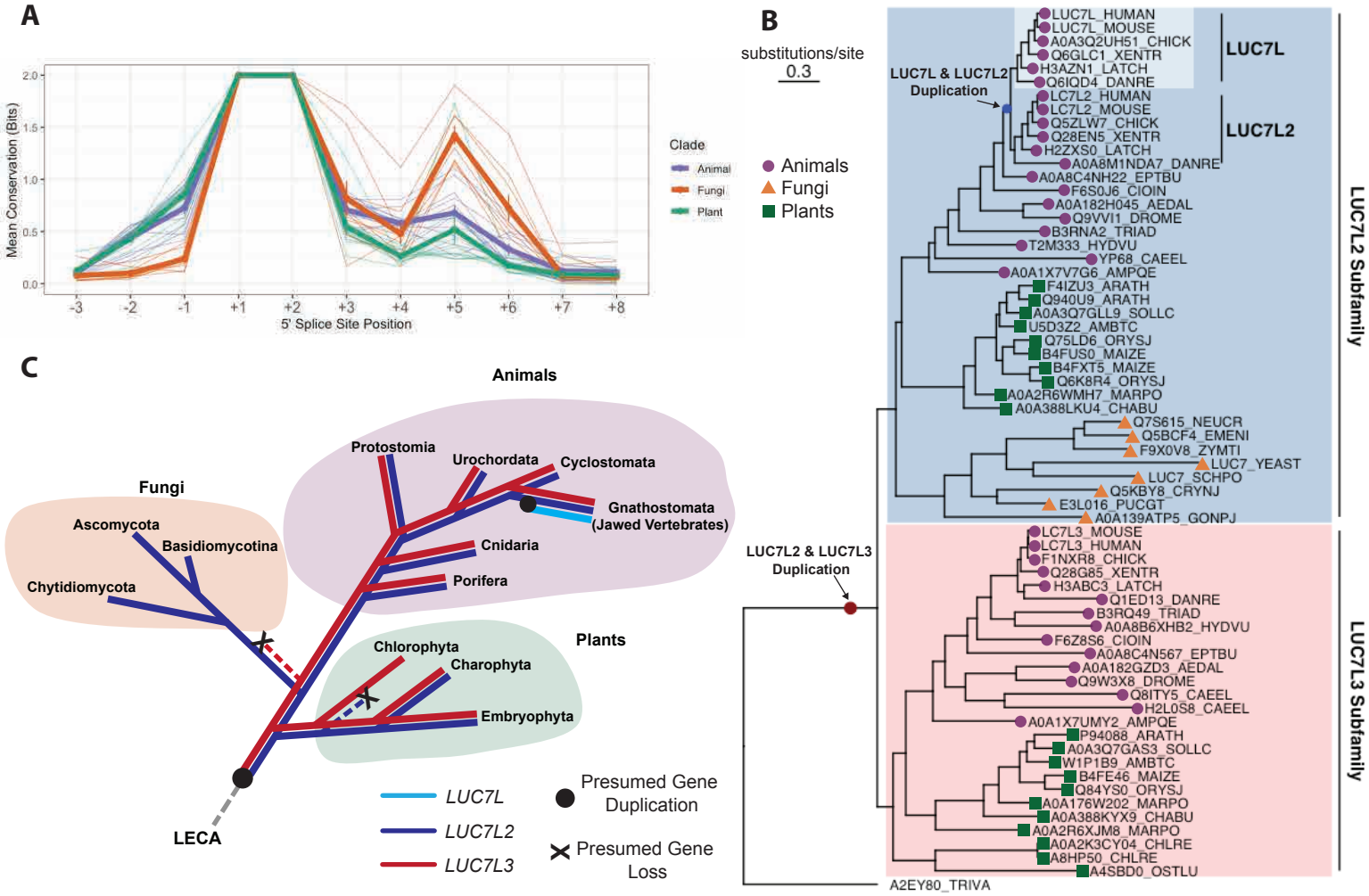
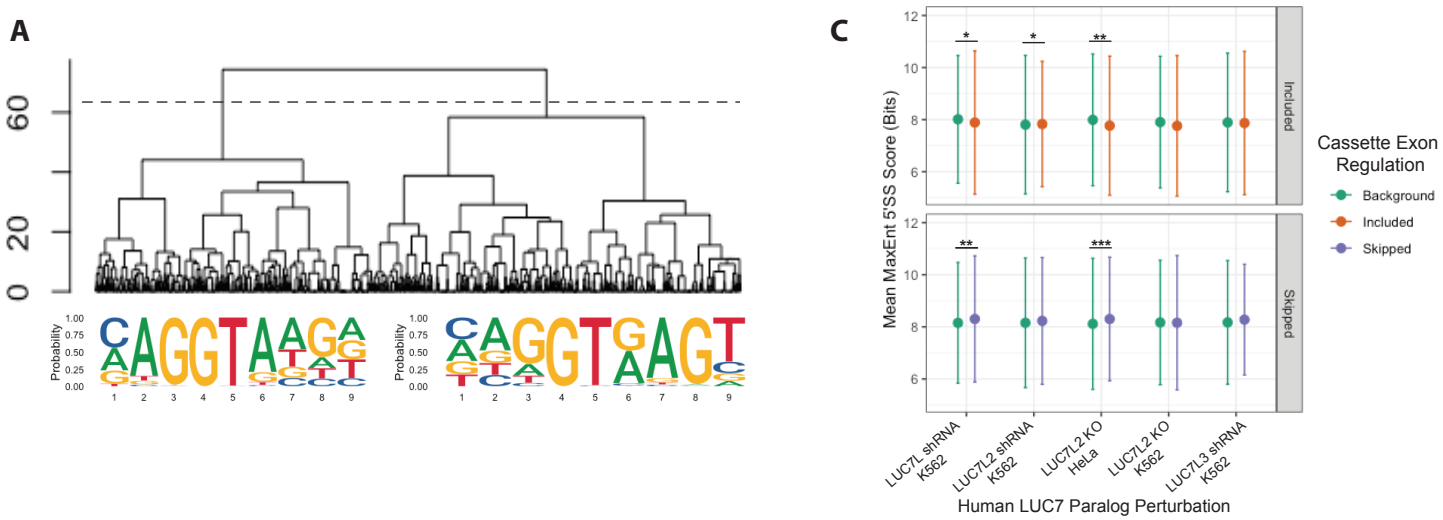
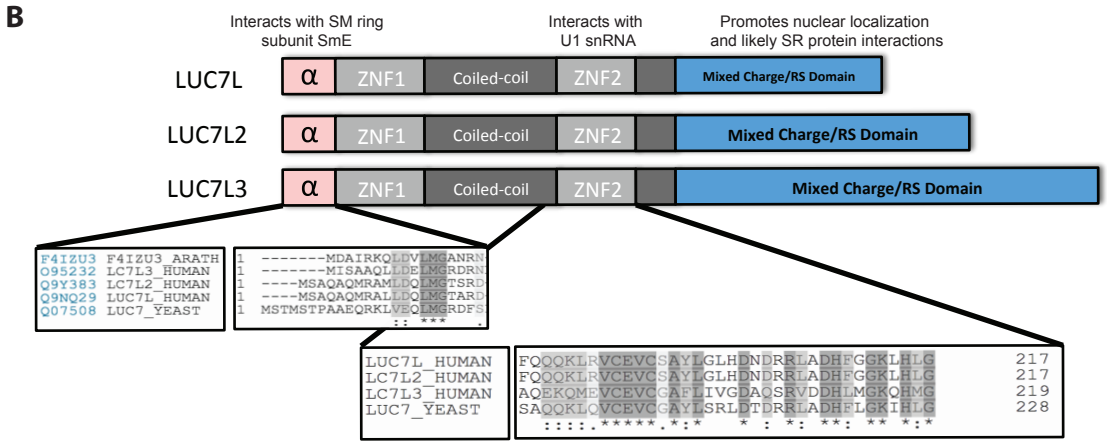

A
Mean Conservation (Bits)
5' Splice Site Position
Clade
Animal
Fungi
Plant
B
substitutions/site
0.3
Animals
Fungi
Plants
LUC7L & LUC7L2 Duplication
LUC7L
LUC7L2
LUC7L2 Subfamily
LUC7L2 & LUC7L3 Duplication
LUC7L3 Subfamily
LUC7L_HUMAN
LUC7L_MOUSE
A0A3Q2UH51_CHICK
Q6GLC1_XENTR
H3AZN1_LATCH
Q6IQD4_DANRE
LC7L2_HUMAN
LC7L2_MOUSE
Q5ZLW7_CHICK
Q28EN5_XENTR
H2ZXS0_LATCH
A0A8M1NDA7_DANRE
A0A8C4NH22_EPTBU
F6S0J6_CIOIN
A0A182H045_AEDAL
Q9VVI1_DROME
B3RNA2_TRIAD
T2M333_HYDVU
YP68_CAEEL
A0A1X7V7G6_AMPQE
F4IZU3_ARATH
Q940U9_ARATH
A0A3Q7GLL9_SOLLC
U5D3Z2_AMBTC
Q75LD6_ORYSJ
B4FUS0_MAIZE
B4FXT5_MAIZE
Q6K8R4_ORYSJ
A0A2R6WMH7_MARPO
A0A388LKU4_CHABU
Q7S615_NEUCR
Q5BCF4_EMENI
F9X0V8_ZYMTI
LUC7_YEAST
LUC7_SCHPO
Q5KBY8_CRYNJ
E3L016_PUCGT
A0A139ATP5_GONPJ
LC7L3_MOUSE
LC7L3_HUMAN
F1NXR8_CHICK
Q28G85_XENTR
H3ABC3_LATCH
Q1ED13_DANRE
B3RQ49_TRIAD
A0A8B6XHB2_HYDVU
F6Z8S6_CIOIN
A0A8C4N567_EPTBU
A0A182GZD3_AEDAL
Q9W3X8_DROME
Q8ITY5_CAEEL
H2L0S8_CAEEL
A0A1X7UMY2_AMPQE
P94088_ARATH
A0A3Q7GAS3_SOLLC
W1P1B9_AMBTC
B4FE46_MAIZE
Q84YS0_ORYSJ
A0A176W202_MARPO
A0A388KYX9_CHABU
A0A2R6XJM8_MARPO
A0A2K3CY04_CHLRE
A8HP50_CHLRE
A4SBD0_OSTLU
A2EY80_TRIVA
C
Animals
Protostomia
Urochordata
Cyclostomata
Gnathostomata
(Jawed Vertebrates)
Cnidaria
Porifera
Fungi
Ascomycota
Basidiomycotina
Chytidiomycota
Plants
Chlorophyta
Charophyta
Embryophyta
LECA
LUC7L
LUC7L2
LUC7L3
Presumed Gene Duplication
Presumed Gene Loss

Figure S1


A
C
Mean MaxEnt 5'SS Score (Bits)
Included
Skipped
Cassette Exon Regulation
Background
Included
Skipped
LUC7L shRNA K562
LUC7L2 shRNA K562
LUC7L2 KO HeLa
LUC7L2 KO K562
LUC7L3 shRNA K562
Human LUC7 Paralog Perturbation


B
Interacts with SM ring subunit SmE
Interacts with U1 snRNA
Promotes nuclear localization and likely SR protein interactions
LUC7L
LUC7L2
LUC7L3
α
ZNF1
Coiled-coil
ZNF2
Mixed Charge/RS Domain

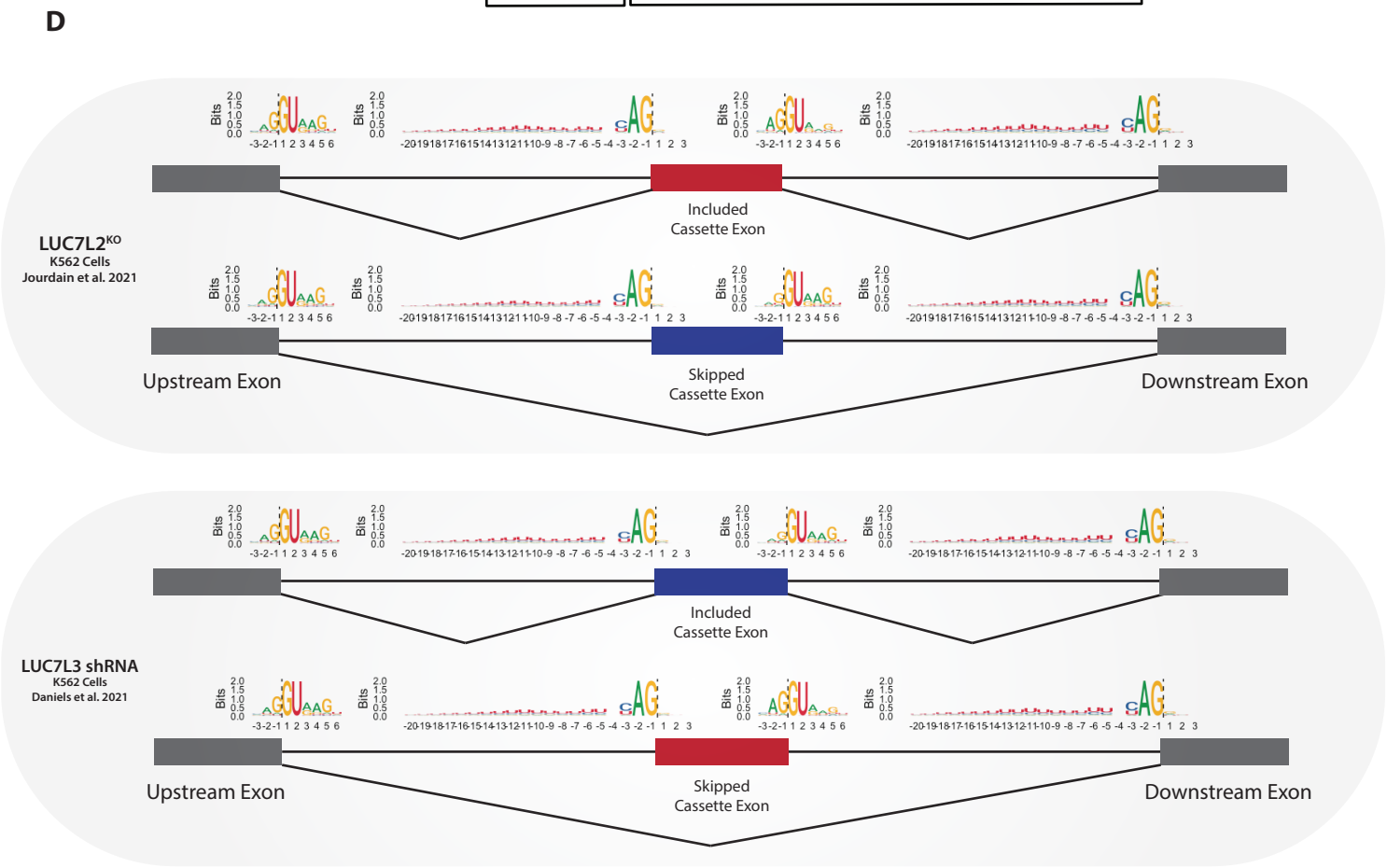

D
LUC7L2KO
K562 Cells
Jourdain et al. 2021
Included Cassette Exon
Skipped Cassette Exon
Upstream Exon
Downstream Exon
LUC7L3 shRNA
K562 Cells
Daniels et al. 2021
Included Cassette Exon
Skipped Cassette Exon
Upstream Exon
Downstream Exon

Figure S2

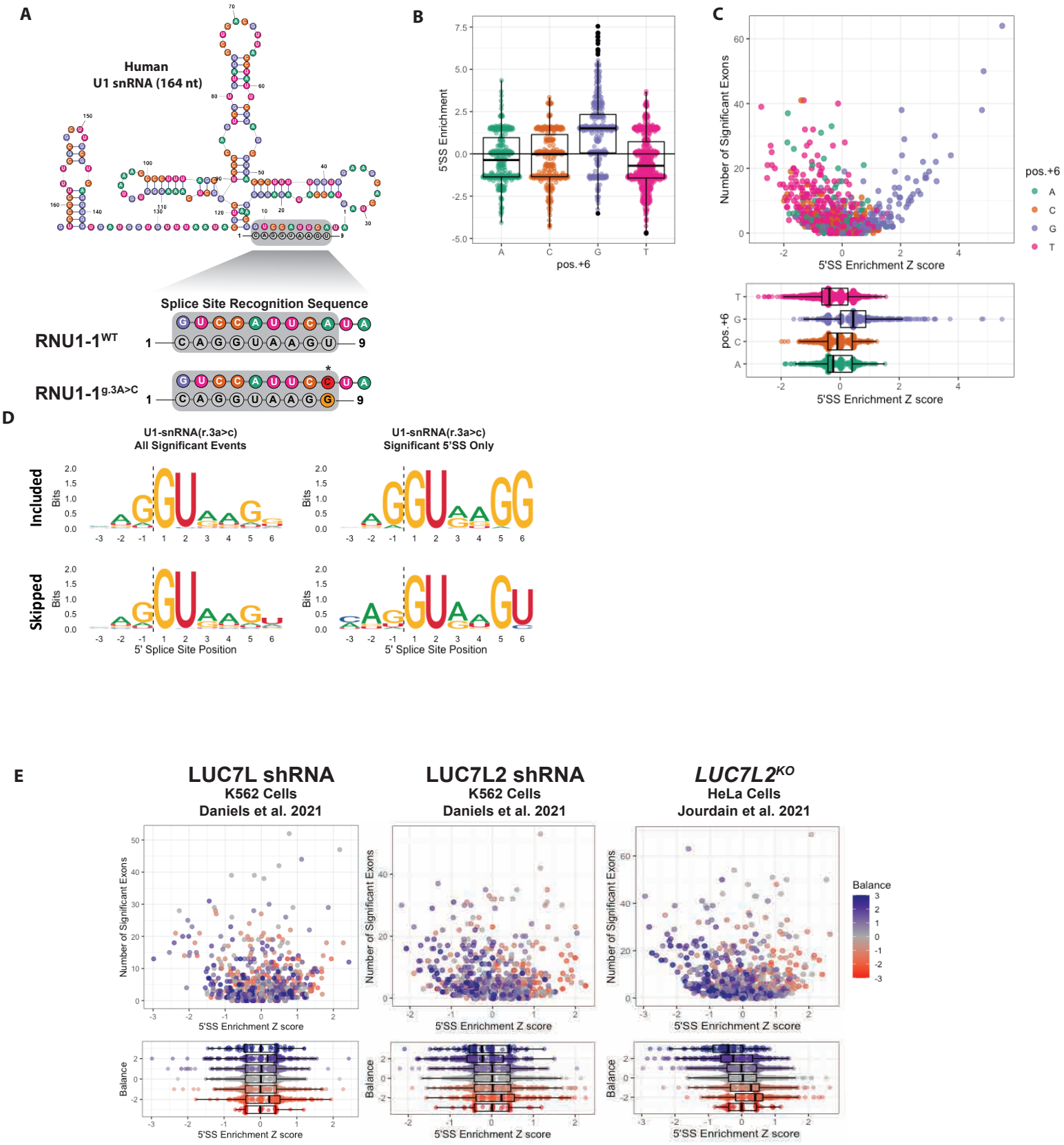
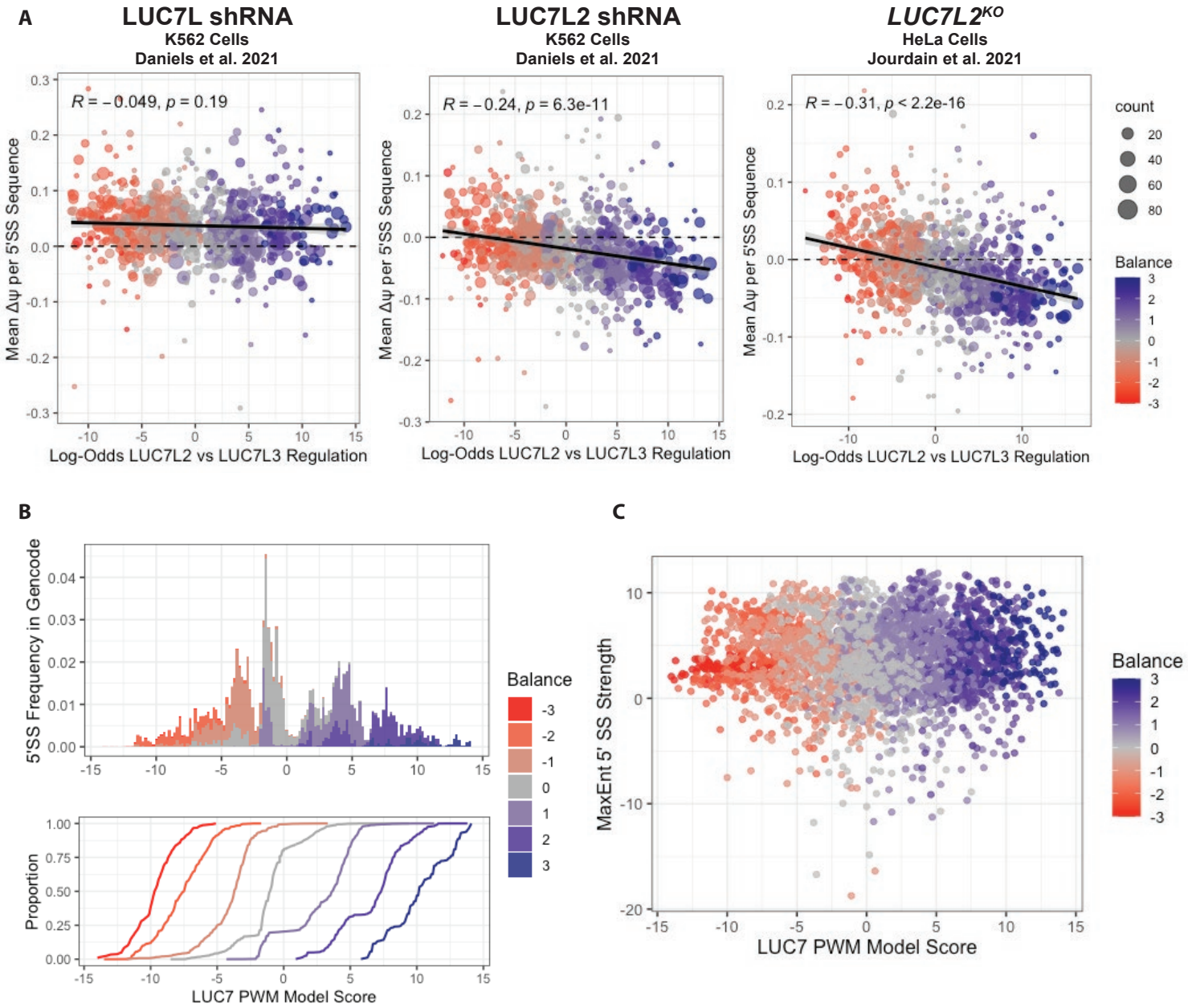

A
Human
U1 snRNA (164 nt)
Splice Site Recognition Sequence
RNU1-1WT
RNU1-1g.3A>C
B
5'SS Enrichment
pos.+6
C
Number of Significant Exons
5'SS Enrichment Z score
pos.+6
D
U1-snRNA(r.3a>c)
All Significant Events
U1-snRNA(r.3a>c)
Significant 5'SS Only
Included
Skipped
Bits
5' Splice Site Position
E
LUC7L shRNA
K562 Cells
Daniels et al. 2021
LUC7L2 shRNA
K562 Cells
Daniels et al. 2021
LUC7L2KO
HeLa Cells
Jourdain et al. 2021
Balance

Figure S3


A
LUC7L shRNA
K562 Cells
Daniels et al. 2021
R = −0.049, p = 0.19
Mean Δψ per 5'SS Sequence
Log-Odds LUC7L2 vs LUC7L3 Regulation
LUC7L2 shRNA
K562 Cells
Daniels et al. 2021
R = −0.24, p = 6.3e-11
Mean Δψ per 5'SS Sequence
Log-Odds LUC7L2 vs LUC7L3 Regulation
LUC7L2KO
HeLa Cells
Jourdain et al. 2021
R = −0.31, p < 2.2e-16
Mean Δψ per 5'SS Sequence
Log-Odds LUC7L2 vs LUC7L3 Regulation
count
20
40
60
80
Balance
3
2
1
0
-1
-2
-3
B
5'SS Frequency in Gencode
Proportion
LUC7 PWM Model Score
Balance
-3
-2
-1
0
1
2
3
C
MaxEnt 5' SS Strength
LUC7 PWM Model Score
Balance
3
2
1
0
-1
-2
-3

Supplemental Methods

Further details on RNA-seq analysis

We use a beta-binomial model to estimate ψ_{ctrl} and $\psi_{kd}$from the corresponding read counts. This amounts to adding pseudocounts to the inclusion and exclusion junction reads, which we determine using empirical Baye's. The effect size in log-odds scale becomes

$$\delta = \log \quad \frac{\psi_{kd}}{\psi_{ctrl}}$$

rMATS evaluates the p-value for each splicing event using a likelihood ratio test, comparing the likelihood of the data assuming a small effect size near zero δ_0 to the likelihood given the most likely effect size δ. Asymptotically, this test statistic $\Lambda = -2\log \quad \frac{L_0}{L_{full}}$ follows a chi-squared distribution with one degree of freedom, and we can use the inverse cumulative distribution function to recover it.

$$\Lambda = CDF_{\chi^2}^{-1}(1-p\,;\; df = 1)$$

Describing the uncertainty about the estimated effect size with a standard deviation σ essentially approximates the likelihood with a normal distribution of mean δ and standard deviation σ. Plugging this in to the likelihood ratio relates the test statistic to the standard deviation of this normal approximation (note that the exponent in the denominator equals zero)

$$\frac{L_0}{L_{full}} = \frac{\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(\delta_0-\delta)^2}{2\sigma^2}\right]}{\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(\delta-\delta)^2}{2\sigma^2}\right]} = \exp\left[-\frac{(\delta_0-\delta)^2}{2\sigma^2}\right]$$

which allows us to rewrite

$$\Lambda = -2\log \quad \frac{L_0}{L_{full}} = \frac{-2(\delta_0-\delta)^2}{-2\sigma^2}$$

and solve for σ as

$$\sigma = \frac{|\delta_0 - \delta|}{\sqrt{\Lambda}}$$

We accomplish shrinkage by passing the effect sizes δ and the associated standard deviations σ of all splicing events to `ashr`. To assume that the fractions of upregulated and downregulated exons may differ, we chose the 'halfnormal' option. This gives us a shrinkage estimator of the effect size δ^* which reframe as an estimate of deltaPSI

$$\psi_{kd}^* = InverseLogit\left(\log \quad \frac{\psi_{ctrl}}{1 - \psi_{ctrl}} + \delta^*\right)$$

$$\Delta\psi^* = \psi_{kd}^* - \psi_{ctrl}$$