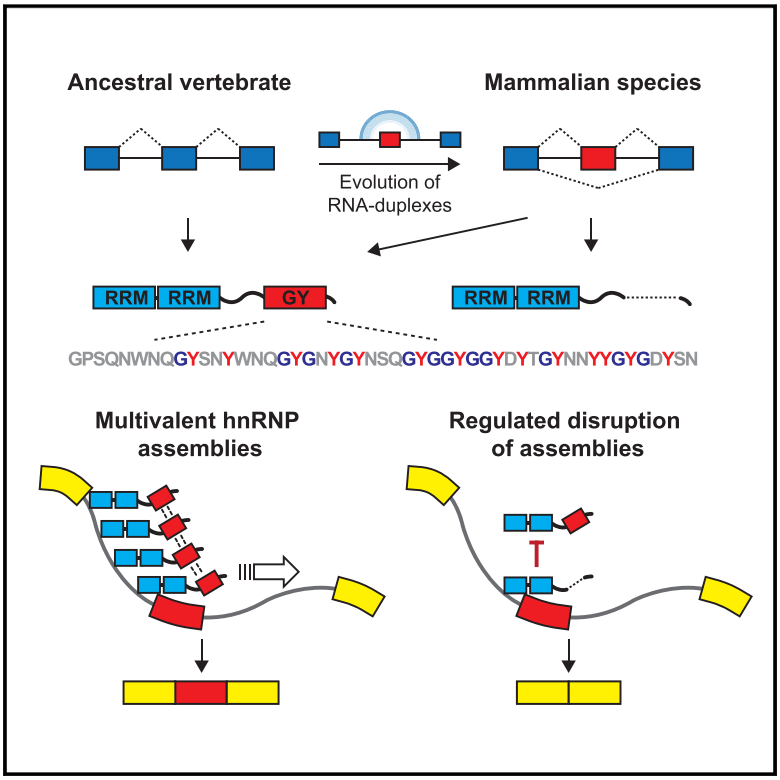
Article

Cell

Regulatory Expansion in Mammals of Multivalent hnRNP Assemblies that Globally Control Alternative Splicing

Graphical Abstract



Highlights

- Mammalian-specific alternative exons impact tyrosine-rich disordered regions
- Long-range RNA duplexes control splicing of these exons in hnRNPs
- The exons promote multivalent assemblies on pre-mRNA to globally control splicing
- Evolution of exon skipping in hnRNPs expanded the regulatory capacity of mammals

Authors

Serge Gueroussov, Robert J. Weatheritt, Dave O'Hanlon, Zhen-Yuan Lin, Ashrut Narula, Anne-Claude Gingras, Benjamin J. Blencowe

Correspondence

serge.gueroussov@utoronto.ca (S.G.), b.blencowe@utoronto.ca (B.J.B.)

In Brief

Mammalian-specific alternative exons control the formation of tyrosine-dependent multi-hnRNP assemblies that, in turn, globally regulate splicing patterns.

Gueroussov et al., 2017, Cell *170*, 324–339
July 13, 2017
http://dx.doi.org/10.1016/j.cell.2017.06.037

CellPress

Cell

Article

Regulatory Expansion in Mammals of Multivalent hnRNP Assemblies that Globally Control Alternative Splicing

Serge Gueroussov,[1,2,5,*] Robert J. Weatheritt,[1,3,5] Dave O'Hanlon,[1] Zhen-Yuan Lin,[4] Ashrut Narula,[1,2] Anne-Claude Gingras,[2,4] and Benjamin J. Blencowe[1,2,6,*]

[1]Donnelly Centre, University of Toronto, Toronto, ON, Canada
[2]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
[3]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge, UK
[4]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON, Canada
[5]These authors contributed equally
[6]Lead Contact
*Correspondence: serge.gueroussov@utoronto.ca (S.G.), b.blencowe@utoronto.ca (B.J.B.)
http://dx.doi.org/10.1016/j.cell.2017.06.037

SUMMARY

Alternative splicing (AS) patterns have diverged rapidly during vertebrate evolution, yet the functions of most species- and lineage-specific splicing events are not known. We observe that mammalian-specific AS events are enriched in transcript sequences encoding intrinsically disordered regions (IDRs) of proteins, in particular those containing glycine/tyrosine repeats that mediate formation of higher-order protein assemblies implicated in gene regulation and human disease. These evolutionary changes impact nearly all members of the hnRNP A and D families of RNA binding proteins. Regulation of these events requires formation of unusual, long-range mammalian-specific RNA duplexes. Differential inclusion of the alternative exons controls the formation of tyrosine-dependent multivalent hnRNP assemblies that, in turn, function to globally regulate splicing. Together, our results demonstrate that AS control of IDR-mediated interactions between hnRNPs represents an important and recurring mechanism underlying splicing regulation. Furthermore, this mechanism has expanded the regulatory capacity of mammalian cells.

INTRODUCTION

A major challenge in biology is to understand which evolutionary changes led to increased cellular and organism complexity. An important mechanism underlying the evolution of complexity is alternative pre-mRNA splicing (AS), the process by which multiple, distinct transcript and protein variants are expressed from a single gene. Recent comparisons of organ transcriptomes from several vertebrate species revealed that AS patterns have diverged rapidly during vertebrate evolution, whereas organ mRNA expression profiles are relatively conserved (Barbosa-Morais et al., 2012; Brawand et al., 2011; Merkin et al., 2012). Moreover, the evolution of complexity in vertebrates is associated with an increased frequency of AS among conserved exons (Barbosa-Morais et al., 2012). Although several studies have demonstrated functional differences between species- and lineage-specific isoforms (Gracheva et al., 2011; Gueroussov et al., 2015; Nicolas et al., 2015), the biological roles and mechanisms underlying the vast majority of these AS events are not known.

Another process associated with the evolution of biological complexity is the expansion of intrinsically disordered regions (IDRs) of proteins. IDRs adopt an ensemble of conformations, rather than defined three-dimensional structures, and frequently embed short linear motifs and sites of post-translational modifications that are important for signaling and protein-protein interactions (van der Lee et al., 2014; Wright and Dyson, 2015; Xue et al., 2012). Because they are not constrained by a specific fold, IDRs are more tolerant to mutations (van der Lee et al., 2014; Wright and Dyson, 2015), which has facilitated both their expansion and rapid evolution (Van Roey et al., 2014; Ward et al., 2004; Xue et al., 2012). Interestingly, alternatively spliced exons, including those subject to tissue- and species- or lineage-specific regulation, are significantly enriched in IDRs (Barbosa-Morais et al., 2012; Buljan et al., 2012; Ellis et al., 2012; Romero et al., 2006). It is therefore possible that the evolution of AS in IDRs has contributed to biological complexity by remodeling protein interaction and signaling networks.

In this study, we investigated the impact on protein function of mammalian lineage-specific AS events—i.e., exons that are alternative in mammalian species but constitutively spliced in non-mammalian vertebrates. We observe that these AS events are enriched in IDRs containing glycine and tyrosine (GY)-rich motifs. Interestingly, IDRs enriched in GY motifs have been implicated in the formation of higher-order protein complexes that can undergo phase separation *in vitro* and assemble into membrane-less organelles and fibrillar-like structures *in vivo* (Kato et al., 2012; Weber and Brangwynne, 2012; Wu and Fuxreiter, 2016). Moreover, aberrant assembly of these structures can lead to the formation of protein aggregates implicated in

multisystem degenerative diseases (Taylor et al., 2016). However, the normal physiological roles of protein assemblies formed by IDRs containing GY and other types of repeat motifs are poorly understood.

Here, we show that mammalian-specific AS events overlapping GY-repeat IDRs are significantly enriched in members of the heterogeneous nuclear ribonucleoprotein (hnRNP) A and D families, which have diverse roles in RNA biology. These AS events arose in the mammalian lineage through the evolution of long-range RNA-RNA interactions that control splice site recognition. Differential inclusion of these mammalian-specific alternative exons dramatically remodels GY-dependent multivalent protein interactions, thereby controlling the assembly of higher-order hnRNP complexes. We demonstrate that formation of these hnRNP assemblies on pre-mRNA is critical for the global regulation of target AS events. Our results thus reveal a recurring mechanism by which AS changes affecting multivalent interactions between hnRNPs have significantly expanded the regulatory complexity of mammalian cells.

RESULTS

Mammalian-Specific AS Events Are Enriched in GY-Rich Low-Complexity Regions Associated with the Formation of Higher-Order Protein Assemblies

Comparative RNA-seq profiling of vertebrate organ transcriptomes has identified hundreds of AS events involving conserved exons that are specific to individual vertebrate species or lineages (Barbosa-Morais et al., 2012; Merkin et al., 2012). These AS events are enriched in exons that overlap IDRs and that preserve reading frame, suggesting that they may often contribute to protein function (Barbosa-Morais et al., 2012). To further investigate these exons, we asked whether they are associated with specific peptide features and gene ontology (GO) categories predictive of specific biological roles. For these analyses, we compiled an expanded set of organ transcriptomes (brain, heart, kidney, liver, lung, skeletal muscle, testis) of four mammals (human, rhesus macaque, mouse, and opossum) and three non-mammalian vertebrates (chicken, frog, and lizard) (Tables S1 and S2). We focused our analysis on exons that were ancestrally constitutive in vertebrates but evolved to become alternatively spliced in the mammalian lineage, referred to below as "mammalian-specific" AS events. These events were compared against three sets of "background" exons: constitutive exons (i.e., >95% spliced in [PSI] across all samples), broadly alternatively spliced exons (i.e., skipped in at least one mammalian and one non-mammalian species), and tissue-regulated alternative exons (i.e., >20 ΔPSI between two or more human tissues) (Figure 1A and Figure S1A).

Confirming previous results (Barbosa-Morais et al., 2012), mammalian-specific AS events are enriched in IDRs (Figure 1B, left; $p < 6.70 \times 10^{-12}$ Wilcoxon rank-sum test). Notably, these IDRs have low sequence complexity (Figure 1B, right; $p < 2.08 \times 10^{-4}$ Wilcoxon rank-sum test), and the protein coding sequences of mammalian-specific exons in these regions are significantly overrepresented in glycine, serine, and proline residues, relative to other classes of exons (Figure 1C; $p < 0.01$; FDR-corrected binomial test). They are also overrepresented in tandem peptide repeats (Figure S1B; $p < 1.67 \times 10^{-3}$ Wilcoxon rank-sum test), of which GY, GP, GS, and AG are the most highly enriched ($p < 0.01$; FDR-corrected binomial test, Figures 1D and S1C). Mammalian-specific AS events containing these sequence features are enriched in genes associated with the GO categories cytoskeleton, proteasome, signaling, RNA binding, and RNA splicing (Figures 1E and S1D; $p < 0.01$; FDR corrected hypergeometric test). Moreover, different types of repeat motifs are associated with distinct GO categories. For example, GP-rich exons are enriched in extracellular matrix genes such as collagen, whereas GY-rich exons are enriched in RNA binding protein (RBP) genes (Figures S1E and S1F; $p < 0.01$, FDR corrected hypergeometric test). Interestingly, GY-rich IDRs in RBPs have recently been implicated as multivalent interaction surfaces that underlie the formation of membrane-less organelle-like structures in cells (Weber and Brangwynne, 2012; Wu and Fuxreiter, 2016) and pathogenic protein aggregates implicated in multisystem degenerative diseases (Taylor et al., 2016). Our observation that mammalian-specific exons are enriched in GY-repeat IDRs thus suggests that they may regulate protein function by controlling the formation of high-order protein assemblies.

Multiple hnRNP Families Contain Mammalian-Specific Alternative Splicing Events Overlapping GY-Repeat-Rich IDRs

Enrichment of GY dipeptides is particularly high among hnRNPs (Figure 2A). The human genome contains 37 hnRNP genes, which comprise distinct subfamilies that have arisen through numerous duplication events (Barbosa-Morais et al., 2006; Busch and Hertel, 2012). All seven members of the related hnRNP A and hnRNP D subfamilies possess GY-rich C-terminal IDRs (Figure 2B). Remarkably, five (HNRNPD, DL, AB, A1, and A2B1) of the six multi-exon genes in these subfamilies contain mammalian-specific AS events within these IDRs (Figure 2B). The effect of these events on the functions of the hnRNP D family (D, DL, and AB) was predicted to be particularly striking, since skipping of the alternative exons would result in the loss of nearly all C-terminal GY motifs (Figure 2B).

To confirm the mammalian-specific patterns of these AS events, we performed reverse-transcription polymerase chain reaction (RT-PCR) assays to assess their inclusion levels across multiple tissues from representative mammalian (human, mouse) and non-mammalian (chicken, frog) species (Figures 2C and S2A). These analyses verified all (4/4) predicted lineage-specific patterns, including those of HNRNPD exon 7 (Barbosa-Morais et al., 2012) and HNRNPA1 exon 8, where the exons are skipped in all analyzed mammalian tissues, and of HNRNPA2B1 exon 9, which is predominantly skipped in testis (Figures 2A, 2C, S2A, and S2B). In addition to being skipped in all mammalian tissues, HNRNPAB exon 7 was found to be skipped in chicken, a species where poor annotation precluded RNA-seq prediction (Figures 2A and S2A). The consistent patterns of AS of paralogous hnRNP exons, and the overlap of these exons with GY-rich IDRs, suggests that they may be regulated by similar mechanisms and that they act broadly across different mammalian cell and tissue types to regulate RNA-associated cellular functions.

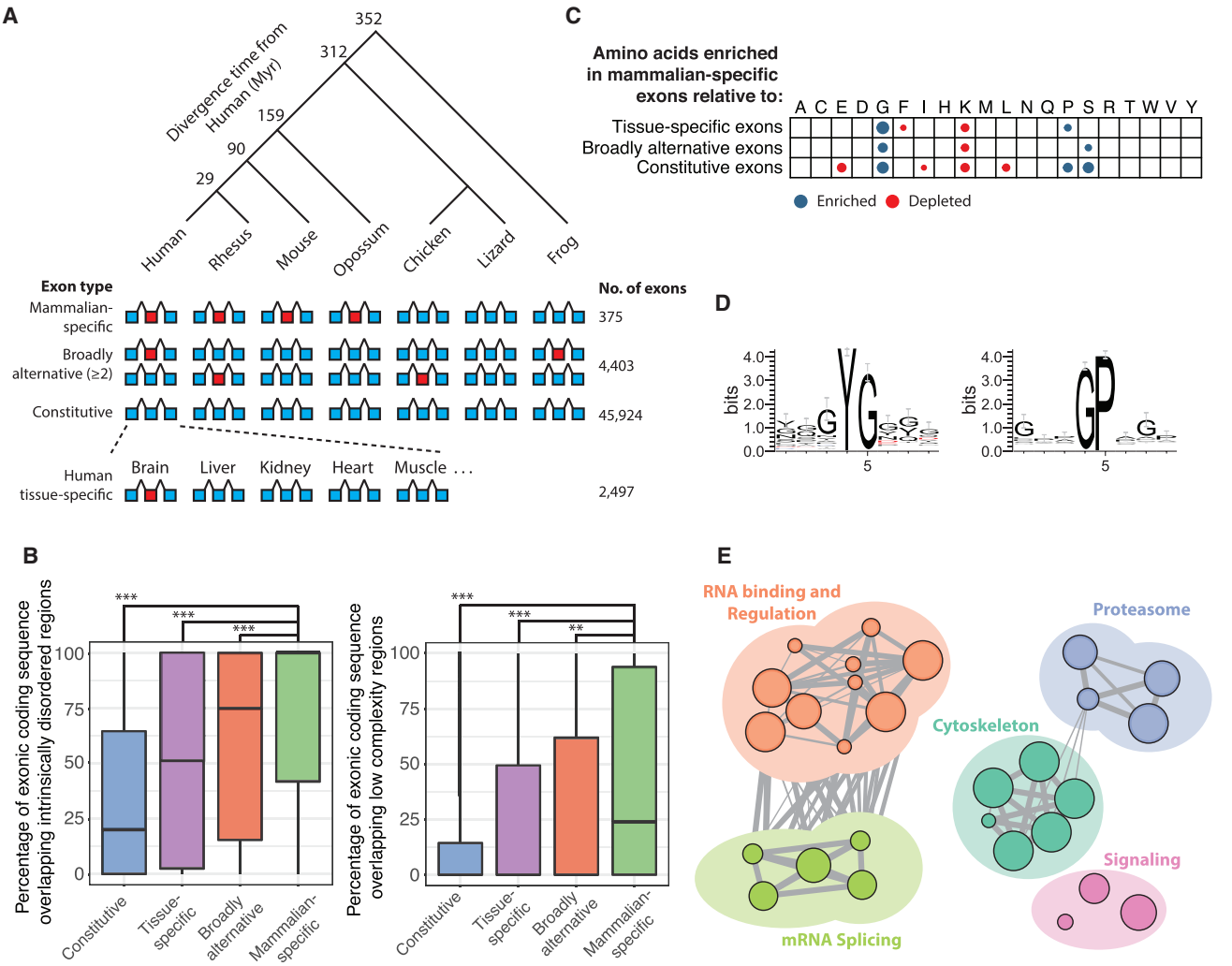


Figure 1. Amino Acid Sequence Features of Mammalian-Specific AS Events

(A) Phylogenetic tree of analyzed species with distance from human in millions of years (Myr). For each splicing event category (refer to main text and STAR Methods), representative splicing patterns and numbers of events detected from analysis of RNA-seq data are shown (blue, constitutive exons; red, alternative exons).

(B) Boxplots comparing the percentages of exonic coding sequence that overlap intrinsically disordered (i.e., predicted to not form stable tertiary structure) (left) and low-complexity (i.e., with low amino acid diversity) (right) regions of proteins, for exons that belong to different classes of splicing events (see A). Boxplots display the interquartile range as a solid box, with vertical thin lines representing 95% confidence intervals and horizontal lines representing median values. $^{**}p < 1 \times 10^{-3}$; $^{***}p < 1 \times 10^{-6}$, Wilcoxon rank-sum tests.

(C) Amino acids enriched (blue dots) or depleted (red dots) in mammalian-specific exons overlapping intrinsically disordered regions relative to exons belonging to the indicated categories ($p < 0.01$; binomial test with Bonferroni correction; size of dots is inversely related to the magnitude of $\log_2$ p value).

(D) Logos of 8-mer amino acid sequences surrounding dipeptides enriched in mammalian-specific alternative exons, compared to tissue-specific alternative exons. The relative height of the amino acid indicates its frequency at a given position, and its total height indicates the amount of information at the position (in bits).

(E) Enrichment map for GO, REACTOME, and KEGG functional categories of genes that contain mammalian-specific AS events, with representative GO terms shown for each sub-network. Node size is proportional to the number of genes associated with the GO category, and edge width is proportional to the number of genes shared between GO categories.

Mammalian-Specific Mechanisms of hnRNP Alternative Splicing

Evolutionary transitions from constitutive to AS are frequently associated with splice site weakening (Xing and Lee, 2006). However, the strength of splice sites associated with the hnRNP mammalian-specific exons does not account for their differential splicing (Figure S3A). To investigate the mechanism underlying mammalian-specific skipping of exon 7 of HNRNPD, we generated splicing minigene reporter constructs containing the orthologous human and chicken genomic sequences surrounding this exon and transfected them into HEK293 cells (Figure 3A; Table S3). Consistent with RNA-seq data, human exon 7 showed

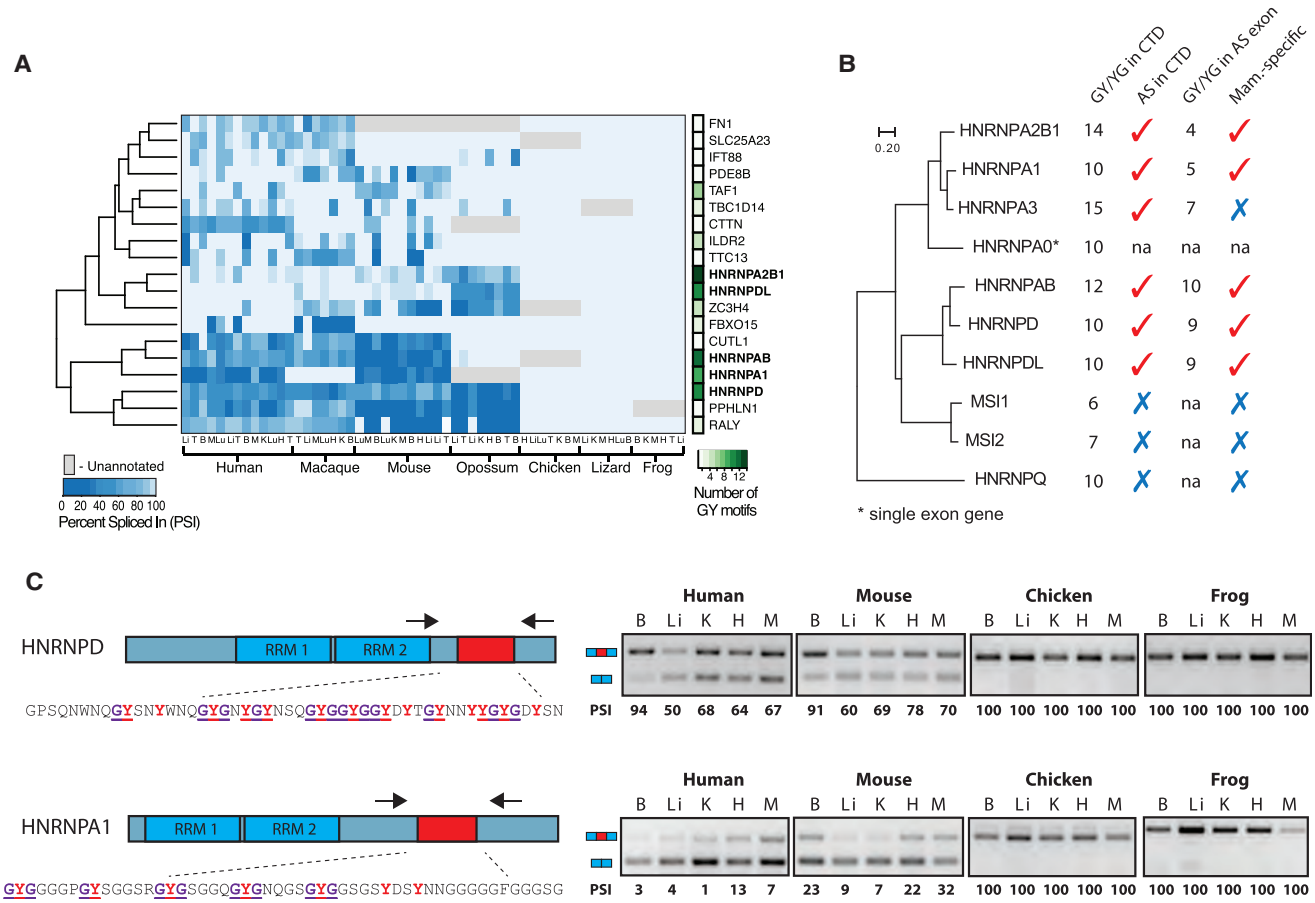


Figure 2. Identification and Characterization of Mammalian-Specific GY-Rich Alternative Exons in hnRNPs

(A) Heatmap showing RNA-seq-derived percent spliced in (PSI) values for mammalian-specific alternative exons and orthologous exons in tissues of different vertebrate species. HNRNP genes are highlighted in bold. Blue-scale, PSI value; green-white scale, numbers of GY/YG dipeptides in the human exon; gray boxes, absence of an Ensembl-annotated orthologous exon. B, brain; H, heart; K, kidney; Li, liver; Lu, lung; M, skeletal muscle; T, testis.

(B) Cladogram showing evolutionary relationships between hnRNP amino acid sequences (see STAR Methods). The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Indicated for each protein are: number of GY/YG dipeptides in low-complexity, intrinsically disordered C-terminal domains (CTD), presence of alternative splicing (AS) in the CTD, number of GY/YG motifs in the exon, and whether the alternative exon is mammalian-specific.

(C) Domain organization of HNRNPD and HNRNPA1. Mammalian-specific alternative exons are indicated in red, with the corresponding amino acid sequence shown below. RT-PCR analyses confirming RNA-seq-derived patterns of splicing for representative mammalian-specific AS events. Percent spliced in (PSI) values are indicated (see STAR Methods). Tyrosine residues are indicated in red, GY/YG motifs are underlined, and glycine residues within the motifs are indicated in purple. Arrows, approximate location of RT-PCR primers; RRM, RNA recognition motif; B, brain; Li, liver; K, kidney; H, heart; M, muscle.

a PSI value of ~40%, whereas the orthologous chicken exon was constitutively included (Figure 3A, Hsa and Gga). To determine which sequences specify constitutive inclusion, we systematically replaced regions of the human minigene with positionally matched chicken sequence. While substitution of the exon itself had little effect (Figure 3A, mt.1), replacing the upstream (I6) or downstream intron (I7) with the corresponding chicken introns resulted in constitutive inclusion (Figure 3A, mt.2 and mt.3). More detailed mutagenesis determined that the critical chicken sequence elements required to promote constitutive splicing are within a 30 nt region at the 3′ end of I6 (mt.5) and a 40 nt region in the 3′ half of I7 (mt.7) (Figures 3B and 3C and Figure S3B). Substitution of either region in the chicken minigene with human sequence did not induce skipping (Figure 3D, mt.8 and mt.9). However, simultaneously introducing both elements resulted in a PSI of ~40%, i.e., equivalent to normal levels of skipping of human exon 7 (Figure 3D, mt.10).

Given the requirement for both elements, we examined these sequences for base-pairing potential. Remarkably, these sequences are predicted to form base pairs at 36 out of 39 possible positions (Figure 3E). The predicted duplex encompasses the branch point site, polypyrimidine tract, and 3′ acceptor sequence adjacent to exon 7, suggesting that skipping results from masking of these critical elements. To confirm the importance of the predicted duplex, we introduced compensatory mutations in I7 to restore base-pairing potential in the context of the mt.5 minigene containing chicken sequence in I6. This change shifted splicing from full inclusion to a PSI of

40% (Figure 3F, mt.12). Moreover, mutating the human I7 sequence to extend the duplex by 20 nt reduced exon 7 PSI to <5% (Figure 3F, mt.11). Finally, to test the importance of base-pairing in the context of endogenous HNRNPD transcripts, we used Cas9 in combination with independent pairs of guide RNAs to delete the I7 element from genomic DNA in a pool of HEK293 cells (Figure 3G). Importantly, the efficiency of genomic deletion in the pool was proportional to the increase in exon 7 inclusion (Figure 3G). Finally, to assess whether duplex formation accounts for mammalian-specific exon 7 skipping, we aligned I6 and I7 sequences from 18 vertebrate genomes (Figure 3H). The I6 and I7 elements characterized above display a striking degree of sequence conservation and conserved base-pairing potential, relative to the surrounding intronic sequence. In contrast, in non-mammalian vertebrates, although the I6 element displays partial sequence conservation, the I7 element is not conserved. These results demonstrate that mammalian-specific skipping of hnRNP D exon 7 arose through the evolution of a long-range RNA duplex in the mammalian lineage that occludes critical splicing elements.

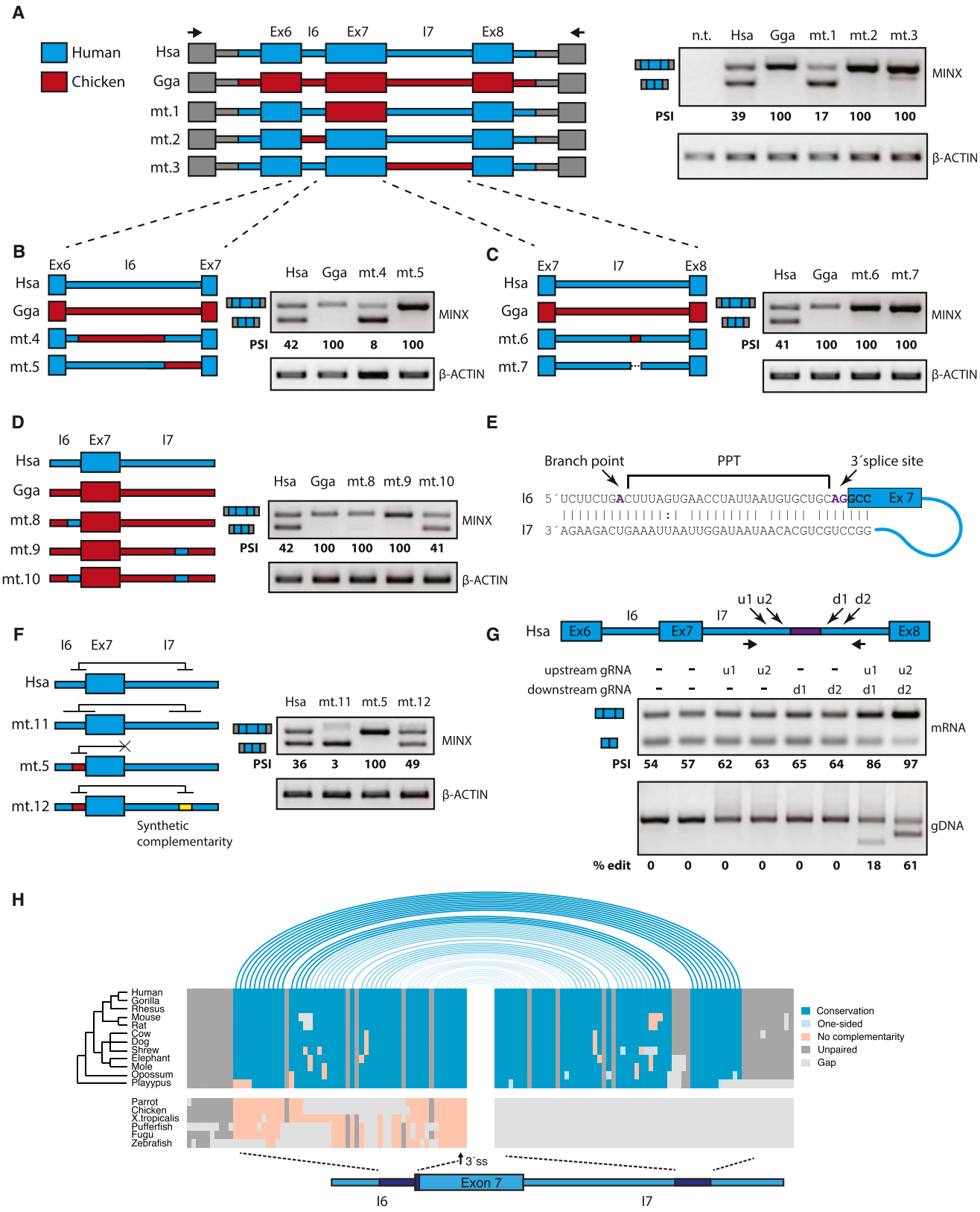
We next investigated whether similar mechanisms may account for evolution of AS of other hnRNP exons. A previous study demonstrated that an intronic RNA element downstream of HNRNPA1 exon 8 can loop back to mask its 5′ splice site, resulting in skipping of the exon (Blanchette and Chabot, 1997). Consistent with its mammalian-specific AS, we observe that this downstream element is located within a sequence that is uniquely conserved in mammals (Figure S3C). Interestingly, intronic sequences flanking HNRNPAB exon 7 show two pronounced regions of conservation. When examining these sequences for possible complementarity, we predict formation of a 28 nt, GC-rich duplex to be conserved in mammals, although poor annotation precluded reliable prediction in the bird lineage (Figure S3D). Regulation of AS via formation of long-range duplexes appears to be rare, as a search for potential duplex-forming sequences flanking all other detected mammalian-specific alternative exons identified only two additional possible examples (in the *RBM10* and *ZMIZ1* genes), and no such sequences were detected in association with a comparable number of more broadly conserved AS events (see STAR Methods). These data thus reveal the emergence of related mechanisms involving RNA duplex formation in the evolution of AS of GY-repeat IDRs of hnRNPs. Independent fixation of these mechanisms in related genes strongly suggests that the mammalian-specific AS events in hnRNPs are functionally important.

Alternative Splicing in GY-Rich IDRs of hnRNPs Remodels Protein Interaction Networks and Affects the Formation of Higher-Order Protein Assemblies

The GY-rich IDRs of hnRNP A/D family members, and of other RBPs, have been reported to act as multivalent protein-protein interaction surfaces that mediate phase separation and formation of organelle-like higher-order protein assemblies (Weber and Brangwynne, 2012; Wu and Fuxreiter, 2016). We therefore hypothesized that an important function of the mammalian-specific AS events is to regulate GY-dependent interactions that mediate higher-order hnRNP assemblies. Skipping of HNRNPD exon 7 is predicted to remove nine of ten GY dipeptides in the C-terminal IDR and therefore have a particularly strong effect on its function (Figures 2B and 2C). To test this, we used two independent proteomic approaches to investigate the impact of exon 7 skipping on protein-protein interactions. HEK293 cell lines were generated that stably express 3xFLAG-tagged isoforms of HNRNPD containing or lacking exon 7 (HNRNPD+Ex7 and HNRNPDΔEx7) at near endogenous levels (Figure S4A), and anti-FLAG immunoaffinity purification (in the absence of nuclease treatment) coupled to mass spectrometry (AP-MS) was used to identify interacting proteins (Figure 4A; Table S4). As a complementary approach to detect interacting or proximal proteins, cell lines were generated that express HNRNPD+Ex7 and HNRNPDΔEx7 fused to a mutant derivative of BirA that promiscuously biotinylates neighboring proteins (Roux et al., 2012) (Figures S4B and S4C; Table S5). Proteins biotinylated by the BirA fusions were then recovered on streptavidin beads and identified by MS (BioID-MS). For comparison and specificity control purposes, we applied these procedures to analyze proteins associated with splice isoforms of the hnRNP PTBP1 (also known as HNRNPI), with and without its mammalian-specific exon (exon 9) (Gueroussov et al., 2015), which overlaps a distinct IDR that lacks GY repeats (Figure S4D).

After clustering the normalized peptide counts of interacting proteins identified by AP-MS, we observe both common and different interaction partners for HNRNPD+Ex7 and HNRNPDΔEx7 (Figure 4B). Importantly, HNRNPD+Ex7 preferentially interacts with multiple other hnRNP proteins including HNRNPA1, A2B1, A3, C, DL, F, H1, K, L, and R (Figures 4B and 4C; $p < 1.51e-07$; Wilcoxon rank-sum test), as well as other GY-rich RBPs, such as KHDRBS1 (Figure 4B; Table S4). Consistent with multivalent, GY-promoted assembly of higher-order complexes, HNRNPD+Ex7-specific interaction partners are significantly enriched in G and Y residues (Figure S4E; $p < 0.01$; binomial test). Furthermore, a computationally identified set of 147 human proteins containing ≥ 4 GY dipeptides concentrated within low-complexity regions is significantly enriched among preys that co-immunoprecipitate with HNRNPD+Ex7 (Figure 4D; $p < 7.35e-11$; Wilcoxon rank-sum test; Table S6). In contrast, proteins interacting with HNRNPDΔEx7 are not enriched in GY dipeptides (Figure S4E; Table S4). Similar results were obtained for proximal proteins detected by BioID-MS (Figures S4F–S4H; Table S5). In contrast to these results, the interaction profiles of PTBP1+Ex9 and PTBP1ΔEx9 are nearly identical to each other and distinct from those of either HNRNPD isoform (Figure 4B). Collectively, these results demonstrate that inclusion of exon 7 of HNRNPD strongly and selectively promotes interactions with many additional hnRNPs as well as other GY-rich proteins.

We next used co-immunoprecipitation assays to investigate the extent to which representative hnRNP isoform-differential homotypic and heterotypic interactions are dependent on RNA and GY-repeat tyrosine residues that overlap the mammalian-specific alternative exons (Figures 4E and 4F and Figures S5A–S5C). Consistent with the AP-MS results, in the absence of nuclease treatment, HNRNPD+Ex7 shows a modest preference for binding HNRNPAB and HNRNPA1 isoforms compared to HNRNPDΔEx7 (Figures 4E and S5A). However, following nuclease treatment, interactions between HNRNPD



(legend on next page)

and HNRNPAB, HNRNPA1, or HNRNPD itself could only be observed when both partners contained their GY-rich alternative exon (Figures 4E, S5A, and S5B). Similar results were observed when assaying interactions between HNRNPAB and HNRNPA1 isoforms (Figure S5C). In contrast to these results, after nuclease treatment, HNRNPA1ΔEx8 can interact with itself, albeit at markedly reduced levels compared to the self-interaction involving HNRNPA1+Ex8 (Figure 4F). The C-terminal IDR of HNRNPA1 is relatively long and, in contrast to exon 7 of HNRNPD, skipping of its mammalian-specific alternative exon removes only five of ten GY motifs, potentially leaving a sufficient number of GY motifs in the IDR to mediate self-interaction (Figure 2B and Figure S5D). Confirming this, mutation of the remaining tyrosine residues to serine abolishes the HNRNPA1 self-interaction and also prevents its interaction with HNRNPAB (Figures 4F, S5C, and S5D). In summary, these protein interaction data collectively show that GY-repeat IDRs mediate interactions with multiple hnRNPs and other RBPs that are enriched in GY motifs and, furthermore, that the mammalian-specific alternative exons overlapping these regions significantly impact these multivalent interactions.

Given these results, and previous studies demonstrating that GY-rich regions of RBPs can promote phase transitions to form droplet-like structures *in vitro*, and membrane-less organelle-like bodies *in vivo* (see Introduction), we also investigated whether the mammalian-specific alternative exons of hnRNPs affect the formation of these higher-order assemblies. In agreement with previous results (Mannen et al., 2016), HNRNPD+Ex7, but not HNRNPDΔEx7, can localize to a small number of subnuclear foci (Figure 5A). Moreover, consistent with protein interaction data described above, these differences are due to the tyrosine residues overlapping exon 7, since HNRNPD+Ex7 with all exon 7 tyrosine residues mutated to serine failed to localize to foci (Figure 5A and Figure S5E). Moreover, consistent with the ability of HNRNPA1 to form homotypic and heterotypic hnRNP interactions (Figure 4) that are less affected by skipping of its mammalian-specific exon, both HNRNPA1 isoforms concentrate in nuclear foci (Figure 5B). However, mutation of the tyrosine residues to serine in the IDR of HNRNPA1 results in the appearance of morphologically distinct foci and a dramatic loss of nuclear-specific localization and dispersal of the protein throughout the cell (Figure 5B). Finally, while solubility issues precluded the biophysical characterization of HNRNPD isoforms, a direct comparison of the propensity of the HNRNPA1 isoforms and tyrosine to serine substitution mutant to undergo phase separation *in vitro* revealed striking differences (Figures 5C–5E and S5F). At moderate salt and protein concentrations, both HNRNPA1+Ex8 and HNRNPA1ΔEx8, but not HNRNPA1ΔEx8-Y8S, form liquid-like droplets (Figure 5D). However, inclusion of the mammalian-specific alternative exon in hnRNPA1 results in the formation of droplets at lower protein and higher salt concentrations than HNRNPA1ΔEx8, whereas HNRNPA1ΔEx8-YS is only able to form droplets in low salt, at the highest protein concentration assayed (Figure 5E). Collectively, these results demonstrate that the mammalian-specific alternative exons of hnRNPs, in addition to controlling multivalent protein interactions with other hnRNPs in a tyrosine-dependent manner, also control the formation of large assemblies *in vitro* and in cells.

Mammalian-Specific hnRNP Isoforms in the Regulation of Alternative Splicing

To further investigate the functional significance of isoform-dependent multivalent protein interactions mediated by GY-rich AS events of hnRNPs, we next determined whether they impact splicing regulation. First, to identify alternative exon targets of hnRNP A and D family members, we systematically compared changes in global AS profiles following depletion of hnRNPs using knockdown RNA-seq data generated by the ENCODE consortium (Sloan et al., 2016; Sundararaman et al., 2016). Interestingly, consistent with the sequence relationships between the C-terminal IDRs of HNRNPD, HNRNPAB, and HNRNPA0 and between HNRNPA1 and HNRNPA2B1, knockdown of these proteins resulted in sub-clusters of more highly correlated AS profiles (Figure 6A). Analysis of the RNA binding specificities of these hnRNPs reveals distinct but overlapping binding preferences that are also consistent with their evolutionary relationships and correlated effects on AS (Figure S6A).

Figure 3. Long-Range Intra-molecular RNA Interactions Control Mammalian-Specific AS Events in hnRNPs

(A) Schematic diagrams of HNRNPD exon 7 (Ex7) minigene reporters containing human (blue) and chicken (red) sequences encompassing the flanking native exons (Ex6 and Ex8) and intervening introns (I6 and I7) (see STAR Methods). Reporter gene constitutive exons and intronic sequences (gray) are derived from adenovirus (MINX exons). Constructs were transfected into HEK293 cells and RT-PCR assays were performed with MINX-specific primers. β-ACTIN was used as a loading control. Reporter percent spliced in (PSI) values are indicated (see STAR Methods). Hsa, *Homo sapiens*; Gga, Gallus gallus; n.t., not transfected; mt, mutation.

(B) Effects of additional substitution mutations in I6 and I7 on Ex7 splicing levels. Reporter backbone and RT-PCR assays as in (A).

(C) Effects of additional substitution mutations in I7 on Ex7 splicing levels. Reporter backbone and RT-PCR assays as in (A).

(D) Effects on splicing levels of introducing human sequence into a chicken HNRNPD exon 7 minigene. Reporter backbone and RT-PCR assays as in (A).

(E) Diagram of predicted RNA-RNA loop that forms between human I6 and I7 sequence elements. Core splicing signals are indicated in bold purple text. PPT, polypyrimidine tract; dots indicate non-Watson-Crick base-pairing.

(F) Effects on splicing of human HNRNPD exon 7 after extending, disrupting, and genetically restoring RNA-RNA loop formation between I6 and I7 sequences. Reporter backbone and RT-PCR assays as in (A).

(G) Role of the endogenous I6-I7 RNA-RNA loop in the control of splicing of human HNRNPD exon 7. CRISPR-Cas9 editing was used to delete the loop-forming element in I7 by employing guide RNAs at the indicated locations. Deletion efficiency was assayed by PCR using primers (arrows) targeted to the flanking genomic sequences. The effect of element deletion on endogenous HNRNPD exon splicing levels was assayed by RT-PCR using primers to flanking constitutive exons.

(H) Arc diagram illustrating conserved base-pairing positions formed by the I6-I7 RNA-RNA duplex. Arcs display complementary nucleotide base-pairings (G:C; A:U, G:U, and vice versa). Blue, conserved complementarity; orange, nucleotide changes resulting in loss of base-pairing; light gray, indels or regions that fail to align; dark gray, lack of complementary between I6 and I7 elements. Schematic diagram displays introns 6, 7, and exon 7 of HNRNPD, with location of base-pairing elements indicated in dark blue. 3′ss, 3′ splice site.

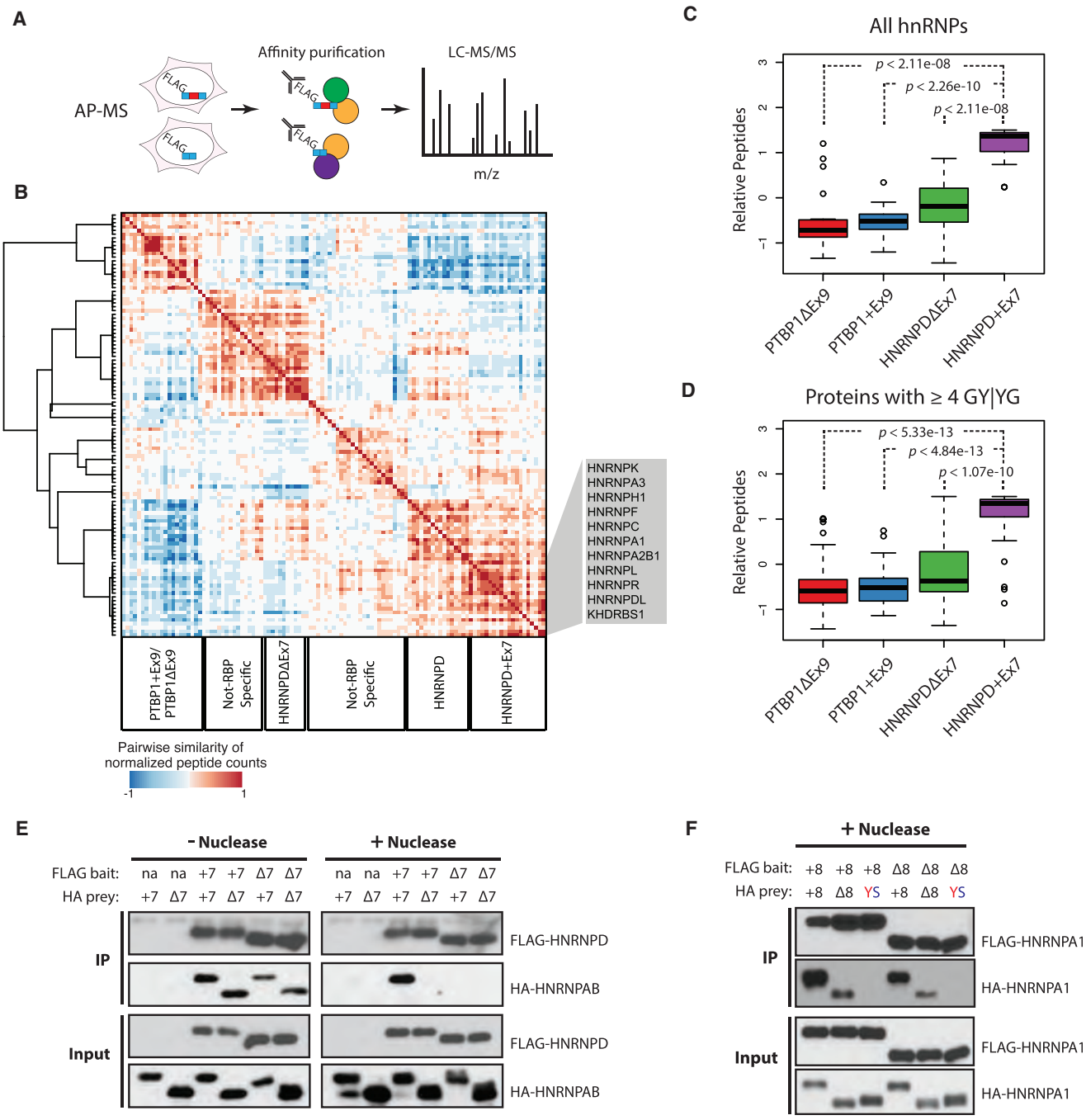


Figure 4. Mammalian-Specific AS of GY-Rich IDRs of hnRNPs Remodels Protein Interaction Networks

(A) Experimental outline for affinity purification followed by mass spectrometry (AP-MS).

(B) Symmetrical heatmap of pairwise correlations of normalized peptide counts from AP-MS of stable cell lines expressing FLAG-tagged HNRNPD+Ex7, HNRNPDΔEx7, PTBP1+Ex9, or PTBP1ΔEx9 proteins. Affinity propagation clustering is based on pairwise similarities (represented as positive and negative correlations) between profiles of detected interaction partners. Sub-clusters identified using exemplar-based agglomerative clustering. Sub-clusters are labeled based on enrichment of peptides interacting with a specific hnRNP isoform or gene. "Not RBP-specific" refers to sub-clusters where there is no clear preference for binding to PTBP1 or HNRNPD. The gray box highlights proteins enriched in the HNRNPD+Ex7-containing cluster.

(C) Boxplots showing normalized peptide distributions of all hnRNP preys identified by AP-MS (see STAR Methods). p values from Wilcoxon signed-rank test. For a description of boxplots, see Figure 1B legend.

(D) Same as in (C) but for all human GY-rich preys identified by AP-MS. A total of 148 GY-rich proteins were identified in the human proteome (see STAR Methods), of which 43 were detected in the AP-MS data.

(legend continued on next page)

These observations indicate that the overlap between target AS events of hnRNPs is related to similarities in both their C-terminal IDR sequences and their RNA binding specificities. Knockdown of HNRNPD, HNRNPAB, HNRNPA1, and HNRNPA2B1, individually or in combination, followed by RT-PCR assays, confirmed AS changes that are unique to the different sub-clusters (Figure 6B). For example, alternative exons in PPIL2 and AFMID transcripts are regulated in an additive manner by HNRNPD and HNRNPAB proteins, whereas alternative exons in CAPN7, C11orf1, and KIF23 transcripts were specifically regulated by HNRNPA1 and HNRNPA2B1 (Figures 6B and S6B). We also validated AS events in ZCCHC10 and PTBP2 transcripts that are regulated following knockdown of any one of the four of the latter hnRNP proteins (Figures 6B and S6C). Thus, consistent with the roles of the GY-dependent IDRs of HNRNPD, HNRNPAB, HNRNPA1, and HNRNPA2B1 proteins in the control of heterotypic and homotypic multivalent hnRNP interactions, these proteins also have overlapping and additive functions in the regulation of target AS events.

To confirm whether the regulation of target AS events is dependent on the inclusion levels of the GY-rich mammalian-specific exons, we next compared the ability of different hnRNP splice isoforms to rescue splicing activity in the context of combined knockdowns of all isoforms of targeted hnRNPs. As a case study, AS levels of exon 10 of PTBP2 were analyzed in response to isoform rescue of different knockdown conditions. This exon displays reduced inclusion levels upon the combined knockdown of HNRNPA1 and HNRNPA2B1, or HNRNPD and HNRNPAB (Figures 6B, 6C, and S6C). Importantly, its inclusion can be rescued by expression of HNRNPD+Ex7 and HNRNPAB+Ex7, but not by HNRNPDΔEx7 or HNRNPABΔEx7 (Figures 6C and S6C). Moreover, this rescue activity is dependent on RNA binding, since mutation of critical phenylalanine residues to aspartic acid (F140D, F142D, F225D, and F227D) in the RRMs of HNRNPD+Ex7 prevents its ability to rescue splicing (Figure 6C and Figure S6D). In contrast, both isoforms of HNRNPA1 had strong enhancing effects on the splicing levels of PTBP2 exon 10 (Figure 6C), reflecting the reduced impact of skipping of exon 8 within the C-terminal IDR of this protein, as described above (Figures 2B and S5D). Yet, as expected from the analysis of HNRNPA1 isoform interactions, progressive substitution of remaining tyrosine residues to serine in the HNRNPA1ΔEx8 IDR resulted in a gradated loss of splicing-rescue capacity (Figures 6D and S5D). In summary, these data reveal semi-redundant and combinatorial functions of hnRNPs in the regulation of AS and, moreover, that the differential inclusion of the GY-rich mammalian-specific alternative exons of these proteins is critical for their splicing regulatory activities.

Regulation of Higher-Order hnRNP Assemblies on Pre-mRNA in the Control of Alternative Splicing

To investigate whether mammalian-specific exons regulate AS by controlling the formation of higher-order protein assemblies on pre-mRNA, we used electrophoretic mobility shift assays (EMSAs) to compare HNRNPD+Ex7 and HNRNPDΔEx7 complexes that form on radiolabeled RNA probes overlapping PTBP2 exon 10 (Figure 7A). We first analyzed binding to downstream intronic sequence, as RBP interactions with sequences in this location are often associated with the stimulation of exon inclusion (Witten and Ule, 2011). Addition of equivalent (Figure S7A) and increasing amounts of each HNRNPD isoform resulted in distinct gel shift patterns. HNRNPD+Ex7 forms higher molecular-weight complexes at lower protein amounts (Figure 7B, compare lane 2 with lanes 9 to 12), and some of these complexes migrate more slowly than those formed at the highest concentrations of HNRNPDΔEx7 (Figure 7B, complex 3). Consistent with these observations, HNRNPD+Ex7 protects more extensive regions of this RNA probe from antisense oligonucleotide-directed RNase H cleavage than does HNRNPDΔEx7 (Figure S7B, compare lanes 7, 15, and 16 with lanes 11, 19, and 20). The binding differences between HNRNPD+Ex7 and HNRNPDΔEx7 were even more pronounced with a longer RNA probe (Figure S7C), where HNRNPDΔEx7 complexes are readily resolved, whereas HNRNPD+Ex7 forms higher-order complexes that remain trapped in the well (Figure S7B).

To investigate whether exon 7 of HNRNPD can differentially regulate formation of higher-order protein assemblies on a transcriptome-wide level, we used PAR-CLIP data (Yoon et al., 2014) to compare the spatial distributions of binding sites of HNRNPD+Ex7 and HNRNPDΔEx7 on intronic sequence in 293 cells (Figure 7C). Remarkably, HNRNPDΔEx7 binding sites more often cluster within shorter sequence distances, whereas HNRNPD+Ex7 binding sites are more often distributed over longer intronic sequences (Figure 7C, $p < 4.31 \times 10^{-123}$; Anderson-Darling k-sample test). Thus, consistent with data in Figures 4 and 5 demonstrating an important role for GY-rich regions overlapping the mammalian-specific exon 7 of HNRNPD in the formation of multivalent hnRNP interactions, these results suggest that skipping of this exon also controls the formation of high-order complexes on pre-mRNA required for the regulation of AS.

A prediction from the results thus far is that hnRNP isoforms lacking IDR mammalian-specific exons may outcompete AS-promoting activities by binding RNA and preventing formation of higher-order protein assemblies. To test this model, we assayed the effects on PTBP2 exon 10 AS following overexpression of HNRNPA1+Ex8 with either HNRNPD+Ex7 or HNRNPDΔEx7 co-expressed. Strikingly, while co-expression of HNRNPD+Ex7 had no effect on the ability of HNRNPA1+Ex8 to promote PTBP2 exon 10 inclusion, overexpression of HNRNPDΔEx7 had a strong dominant-negative effect, preventing the stimulation of exon 10 splicing (Figure 7D). To assess whether this negative effect is due to HNRNPDΔEx7 competing for binding sites on pre-mRNA and blocking the recruitment of additional hnRNPs, we overexpressed HNRNPDΔEx7 harboring

(E) Co-immunoprecipitation western blot experiments analyzing interactions between FLAG-HNRNPD+Ex7 or FLAG-HNRNPDΔEx7 and HA-HNRNPAB+Ex7 or HA-HNRNPABΔEx7. Presence or absence of nuclease (RNaseA1 and benzonase) treatment is indicated.

(F) Same as in (E) but analyzing interactions between FLAG-HNRNPA1+Ex8 or FLAG-HNRNPA1ΔEx8 and HA-HNRNPA1+Ex8, HA-HNRNPA1ΔEx8, or an HA-HNRNPA1ΔEx8–YS mutant in which all tyrosines within C-terminal intrinsically disordered region (IDR) are substituted with serines (see Figure S5D).

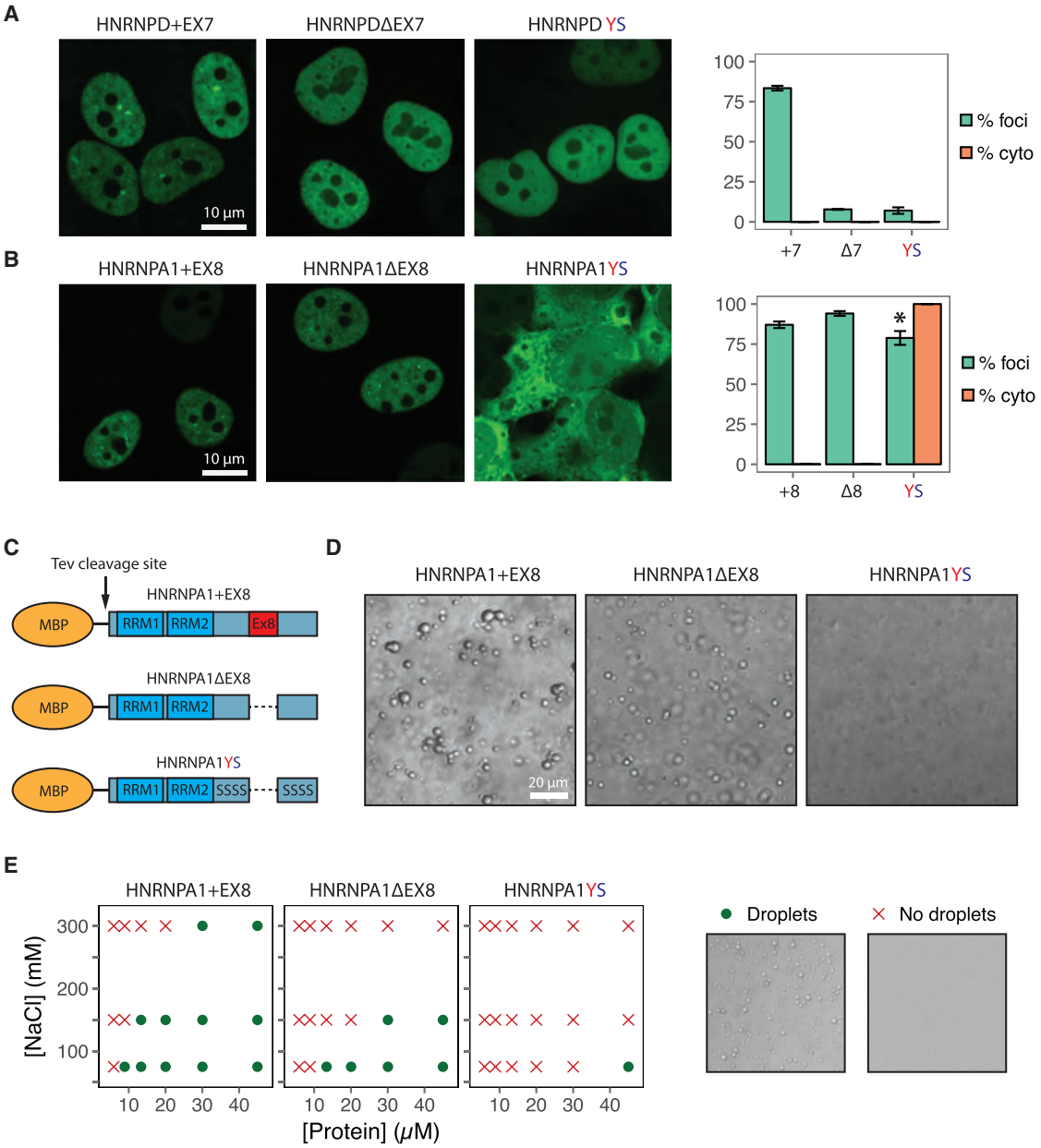


Figure 5. Dependence on IDR Mammalian-Specific Alternative Exons and Tyrosine Residues for *In Vitro* Phase-Separation and Sub-cellular Localization of hnRNPs

(A) Confocal fluorescence microscopy of HeLa cells showing localization of N-terminally GFP-tagged HNRNPD proteins, with or without IDR mammalian-specific alternative exons and substitution of tyrosine residues (see Figure S5E). Quantification of foci and cytoplasmic localization from three independent biological replicate experiments, with 50–100 cells scored per replicate. Error bars correspond to standard error.

(B) Same as (A) but for HNRNPA1 isoforms (see also Figure S5D). * indicates scoring of morphologically distinct (i.e., less spherical) nuclear foci formed by the mutant HNRNPA1 in which its tyrosines within C-terminal intrinsically disordered region (IDR) are substituted with serines.

(C) Domain diagrams of N-terminally tagged maltose binding protein (MBP)-HNRNPA1 isoforms used in phase separation assays in (D) and (E). RRM, RNA recognition motif.

(D) Purified recombinant MBP-HNRNPA1 isoforms were concentrated and cleaved from MBP, and phase separation was induced by addition of ficoll to a final concentration of 100 mg/mL (see STAR Methods). The final protein and NaCl concentrations in the reactions were 45 μM and 150 mM, respectively. Turbid HNRNPA1 solutions containing liquid-like protein droplets observed by differential interference contrast microscopy.

(E) Phase separation experiments performed as in (D) and quantified for the presence (green circles) or absence (red crosses) of protein droplets at varying NaCl (75, 150, or 300 mM) and protein (5.9, 8.9, 13.3, 20, 30, 45 μM) concentrations. Representative images are shown (right).

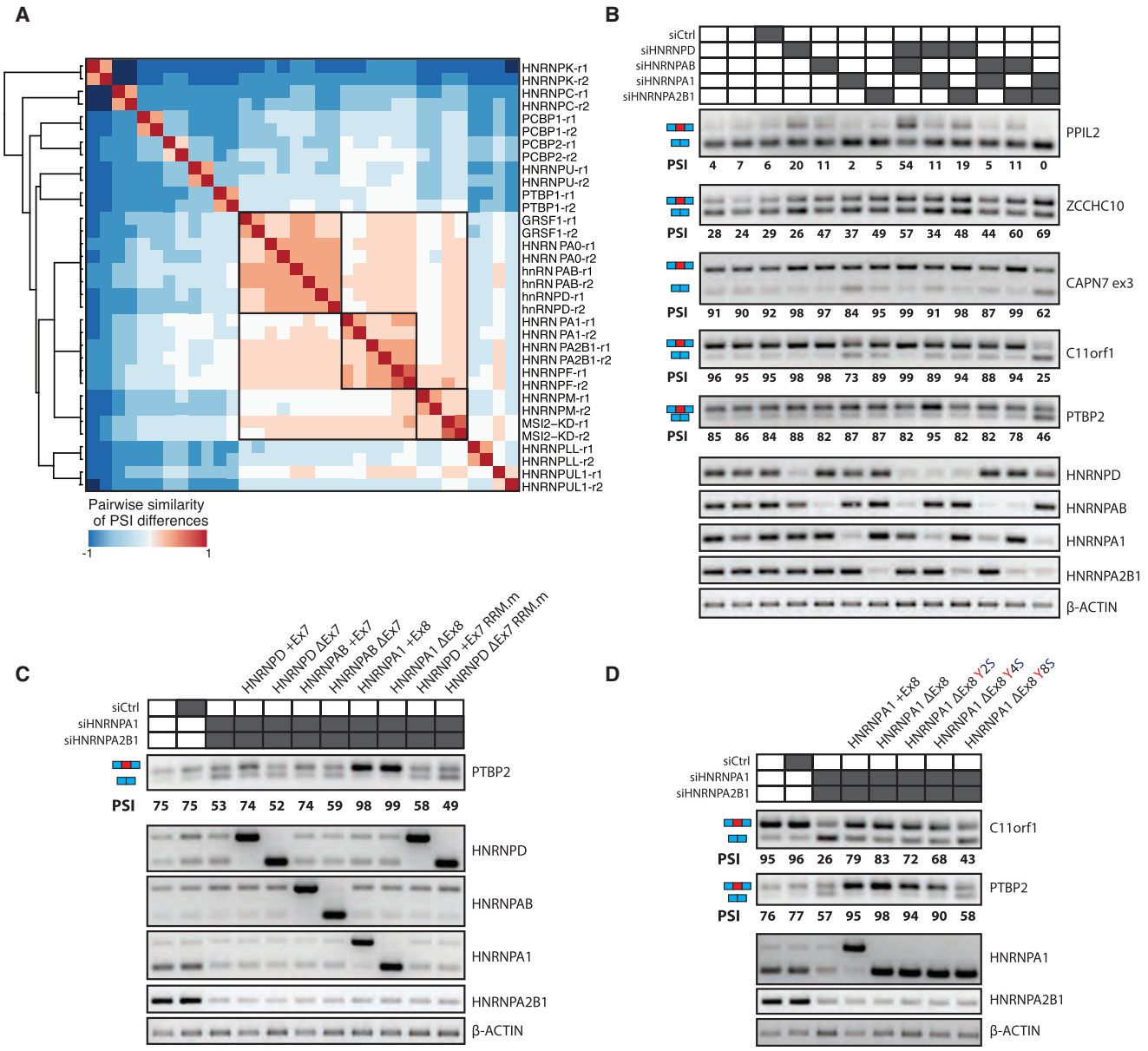


Figure 6. Mammalian-Specific Alternative Exon- and Tyrosine-Dependent Combinatorial Regulation of AS by hnRNPs

(A) Symmetrical heatmap of pairwise correlations of differential RNA-seq-detected AS change profiles as a consequence of knockdown of individual hnRNPs in HepG2 cells. Affinity propagation clustering of the pairwise similarities between the knockdown-dependent AS change profiles, indicated in the scale as correlation of percent spliced in (PSI) differences between knockdown and control conditions. Boxes highlight highly correlated knockdown change profiles. Red, strong correlation between the AS change profiles of two knockdowns; blue, poor or no correlation. r1, replicate 1; r2, replicate 2.

(B) RT-PCR assays analyzing hnRNP-dependent regulation of target alternative exons identified from the RNA-seq analysis in (A). hnRNPs were knocked down individually or in combination using siRNAs, as indicated by shaded boxes. β-ACTIN was used as a loading control. Percent spliced in (PSI) values are indicated (see STAR Methods).

(C) RT-PCR assays showing hnRNP-isoform-dependent regulation of PTBP2 exon 10. HNRNPA1 and HNRNPA2B1 were knocked down in combination. Rescue was performed with pairs of splice isoforms of HNRNPD, HNRNPAB, or a siRNA-resistant HNRNPA1 construct. Additional rescue assays were performed with HNRNPD splice isoforms harboring mutant, inactive RNA Recognition Motifs (RRM.m).

(D) RT-PCR assays showing that splicing regulation by HNRNPA1 is dependent on tyrosine residues in its C-terminal IDR. HNRNPA1 and HNRNPA2B1 were knocked down in combination. Rescue was performed with siRNA-resistant HNRNPA1 constructs as in (C).

the aforementioned RRM-disrupting mutations and found that this mutant lacks the ability to interfere with the splicing stimulatory activity of HNRNPA1+Ex8 (Figure 7D).

Finally, we asked whether competition between hnRNP splice isoforms may have contributed to the evolution of differential AS of hnRNP target exons in mammals. Accordingly, using RNA-seq

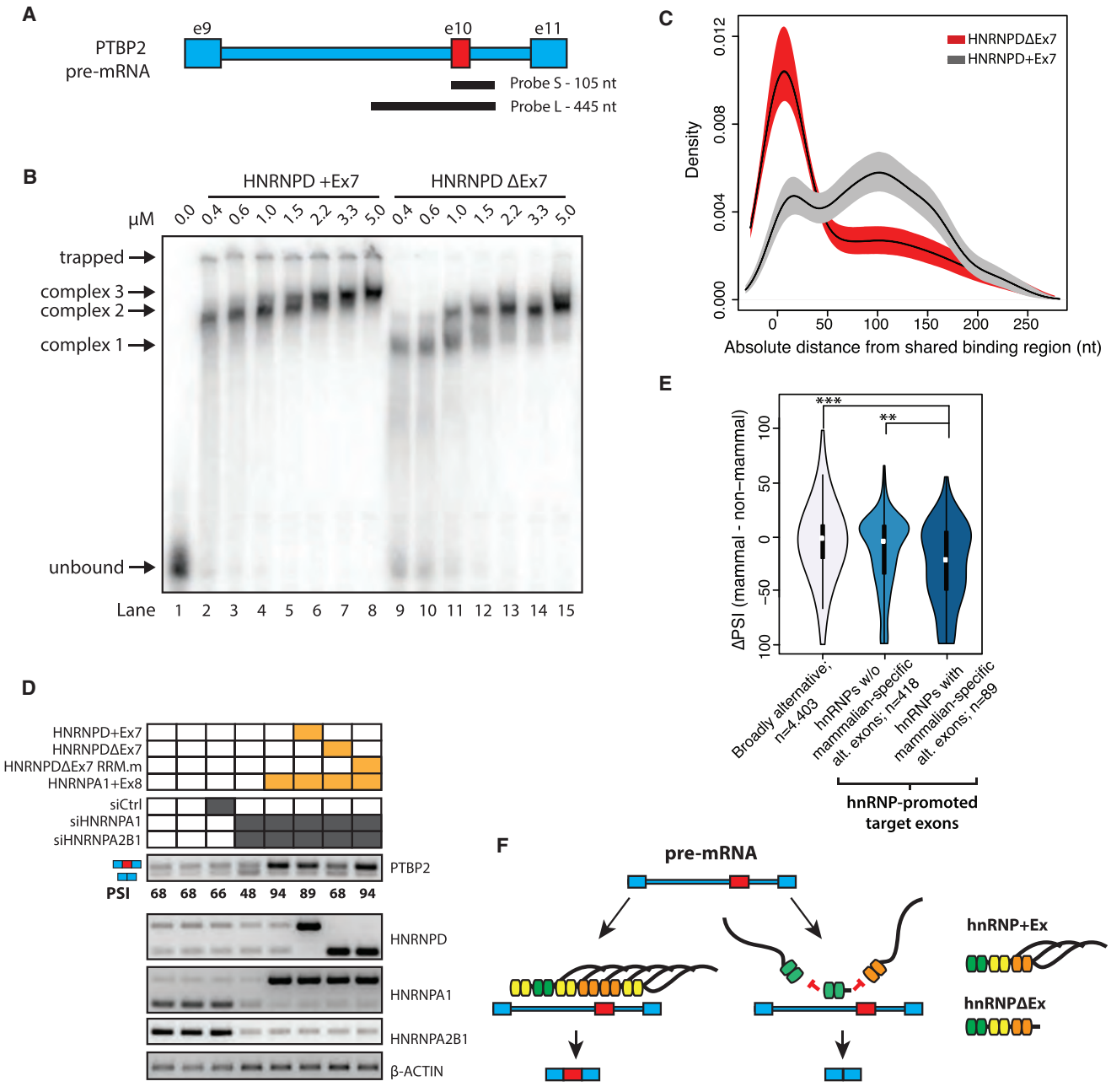


Figure 7. Higher-Order hnRNP Assemblies on Target Pre-mRNA

(A) Gene diagram of PTBP2 exons 9 to 11 and intervening introns. Probes (105 nt and 445 nt) encompassing alternative exon 10 used in electrophoretic mobility shift assays (EMSAs) are shown below.

(B) Electrophoretic mobility shift assay (EMSA) showing binding of recombinant MBP- and GST-tagged HNRNPD splice isoforms to the 105 nt probe. Protein concentrations in binding reactions are shown. Arrows indicate distinct protein-RNA complexes.

(C) Density plots of HNRNPD+Ex7 and HNRNPDΔEx7 binding sites identified by PAR-CLIP on transcripts that contain binding sites for both proteins (see STAR Methods), with 95% confidence intervals displayed in the shaded regions. Absolute distance from an overlapping binding site to adjacent sites within 250 nt windows are shown.

(D) RT-PCR assays showing dominant-negative effects of expression of HNRNPDΔEx7 on HNRNPA1-dependent rescue of PTBP2 exon 10. HNRNPA1 and HNRNPA2B1 were knocked down in combination (dark gray boxes). Rescues were performed with siRNA-resistant, HNRNPA1+Ex8 alone, or in combination with the indicated isoform of HNRNPD (golden boxes). β-ACTIN was used as a loading control.

(E) Violin plots showing percent spliced in (PSI) differences for exons that are conserved between mammals and non-mammals and representing three classes of events: exon targets whose splicing levels are promoted by hnRNPs with mammalian-specific AS events, as defined by the analysis of knockdown RNA-seq data (see Figure 6A), exon targets promoted by hnRNPs that lack mammalian-specific alternative exons (also defined from the data in Figure 6A), and a set of broadly

(legend continued on next page)

data (Table S1), we compared the splicing levels of orthologous exons across mammalian and non-mammalian tissues corresponding to the target exons of hnRNPs defined in human cells (Figure 6A). For this, we compared the levels of target exon inclusion of hnRNPs that possess mammalian-specific AS events versus those that do not. Remarkably, consistent with a global-scale, dominant-negative effect of hnRNP isoforms that exclude their mammalian-specific alternative exons, the target exons of these hnRNPs (where the targets are normally enhanced by the hnRNP) showed significantly decreased inclusion levels in mammalian tissues, as compared to the orthologous exons in the equivalent tissues in non-mammalian vertebrates and also as compared to target AS events enhanced by other hnRNPs that do not contain mammalian-specific alternative exons (Figures 7E and S7D; $p < 0.01$ Wilcoxon rank-sum test). Collectively, these data and the results described above provide evidence that mammalian-specific alternative exons located within the C-terminal, GY-rich IDRs of hnRNPs promote the formation of higher-order assemblies on target pre-mRNAs that function in the global regulation of AS.

DISCUSSION

In this study, we show that vertebrate-conserved exons that are specifically alternatively spliced in mammalian species are concentrated in the low-complexity, GY-rich IDRs of hnRNPs and other RBPs. These mammalian-specific alternative exons function in the regulation of multivalent hnRNP interactions that globally control AS in different cell and tissue types, at least in part by controlling the availability of IDR GY motifs. In particular, inclusion of the GY-rich exons appears to promote spreading of hnRNP interactions across introns by forging protein-protein interactions on RNA. The multivalency of the hnRNP protein-protein interactions, together with their distinct and overlapping RNA binding specificities, thus facilitates the coordinated and combinatorial regulation of AS. Furthermore, skipping of the mammalian-specific exons, and likely the differential phosphorylation of the overlapping GY repeats (Han et al., 2012), prevent the assembly of these multivalent complexes to control the splicing of target exons (Figure 7F). Collectively, these results provide new insight into the control and function of higher-order protein complex formation in gene regulation, as well as the evolutionary processes that have contributed to increased regulatory complexity in mammalian cells.

Previous studies have shown that the emergence of species- and lineage-specific AS events is associated with rapid change in linear *cis*-regulatory motifs that function as recognition sites for protein *trans*-acting factors (Barbosa-Morais et al., 2012; Lev-Maor et al., 2007; Merkin et al., 2012; Brooks et al., 2011; Jelen, et al., 2007). In the present study, we demonstrate the role of evolution of long-range, intra-molecular RNA duplexes that resulted in ancestral constitutive exons becoming skipped in diverse tissues in a mammalian lineage-specific manner. The multiple independent examples of this mechanism in related hnRNPs, and its conservation across mammalian species, strongly suggest that regulation of the number of GY motifs in the IDRs of hnRNPs confers a substantial fitness benefit. In addition to evolving to facilitate the dynamic regulation of multi-protein assemblies that have expanded gene regulatory complexity in mammals, it is also interesting to consider that this mechanism may have arisen to reduce the propensity of hnRNPs to form pathogenic aggregates. Consistent with this view, our data show that the mammalian-specific AS events of hnRNPs can influence the appearance and morphology of nuclear foci in cells and the propensity of hnRNPs to undergo phase separation to form liquid-like droplets *in vitro*. However, it is important to note that while the properties of hnRNPs that promote the formation of multivalent interactions required for the regulation of AS are similar to those required for the formation of foci and droplets, whether these microscopic structures have a direct role in splicing regulation is unclear, since they are not detected in all cell types in which hnRNPs are functionally active (data not shown) (Chen et al., 1999). It is also noteworthy that the evolution of exon skipping in hnRNP IDRs, in addition to expanding the splicing regulatory capacity of hnRNPs, likely arose to "liberate" these proteins from nuclei so as to allow their function in other processes. For example, skipping of the mammalian-specific alternative exon in HNRNPD defined in the present study facilitates the nuclear-cytoplasmic shuttling and regulation of translation and RNA decay by HNRNPD in the cytoplasm (White et al., 2013).

Previous work has shown that protein-protein interactions underlie various mechanisms of hnRNP-dependent AS control (Martinez-Contreras et al., 2007). For example, binding of an hnRNP protein to a high-affinity site followed by cooperative spreading on pre-mRNA has been shown to occlude binding of SR proteins to repress splicing (Okunola and Krainer, 2009; Zhu et al., 2001). Furthermore, interactions between distally bound hnRNPs can loop out intervening intronic or exonic sequences leading to exon inclusion or exclusion, respectively (Chen and Manley, 2009; Martinez-Contreras et al., 2006). The importance of higher-order complex formation on pre-mRNA is further supported by recent results showing that the tissue-specific splicing regulator Rbfox1 controls AS through the recruitment of LASR, a complex that contains a distinct set of hnRNPs from those studied here (Damianov et al., 2016). Our work demonstrates the functional importance of the regulation of GY

alternatively spliced exons, as defined in Figure 1A (refer to main text). *$p < 0.05$; **$p < 0.005$; ***$p < 1x10^{-5}$; all p values calculated using Wilcoxon-rank sum test. Violin plots show the distribution of the data (density plot), with the box inside the density plot representing the interquartile range, the vertical thin lines representing the 5–95 percentile range, and the horizontal line representing the median. w/o, without.

(F) Mechanistic model for the coordinated regulation of alternative exons by hnRNPs in the A and D families (represented by different colors). Homotypic and heterotypic cooperative binding of hnRNP isoforms containing GY-repeat-rich mammalian-specific alternative exons can promote target exon inclusion (and on other transcripts exon skipping). In contrast, hnRNP isoforms lacking these exons can bind pre-mRNA but lack the ability to recruit additional hnRNPs (as well as other RBPs containing GY repeats), thereby preventing the formation of higher-order assemblies required for exon inclusion. Differences in the expression levels of hnRNP isoforms with and without the mammalian-specific exons between cell/tissue types or conditions can thus globally regulate splicing patterns of target exons. Alternative and constitutive exons represented in red and blue, respectively.

repeats underlying the formation of multi-protein assemblies. Elimination of these repeats through exon skipping results in dominant-negative hnRNPs that bind RNA but that lack the capacity to form multimeric complexes and thereby impart distinct patterns of AS. In summary, we demonstrate that the recurring evolution of RNA-duplex-controlled AS in low-complexity GY-rich regions of hnRNPs, through its capacity to dynamically remodel high-order protein assemblies, has played an important role in diversifying the regulatory capabilities of mammalian cells.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - HeLa cells
  - Flp-In-293 cells
  - E. coli
- METHOD DETAILS
  - Molecular cloning
  - Generation of stable Flp-In-293 cell lines
  - Co-immunoprecipitation experiments
  - Immunoblotting
  - Protein expression and purification
  - *In vitro* phase separation assays
  - Immunofluorescence
  - RT-PCR assays
  - *In vitro* transcription
  - Electrophoretic mobility shift assays
  - RNase H protection assays
  - Mass spectrometry
  - Mass spectrometry data analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Sequence annotation
  - Detection and quantification of alternative splicing
  - Orthology definition
  - Definition of mammalian-specific alternative splicing events
  - Definition of other alternative splicing categories
  - Functional analysis
  - Phylogenetic Analysis
  - hnRNP phylogeny construction
  - Crosslinking immunoprecipitation RNA Sequencing (CLIP-seq) analysis
  - Peptide Analysis
  - Analysis of GY-rich protein enrichment
  - Protein feature analysis
  - Duplex analysis
  - RNA binding motifs and splice site strength calculations
  - Mammalian versus non-mammalian hnRNP-regulated alternative splicing events
  - Statistical Tests
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and seven tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2017.06.037.

AUTHOR CONTRIBUTIONS

Conceptualization, S.G., R.J.W, and B.J.B.; Formal Analysis, R.J.W., S.G., and A.-C.G.; Investigation, S.G., D.O., A.N., and Z.-Y.L; Resources, B.J.B. and A.-C.G.; Writing – Original Draft, S.G., R.J.W., and B.J.B.; Writing – Review & Editing, S.G., R.J.W., A.-C.G., and B.J.B.; Supervision, B.J.B., S.G., and A.-C.G.; Funding Acquisition, B.J.B. and A.-C.G.

ACKNOWLEDGMENTS

We gratefully acknowledge T. Sterne-Weiler for sharing Whippet software in advance of publication. We also thank D. Black for sharing unpublished results and M. Babu, E. Sharma, U. Braunschweig, B. Harpur, T. Gonatopoulos-Pournatzis, and K. Ha for helpful discussions and critical review of the manuscript. This work was supported by CIHR grants to A.-C.G. and B.J.B. S.G. was supported by NSERC and OGS scholarships. R.J.W. was supported by CIHR postdoctoral and Marie Curie IOF fellowships. B.J.B holds the University of Toronto Banbury Chair in Medical Research.

Received: December 5, 2016
Revised: February 24, 2017
Accepted: June 23, 2017
Published: July 13, 2017

REFERENCES

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Barbosa-Morais, N.L., Carmo-Fonseca, M., and Aparício, S. (2006). Systematic genome-wide annotation of spliceosomal proteins reveals differential gene family expansion. Genome Res. *16*, 66–77.

Barbosa-Morais, N.L., Irimia, M., Pan, Q., Xiong, H.Y., Gueroussov, S., Lee, L.J., Slobodeniuc, V., Kutter, C., Watt, S., Colak, R., et al. (2012). The evolutionary landscape of alternative splicing in vertebrate species. Science *338*, 1587–1593.

Blanchette, M., and Chabot, B. (1997). A highly stable duplex structure sequesters the 5′ splice site region of hnRNP A1 alternative exon 7B. RNA *3*, 405–419.

Bodenhofer, U., Kothmeier, A., and Hochreiter, S. (2011). APCluster: an R package for affinity propagation clustering. Bioinformatics *27*, 2463–2464.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene expression levels in mammalian organs. Nature *478*, 343–348.

Brooks, A.N., Yang, L., Duff, M.O., Hansen, K.D., Park, J.W., Dudoit, S., Brenner, S.E., and Graveley, B.R. (2011). Conservation of an RNA regulatory map between Drosophila and mammals. Genome Research *21*, 193–202.

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G.P., Fuxreiter, M., Bateman, A., and Babu, M.M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol. Cell *46*, 871–883.

Busch, A., and Hertel, K.J. (2012). Evolution of SR protein and hnRNP splicing regulatory factors. Wiley Interdiscip. Rev. RNA *3*, 1–12.

Chen, M., and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat. Rev. Mol. Cell Biol. *10*, 741–754.

Chen, T., Boisvert, F.M., Bazett-Jones, D.P., and Richard, S. (1999). A role for the GSG domain in localizing Sam68 to novel nuclear structures in cancer cell lines. Mol. Biol. Cell *10*, 3015–3033.

Chen, B., Yun, J., Kim, M.S., Mendell, J.T., and Xie, Y. (2014). PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. Genome Biol. *15*, R18.

Couzens, A.L., Knight, J.D., Kean, M.J., Teo, G., Weiss, A., Dunham, W.H., Lin, Z.Y., Bagshaw, R.D., Sicheri, F., Pawson, T., et al. (2013). Protein interaction network of the mammalian Hippo pathway reveals mechanisms of kinase-phosphatase interactions. Sci. Signal. *6*, rs15.

Damianov, A., Ying, Y., Lin, C.H., Lee, J.A., Tran, D., Vashisht, A.A., Bahrami-Samani, E., Xing, Y., Martin, K.C., Wohlschlegel, J.A., and Black, D.L. (2016). Rbfox proteins regulate splicing as part of a large multiprotein complex LASR. Cell *165*, 606–619.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433–3434.

Ellis, J.D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M., et al. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. Mol. Cell *46*, 884–892.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44*(D1), D279–D285.

Gracheva, E.O., Cordero-Morales, J.F., González-Carcacía, J.A., Ingolia, N.T., Manno, C., Aranguren, C.I., Weissman, J.S., and Julius, D. (2011). Ganglion-specific splicing of TRPV1 underlies infrared sensation in vampire bats. Nature *476*, 88–91.

Gueroussov, S., Gonatopoulos-Pournatzis, T., Irimia, M., Raj, B., Lin, Z.Y., Gingras, A.C., and Blencowe, B.J. (2015). An alternative splicing event amplifies evolutionary differences between vertebrates. Science *349*, 868–873.

Han, T.W., Kato, M., Xie, S., Wu, L.C., Mirzaei, H., Pei, J., Chen, M., Xie, Y., Allen, J., Xiao, G., and McKnight, S.L. (2012). Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. Cell *149*, 768–779.

Hedges, S.B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. Mol. Biol. Evol. *32*, 835–845.

Hornbeck, P.V., Chabra, I., Kornhauser, J.M., Skrzypek, E., and Zhang, B. (2004). PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics *4*, 1551–1561.

Irimia, M., Weatheritt, R.J., Ellis, J.D., Parikshak, N.N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallières, M., Tapial, J., Raj, B., O'Hanlon, D., et al. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell *159*, 1511–1523.

Jelen, N., Ule, J., Zivin, M., and Darnell, R.B. (2007). Evolution of Nova-dependent splicing regulation in the brain. PLoS Genetics *3*, 1838–1847.

Kato, M., Han, T.W., Xie, S., Shi, K., Du, X., Wu, L.C., Mirzaei, H., Goldsmith, E.J., Longgood, J., Pei, J., et al. (2012). Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. Cell *149*, 753–767.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. *33*, 1870–1874.

Lai, D., Proctor, J.R., Zhu, J.Y., and Meyer, I.M. (2012). R-CHIE: a web server and R package for visualizing RNA secondary structures. Nucleic Acids Res. *40*, e95.

Lambert, J.P., Ivosev, G., Couzens, A.L., Larsen, B., Taipale, M., Lin, Z.Y., Zhong, Q., Lindquist, S., Vidal, M., Aebersold, R., et al. (2013). Mapping differential interactomes by affinity purification coupled with data-independent mass spectrometry acquisition. Nat. Methods *10*, 1239–1245.

Lev-Maor, G., Goren, A., Sela, N., Kim, E., Keren, H., Doron-Faigenboim, A., Leibman-Barak, S., Pupko, T., and Ast, G. (2007). The "alternative" choice of constitutive exons throughout evolution. PLoS Genet. *3*, e203.

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Mannen, T., Yamashita, S., Tomita, K., Goshima, N., and Hirose, T. (2016). The Sam68 nuclear body is composed of two RNase-sensitive substructures joined by the adaptor HNRNPL. J. Cell Biol. *214*, 45–59.

Martinez-Contreras, R., Fisette, J.F., Nasim, F.U., Madden, R., Cordeau, M., and Chabot, B. (2006). Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. PLoS Biol. *4*, e21.

Martinez-Contreras, R., Cloutier, P., Shkreta, L., Fisette, J.F., Revil, T., and Chabot, B. (2007). hnRNP proteins and splicing control. Adv. Exp. Med. Biol. *623*, 123–147.

Merico, D., Isserlin, R., Stueker, O., Emili, A., and Bader, G.D. (2010). Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS ONE *5*, e13984.

Merkin, J., Russell, C., Chen, P., and Burge, C.B. (2012). Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science *338*, 1593–1599.

Nicolas, M., Rodríguez-Buey, M.L., Franco-Zorrilla, J.M., and Cubas, P. (2015). A recently evolved alternative splice site in the BRANCHED1a gene controls potato plant architecture. Curr. Biol. *25*, 1799–1809.

Okunola, H.L., and Krainer, A.R. (2009). Cooperative-binding and splicing-repressive properties of hnRNP A1. Mol. Cell. Biol. *29*, 5620–5631.

Pellegrini, M., Renda, M.E., and Vecchio, A. (2012). Ab initio detection of fuzzy amino acid tandem repeats in protein sequences. BMC Bioinformatics *13*(*Suppl 3*), S8.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. *8*, 2281–2308.

Ray, D., Kazan, H., Cook, K.B., Weirauch, M.T., Najafabadi, H.S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. Nature *499*, 172–177.

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., and Vilo, J. (2016). g:Profiler-a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res. *44*(W1), W83–W89.

Romero, P.R., Zaidi, S., Fang, Y.Y., Uversky, V.N., Radivojac, P., Oldfield, C.J., Cortese, M.S., Sickmeier, M., LeGall, T., Obradovic, Z., and Dunker, A.K. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. Proc. Natl. Acad. Sci. USA *103*, 8390–8395.

Roux, K.J., Kim, D.I., Raida, M., and Burke, B. (2012). A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. J. Cell Biol. *196*, 801–810.

Shteynberg, D., Deutsch, E.W., Lam, H., Eng, J.K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R.L., Aebersold, R., and Nesvizhskii, A.I. (2011). iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. Mol. Cell. Proteomics *10*, M111.007690.

Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., et al. (2016). ENCODE data at the ENCODE portal. Nucleic Acids Res. *44*(D1), D726–D732.

Sterne-Weiler, T., Weatheritt, R.J., Best, A., Ha, K.C.H., and Blencowe, B.J. (2017). Whippet: an efficient method for the detection and quantification of alternative splicing reveals extensive transcriptomic complexity. bioRxiv. http://dx.doi.org/10.1101/158519.

Sundararaman, B., Zhan, L., Blue, S.M., Stanton, R., Elkins, K., Olson, S., Wei, X., Van Nostrand, E.L., Pratt, G.A., Huelga, S.C., et al. (2016). Resources for the comprehensive discovery of functional rna elements. Mol. Cell *61*, 903–913.

Taylor, J.P., Brown, R.H., Jr., and Cleveland, D.W. (2016). Decoding ALS: from genes to mechanism. Nature *539*, 197–206.

Teo, G., Liu, G., Zhang, J., Nesvizhskii, A.I., Gingras, A.C., and Choi, H. (2014). SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. J. Proteomics *100*, 37–43.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. USA *99*, 6567–6572.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., et al. (2014). Classification of intrinsically disordered regions and proteins. Chem. Rev. *114*, 6589–6631.

Van Roey, K., Uyar, B., Weatheritt, R.J., Dinkel, H., Seiler, M., Budd, A., Gibson, T.J., and Davey, N.E. (2014). Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem. Rev. *114*, 6733–6778.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. *337*, 635–645.

Weber, S.C., and Brangwynne, C.P. (2012). Getting RNA and protein in phase. Cell *149*, 1188–1191.

White, E.J., Brewer, G., and Wilson, G.M. (2013). Post-transcriptional control of gene expression by AUF1: mechanisms, physiological targets, and regulation. Biochim. Biophys. Acta *1829*, 680–688.

Witten, J.T., and Ule, J. (2011). Understanding splicing regulation through RNA splicing maps. Trends Genet. *27*, 89–97.

Wootton, J.C., and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. Methods Enzymol. *266*, 554–571.

Wright, P.E., and Dyson, H.J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. Nat. Rev. Mol. Cell Biol. *16*, 18–29.

Wu, H., and Fuxreiter, M. (2016). The structure and dynamics of higher-order assemblies: amyloids, signalosomes, and granules. Cell *165*, 1055–1066.

Xing, Y., and Lee, C. (2006). Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. Nat. Rev. Genet. *7*, 499–509.

Xue, B., Dunker, A.K., and Uversky, V.N. (2012). Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J. Biomol. Struct. Dyn. *30*, 137–149.

Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. Nucleic Acids Res. *44*(D1), D710–D716.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394.

Yoon, J.H., De, S., Srikantan, S., Abdelmohsen, K., Grammatikakis, I., Kim, J., Kim, K.M., Noh, J.H., White, E.J., Martindale, J.L., et al. (2014). PAR-CLIP analysis uncovers AUF1 impact on target RNA fate and genome integrity. Nat. Commun. *5*, 5248.

Zhu, J., Mayeda, A., and Krainer, A.R. (2001). Exon identity established through differential antagonism between exonic splicing silencer-bound hnRNP A1 and enhancer-bound SR proteins. Mol. Cell *8*, 1351–1361.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse monoclonal anti-FLAG M2	Sigma-Aldrich	F3165; RRID: AB_259529
Rat monoclonal anti-HA 3F10	Sigma-Aldrich	11867423001; RRID: AB_390918
Rabbit polyclonal anti-hnRNPD	Abcam	Ab50692; RRID: AB_880476
Mouse anti-tubulin	Sigma-Aldrich	T6199; RRID: AB_477583
Chemicals, Peptides, and Recombinant Proteins		
Horseradish peroxidase streptavidin	Vector laboratories	SA-5014
MBP-HNRNPA1+Ex8	This paper	N/A
MBP-HNRNPA1ΔEx8	This paper	N/A
MBP-HNRNPA1ΔEx8 YS	This paper	N/A
GST-HNRNPD+Ex7-MBP	This paper	N/A
GST-HNRNPDΔEx7-MBP	This paper	N/A
Ficoll PM 400	Sigma-Aldrich	F4375
Critical Commercial Assays		
OneStep RT-PCR Kit	QIAGEN	Cat#: 210210
Deposited Data		
Affinity-purification mass spectrometry (AP-MS) data	This paper	ProteomeXchange: PXD005476
Proximity ligation proteomics (BioID) data	This paper	ProteomeXchange: PXD005475
RNA-seq datasets used to identify mammalian-specific AS events; see Table S1	Gene Expression Omnibus - NCBI	N/A
siRNA knockdown RNA-seq datasets used to identify hnRNP-regulated AS events; see Table S1	(Sundararaman et al., 2016)	https://www.encodeproject.org/
Ensembl Human Reference Genome: GrCh37	Ensembl	ensembl.org
NCBI human and adenovirus RefSeq databases v57	REFSeq	https://www.ncbi.nlm.nih.gov/refseq
hnRNP RNA binding motifs used in Figure S6A	(Ray et al., 2013)	http://cisbp-rna.ccbr.utoronto.ca/
hnRNP RNA binding motifs used in Figure S6A	(Sundararaman et al., 2016)	https://www.encodeproject.org/
RNA-seq datasets used to define tissue-specific AS events in Figures 1 and 7; see Table S1	Gene Expression Omnibus - NCBI	N/A
CLIP-seq datasets used in Figure 7C; see Table S1	(Yoon et al., 2014)	SRP033497
Experimental Models: Cell Lines		
Human: HeLa cells	ATCC	N/A
Human: Flp-In-293 Cell Line	Invitrogen	R75007
Flp-In-293 Cell Line with stable integration of pcDNA 5/FRT/TO – N-3xFLAG HNRNPD+Ex7	This paper	N/A
Flp-In-293 Cell Line with stable integration of pcDNA 5/FRT/TO – N-3xFLAG HNRNPDΔEx7	This paper	N/A
Flp-In-293 Cell Line with stable integration of pcDNA 5/FRT/TO – N-BirA-3xFLAG HNRNPD+Ex7	This paper	N/A
Flp-In-293 Cell Line with stable integration of pcDNA 5/FRT/TO – N-BirA-3xFLAG HNRNPDΔEx7	This paper	N/A
Flp-In-293 Cell Line with stable integration of pcDNA 5/FRT/TO – N-3xFLAG PTBP1+Ex9	This paper	N/A
Flp-In-293 Cell Line with stable integration of pcDNA 5/FRT/TO – N-3xFLAG PTBP1ΔEx9	This paper	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Flp-In-293 Cell Line with stable integration of pcDNA 5/FRT/TO – N-BirA-3xFLAG PTBP1+Ex9	This paper	N/A
Flp-In-293 Cell Line with stable integration of pcDNA 5/FRT/TO – N-BirA-3xFLAG PTBP1ΔEx9	This paper	N/A
Experimental Models: Organisms/Strains		
Escherichia coli OverExpress C41(DE3)	Lucigen	60442-1
Oligonucleotides		
Cas9 guide RNA upstream 1: GTAGCAACTGTTGAGT CGTC-AGG	This paper	N/A
Cas9 guide RNA upstream 2: TCCAAGAACTGGTCTA ACCA-AGG	This paper	N/A
Cas9 guide RNA downstream 1: ATTGGGGCCCTTTT CTTAGA-AGG	This paper	N/A
Cas9 guide RNA downstream 2: GAGATACTAAGCA CTGATTG-TGG	This paper	N/A
EMSA probe S: ATGGTTACGCCCCAAAGTCTGTT TACCCTCTTCGgtatgttattgttagcactatacttttattattgat ttgatttttgtttcaccttaattcttatttgtagc	This paper	N/A
EMSA probe L: ctgcattgctgtttcccttccccatttcatccttttccctgtgtgttcaccttcccttttccttgtctttttcccaaatgcccattccttccctgtcttatcctttatttttccttgtccttgtcttcattccctgtctccattccctatgttcatgcttctgtgcttgaacaaatgttcctcggaccaacttgccccaattaaccgccttgaaccatgatccatgaccacctcaccattctgcgggaaccacccttcgttatggatgatctgttcatctccgctcttcctcgactcttctctcttcttgtcttacgctgcttgctcttctcctcttctaaagATGGTTACGCCCCAAAGTCTGTTTACCCTCTTCGgtatgttattgttagcactatacttttattattgatttgatttttgtttcaccttaattcttatttgtagc	This paper	N/A
Primers for Figures 2, 3, 6, and 7, see Table S7	This paper	N/A
Recombinant DNA		
Splicing minigene sequences in pET01 exon trap vector: see Table S3	pET01 vector: MoBiTec; custom sequences: this paper	Backbone vector order#: PET01
pSpCas9(BB)-2A-Puro (PX459) vector	(Ran et al., 2013)	Addgene: #48139
pcDNA 5/FRT/TO – N-3xFLAG HNRNPD+Ex7	This paper	N/A
pcDNA 5/FRT/TO – N-3xFLAG HNRNPDΔEx7	This paper	N/A
pcDNA 5/FRT/TO – N-3xFLAG HNRNPD+Ex7 RRM mutant	This paper	N/A
pcDNA 5/FRT/TO – N-3xFLAG HNRNPDΔEx7 RRM mutant	This paper	N/A
pcDNA 5/FRT/TO – N-3xFLAG PTBP1+Ex9	This paper	N/A
pcDNA 5/FRT/TO – N-3xFLAG PTBP1ΔEx9	This paper	N/A
pcDNA 5/FRT/TO – N-3xFLAG HNRNPAB+Ex7	This paper	N/A
pcDNA 5/FRT/TO – N-3xFLAG HNRNPABΔEx7	This paper	N/A
pcDNA 5/FRT/TO – N-3xFLAG HNRNPA1+Ex8	This paper	N/A
pcDNA 5/FRT/TO – N-3xFLAG HNRNPA1ΔEx8	This paper	N/A
pcDNA 5/FRT/TO – N-3xFLAG HNRNPA1ΔEx8 Y2S mutant	This paper	N/A
pcDNA 5/FRT/TO – N-3xFLAG HNRNPA1ΔEx8 Y4S mutant	This paper	N/A
pcDNA 5/FRT/TO – N-3xFLAG HNRNPA1ΔEx8 Y8S mutant	This paper	N/A
pcDNA 5/FRT/TO – N-BirA-3xFLAG HNRNPD+Ex7	This paper	N/A
pcDNA 5/FRT/TO – N-BirA-3xFLAG HNRNPDΔEx7	This paper	N/A
pcDNA 5/FRT/TO – N-BirA-3xFLAG PTBP1+Ex9	This paper	N/A
pcDNA 5/FRT/TO – N-BirA-3xFLAG PTBP1ΔEx9	This paper	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
pcDNA 5/FRT/TO – N-HA HNRNPD+Ex7	This paper	N/A
pcDNA 5/FRT/TO – N-HA HNRNPDΔEx7	This paper	N/A
pcDNA 5/FRT/TO – N-HA HNRNPAB+Ex7	This paper	N/A
pcDNA 5/FRT/TO – N-HA HNRNPABΔEx7	This paper	N/A
pcDNA 5/FRT/TO – N-HA HNRNPA1+Ex8	This paper	N/A
pcDNA 5/FRT/TO – N-HA HNRNPA1ΔEx8	This paper	N/A
pcDNA 5/FRT/TO – N-HA HNRNPA1ΔEx8 Y8S mutant	This paper	N/A
pcDNA 5/FRT/TO – N-eGFP HNRNPD+Ex7	This paper	N/A
pcDNA 5/FRT/TO – N-eGFP HNRNPDΔEx7	This paper	N/A
pcDNA 5/FRT/TO – N-eGFP HNRNPD+Ex7 YS mutant	This paper	N/A
pcDNA 5/FRT/TO – N-eGFP HNRNPA1+Ex8	This paper	N/A
pcDNA 5/FRT/TO – N-eGFP HNRNPA1ΔEx8	This paper	N/A
pcDNA 5/FRT/TO – N-eGFP HNRNPA1ΔEx8 Y8S mutant	This paper	N/A
pMAL-c2X TEV N-MBP HNRNPA1+Ex8	This paper	N/A
pMAL-c2X TEV N-MBP HNRNPA1ΔEx8	This paper	N/A
pMAL-c2X TEV N-MBP HNRNPA1ΔEx8 Y8S mutant	This paper	N/A
pGEX4T1 N-GST C-MBP HNRNPD+Ex7	This paper	N/A
pGEX4T1 N-GST C-MBP HNRNPDΔEx7	This paper	N/A
Software and Algorithms		
VAST-TOOLS V1.0	(Irimia et al., 2014)	https://github.com/vastgroup/vast-tools
ProHits v4	Anne-Claude Gingras' lab	http://prohitsms.com/Prohits_download/list.php
Trans-Proteomic Pipeline v4.7	Institute for Systems Biology, Seattle	tools.proteomecenter.org/wiki/index.php?title = Software:TPP
SAINTexpress v3.6.1	Hyungwon Choi's lab	http://saint-apms.sourceforge.net/
ProteoWizard v3.0.4468	ProteoWizard Software Foundation	proteowizard.sourceforge.net
AB SCIEX MS Data Converter v1.3 beta	AB SCIEX	See AB SCIEX
Mascot v2.3.02	Matrix Science	http://www.matrixscience.com/
Comet v2014.02 rev.2	Mike MacCoss lab	comet-ms.sourceforge.net/
Whippet v0.5	Tim Sterne-Weiler	https://github.com/timbitz/Whippet.jl
Pamr v1.55	(Tibshirani et al., 2002)	cran.r-project.org/web/packages/pamr
Diff (differential splicing analysis) v1.0	Tim Sterne-Weiler	https://github.com/vastgroup/diffR
G:Profiler. v.0.3.4	(Reimand et al., 2016)	http://biit.cs.ut.ee/gprofiler/
Cytoscape v3.3.0	Cytoscape	cytoscape.org
Enrichment Map	(Merico et al., 2010)	baderlab.org/Software/EnrichmentMap
TimeTree v4.0	(Hedges et al., 2015)	timetree.org
MEGA7	(Kumar et al., 2016)	MEGA7
STAR RNA-seq aligner v020201	(Dobin et al., 2013)	https://github.com/alexdobin/STAR
PIPE-CLIP v1.1.0	(Chen et al., 2014)	https://github.com/QBRC/PIPE-CLIP
DESeq2 v1.14.1	(Love et al., 2014)	bioconductor.org/packages/release/bioc/html/DESeq2.html
APCluster v1.4.3	(Bodenhofer et al., 2011)	http://cran.r-project.org/web/packages/apcluster
IUPred	(Dosztányi et al., 2005)	iupred.enzim.hu
SEG	(Wootton and Federhen, 1996)	SEG
Pfam v29.0-30	(Finn et al., 2016)	V29.0-30
PTRStalker	(Pellegrini et al., 2012)	http://bioalgo.iit.cnr.it/index.php?pg=ptrs

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Phosphosite	(Hornbeck et al., 2004)	http://www.phosphosite.org/
R-CHIE	(Lai et al., 2012)	e-rna.org/r-chie/
ViennaRNA package v2.0	(Lorenz et al., 2011)	https://www.tbi.univie.ac.at/RNA/
WebLogo V2.8.2	WebLogo	weblogo.berkeley.edu/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for reagents may be directed to, and will be fulfilled by the lead author Benjamin Blencowe (b.blencowe@utoronto.ca).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

HeLa cells

Originally derived from female donor. HeLa cells were obtained from American Type Culture Collection (ATCC) and were maintained at 37°C and 5% CO_2 in DMEM supplemented with 10% fetal bovine serum.

Flp-In-293 cells

Parental HEK293 cells originally derived from female fetus. Flp-In-293 cells were obtained from ThermoFisher and were maintained at 37°C and 5% CO_2 in DMEM supplemented with 10% fetal bovine serum. Parental cell line with no integrated trans-genes was maintained with 3μg/mL Blasticidin S and 100 μg/mL Zeocin.

E. coli

A BL21(DE3)-derivative strain, OverExpress C41, was obtained from Lucigen and was grown in standard LB media supplemented with 0.2% weight by volume D-glucose.

Subcloning Efficiency DH5α Competent Cells obtained from ThermoFisher were used for all cloning performed in this study. Cells were grown in standard LB media.

METHOD DETAILS

Molecular cloning

Splicing mini-gene vectors

A series of constructs were generated to assess regulation of human and chicken HNRNPD exon 7 alternative splicing using the Exontrap Cloning Vector pET01 base vector (http://www.mobitec.com/cms/products/bio/04_vector_sys/exontrap.html). The general strategy was to introduce the genomic region including the alternative exon, its flanking introns, the up and downstream constitutive exons, and an additional up and downstream 300 intronic nucleotides (total sequence of ~2200 nucleotides) between the 5′ and 3′ exons of the base vector. The sequence was PCR amplified from human and chicken genomic DNA and was cloned into the vector using 5′ SalI and 3′ BamHI restriction sites. The specific sequences flanked by the SalI and BamHI restriction sites that were introduced into the 24 constructs described in this study are provided in Table S3. Note that while human sequence corresponds perfectly to the human reference genome, hg19, the genomic chicken sequence amplified from DT40 genomic DNA differed slightly from the reference galGal4 genome at several intronic loci. The DT40 sequence in this region is listed under the "Gga" construct in Table S3.

Cas9 gRNA vectors

Guide RNA (gRNA) sequences of 20 nucleotides that flank the duplex-forming element downstream of HNRNPD exon 7 were designed to have minimum off-target effects using the online CRISPR design tool provided by the Zhang laboratory (Ran et al. 2013). Two independent upstream and downstream guide sequences were selected. They are listed with their associated NGG sequence:

Upstream #1 GTAGCAACTGTTGAGTCGTC AGG
Upstream #2 TCCAAGAACTGGTCTAACCA AGG
Downstream #1 ATTGGGGCCCTTTTCTTAGA AGG
Downstream #2 GAGATACTAAGCACTGATTG TGG

The guide sequences were cloned individually into the pSpCas9(BB)-2A-Puro (PX459) vector (Addgene #48139) by annealing forward and reverse sequences with appropriate restriction enzyme overhangs.

hnRNP expression vectors

HNRNPD, HNRNPAB, HNRNPA1, and PTBP1 genes were cloned for mammalian expression in this study. HNRNPD+Ex7 (BC002401), HNRNPDΔEx7 (BC023977), PTBP1+Ex9 (BC002397), and PTBP1ΔEx9 (BC004383) were obtained from a Mammalian Gene Collection depository. HNRNPAB+Ex7, HNRNPABΔEx7, HNRNPA1+Ex8, and HNRNPA1ΔEx8 were cloned from human complementary DNA. All genes were amplified with Gateway (ThermoFisher) compatible overhangs to include the stop codon and omit the starting ATG codon. Amplicons were cloned into the pDONR223 Gateway compatible entry vector. Mutagenesis of RNA recognition motifs and hnRNP tyrosine residues to serine was performed using the Q5 mutagenesis kit (NEB). Briefly, mutagenesis primers were used to amplify the plasmid of interest, leading to incorporation of the desired mutations. The amplified linear product was separated by agarose gel electrophoresis and purified using a gel extraction kit (ThermoFisher). 1 μL of purified DNA was used as input for the Kinase, Ligase & DpnI (KLD) reaction containing 2.5 μL of 2X KLD Reaction Buffer, 0.5 μL of 10X KLD Enzyme Mix and 1.0 μL of nuclease-free water. The reaction was allowed to proceed for 20 min at room temperature, after which the reactions were cooled on ice and transformed into DH5α competent cells. Entry clones were then cloned into pcDNA5/FRT/TO-derived gateway destination vectors containing either 3xFLAG, HA, BirA-3xFLAG, or GFP amino-terminal tags. Vectors were used both for generating stable Flp-In-293 cell lines and for transient transfections.

hnRNP bacterial expression vectors

Constructs for recombinant proteins that were used for *in vitro* liquid-liquid phase separation assays were generated by cloning the indicated hnRNPs into the pMAL-c2X vector containing an N-terminal maltose binding protein (MBP) tag. The factor Xa protease site in this vector was substituted for a tobacco etch virus (TEV) protease site, with the amino acid sequence ENLYFQS, to avoid predicted off-target cleavage.

Constructs for recombinant proteins that were used in electrophoretic mobility shift assays were generated by cloning HNRNPD+Ex7 and HNRNPDΔEx7 into the pGEX4T1 vector (Sigma-Aldrich) containing an N-terminal glutathione S-transferases (GST) tag and C-terminal MBP tag.

Generation of stable Flp-In-293 cell lines

Doxycycline-inducible stable Flp-in T-REx 293 cells were generated by transfecting 200 ng of pcDNA5/FRT/TO-based plasmid with 2 μg of plasmid encoding pOG44 recombinase. Cell lines with stably integrated genes were selected and maintained with 3 μg/mL Blasticidin S and 200 μg/mL Hygromycin B. Transgene expression was induced by addition of 1 μg/mL Doxycycline.

Co-immunoprecipitation experiments

3xFLAG and HA tagged constructs were transiently transfected into Flp-In-293 cells grown in 6-well plate format. After 48 hr, cells were harvested in cold phosphate buffered saline (PBS) and pellets were flash-frozen in liquid nitrogen. Frozen pellets were resuspended in 300 μL of lysis buffer (50 mM HEPES-KOH pH 8.0, 100 mM KCl, 0.5 mM EDTA, 10% glycerol, 0.1% NP-40, 1mM DTT, and 1mM PMSF). Cells were gently lysed with two freeze/thaw cycles on dry ice (5 min on ice followed by ~30 s in a 37°C water bath). For harsh lysis, lysates were additionally subject to sonication (20 1 s pulses with 1 s in between at 30% power). For nuclease digestion 15 μg RNaseA1 and 75 Units of benzonase were added and lysates were incubated at 37°C with shaking for 10 min. Lysates were cleared in a microcentrifuge by spinning at max speed for 20 min at 4°C. Anti-flag immunoprecipitation was performed using magnetic Dynabeads protein G (ThermoFisher) complexed with anti-Flag M2 antibody (Sigma-Aldrich). For a pellet corresponding to one well in a 6-well plate, 10 μL of bead slurry was combined with 2 μg of antibody per IP. Prior to binding, the slurry was washed three times with 500 μL of PBS. Antibody was bound in bulk, by re-suspending the washed slurry with the desired amount of antibody in a final volume of 500 μL of PBS. Antibody was bound for 1 hr at 4°C. The bound complex was washed twice with PBS, then once with lysis buffer, followed by incubation with cleared lysate for 2 hr at 4°C with rotation. Following the incubation step, the complexes were washed 2 times with 500 μL PBS, transferred to new tubes and washed once more. Elution was performed by boiling in sample buffer at 95°C for 5 min.

Immunoblotting

Cell lysates and co-immunoprecipitation samples were heated with NuPage LDS Sample Buffer (Life Technologies) and NuPage Sample Reducing Agent (Life Technologies) at 95°C for 5 min, separated on variable percent SDS-PAGE gels, and transferred to PVDF membranes. Blots were incubated overnight with the following primary antibodies at the specified dilutions:

- Mouse anti-Flag M2 (Sigma-Aldrich) at 1:10,000
- Mouse anti-tubulin (Sigma-Aldrich) at 1:10,000
- Rat anti-HA (Roche) at 1:10,000
- Rabbit anti-HNRNPD (Abcam ab50692) at 1:500
- Streptavidin-HRP (Vector laboratories SA-5014) at 1:1000

Protein expression and purification

C41 *E. coli* (Lucigen) bacteria were transformed with pMAL-c2X or pGEX4T1-derived constructs. Cultures were expanded to 1.0 L of LB media with 0.2% D-glucose and grown to an OD600 of 0.5-0.6, at which point they were induced with 0.8 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). They were grown for 2 hr at 37 degrees, spun down and the pellets were flash frozen in liquid nitrogen. Pellets were lysed in 50 mM HEPES-NaOH pH 7.5, 1.0 M NaCl, 1 mM EDTA, 10 mM 2-Mercaptoethanol, and 1 mM Phenylmethylsulfonyl fluoride (PMSF) to raise the total volume to 32 mL. 280 μL of 50 mg/mL lysozyme was added to the resuspended pellet followed by a 20 min incubation on ice. Cells were lysed by sonication after which lysates were clarified by centrifugation at 15,000 rpm for 20 min at 4°C. Purification was performed using 4.5 mL of amylose slurry for 1 L of bacterial culture. Cleared lysate was mixed with washed amylose resin and binding was allowed to proceed for 2 hr at 4°C with rotation. Resin was then washed with lysis buffer, followed by subsequent washes with reduced-salt lysis buffer (0.5 M and 0.25 M NaCl). Proteins were eluted with 10 mM maltose in 15 mL of elution buffer consisting of 20 mM Tris-HCl (pH 7.5), 150 mM NaCl, and 1 mM DTT. Concentration of eluted protein was determined by running a Coomassie gel with a bovine serum albumin (BSA) standard curve.

In vitro phase separation assays

Purified protein was concentrated to the indicated concentration using centrifugal protein concentrators with a 10,000 molecular weight cutoff. Following concentration, proteins were digested with TEV protease at room temperature for roughly 2 hr. To induce droplet formation, proteins were diluted to various concentrations and mixed with equal volume of 200 mg/mL Ficoll PM 400 with varying amounts of NaCl for a final Ficoll concentration of 100 mg/mL with variable salt concentration. At several conditions, solutions were observed to become turbid. Solutions were transferred to a 96-well microscopy plate and occurrence of phase separation was determined by the presence or absence of visible liquid droplets in the solution.

Immunofluorescence

HeLa cells were transfected with the indicated hnRNP constructs N-terminally tagged with eGFP for 48 hr on size 1.5 microscopy slides from Zeiss. Cells were fixed, permeabilized and DNA was labeled with Hoechst stain to indicate nuclei. Imaging was performed on a Zeiss spinning disc confocal microscope. Image capture and analysis was performed using ZEN software (Zeiss). Between 50 and 100 cells were quantified for each of three biological replicates and cells were scored for the presence of discernable subnuclear foci as well as whether protein localization was restricted to the nucleus. Foci quantification was performed manually, by visualizing multiple Z stacks for each field of view, as foci could often be observed in different focal planes. A Z stack interval of 0.4 μm was used during image capture.

RT-PCR assays

To assess inclusion of an alternative exon, forward and reverse primers were designed to anneal to the constitutively included exons upstream and downstream of the alternative exon, respectively. RNA from cell lines was extracted using the QIAGEN RNeasy Mini Kit. In brief, cells were lysed directly in the growth plates following washing with PBS. 350 μL of Buffer RLT was added, and the lysate was transferred to a QIAshredder spin column for lysis by centrifugation. An equal volume of 70% ethanol was added to the lysed cells and the 700 μL combined mixture was bound to an RNeasy spin column by centrifugation. Bound RNA was washed with 350 μL of Buffer RW1, treated with 80 μL DNaseI reaction mix (10 μL DNaseI stock with 70 μL Buffer RDD per tube; QIAGEN) on the membrane and washed again with 350 μL of Buffer RW1. Two additional washes with 500 μL of Buffer RPE were performed and the RNA was eluted with 50 μL of nuclease-free water into a fresh micro-centrifuge tube. Splicing assays were carried out using the QIAGEN OneStep RT-PCR Kit, using modified reaction conditions. Each 10 μL splicing reaction contained 4.8 μL of nuclease-free water, 2 μL of 5x reaction buffer, 1.2 μL of 5μM forward and reverse primer mix, 0.4 μL of 10 mM dNTP mix, μ 0.5 μL of enzyme mix, 0.1 μL Ribolock RNase inhibitor, and 10 ng of the appropriate RNA sample in 1.0 μL of nuclease-free water. Cycling was performed as follows, Reverse transcription: 30 min 50°C; Initial PCR activation step: 15 min 95°C; 3-step cycling: 1. Denaturation: 1 min 94°C, 2. Annealing: 1 min 58-60°C, 3. Extension: 1 min 72°C; Final extension: 10 min 72°C. The number of amplification cycles ranged from 21-34 depending on the transcripts analyzed. Reaction products were separated on 1%–2.5% agarose gels for imaging.

Percent spliced in (PSI) values were calculated using ImageJ software. First the exon-included and exon-excluded band intensities were corrected by subtracting background. Then, intensity of the exon-included band was divided by the sum of the exon-included and exon-excluded bands. The result was multiplied by 100% to obtain PSI value, which was rounded to the nearest whole integer.

In vitro transcription

In vitro transcribed RNAs were synthesized from double-stranded linear DNA templates amplified from human genomic DNA. For template design, a T7 promoter was added by incorporating the promoter sequence along with flanking sequence to the 5′ end of the forward primer. Full sequence added: 5′-GAAAT(TAATACGACTCACTATAG)GGAGA-template specific sequence, with the minimum T7 promoter in parentheses, the upstream sequence designed to aid T7 binding and the downstream nucleotides to aid transcription efficiency. The full primer sequences, including gene-specific sequence used for probe amplification were:

Probe L-Fwd GAAATTAATACGACTCACTATAGGGAGActgcattgctgtttcccttc
Probe L-Rev gctacaaataagaattaaggtgaaaca

Probe S-Fwd GAAATTAATACGACTCACTATAGGGAGAatggttacgccccaaagt
Probe S-Rev gctacaaataagaattaaggtgaaaca (same as Probe-L)

The amplified DNA fragments were purified on a 1% agarose gel for use in the *in vitro* transcription reaction. Radiolabeling *in vitro* transcription was performed using the MEGAshortscript T7 Transcription Kit (Life technologies) with modified reaction conditions. For a 20 μL reaction, 2 μL of 10x reaction buffer, 2 μL of 100 mM DTT, 2 μL of Ribolock RNase inhibitor, 400 ng of template DNA, 2 μL of T7 polymerase enzyme, 1 μL of UTP-reduced nucleotide mix (10 mM AGC, 0.5 mM U) and 5 μL P32-UTP (3000 curies /mmol; 10 μCi/μL) with the appropriate amount of nuclease-free water were incubated for 4 hr at 37°C. Following transcription, the reaction mix was treated with 2.5 μL of TURBO DNase, purified by phenol/chloroform extraction and resuspended in 40 μL of nuclease-free water.

Electrophoretic mobility shift assays

20 μL binding reactions used for gel shift assays contained 12 mM HEPES-KOH pH 7.9, 60 mM KCl, 0.12 mM EDTA, 6% glycerol, 0.02% NP-40, 1 mM DTT, and 100 ng/uL of tRNA. Recombinant MBP+GST-tagged HNRNPD+Ex7 and HNRNPDΔEx7 were added at indicated concentrations and incubated for 8 min at 30°C. 1 μL of radio-labeled RNA was then added (approximate final concentration of 0.1 μM) and each sample was incubated for 15 min at 30°C to allow complex formation. 4μL of gel loading dye (50% glycerol, 62.5 mM EDTA pH 8.0, and bromophenol blue) was added and the samples were separated for 4-5 hr on a 4% native polyacrylamide gel in 0.5x TBE buffer. Gels were transferred to Whatman filter paper, dried and imaged using a Typhoon scanner (GE Healthcare).

RNase H protection assays

RNP reconstitution was performed under the same conditions as electrophoretic mobility shift assays, but in a 15μL volume. After binding was allowed to occur, a 5 μL mix containing 2μL of 10x RNaseH reaction buffer, 1 U of RNaseH and 8 ng of anti-sense DNA were added, and reactions were incubated for 30 min at 30°C. Reactions were stopped by addition of 1μL of 10% SDS and 1μL of 20 mg/mL Proteinase K followed by a 15 min incubation at 37°C. RNA was extracted with phenol-chloroform and separated on a denaturing gel in 1.0x TBE buffer. Gels were transferred to Whatman filter paper, dried and imaged using a Typhoon scanner (GE Healthcare).

Mass spectrometry

FLAG Affinity Purification Coupled with Mass Spectrometry (AP-MS) and BioID analysis

For all experiments, 3 biological replicates were acquired for each bait, alongside three replicates of the negative control. For FLAG AP-MS, cell pellets from two 150 mm plates, induced with 1μg/mL doxycycline for 24 hr, were lysed in 50 mM HEPES-KOH (pH 8.0), 100 mM KCl, 2 mM EDTA, 0.1% NP-40, and 10% glycerol and affinity-purified with M2-FLAG magnetic beads and on-bead trypsin digest essentially as described, except that 1 μg trypsin was first applied for 4 hr, followed by a spike-in of 0.5 μg trypsin overnight (Couzens et al., 2013). For BioID, cell pellets from two 150 mm plates, induced with 1μg/mL doxycycline and 50 μM biotin for 24 hr, were harvested and lysed for 1 hr at 4°C in 10 mL RIPA buffer (50 mM Tris-HCL pH 7.5, 150 mM NaCl, 1% NP-40, 1mM EDTA, 1 mM EGTA, 0.1%SDS, Sigma protease inhibitors P8340 1:500, and 0.5% Sodium deoxycholate). After incubation, 1 μl of benzonase (250U) was added to each sample and the lysates were sonicated on ice. Cleared lysates were used for affinity purification on streptavidin-Sepharose and on-bead trypsin digest as previously described, with the exception that a total of 1.5 μg of trypsin was used (Couzens et al., 2013).

Samples were analyzed on the AB SCIEX 5600 TripleTOF in Data Dependent Acquisition (DDA) mode. Digested samples (5 μl) were analyzed on a 5600 TripleTOF using a Nanoflex cHiPLC system at 200 nL/min (Eksigent ChromXP C18 3 μm × 75 μm × 15 cm column chip; FLAG AP-MS) or a home-packed emitter column (Dr. Maish Reprosil C18 3 μm × 75 μm × 10 cm; BioID) as previously described (Lambert et al., 2013). The column was coupled to an Eksigent Nano LC system with 0.1% formic acid in water as buffer A and 0.1% formic acid in acetonitrile as buffer B. Samples were loaded on the column at 400 nL/min (2% buffer B). The flow rate was then decreased to 200 nL/min and the HPLC delivered an acetonitrile gradient over 120 min (2%–35% buffer B over 90 min, 40%–60% buffer B over 5 min, hold buffer B at 80% 5 min, and return to 2% B over 5 min and re-equilibrate the column for 15 min at 2% B). The DDA parameters for acquisition on the TripleTOF 5600 were 1 MS scan (250 ms; mass range 400–1250) followed by up to 20 MS/MS scans (50 ms each). Candidate ions between charge states 2–5 and above a minimum threshold of 200 counts per second were isolated using a window of 0.7 amu. Previous candidate ions were dynamically excluded for 20 s with a 50 mDa window.

Mass spectrometry data analysis

The raw mass spectrometry files were stored, searched and analyzed using the ProHits laboratory information management system (LIMS) (http://prohitsms.com/Prohits_download/list.php). The WIFF data files were converted to MGF format using WIFF2MGF and subsequently converted to an mzML format using ProteoWizard (3.0.4468) and the AB SCIEX MS Data Converter (V1.3 beta). The mzML files were searched using Mascot (v2.3.02) and Comet (2014.02 rev.2). The results from each search engine were jointly analyzed through TPP (the Trans-Proteomic Pipeline, (tools.proteomecenter.org/wiki/index.php?title = Software:TPP) (v4.7) via

the iProphet pipeline (Shteynberg et al., 2011). The spectra were searched against a total of 72,230 proteins consisting of the NCBI human RefSeq database (v57, May28th, 2014, forward and reverse sequences), adenovirus sequences and supplemented with "common contaminants" from the Max Planck Institute (http://141.61.102.106:8080/share.cgi?ssid=0f2gfuB) and the Global Proteome Machine (GPM; http://www.thegpm.org/crap/index.html).

The database parameters were set to search for tryptic cleavages, allowing up to 2 missed cleavage sites per peptide, MS1 mass tolerance of 40 ppm with charges of 2+ to 4+ and an MS2 mass tolerance of ± 0.15 amu. Asparagine/glutamine deamidation and methionine oxidation were selected as variable modifications. Only proteins with two unique peptide ions and a minimum iProphet probability of 0.95 were used for further analysis.

QUANTIFICATION AND STATISTICAL ANALYSIS

Sequence annotation

Full genomic and transcriptomic sequences, as well as gene transfer format (GTF) files for all 7 species analyzed, were downloaded from the Ensembl database (assemblies listed in Table S1) (Yates et al., 2016). Exon annotations (including genomic annotations) were downloaded from Ensembl using BioMart (http://grch37.ensembl.org/biomart/martview/a618c55f6e41dfb700bd0b2aa7423a13).

Detection and quantification of alternative splicing

To comprehensively detect and quantify AS events, we used the VAST-TOOLS multi-module analysis pipeline, as previously described (Irimia et al., 2014), as well as *Whippet*, a new lightweight algorithm for event detection and quantification (Sterne-Weiler et al., 2017).

VAST-TOOLS was used to detect and quantify regulated target exons in Encode data in Figure 6. Reads were initially mapped to genome assemblies using Bowtie, using –m 1 –c 2 parameters and discarded for AS quantification. Unique EEJ (exon-exon junction) libraries were generated to derive measurements of exon inclusion levels using the 'Percent Spliced In' (PSI) metric. This utilized all hypothetically possible EEJ combinations from annotated and de novo splice sites, including cassette, mutually exclusive, and microexon events (Irimia et al., 2014).

Whippet was used to analyze RNA-Seq data (see Table S1 for datasets) employed for the identification of mammalian-specific AS events in Figure 1. To create the splice graphs required for Whippet splicing quantification, genome annotation files were extracted from the Ensembl (assemblies listed in Table S1). Whippet (v0.5) was run using default settings. Whippet was used to quantify all combinations of EEJs, including cassette, mutually exclusive, and microexon events.

Orthology definition

The orthology relationships between exons for cross-species comparative AS analyses were obtained by converting the genomic coordinates between genomes using the Lift-Over tool from UCSC (genome.ucsc.edu) with at least a 0.75 minimum ratio of bases that must remap (-minMatch). Only exons annotated within Ensembl GTF files were used (Yates et al., 2016). Only orthologous exons mapped to at least 3 mammalian-species and at least 2 non-mammalian species with adequate read coverage were included in the analysis.

Definition of mammalian-specific alternative splicing events

The selection of mammalian-specific AS events was performed with the R package *pamr*, which implements the "nearest shrunken centroids' method (Tibshirani et al., 2002). We applied it to a matrix of PSI values from the 57 samples from human, rhesus macaque, mouse, chicken, lizard and frog in order to identify the set of AS events that best distinguish mammalian-species from non-mammals. To remove NA values from the matrix, AS events with insufficient read coverage were assigned a median PSI value calculated using data from other tissues within the same species. For this classification, the "shrinkage" threshold was chosen to yield the maximum number of specific events in the 95th percentile of the false discovery rate, calculated by conducting 1000 permutations using the pamr.fdr package (Tibshirani et al., 2002). Only events that are constitutive in non-mammals (> 95 PSI) and selected using the nearest shrunken centroids method were considered mammalian-specific events.

Definition of other alternative splicing categories

An exon was defined as tissue-specific based on a survey of 26 human tissues with at least two replicates for each tissue (see Table S1). Three criteria must be met for an event to be considered tissue-specific. First, at least 20 mapped reads must support the AS event. Second, the differential splicing tool *diff* (https://github.com/vastgroup) must identify a statistically robust PSI difference of > 20 between the respective tissue and all other tissue samples. Third, the event must be conserved in mouse samples.

An event is defined as broadly alternative if the following three criteria are met. First, the exon is identified as conserved across at least three mammalian-species and at least two non-mammalian species (as per the Orthology definition above). Second, the exon is identified as alternatively spliced (i.e., PSI < 90) in at least one of the mammalian species analyzed (i.e human, rhesus, mouse, or opossum). Third, the exon is identified as alternatively spliced (i.e., PSI < 90) in at least one of the non-mammalian species tested (i.e., frog, lizard, or chicken).

Constitutive events are defined by two criteria. First, the exon is identified as conserved across at least 3 mammalian-species and at least 2 non-mammalian species (as per the Orthology definition above). Second, the exon must consistently have a PSI > 95 across all analyzed samples.

In the analysis in Figure 1, only those exons within a coding sequence and likely to contribute to protein function by not disrupting reading frame (i.e., length of exon is divisible by 3) are included.

Functional analysis

Functional enrichment analysis was performed using the g:Profiler (biit.cs.ut.ee/gprofiler) tool. Genes identified as containing mammalian-specific splicing events were compared to a background of multi-exon genes conserved within vertebrates. Structured controlled vocabularies from Gene Ontology organization, as well as information from the curated KEGG and Reactome databases were included in the analysis. Only functional categorizes with more than five members and fewer than 2,000 members were included in the analysis. Significance was assessed using the hypergeometric test. p values were corrected for multiple testing using the method of Benjamini-Hochberg. The Cytoscape application EnrichmentMap (baderlab.org/Software/EnrichmentMap) was used to visualize functional enrichment.

Phylogenetic Analysis

The divergence times indicated in Figure 1 were extracted from the TimeTree database (http://www.timetree.org/), which calculates the estimated pairwise divergence time based on the median of multiple published studies.

hnRNP phylogeny construction

Protein sequences for the canonical isoform of all hnRNPs were aligned using Clustal (https://www.ebi.ac.uk/Tools/msa/clustalo/) with MEGA7 software (Kumar et al., 2016). All members of the hnRNPA and hnRNPD families were selected, along with MSI1 and MSI2, which aligned with the hnRNPD family. HNRNPR was included as an outgroup member. The alignment was used to construct a phylogenetic tree using the Maximum Likelihood algorithm with default parameters. The tree with the highest log likelihood (−3896.2152) is shown. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site.

Crosslinking immunoprecipitation RNA Sequencing (CLIP-seq) analysis

41 nt reads were initially trimmed from the 3′ end to remove adapters. Remaining reads longer than 25 nt were mapped to the Hg19 human genome build using the universal RNA-Seq aligner STAR (https://github.com/alexdobin/STAR). To prevent false assignments of reads within repetitive regions, an "outFilterMismatchNoverLmax" of 0.1 was used. To identify enriched peaks within the aligned data, the PAR-CLIP analysis tool, PIPE-CLIP, was used with an FDR cutoff of 0.05 and with 4SU detection enabled (Chen et al., 2014). Analysis of the PAR-CLIP data identified: 7,666 HNRNPDΔEx7 clusters; 6,746 HNRNPD+Ex7 clusters; 2,342,032 HNRNPDΔEx7 mapped reads; and 1,826,356 HNRNPD+Ex7 mapped reads. Only reads mapped to intronic regions were used for downstream analysis. Significant peaks in the HNRNPD+Ex7 and HNRNPDΔEx7 datasets, identified within 250 nt windows in which there is at least one significant overlapping peak for both isoforms, were analyzed further. To compare the distributions of the significant peaks representing the isoforms within the 250 nt windows, density plots were generated. Significant differences in the distributions of peaks were assessed with the null hypothesis that distributions are identical, using the non-parametric Anderson-Darling k-sample test with 10,000 trials. 95% confidence intervals were calculated using a bootstrap sub-sampling approach (n = 10,000).

Peptide Analysis

Analysis was performed on total peptide tables obtained by AP-MS and BioID from 3 biological replicates for PTBP1+Ex9, PTBP1ΔEx9, HNRNPD+Ex7, and HNRNPDΔEx7. Only proteins identified across at least two of the three replicates were used for differential peptide analysis. In order to assess biological variation between replicates, the differential protein abundance was tested against the null hypothesis that the difference is zero. This was done by modeling the peptides as compared to control samples based on a negative binomial distribution (Anders and Huber, 2010). Only proteins with a minimum of ten peptides were included in the analysis. p values were corrected for multiple testing using the method of Benjamini-Hochberg. Proteins with an adjusted p value of max. 0.01 and a 1.5 $\log_2$ fold change over control were included in further analysis. Prior to clustering, the peptide counts of differentially interacting proteins were normalized by subtracting the mean of the peptides across all conditions, and dividing by the standard deviation. Affinity propagation was used to cluster (cran.r-project.org/web/packages/apcluster) normalized peptide counts using pairwise correlation values. Sub-clusters were identified using exemplar-based agglomerative clustering and annotated manually.

Analysis of GY-rich protein enrichment

To assess the abundance of GY-rich or hnRNP proteins in each condition the relative abundance of peptides for each protein of interest in the data was quantified after normalization using the procedure described above. The boxplots show the distribution of normalized values for all hnRNPs or GY-rich proteins. p values comparing distributions of abundance were calculated using the Wilcoxon rank-sum test.

Protein feature analysis

For all positions in a protein a score for intrinsic disorder was computed using IUPred (iupred.enzim.hu). Amino acid residues with a score greater than 0.4 were considered disordered. For each coding exon the fraction of disordered residues was estimated.

For all positions in a protein, low-complexity regions were calculated using SEG (http://www.dbbm.fiocruz.br/cgc/seg.html). Only amino acids not located within ordered Pfam annotated protein domains (pfam.xfam.org), putative transmembrane domains, signal peptides and coiled coil regions, were considered as low-complexity regions. For each coding exon, the fraction of amino acids annotated as within a low-complexity region was estimated. Tandem protein repeat regions within low-complexity regions were identified using the PTRStalker algorithm for de-novo detection of fuzzy tandem repeats (Pellegrini et al., 2012). Tyrosine and serine phosphorylation sites were extracted from the Phosphosite database (phosphosite.org).

To calculate the enrichment of single amino acid residues and dipeptides within exonic coding sequences representing disordered regions identified in Figure 1A, their occurrence relative to all residues or dipeptides within the same regions was calculated. Over-representation was assessed using a binomial test. p values were corrected for multiple testing using the method of Bonferroni. Residues or dipeptides with a corrected p value less than 0.01 were regarded as enriched within disordered regions.

For amino acid logos, the composition of residues around the enriched dipeptide is reported only for disordered regions. All subsequences in a window of four amino acids before and after a repetitive dipeptide are indicated. The height of each amino acid is proportional to the distribution of the amino acids at the same position in the exon coding sequence. All logos were created using weblogo (weblogo.berkeley.edu).

Three criteria were applied for identifying proteins enriched in GY motifs. First, motifs had to be found in the disordered regions of proteins and excluded from known functional features such as coiled coils and signal peptide sequences. Second, the motifs had to be found within low-complexity domains. Third, more than three motifs must be clustered together within a hundred amino acid window.

Duplex analysis

18 species representing multiple clades of vertebrate evolution were extracted from the UCSC vertebrate 100 species multiz alignment (genome.ucsc.edu) and manually checked using the Jalview multiple sequence editor. Arc diagrams were created using the R-chie and R4RNA packages (Lai et al., 2012).

To investigate the occurrence of additional RNA duplexes, we used a sliding window approach to scan 1000 nt of intronic sequence flanking each mammalian-specific alternative cassette exon for potential base pairing regions, after removal of repetitive sequences. 100 nucleotide windows were shifted 25 nt at a time. Putative base pairing interactions between the sequences were performed using the ViennaRNA package (https://www.tbi.univie.ac.at/RNA/; (Lorenz et al., 2011)). An initial pre-filter was applied using the program RNAduplex (Lorenz et al., 2011) to select only predicted duplexes with a free energy of less than −70 kcal/mol. To account for potential competition from intra-molecular base-pairing within the individual sliding window sequences, RNAup (Lorenz et al., 2011) was applied. This algorithm determined an overall estimated free-energy value for putative base-pairing interactions flanking each exon, after taking into account possible competing intra-molecular base pairing interactions in each individual window. To be considered as a candidate, an estimated free-energy of −20 kcal/mol was required. To further assess relevance of detected base-pairing interactions, UCSC PhyloP conservation scores (genome.ucsc.edu) from whole genome alignments of 100 vertebrates were included, and only putative base-pairing regions that displayed conservation were retained. This procedure was performed both within and between the flanking intronic sequences.

RNA binding motifs and splice site strength calculations

For all hnRNPs analyzed in Figure 6A, RNA binding motifs were extracted from ENCODE Bind-n-Seq data (https://www.encodeproject.org/), RNACompete data (http://cisbp-rna.ccbr.utoronto.ca/), or from independent CLIP-Seq analyses (Yoon et al., 2014); see Table S1 for datasets used. To calculate similarities between the RNA binding motifs of hnRNPs, pairwise distance measurements were calculated using Euclidean distance. A distance matrix was thus obtained, facilitating the construction of a tree using a minimum evolution approach with the MEGA7 software (http://www.megasoftware.net/).

MaxEntScan (Yeo and Burge, 2004) was used to estimate the splice site strength of both the 3′ splice sites and 5′ splice sites. 5′ splice site strength was assessed using sequences spanning −3 to +6 of the 5′ splice site boundary. 3′ splice site strength was assessed using sequences spanning −20 to +3 of the 3′ splice site boundary.

Mammalian versus non-mammalian hnRNP-regulated alternative splicing events

Alternative splicing events regulated by hnRNPA and D family members with mammalian-specific isoforms were identified using ENCODE knockdown data (www.encodeproject.org) (Figure 6A). For comparison, two additional groups were defined: a set of exons regulated by hnRNPs that lack mammalian-specific isoforms (also from ENCODE (https://www.encodeproject.org/)), and a set of broadly alternative exons, as defined in Figure 1. Only exons conserved across mammalian and non-mammalian species were included. To determine the difference in splicing regulation for each set of AS events, the difference in their average, non-constitutive PSI values between mammalian and non-mammalian tissues, was calculated.

Statistical Tests

The nearest shrunken centroids false discovery rate was estimated using a permutation test and only events within the 95th percentile were selected. For gene function enrichment analysis, significance was assessed using the hypergeometric test with multiple testing correction using the method of Benjamini and Hochberg. The Wilcoxon rank-sum test was used for comparing distributions. A binomial-test with Bonferroni multiple testing correction was used to calculate enrichment of amino acids and residues compared to background. PIPE-CLIP used zero-truncated negative binomial (ZTNB) likelihoods to identify enriched clusters of CLIP reads. A negative binomial test with p values corrected for multiple testing by the method of Benjamini-Hochberg was used to identify differential peptide counts. Affinity propagation clustering of either pairwise correlation (Pearson) values, or of negative Euclidean distance, was used to create heatmaps. When applicable, sub-clusters were identified using exemplar-based agglomerative clustering and annotated manually.

DATA AND SOFTWARE AVAILABILITY

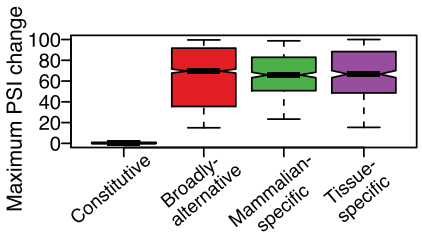
All scripts used for data processing and statistical analysis were written in Python, Perl, or R and are available upon request. VAST-TOOLS is available via its github portal (https://github.com/vastgroup). *Whippet* software is available via https://github.com/timbitz/Whippet.jl.

All data with reference accession numbers are annotated within Table S1.

Raw mass spectrometry files, peak lists and result files generated in this study are deposited at MassIVE (http://massive.ucsd.edu/ProteoSAFe/static/massive.jsp), alongside the complete Significance Analysis of INTeractome (SAINT) outputs (Teo et al., 2014) and protein identification tables. The MassIVE ID numbers are MSV000080370 (BioID) and MSV000080371 (FLAG AP-MS) and the MassIVE links for download are ftp://MSV000080370@massive.ucsd.edu and ftp://MSV000080371@massive.ucsd.edu, respectively. ProteomeXchange numbers are PXD005475 and PXD005476, respectively.

Cell

A

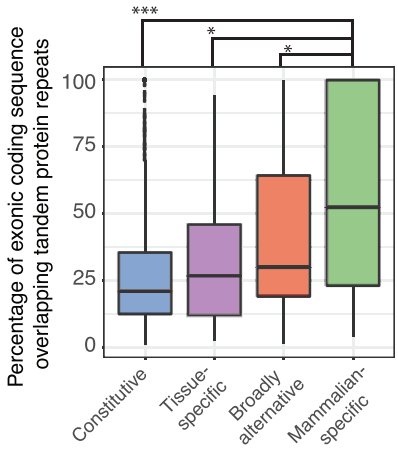


C

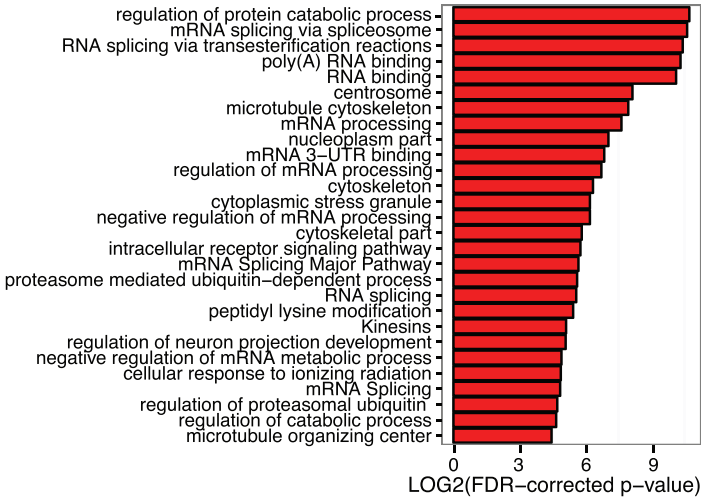
Dipeptide motifs enriched in mammalian-specific exons relative to:

Dipeptide	P-value		
	Tissue-specific	Constitutive	Broadly AS
AG	1.63 x 10-9	3.78 x 10-7	2.57 x 10-03
EK	NS	1.59 x 10-3	NS
GA	1.64 x 10-7	1.53 x 10-6	6.93 x 10-04
GP	4.13 x 10-16	3.33 x 10-9	1.38 x 10-04
GS	4.20 x 10-4	4.46 x 10-3	NS
GY	2.64 x 10-6	2.60 x 10-6	2.67 x 10-06
PG	1.73 x 10-16	1.89 x 10-11	2.03 x 10-04
YG	3.95 x 10-14	1.16 x 10-5	2.47 x 10-07

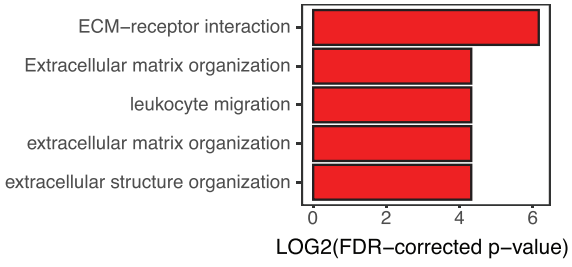
B



D



E



F

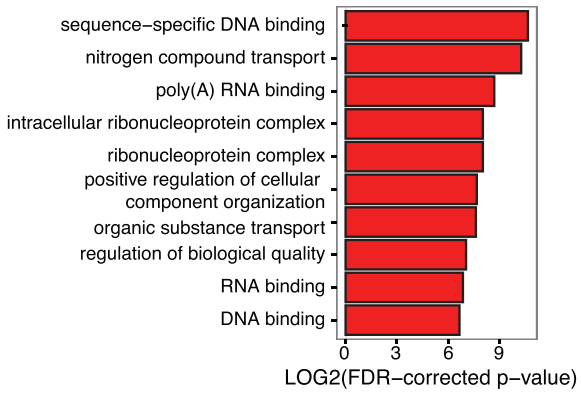


Figure S1. Properties of Mammalian-Specific Alternative Exons, Corresponding Coding Regions, and Host Genes, Related to Figure 1

(A) Boxplots showing the distribution of maximum percent spliced in (PSI) changes (maximum PSI across samples minus minimum PSI across samples) for exons categorized as constitutive (blue), broadly alternative (red), mammalian-specific (green), and tissue-specific (purple). For a description of boxplots see Figure 1B legend.

(B) Boxplot showing the percentages of exonic coding sequence that overlap tandem protein repeats (i.e., region when a pattern of one or more amino acids is repeated) in disordered regions. Exons analyzed for overlap are classified as constitutive (blue), broadly alternative (red), mammalian-specific (green), and tissue-specific (purple). * = $p < 0.05$; ** = $p < 1 \times 10^{-3}$; *** = $p < 1 \times 10^{-6}$, Wilcoxon rank-sum tests. For a description of boxplots see Figure 1B legend.

(C) Table of dipeptides over-represented in mammalian-specific exons compared to tissue-specific, constitutive, and broadly alternatively spliced (AS) exons. Significance of over-representation of dipeptides assessed using the binomial test. p values are adjusted for multiple hypothesis testing using Bonferroni correction. Dipeptides with a corrected p value < 0.01 are regarded as significantly enriched within mammalian-specific exons. NS, not significant.

(legend continued on next page)

(D) Bar plot for GO, REACTOME, and KEGG functional categories of genes that contain mammalian-specific AS events visualized by the enrichment map in Figure 1E. FDR, false detection rate; LOG2, absolute logarithm of 2.

(E and F) Bar plots for GO, REACTOME, and KEGG functional categories of genes with (E) GP/PG or (F) GY/YG dipeptide-containing mammalian-specific AS exons, as compared to the background of all expressed genes with mammalian-specific events. See (D) for details.

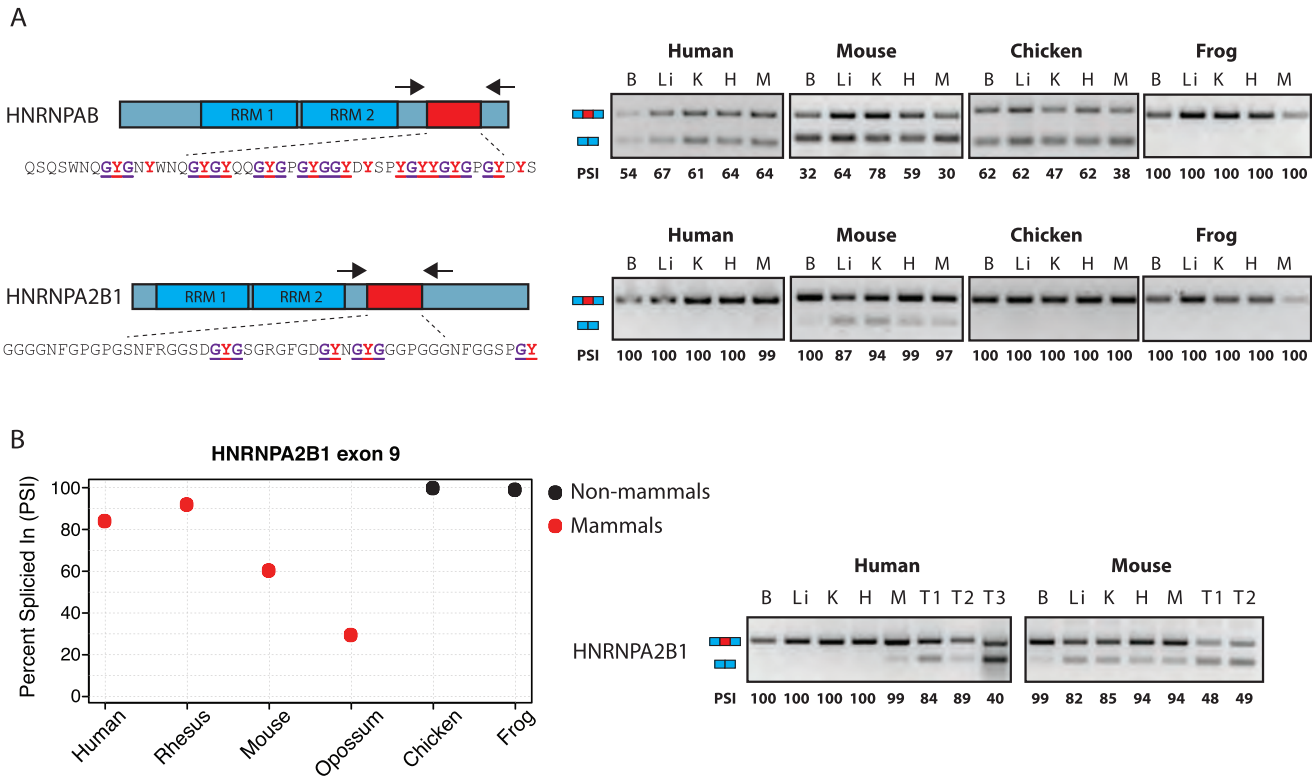


Figure S2. Protein Sequence Features and Regulation of Mammalian-Specific Alternative Exons, Related to Figure 2

(A) Domain organization of HNRNPAB and HNRNPA2B1. Mammalian-specific alternative exons are indicated in red, with the corresponding amino acid sequence shown below. RT-PCR analyses confirming RNA-seq-derived patterns of splicing for representative mammalian-specific AS events. Percent Spliced In (PSI) values are indicated (see STAR Methods). Tyrosine residues are indicated in red, GY/YG motifs are underlined, and glycine residues within the motifs are indicated in purple. Arrows, approximate location of RT-PCR primers; RRM, RNA recognition motif; B, brain; Li, liver; K, kidney; H, heart; M, muscle.

(B) Testis-specific skipping of HNRNPA2B1 alternative exon in mammals. Dot plot (left) shows percent spliced in (PSI) values from RNA-seq data for testis across mammalian (red) and non-mammalian species (black). RT-PCR gels (right) show testis-specific skipping in human and mouse organ samples. Percent spliced in (PSI) values are indicated (see STAR Methods). B, brain; Li, liver; K, kidney; H, heart; M, muscle; T1, T2, and T3, independent testis samples.

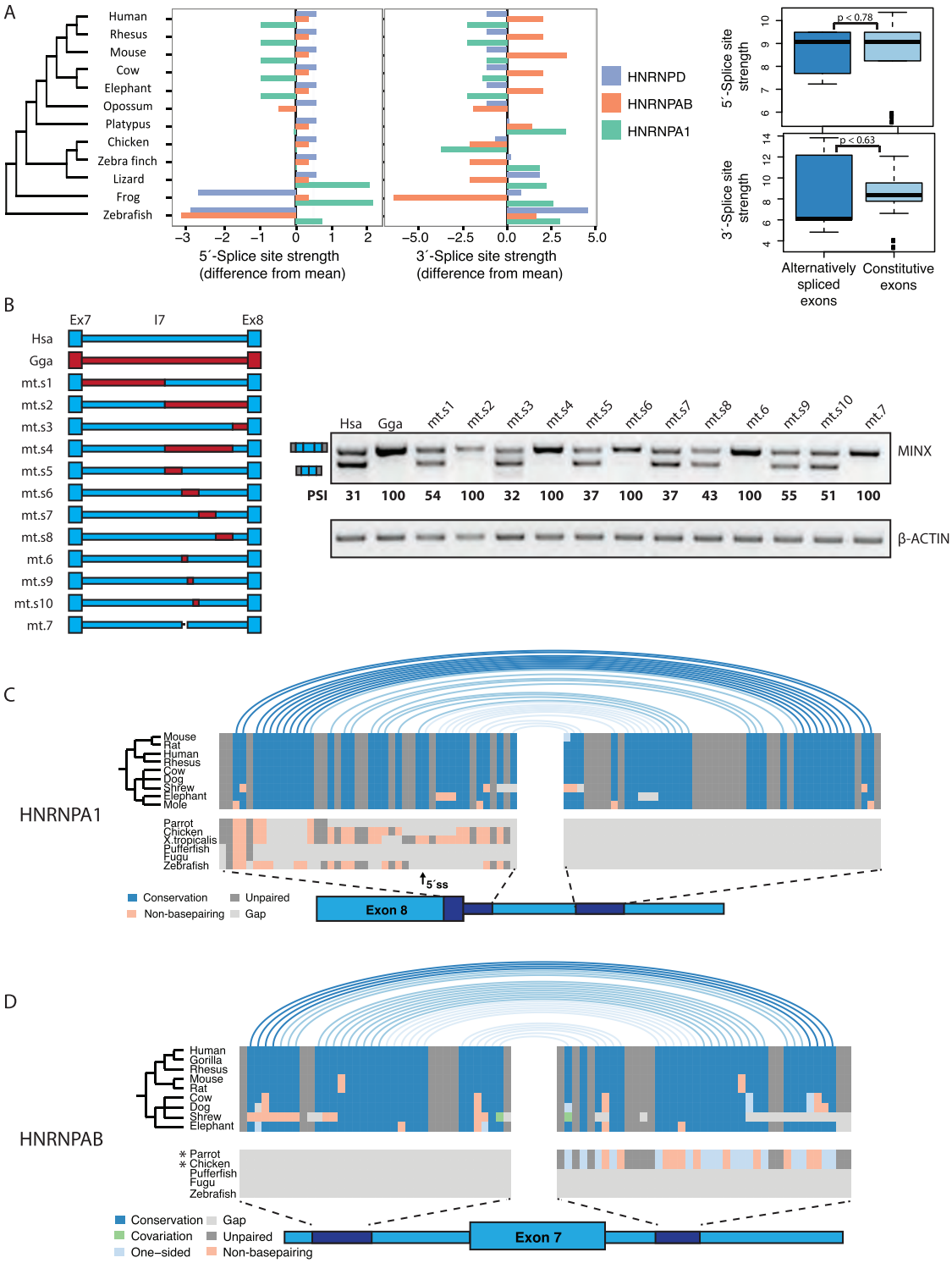


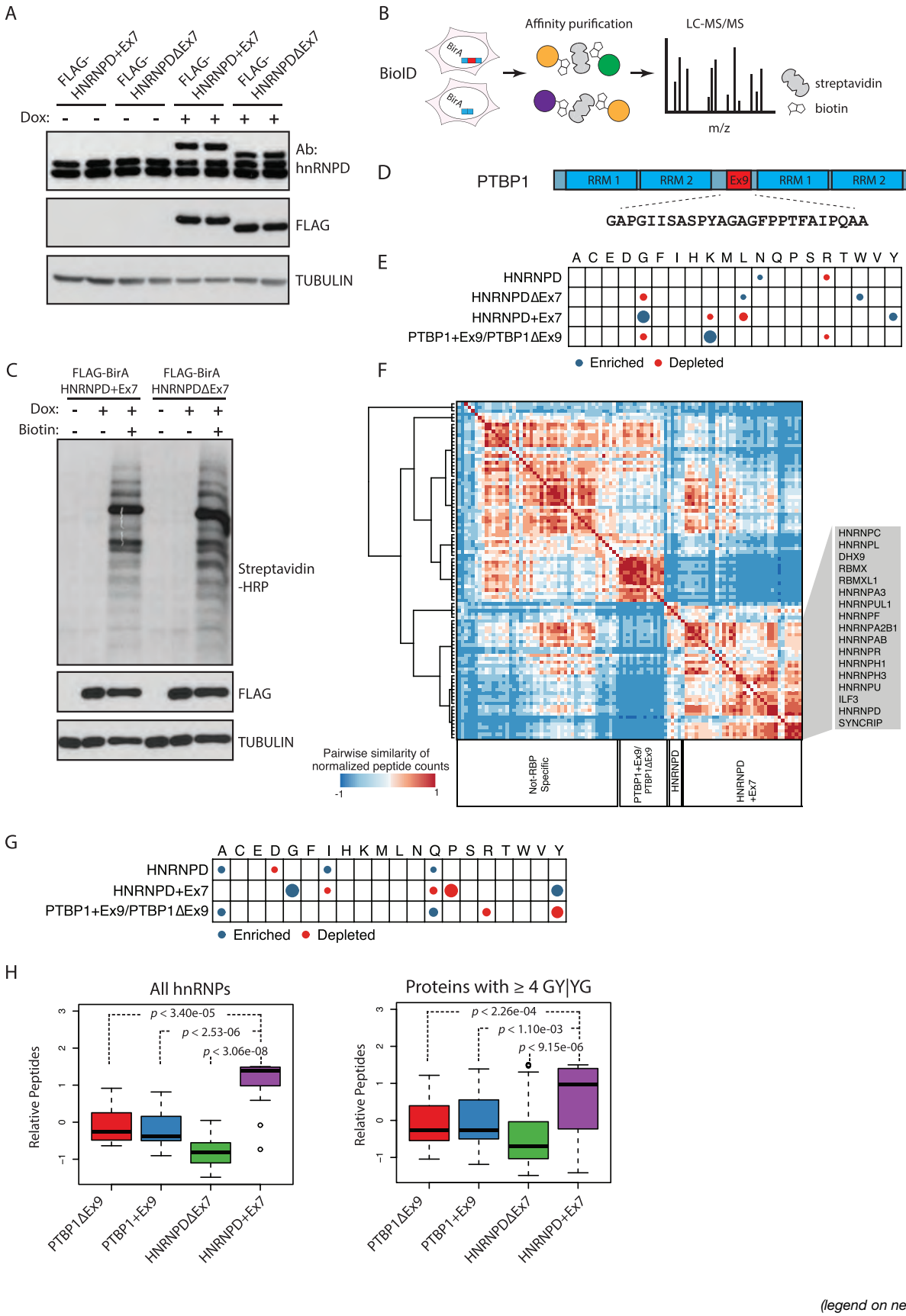
Figure S3. RNA Sequence Features of Mammalian-Specific Alternative Exons, Orthologous Exons, Surrounding Sequences, and Reporter Constructs, Related to Figure 3

(A) Bar plot (left) showing difference from the mean of 3′and 5′ splice-site strength scores calculated using Maximum Entropy Model for mammalian-specific exons in HNRNPD, HNRNPA1, and HNRNPAB (Yeo and Burge, 2004). Cladogram displays phylogenetic relationships between species. Boxplots (right) show distribution of splice site scores for the same exons, divided by species where they are alternatively or constitutively spliced. p value calculated using Wilcoxon rank-sum test.

(legend continued on next page)

(B) Extension of Figure 3C. Schematic diagrams of HNRNPD exon 7 (Ex7) minigene reporters containing human (blue) and chicken (red) sequences encompassing the flanking native exons (Ex6 and Ex8) and intervening introns (I6 and I7) (see STAR Methods). Heterologous flanking constitutive exons and intronic sequence (gray) are derived from adenovirus sequence (MINX). Constructs were transfected into HEK293 cells and RT-PCR assays were performed with MINX exon-specific primers. β-ACTIN was used as a loading control. Reporter percent spliced in (PSI) values are indicated (see STAR Methods). Hsa, *Homo sapiens*; Gga, Gallus gallus; mt, mutation.

(C and D) Arc diagrams visualizing conserved base-pairing positions formed by RNA-RNA duplex within (C) HNRNPA1 and (D) HNRNPAB. Arcs display complementary nucleotide base-pairings (G:C; A:U, G:U, and vice versa). Blue, conserved complementarity; orange, nucleotide changes resulting in loss of base-pairing; light gray, indels or regions that fail to align; dark gray, lack of complementary between up and downstream elements; green, covariation that maintains base-pairing, cyan, one-sided mutation that retains base-pairing. Diagrams below indicate locations of complementary base-pairing within exon and intronic sequence in dark blue. 5′ss, 5′ splice site. *, notable differences in intron length upstream of HNRNPAB exon 8 in bird lineage.



(legend on next page)

Figure S4. Affinity-Purification Mass Spectrometry and BioID Data, Related to Figure 4

(A) Western blot showing expression of HNRNPD splice isoforms in HEK293 cell lines used for AP-MS experiment. Western blot with anti-HNRNPD antibody shows expression relative to endogenous protein. Tubulin is used as a recovery and loading control.

(B) Experimental outline for promiscuous biotin labeling (BioID) experiments.

(C) Western blots showing samples used for BioID experiments. Cells were grown as in (A) but additionally supplemented with biotin. Blot with streptavidin-HRP shows specific proteome biotinylation in the presence of both doxycycline and biotin.

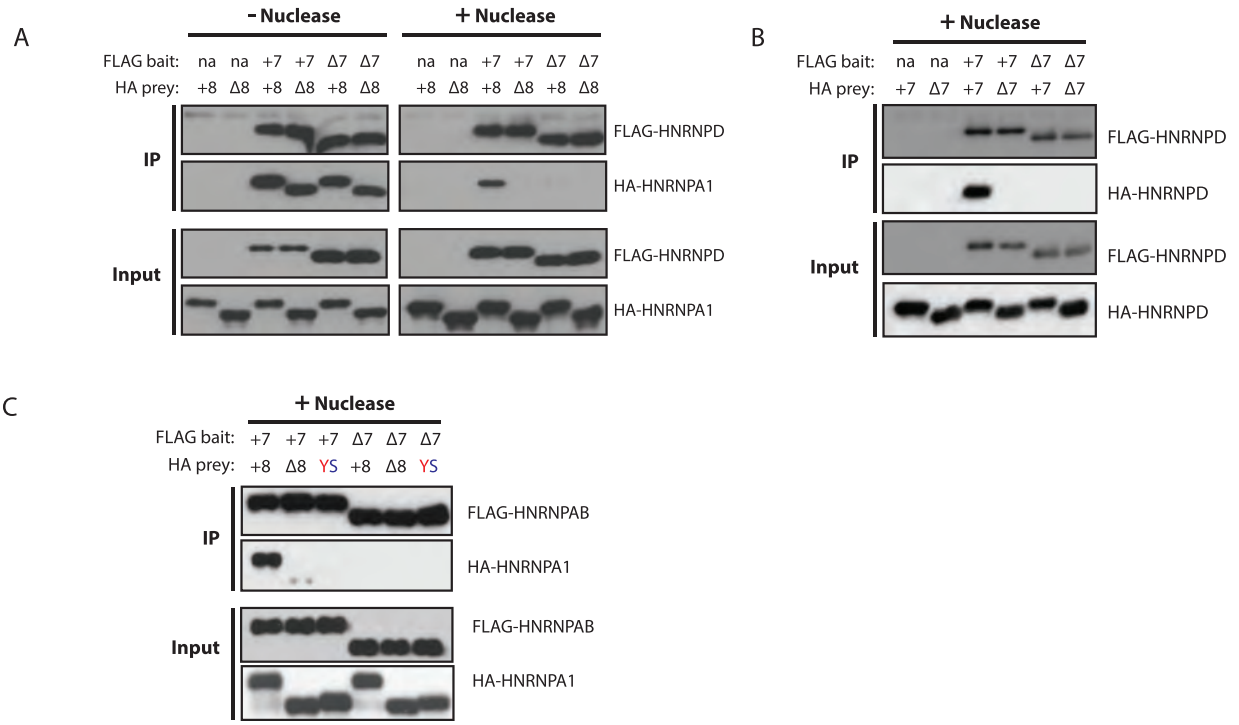
(D) Domain organization of PTBP1 showing location of RNA recognition motifs (blue) and location of mammalian-specific alternative exon 9 (in red), with the corresponding sequence indicated below.

(E) Amino acids enriched (blue dots) or depleted (red dots) within proteins identified by AP-MS in Figure 4B clusters, as compared to all detected proteins. Only amino acids with significant differential enrichment are shown ($p < 0.01$; binomial test with Bonferroni correction; size of dots is inversely related to the magnitude of $\log_2$ p value.

(F) Symmetrical heatmap of pairwise correlations of normalized counts of peptides from proteins enriched using BioID and detected by mass spectrometry. BioID samples were prepared from stable cell lines expressing BirA-tagged HNRNPD+Ex7, HNRNPDΔEx7, PTBP1+Ex9, or PTBP1ΔEx9 proteins. Affinity propagation clustering is based on pairwise similarities (represented as positive and negative correlations) between profiles of detected interaction partners. Sub-clusters were identified using exemplar-based agglomerative clustering. Sub-clusters are labeled based on enrichment of peptides interacting with a specific hnRNP isoform or pair of isoforms (i.e., ''HNRNPD'' indicates peptides that do not show preferential interaction with either HNRNPD isoform). ''Not RBP-specific'' refers to sub-clusters where there is no clear preference for binding to PTBP1 or HNRNPD. The Grey box highlights proteins enriched in the HNRNPD+Ex7-containing cluster.

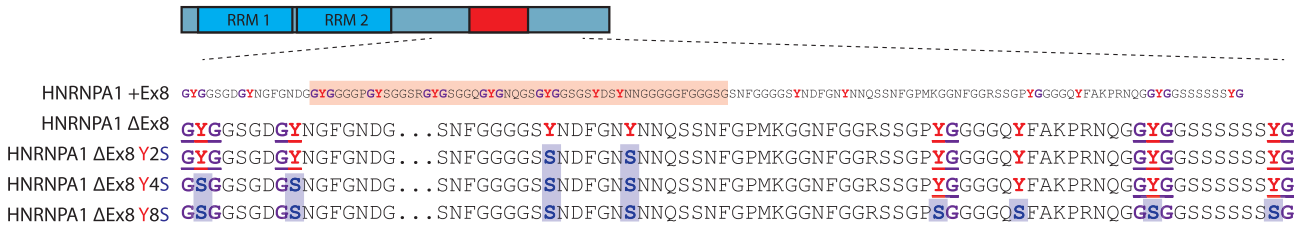
(G) Amino acids enriched (blue dots) or depleted (red dots) within proteins identified in (F) clusters, as compared to all detected proteins. Only amino acids with significant differential enrichment are shown ($p < 0.01$; binomial test with Bonferroni correction; size of dots is inversely related to the magnitude of $\log_2$ p value.

(H) Boxplots showing normalized peptide distributions of all GY-rich or hnRNP preys identified by BioID. p values correspond to Wilcoxon signed-rank test. A total of 148 GY proteins were identified computationally in the human proteome (see STAR Methods), and 38 of these were detected in the BioID data. For description of boxplots see Figure 1B legend.

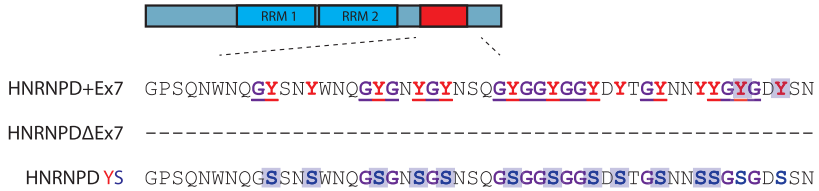


D

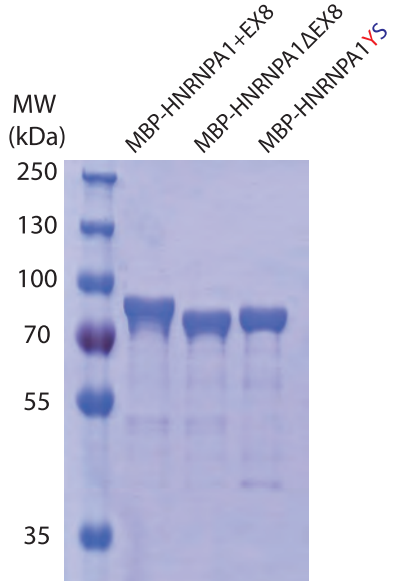
HNRNPA1 Tyrosine to Serine mutants



E



F



(legend on next page)

Figure S5. Co-immunoprecipitation-Western Blotting Analyses of hnRNP Isoform Interactions and Description of hnRNP Tyrosine-Serine Mutants, Related to Figures 4 and 5

(A) Co-immunoprecipitation western blot experiments analyzing interactions between FLAG-HNRNPD+Ex7 or HNRNPDΔEx7, and HA-HNRNPA1+Ex8 or HNRNPA1ΔEx8. Presence or absence of nuclease (RNaseA1 and benzonase) treatment is indicated.

(B) Same as in (A), but analyzing interactions between reciprocally FLAG- and HA-tagged HNRNPD+Ex7 and HNRNPDΔEx7 in the presence of nuclease (RNaseA1 and benzonase) treatment.

(C) Same as in (A), but analyzing interactions between FLAG-HNRNPAB+Ex7 or HNRNPABΔEx7 and HA-HNRNPA1+Ex8, HNRNPA1ΔEx8, or an HA-HNRNPA1ΔEx8–YS mutant in which all tyrosines within the C-terminal intrinsically disordered region (IDR) are substituted with serines (see D).

(D) Domain organization of HNRNPA1 showing protein sequences overlapping and surrounding exon 8. RNA recognition motifs (RRMs) shown in light blue and mammalian-specific AS exon in red with corresponding protein sequence boxed in red below. Tyrosine residues that were mutated to serine are indicated. With the exception of Figure 6D, the tyrosine to serine mutant of HNRNPA1 refers to the Y8S mutant, with all exonic tyrosines substituted with serine. Tyrosine residues are indicated in red, GY/YG motifs are underlined, and glycine residues within the motifs are indicated in purple.

(E) Domain organization of HNRNPD showing amino acid sequence of exon 7. RNA recognition motifs (RRMs) shown in light blue and mammalian-specific AS exon in red with corresponding amino acid sequence indicated below. Exon 7 sequences of mutant constructs in which tyrosine residues were mutated to serine are also indicated. Tyrosine residues are indicated in red, GY/YG motifs are underlined, and glycine residues within the motifs are indicated in purple.

(F) Coomassie-stained protein gel showing recombinant human N-MBP HNRNPA1+Ex8, HNRNPA1ΔEx8, and HNRNPA1ΔEx8 Y8S tyrosine to serine mutant.

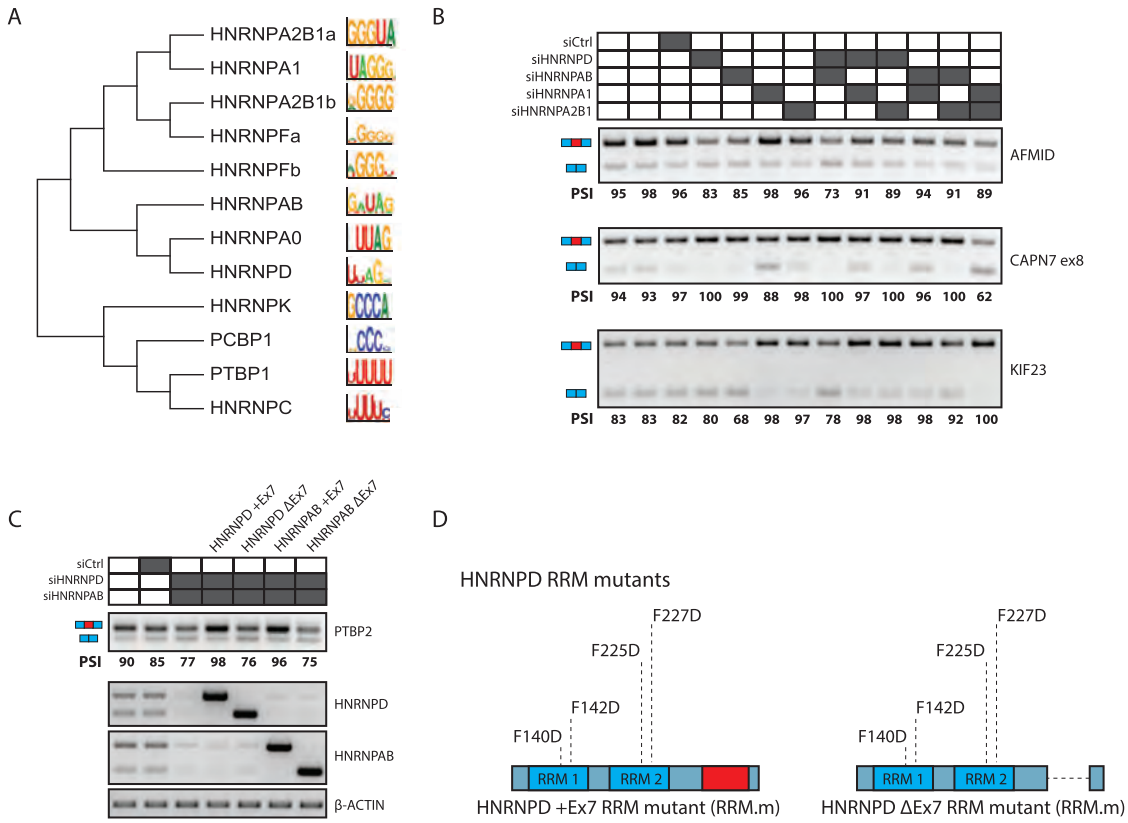


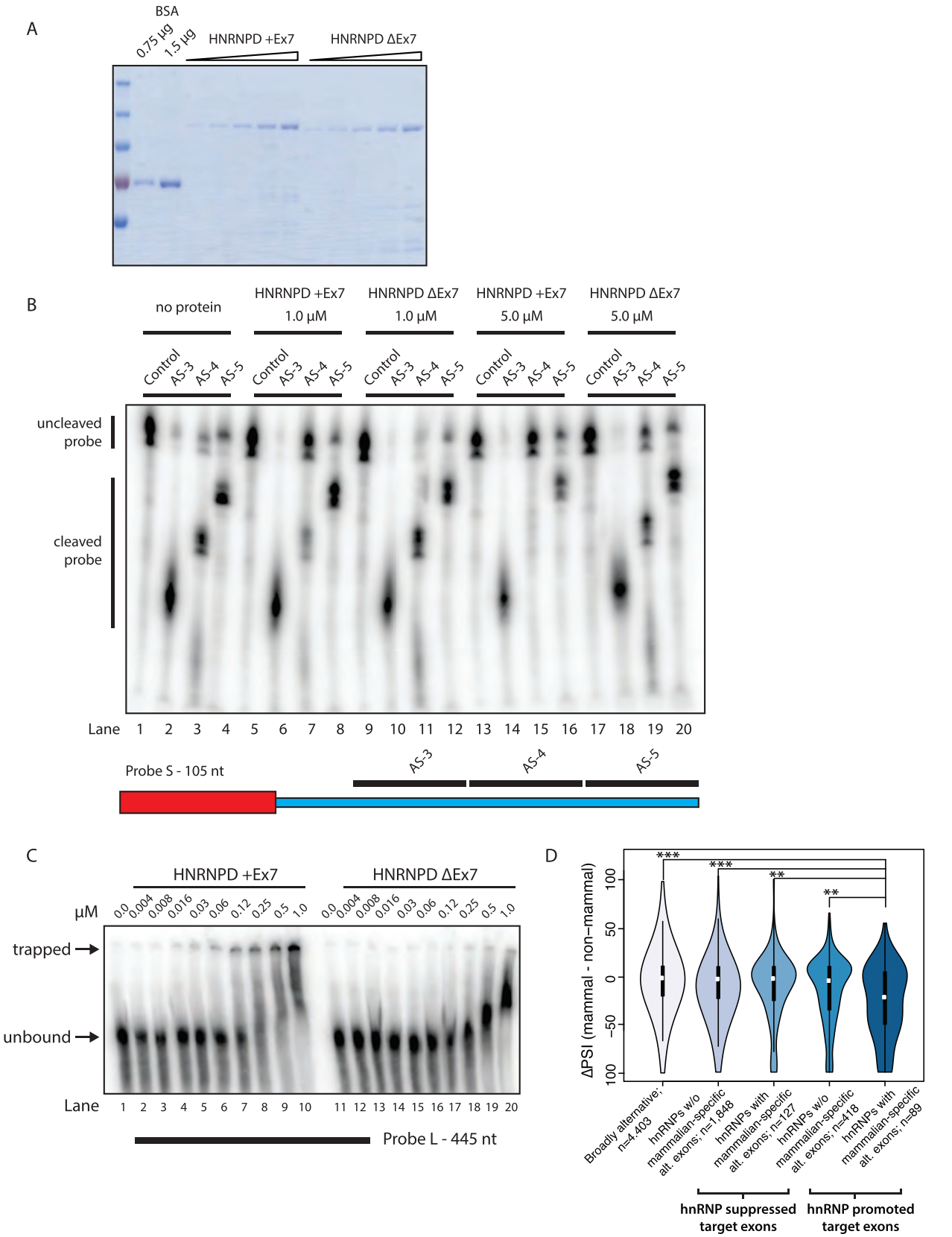
Figure S6. RNA Binding Specificity, Regulatory Properties, and RRM-Inactivating Mutants of hnRNPs, Related to Figure 6

(A) Dendrogram showing relationships between selected hnRNPs based on similarity of their respective RNA binding motifs. RNA binding motifs displayed using sequence logos with height of nucleotide representing enrichment at a particular position.

(B) RT-PCR assays showing the hnRNP-dependent regulation of additional alternative exons identified from RNA-seq analysis in Figure 6A. hnRNPs were knocked down using siRNAs against individual proteins or in combination, as indicated by shaded boxes.

(C) RT-PCR assays showing isoform-dependent regulation of PTBP2 exon 10. Combined knockdown of HNRNPD and HNRNPAB was performed as indicated. Rescue was performed with pairs of splice isoforms of HNRNPD or HNRNPAB. β-ACTIN was used as a loading control.

(D) Domain diagrams showing the mutations introduced into HNRNPD to inactivate RNA recognition motifs (RRMs).



(legend on next page)

Figure S7. RNaseH Protection Assays, EMSA Assays, and Global Regulatory Properties of hnRNPs, Related to Figure 7

(A) Recombinant human N-terminal GST and C-terminal MBP-tagged HNRNPD+Ex7 and HNRNPDΔEx7 were purified from *E. coli*. Purity and relative amounts were assessed by SDS-PAGE and staining with Coomassie Brilliant Blue, with known quantities of bovine serum albumin (BSA) used to estimate concentrations.

(B) RNaseH protection assay showing greater protection of 105 nt probe 'S' (Figure 7A) by HNRNPD+Ex7 than HNRNPDΔEx7. Proteins at indicated concentrations were bound to RNA after which indicated 20 nt-long antisense oligonucleotides were added along with RNaseH, which specifically degrades DNA-RNA hybrids. Protein binding interferes with DNA annealing and thus inhibits degradation.

(C) Electrophoretic mobility shift assay showing binding of recombinant HNRNPD splice isoforms to 445 nt probe 'L' (Figure 7A). Protein concentrations in the binding reaction are shown. Arrows indicate distinct protein-RNA complexes.

(D) Extension of Figure 7E. Violin plots showing percent spliced in (PSI) differences for exons that are conserved between mammals and non-mammals and representing five sets of events: exon targets of hnRNPs with mammalian-specific AS events (MS-AS) as defined by the analysis of knockdown RNA-seq data (see Figure 6A), exon targets regulated by hnRNPs that lack mammalian-specific alternative exons (also as defined from data in Figure 6A), and a set of broadly alternatively spliced exons, as defined in Figure 1A (refer to main text). The first two sets are further sub-divided into events that are enhanced or suppressed by hnRNP expression. Violin plots display the median of the data, a box indicating the interquartile range, and a visualization of the full distribution of the data. *$p < 0.05$; **$p < 0.005$; ***$p < 1x10^{-5}$; all p values calculated using Wilcoxon rank-sum test. For description of violin plots see Figure 7E.