A Large-Scale Binding and Functional Map of Human RNA Binding Proteins

Eric L Van Nostrand[§,1,2], Peter Freese[§,3], Gabriel A Pratt[§,1,2,4], Xiaofeng Wang[§,5], Xintao Wei[§,6], Rui Xiao[§,1,2], Steven M Blue[1,2], Daniel Dominguez[7], Neal A.L. Cody[5], Sara Olson[6], Balaji Sundararaman[1,2], Lijun Zhan[6], Cassandra Bazile[7], Louis Philip Benoit Bouvrette[5,9], Jiayu Chen[1,2], Michael O Duff[6], Keri E. Garcia[1,2], Chelsea Gelboin-Burkhart[1,2], Abigail Hochman[7], Nicole J Lambert[7], Hairi Li[1,2], Thai B Nguyen[1,2], Tsultrim Palden[7,8], Ines Rabano[1,2], Shashank Sathe[1,2], Rebecca Stanton[1,2], Julie Bergalet[5], Bing Zhou[1,2], Amanda Su[7], Ruth Wang[1,2], Brian A. Yee[1,2], Ashley L Louie[1,2], Stefan Aigner[1,2], Xiang-dong Fu[*,1,2], Eric Lecuyer[*,5,9,10], Christopher B. Burge[*,3,7,8] , Brenton R. Graveley[*,6], Gene W. Yeo[*,1,2,11,12]

1. Department of Cellular and Molecular Medicine, University of California at San Diego, La Jolla, CA
2. Institute for Genomic Medicine, University of California at San Diego, La Jolla, CA
3. Program in Computational and Systems Biology, MIT Cambridge MA
4. Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA
5. Institut de Recherches Cliniques de Montréal (IRCM), Montreal, Canada
6. Department of Genetics and Genome Sciences, Institute for Systems Genomics, UConn Health, Farmington, CT
7. Department of Biology, MIT, Cambridge MA
8. Department of Biological Engineering, MIT Cambridge MA
9. Department of Biochemistry and Molecular Medicine, Université de Montréal, Montreal, Canada
10. Division of Experimental Medicine, McGill University, Montreal, Canada
11. Department of Physiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
12. Molecular Engineering Laboratory. A*STAR, Singapore

[§]Equal contribution

[*]Correspondence and requests for materials should be addressed to Brenton Graveley (graveley@uchc.edu), Chris Burge (cburge@mit.edu), Xiang-dong Fu (xdfu@ucsd.edu), Eric Lecuyer (eric.lecuyer@ircm.qc.ca) and Gene Yeo (geneyeo@ucsd.edu).

Abstract

Genomes encompass all the information necessary to specify the development and function of an organism. In addition to genes, genomes also contain a myriad of functional elements that control various steps in gene expression. A major class of these elements function only when transcribed into RNA as they serve as the binding sites for RNA binding proteins (RBPs) which act to control post-transcriptional processes including splicing, cleavage and polyadenylation, RNA editing, RNA localization, translation, and RNA stability. Despite the importance of these functional RNA elements encoded in the genome, they have been much less studied than genes and DNA elements. Here, we describe the mapping and characterization of RNA elements recognized by a large collection of human RBPs in K562 and HepG2 cells. These data expand the catalog of functional elements encoded in the human genome by addition of a large set of elements that function at the RNA level through interaction with RBPs.

Introduction

RBPs have emerged as critical players in regulating gene expression, controlling when, where, and at what rate RNAs are processed, trafficked, translated, and degraded within the cell. They represent a diverse class of proteins involved in co- and post-transcriptional gene regulation[1,2]. RBPs interact with RNA to form ribonucleoprotein complexes (RNPs), governing the maturation and fate of their target RNA substrates. Indeed, they regulate numerous aspects of gene expression including pre-mRNA splicing, cleavage and polyadenylation, RNA stability, RNA localization, RNA editing, and translation. In fact, many RBPs participate in more than one of these processes. For example, studies on the mammalian RBP Nova using a combination of crosslinking and immunopreciptation (CLIP)-seq and functional studies revealed that Nova not only regulates alternative splicing, but also modulates poly(A) site usage[3]. Moreover, in contrast to regulation at the transcriptional level, post-transcriptional regulatory steps are often carried out in different sub-cellular compartments of the nucleus (e.g. nucleoli, nuclear speckles, paraspeckles, coiled bodies, etc.) and/or cytoplasm (e.g. P-bodies, endoplasmic reticulum, etc.) by RBPs that are enriched within these compartments. These regulatory roles are essential for normal human physiology, as defects in RBP function are associated with diverse genetic and somatic disorders, such as neurodegeneration, auto-immune defects, and cancer[4-10].

Traditionally, RBPs were identified by affinity purification of single proteins[11,12]. However, several groups have recently used mass spectrometry-based methods to identify hundreds of proteins bound to RNA in human and mouse cells[13-16]. Recent censuses conducted by us and others indicate that the human genome may contain between 1,072[(ref. 17)] and 1,542[(ref. 1)] RBP-encoding genes. This large repertoire of RBPs likely underlies the high complexity of post-transcriptional regulation, motivating concerted efforts to systematically dissect the binding properties, RNA targets and functional roles of these proteins.

The dissection of RBP-RNA regulatory networks therefore requires the integration of multiple data types, each viewing the RBP through a different lens. *In vivo* binding assays such as CLIP-seq provide a set of candidate functional elements directly bound by each RBP. Assessments of *in vitro* binding affinity help understand the mechanism driving these interactions, and (as we show) improve identification of functional associations. Functional assays that identify targets whose expression or alternative splicing is responsive to RBP perturbation can then fortify evidence of function. For example, observation of protein binding by CLIP-seq within introns flanking exons whose splicing is sensitive to RBP levels in RNA-seq provides support for the RBP as a splicing factor and for the binding sites as splicing regulatory

elements. *In vivo* interactions of RBPs with chromatin can also be assayed to provide insight into roles of some RBPs as transcription regulators and can provide evidence for co-transcriptional deposition of RBPs on target RNA substrates. The regulatory roles of RBPs are also informed by the subcellular localization properties of RBPs and of their RNA substrates. Furthermore, these data resources comprised of multiple RBPs profiled using the same methodology may be integrated to enable the identification of factor-specific regulatory modules, as well as a factor's integration into broader cellular regulatory networks, through novel integrated analyses.

Results

Overview of data and processing

To develop a comprehensive understanding of the binding and function of the human RBP repertoire, we used five assays to produce 1,076 replicated datasets for 352 RBPs (Fig. 1, Supplementary Table 1). The RBPs characterized by these assays have a wide diversity of sequence and structural characteristics and participate in diverse aspects of RNA biology (Fig. 1). Functionally, these RBPs are most commonly known to play roles in the regulation of RNA splicing (96 RBPs, 27%), RNA stability and decay (70, 20%), and translation (69, 20%), with 161 RBPs (46%) having more than one function reported in the literature. However, 82 (23%) of the characterized RBPs have no known function in RNA biology other than being annotated as involved in RNA binding (Fig. 1). Although 57% of the RBPs surveyed contain well-characterized RNA binding domains (RNA recognition motif (RRM), hnRNP K homology (KH), zinc finger, RNA helicase, ribonuclease, double-stranded RNA binding (dsRBD), or pumilio/FBF domain (PUM-HD)), the remainder possess either less well studied domains or lack known RNA-binding domains altogether (Fig. 1). Each of the five assays used focused on a distinct aspect of RBP activity:

Transcriptome-wide RNA binding sites of RBPs: We identified and validated hundreds of immunoprecipitation-grade antibodies that recognize human RBPs[17] and developed enhanced CLIP (eCLIP) followed by sequencing to efficiently identify RNA targets bound by these RBPs at scale[18]. We performed eCLIP to profile 97 RBPs in K562 cells and 84 RBPs in HepG2 cells, for a total of 126 RBPs (of which 55 were characterized in both cell types). This effort identified 717,765 significantly enriched peaks (relative to size-matched input controls for each RBP) that cover 16.5% of the annotated mRNA transcriptome and 2.4% of the pre-mRNA transcriptome (Fig. 2d, Extended Data Fig. 2d).

RBP-responsive genes and alternative splicing events: To obtain insight into the functions of the RBP binding sites identified by eCLIP, we used shRNA- or CRISPR-mediated depletion followed by RNA-seq of 224 RBPs in K562 and 228 RBPs in HepG2 cells, for a total of 251 RBPs (of which 200 were characterized in both cell types). This effort identified 223,118 instances of RBP-mediated differential gene expression involving 14,281 genes affected upon knockdown of at least one RBP as well as 205,919 cases of RBP-mediated alternative splicing

events involving 37,829 alternatively spliced events impacted upon knockdown of at least one RBP.

In vitro RBP binding motifs: To identify the *in vitro* RNA sequence and structural binding preferences of RBPs, we developed a high-throughput version of RNA Bind-N-Seq (RBNS)[19] that assays binding of purified RBPs to pools of random RNA oligonucleotides. This effort identified the binding specificities of 78 RBPs. Short oligonucleotides of length *k*=5 (*k*mers) with high RBNS affinity clustered into a single motif for about half of the RBPs assayed (37/78). The remaining RBPs had more complex patterns of binding, best described by two motifs (32/78), or even three or more motifs (9 RBPs). These data also indicate that many RBPs are sensitive to the sequence and RNA structural context in which motifs are embedded.

RBP sub-cellular localization: Post-transcriptional gene regulation occurs in different intracellular compartments. For instance, rRNA maturation and pre-mRNA splicing primarily occur in sub-regions of the nucleus, whereas mRNA translation and default mRNA decay pathways operate in the cytoplasm. To illuminate functional properties of RBPs in intracellular space, we took advantage of our validated antibody resource[17] to conduct systematic immunofluorescence (IF) imaging of 274 RBPs in HepG2 cells and 268 RBPs in HeLa cells, in conjunction with a dozen markers for specific organelles and sub-cellular structures. These data, encompassing ~230,000 images and controlled vocabulary localization descriptors, have been organized within the RBP Image Database (http://rnabiology.ircm.qc.ca/RBPImage/).

RBP association with chromatin: Recent work has suggested that RBP association with chromatin may play roles in transcription and co-transcriptional splicing[20,21]. To generate a large-scale resource of chromatin association properties for RBPs, we identified the DNA elements associated with 58 RBPs in HepG2 cells and 45 RBPs in K562 cells by ChIP-seq, for a total of 63 RBPs (of which 34 RBPs were characterized in both cell types). These experiments identified 622,443 binding locations covering 3.1% of the genome.

To facilitate integrated analyses, all data for each data type were processed by the same data processing pipeline, and consistent, stringent quality control metrics and data standards were uniformly applied to all experiments. Although only 8 RBPs were investigated using all five assays, 249 of the 352 RBPs (71%) were studied using at least two different assays and 129 (37%) were subjected to at least three different assays, providing opportunities for integrated

analysis using multiple datasets. As an example of how these complementary datasets provide distinct insights into RNA processing regulation, we considered the Vascular Endothelial Growth Factor A gene (*VEGFA*) (Fig. 2a). Although the RBPs IGF2BP1 and IGF2BP3 both showed significant binding to the *VEGFA* 3'UTR, knockdown of IGF2BP1 increased *VEGFA* mRNA levels, while knockdown of IGF2BP2 decreased them, pointing to opposing regulatory effects. Knockdown of HNRNPK also yielded decreased *VEGF* mRNA, as well as a significant decrease in inclusion of exon 6. This splicing event is likely directly regulated by HNRNPK, as the flanking introns contain multiple hnRNP K eCLIP peaks, some of which contain the preferred binding motif for HNRNPK (GCCCA) identified in RBNS experiments. Similar integrated analysis can provide insight into mechanisms of cryptic exon repression, illustrated by HNRNPL binding to a region downstream of a *GTPBP2* cryptic exon that contains repeats of HNRNPL's top RBNS motif, contributing to production of *GTPBP2* mRNA with a full-length open reading frame (Extended Data Fig. 1).

The scale of data available enabled us to query the degree to which we have saturated the discovery of RBP binding sites on RNA and RBP-associated RNA processing events. In total, 14,281 genes were differentially expressed in at least one knockdown experiment (Fig. 2b), including 72.8% of genes expressed in both cell types and 71.2% of those expressed in at least one of the two (Extended Data Fig. 2a-b). Similarly, 11,211 genes were bound in at least one eCLIP dataset, representing 86.1% of genes expressed in both cell types and 75.6% of those expressed in at least one (Fig. 2b, Extended Data Fig. 2a-b). However, only 2,827 genes were both bound by and responsive to knockdown of the same RBP, suggesting that a large fraction of knockdown-responsive expression changes result from indirect effects and that many binding events do not directly affect gene expression levels in the conditions assayed here (Fig. 2b, Extended Data Fig. 2a-b). Similar analysis of alternative splicing changes revealed that differentially spliced events were saturated to a lesser degree than differentially expressed genes. (Greater variability in the cumulative number of unique events resulted from the large number of splicing changes in one knockdown dataset, the RNA helicase and spliceosomal protein AQR[22] (Extended Data Fig. 2c).)

Considering RBP binding, we observed a total of 23.4 Mb of annotated pre-mRNA transcripts covered by at least one reproducible RBP binding site, representing 9.1 Mb of exonic and 13.4 Mb of intronic sequence (Fig. 2c). This total represents only 1.5% of annotated intronic sequence (1.1% of distal intronic, 2.5% of proximal intronic, and 8.9% of splice site), but 16.5% of annotated exonic sequences (22.9% of 5' UTR, 19.3% of CDS, and 11.4% of 3' UTR, respectively) were covered by at least one peak (Fig. 2d). Restricting our analysis to only genes

expressed (TPM>1) in both cell types resulted in substantial increases in these percentages (Extended Data Fig. 2d). We found that profiling a new RBP often resulted in greater increases in covered bases of the transcriptome than did re-profiling the same RBP in another cell line, with the marginal 181st dataset averaging a 0.38% (for newly profiled RBPs) or 0.32% (for RBPs profiled in a second cell type) increase (Extended Data Fig. 2e-g). We additionally observed an average 1.4% increase in covered bases when we added eCLIP datasets from H1 and H9 embryonic stem cells for RBFOX2 and IGF2BP1-3 (all of which were also profiled in either K562 or HepG2), suggesting that substantial numbers of additional RBP binding sites remain to be detected in cell-types distinct from K562 and HepG2 (Extended Data Fig. 2g).

Although these results correspond well with previous work suggesting that RNAs are often densely coated by RBPs[23], it remains to be seen what fraction of these peaks mark alternative regulatory interactions versus constitutive RNA processing. Indeed, many may mark relative association of proteins which coat or broadly interact with RNAs as part of their basic function (such as association of RNA Polymerase II component POLR2G with pre-mRNAs, or spliceosomal component association with splice sites).

In vivo **binding is largely determined by** ***in vitro*** **binding specificity**

Binding of an RBP to RNA *in vivo* is determined by the combination of the protein's intrinsic RNA binding specificity and other influences such as RNA structure and protein cofactors. To compare the binding specificities of RBPs *in vitro* and *in vivo*, we calculated the enrichment or *R* value of each 5mer in RBNS-bound sequences relative to input sequences and compared it to the corresponding enrichment of the 5mer in eCLIP peaks relative to randomized locations in the same genes (R_{eCLIP}). Significantly enriched 5mers *in vitro* and *in vivo* were mostly in agreement, with 17 of the 26 RBPs having significant overlap in the 5mers that comprise their motif logos (Fig. 3a, left). The top RBNS 5mer for an RBP was almost always enriched in eCLIP peaks (Fig. 3a, center). In most cases, similar degrees of enrichment and similar motif logos were observed in eCLIP peaks located in coding, intronic or UTR regions, suggesting that RBPs have similar binding determinants in each of these transcript regions (Fig. 3a, center; Extended Data Fig. 3a). Strikingly, the single most enriched RBNS 5mer occurred in 30% or more of the peaks for several RBPs including SRSF9, TRA2A, RBFOX2, PTBP3, TIA1, and HNRNPC. For most RBPs, at least half of eCLIP peaks contained one of the top five RBNS 5mers. Therefore, instances of these 5mers provide candidate nucleotide-resolution binding locations for the RBP

(Fig. 3a, right). Such precise binding locations have a number of applications, e.g., they can be intersected with databases of genetic variants to identify those likely to alter function at the RNA level. When two or more distinct motifs were enriched in both RBNS and eCLIP, the most enriched motif *in vitro* was usually also the most enriched *in vivo* (5 out of 7 cases). These observations are consistent with the idea that intrinsic binding specificity observed *in vitro* explains a substantial portion of *in vivo* binding preferences for most RBPs studied to date.

For a minority of RBPs (12/26), the top five RBNS 5mers explained less than half of eCLIP peaks. Some of these factors appear to have affinities to RNA structural features or to more extended RNA sequence elements not well represented by 5mers (Dominguez et al., in preparation), while the sequence-specific binding of others may be driven to a large extent by cofactors. In some cases, RBNS revealed affinity to a subset of the motifs that were enriched in eCLIP peaks. For example, C-rich 6mers were most enriched in RBNS data for PCBP2 and also in PCBP2 eCLIP peaks (Fig. 3b). In this example, and in several others, a subset of similar eCLIP-enriched *k*mers were not enriched at all by RBNS (e.g., the G-rich 6mers in Fig. 3b). Such “eCLIP-only” motifs, which were often G-, GC-, or GU-rich (Extended Data Fig. 3b), may represent RNA binding of cofactors in complex with the targeted RBP (e.g., G-rich motifs enriched near RBFOX2 peaks may represent sites bound by HNRNPF, HNRNPH and HNRNPM in complex with RBFOX2)[24,25], or could represent biases in crosslinking or in genomic sequences near eCLIP peaks[26,27].

The extent to which strength and mode of binding correlates with eCLIP read density and regulatory activity is not well understood. We focused on regulation of splicing because a large proportion of the available cell type/RBP combinations that included KD, eCLIP, and RBNS data involved RBPs with known roles in splicing, and splicing changes could be readily detected in the KD/RNA-seq data. For most datasets involving KD of known splicing RBPs (18/28), eCLIP binding to one or more specific regions near alternative exons was associated with increased splicing changes upon KD of the RBP. In contrast, this association was observed for only one of the seven datasets involving RBPs that lacked known splicing functions (Fisher exact $P<0.05$, Extended Data Fig. 4a). To explore the relationship between sequence-specific binding and regulation, we classified eCLIP peaks as RBNS+ or RBNS– depending on whether they contained the highest-affinity RBNS motif (Supp. Methods). We then asked whether these classes of peaks differed in their association with splicing regulation. Examining exon-proximal regions commonly associated with splicing regulation, we found that RBNS+ eCLIP peaks conferred stronger regulation of exon skipping – with an average ~25% increase in $|\Delta\Psi|$ – than did RBNS– peaks (Fig. 3c). Thus, sequence-specific binding appears to confer stronger

regulation than non-sequence-specific binding, and RBNS motifs can be used to distinguish a subset of eCLIP peaks that has greater regulatory activity. The *in vitro* data were needed to make this distinction, because a similar analysis of eCLIP peaks classified by presence/absence of the top eCLIP-only 5mer exhibited minimal differences in splicing regulatory activity (Extended Data Fig. 4b). In general, RNA binding directed by intrinsic RNA affinity may involve longer-duration interactions that more consistently impact recruitment of splicing machinery.

Functional Characterization of the RBP Map

Analysis of the knockdown/RNA-seq data enables inference of the function of some RNA elements identified by eCLIP and RBNS. Regulation of RNA stability, which alters steady-state mRNA levels, can be observed by an increase or decrease in mRNA expression upon knockdown of an RBP. We compared differentially expressed genes upon RBP knockdown with eCLIP binding to three regions of mRNAs: 5'UTR, CDS, and 3'UTRs. We observed that binding of 17 RBPs correlated with increased expression upon knockdown, including RBPs with previously identified roles in induction of RNA decay (such as UPF1, XRN2, and DDX6) (Fig. 4a, Extended Data Fig. 5a) as well as previously unknown candidates including METAP2, a methionyl aminopeptidase which has been co-purified with polyA-selected RNA but has no previously known RNA processing roles[14]. Although METAP2 was highly bound throughout the transcriptome, CDS regions were on average 3.4-fold enriched in METAP2 eCLIP, well above the 2.4-fold average enrichment of 3'UTR and 1.4-fold depletion of intronic regions (Extended Data Fig. 5b-c). We further observed a trend in which increasing METAP2 eCLIP fold-enrichment correlated with progressively stronger increases in RNA expression upon knockdown, supporting an RNA regulatory role (Fig. 4b).

Additionally, we observed 11 RBPs for which binding correlated with decreased RNA levels following knockdown (Fig. 4a), including the stress granule component TIA1. Surprisingly, although our transcriptome-wide analysis indicated that transcripts with 3' UTR binding of TIA1 decreased upon knockdown in K562 cells (suggesting a globally stabilizing role for TIA1) (Fig. 4c), little to no stabilization activity was observed for 3' UTR-bound mRNAs in HepG2 cells (Extended Data Fig. 5d). Using TIA1 RBNS motif content in 3' UTRs rather than eCLIP binding sites, we additionally observed cell-type specific enrichment of TIA1 motifs in destabilized transcripts upon KD in K562, with no significant effect (though a slight motif enrichment in stabilized genes upon KD) in HepG2 (Extended Data Fig. 5e-g). This distinction is reminiscent of previous studies, which indicate that TIA1 can either induce RNA decay when tethered to a 3'

UTR[28], or stabilize target mRNA levels through competition with other RBPs including HuR[29]. Thus, our results provide further evidence that TIA1 can regulate mRNA stability through dynamic cell type-specific interactions.

RBP association with splicing regulation

Next, we considered how localized RBP binding was associated with splicing regulation. To do this, we generated a 'splicing map' for each RBP[30], which depicts the average RBP binding at relative positions in the regulated exon and flanking introns of exons that show RBP-responsive splicing changes, relative to binding at non-responsive alternative exons (Fig. 5a, Extended Data Fig. 6). Considering 57 RBPs with eCLIP and knockdown/RNA-seq performed in the same cell line, we observed a wide variety of RNA maps (Fig. 5b). Binding of SR proteins typically correlated with decreased inclusion upon knockdown and hnRNP protein binding correlated with increased inclusion upon knockdown, consistent with classical models of antagonistic effects of SR and hnRNP proteins on splicing[31] (Extended Data Fig. 7a-b).

Surprisingly, in addition to canonical alternative splicing regulators such as RBFOX2 and PTBP1, we observed significant differential association of spliceosomal components near knockdown-sensitive alternative splice sites (Fig. 5b, Extended Data Fig. 7c-d). For example, we observed that comparing cassette exons to constitutive exons revealed strikingly different binding patterns, with cassette exons characterized by increased association of 5' splice site machinery such as PRPF8, EFTUD2, and RBM22 at the upstream 5' splice site but depletion at the cassette exon 5' splice site (Extended Data Fig. 7c). Branch point recognition factors such as SF3B4 and SF3A3 showed similar depletion for the cassette exon branch point and enrichment at the downstream branch point, leading to a distinctive pattern of spliceosomal association at the cassette exon (Extended Data Fig. 7c). Moreover, when considering non-spliceosomal RBPs we similarly observed that while RBP association was higher at cassette-bordering proximal intron regions relative to constitutive exons, the upstream 5' splice site and proximal intron showed an even greater enrichment (Extended Data Fig. 7d). Together, these observations for non-spliceosomal RBPs suggest that the upstream 5' splice site of alternative exons represent a greater source of regulatory RBP binding than previously believed.

Splicing maps were also constructed for alternative 5' (A5SS) and alternative 3' splice site (A3SS) events (Fig. 5c). Again, we noticed differential association of spliceosomal components (Fig. 5c, Extended Data Fig. 8a). Focusing on A3SS events, we noted a particularly prominent pattern of association for branch point factors SF3B4 and SF3A3, which typically bind

to the branch point region ~50 nt upstream of the 3' splice site. For both, we observed that enrichment was shifted downstream towards the 3' splice site for the set of A3SS events where the distal (downstream) 3' splice site is preferred upon knockdown (Fig. 5c-d, Extended Data Fig. 8b). These genome-level results generalize previous mini-gene studies showing that 3' splice site scanning and recognition originates from the branch point and can be blocked if the branch point is moved close to the 3' splice site AG[32], and suggest that regulated branch point recognition plays a key role in A3SS regulation by restricting recognition by the 3' splice site machinery[33] (Extended Data Fig. 8c).

In summary, the RBPs we have surveyed that participate in alternative splicing display a wide diversity of regulatory modes. Moreover, although the target splicing events differ, the splicing map of a given RBP is highly consistent between cell types (Fig. 5b,c, Extended Data Fig. 7a,b, Extended Data Fig. 8a). Thus, performing eCLIP and KD/RNA-seq in a single cell type may be sufficient to elucidate the splicing map for an RBP, but multiple cell types must be surveyed to identify the full repertoire of direct regulatory events.

RBP Association with Chromatin

Increasing evidence suggests that regulatory RNAs, both coding and non-coding, are broadly involved in gene expression through their interaction with chromatin[34,35]. Based on the expectation that these regulatory events must enlist specific RBPs participating in co-transcriptional RNA processing, we surveyed 58 RBPs in HepG2 and 45 RBPs in K562 cells for their association with chromatin by ChIP-seq (Supplementary Table 1). We selected RBPs for analysis based on their complete or partial localization in the nucleus and on the availability of antibodies that could efficiently immunoprecipitate each factor. These RBPs belong to a wide range of functional categories, including SR and hnRNP proteins, spliceosomal components and RBPs that have been generally considered to function as transcription factors, such as POLR2G and GTF2F1. Interestingly, 30 of 58 RBPs (52%) in HepG2 cells and 29 of 45 RBPs (64%) in K562 cells showed specific interactions with chromatin, with several hundred to more than ten thousand specific binding peaks identified for each RBP.

We next characterized the RBP-chromatin interactions with respect to established chromatin activities, including DNase I hypersensitive sites and various histone marks (Fig. 6a). This analysis revealed a general preference of RBPs for euchromatic relative to heterochromatic regions, especially gene promoters, although individual RBPs showed distinct preferences. Collectively, even this moderately sized set of RBPs occupied ~30% of all DNase

hypersensitive or open chromatin regions and ~70% of annotated gene promoters, which was similar between the two cell types, suggesting broad involvement of RBPs in chromatin activities in the human genome. These data provide a foundation for future investigation of direct roles of specific RBPs in transcriptional control, exemplified by recent studies on SR proteins[21] and RBFOX2[36].

It is well established that a variety of RNA processing events are co-transcriptionally coupled[20]. It is therefore possible that some RBP-chromatin association events are coupled with their direct RNA binding activities in cells. To explore this possibility, we intersected ChIP-seq peaks with eCLIP peaks for individual RBPs in HepG2 cells, revealing specific RBPs with relatively high degrees of overlap in these two types of interactions (Fig. 6b). Interestingly, the distribution of the overlapped regions varies among individual RBPs. For example, the chromatin and RNA binding activities of SR proteins, as exemplified by SRSF1, are predominantly coincident near gene promoters, while most other RBPs show such overlapping activities further into the gene body (Fig. 6c). These data revealed coordinated actions of RBPs at the chromatin and RNA levels, suggesting that a fraction of their RNA binding events are chromatin-associated. We next sought to quantify the similarity of each RBP pair's chromatin and RNA binding activities by computing the overlap in ChIP-seq and eCLIP-seq signal in non-promoter regions (Fig. 6d). Clustering of the data indicated that many RBPs might function in conjunction with other RBP(s) to coordinate their chromatin and RNA binding activities (red signals in Fig. 6d). Particularly interesting is the high correlation between poly(rC) binding proteins HNRNPK and PCBP1/2, which share a common evolutionary history and domain composition yet perform diverse functions[37] (red box in Fig. 6d). To illustrate the relationship between RNA and chromatin interactions, we plotted the ChIP-seq density of these three RBPs relative to PCBP2 and HNRNPK eCLIP peaks in non-promoter regions. We found that chromatin binding signals were typically centered around eCLIP peaks, although HNRNPK and PCBP1 to a lesser degree were slightly biased for chromatin binding upstream of RNA binding indicative of a specific topological arrangement of these potential RBP complexes on chromatin in a manner dependent on the direction of transcription (Fig. 6e, top panels). We next asked whether the signal overlapping eCLIP peaks was specific to the DNA binding activity of these RBPs or if it also held for related RNA binding activities. For this purpose, we plotted the eCLIP density of multiple RBPs relative to another RBP's eCLIP peaks. This analysis revealed a significant degree of overlap among the RNA binding activities of HNRNPK and PCBP2, which contrasts to a randomly selected RBP for comparison (Fig. 6e, bottom panels). Combined,

these data strongly reinforce coordinated chromatin and RNA binding activities of specific RBPs.

To investigate whether there is correlation between gene expression and RBP association, we clustered the probability of each RBP's association with genes binned by increasing expression levels, revealing a pervasive positive correlation between gene expression and RBP association for the majority of RBPs with the exception of SAFB2 and SFPQ which have only a few binding sites (Fig. 6f, left panels). These data implicated regulatory roles of RBPs in gene expression through chromatin association. To pursue this point, we compared the frequency that a gene is differentially expressed upon RBP knockdown depending on whether or not the RBP was chromatin-associated at that gene, controlling for the different expression levels of these two groups. We found that chromatin association of HNRNPLL, HNRNPL, HNRNPK, and AGO1 correlated with a significantly increased chance of gene expression change upon knockdown (Fig. 6f, right panels). To explore the connection between alternative splicing regulation and RBP-chromatin association, we calculated similar ratios for each RBP and found that RBM22 chromatin association was associated with a significant increase in alternative splicing changes upon knockdown (Fig. 6f, right panels). Together, these data suggest that chromatin-association of RBPs affects RNA processing and gene expression.

RBP regulatory features in subcellular space

The systematic imaging screen revealed that RBPs display a broad diversity of localization patterns (Fig. 7a), with most factors exhibiting targeting to multiple structures in the nucleus and cytoplasm (Fig. 7b). Next, we integrated RBP localization data with other datasets generated here. First, we considered the nuclear relative to cytoplasmic ratios for each RBP. As expected, we observed a significant shift towards increased binding to unspliced transcripts for nuclear RBPs, whereas cytoplasmic RBPs were enriched for binding to spliced transcripts (Extended Data Fig. 9a-b). Next, we identified a collection of 80 RBPs that exhibit robust localization to SC35 (SRSF2)-labeled nuclear speckles, a class of subnuclear structures enriched for proteins involved in pre-mRNA splicing[38]. Consistent with the established function of speckles in splicing, analysis of splicing changes associated with RBP depletion revealed that speckle-localized RBPs impact larger numbers of splicing events compared to non-speckle proteins (Fig. 7c). Notably, the top 12 (and 19 of the top 20) RBP knockdowns with the strongest impact on splicing (in terms of number of altered splicing events) corresponded to speckle-localized RBPs, including spliceosomal components U2AF1, U2AF2, SF3B4, and SF3A1 as

well as splicing regulators HNRNPK and SRSF1 (Fig. 7c). By contrast, nucleolar RBPs had significantly less impact on pre-mRNA splicing than non-nucleolar RBPs, as expected (Fig. 7c). Furthermore, when we queried the enrichment for RNA binding modalities in eCLIP data, we observed similarly striking enrichment for speckle-localized RBPs binding to proximal introns as well as snRNAs (notably *RNU2* and *RNU6*) (Extended Data Fig. 9c), again consistent with nuclear speckles playing a key role in pre-mRNA splicing.

Focusing on localization to specific cytoplasmic organelles, we noted that 42 RBPs exhibited localization to mitochondria, an organelle with unique transcriptional and RNA processing regulation[39]. We observed numerous RBPs with significant binding to the mitochondrial genome in either Heavy (H)-strand (TBRG4), Light (L)-strand (GRSF1), or both (FASTKD2, DHX30) (Fig. 7d-e). Validating these observations, mitochondrial localization by immunofluorescence was associated with significantly increased eCLIP binding to mitochondrial RNA (Extended Data Fig. 9d). DHX30 is essential for proper mitochondrial ribosome assembly and oxidative phosphorylation[40]. Intriguingly, in addition to widespread association with many mitochondrial genes that was consistent with previous RIP-seq findings[40] (Extended Data Fig. 9e), we observed dramatic enrichment at an unannotated region on the mitochondrial H-strand downstream of all annotated genes and just upstream of the replication D loop, which has strong potential to form a stem-loop structure (Fig. 7f). We further observed that DHX30 knockdown resulted in increased expression of nearly all H-strand transcripts, but decreased expression of L-strand transcript ND6 (Fig. 7g). Identification of the termination signal for mitochondrial H-strand transcription has remained elusive; it is tempting to speculate this site of DHX30 association could mark such signal. These examples illustrate how intracellular localization of RBPs can be used synergistically with binding and loss-of-function studies to infer aspects of post-transcriptional regulation that occur in different cellular compartments and organelles.

Preservation of RBP regulation across cell types

Next we evaluated whether RBP regulation is preserved across cell types. Analyzing the 55 RBPs profiled by eCLIP in both K562 and HepG2 cell lines, we found that only a small portion (averaging 11.0%) of peaks were found in genes with cell type-specific expression (TPM ≥ 1 in one cell type and TPM < 0.1 in the other) (Fig. 8a, Extended Data Fig. 10a). An average of 70.6% of peaks were located within genes expressed in both cell types (TPM ≥ 1 in both), although only 10.3% were present in genes that were unchanged between K562 and HepG2 (defined as fold-difference ≤ 1.2), and 31.6% of peaks were within genes that changed by at

least two-fold (Fig. 8a, Extended Data Fig. 10a). To illustrate, we observed that RBFOX2 eCLIP peaks at least 8-fold enriched in HepG2 were typically also enriched in K562 (average enrichment of 6.1-fold) if the bound gene was expressed within five-fold of the level in HepG2 cells (Covering the “Unchanged”, “Weakly differential”, and “Moderately differential” classes in Fig. 8a). In contrast, 89.8% of HepG2 peaks in genes with cell type-specific expression were not enriched in K562 (Fig. 8a). Indeed, 49.7% of RBFOX2 HepG2 peaks that were not enriched in K562 occurred in genes with cell type-specific expression whereas only 5.5% occurred in genes within a two-fold change.

Expanding this analysis to all RBPs profiled in both cell types, we observed similar results: an average of 84.3% of peaks in genes with cell type-specific expression were not enriched in the other cell type, whereas 68.0%, 65.5%, and 67.6% of peaks in unchanging, weakly, or moderately differentially expressed genes were enriched by at least 4-fold in the second cell type, respectively (Fig. 8b). Requiring the strict IDR thresholds for peak identification, an average of 44.1% of peaks in genes with similar expression were preserved across cell types (Extended Data Fig. 10c). 48.9% of all peaks that showed no enrichment in the second cell type occurred in genes with cell type-specific expression (a 4.4-fold enrichment), whereas only 21.0% occurred in unchanging, weakly, or moderately differentially expressed genes, respectively (a 3-fold depletion) (Extended Data Fig. 10c). Thus, these results suggest that most RBP binding is preserved across cell types for similarly expressed genes, and that much of differential eCLIP signal between HepG2 and K562 likely reflects underlying gene expression differences more than differential binding.

Next, we asked whether an RBP's positional pattern of splicing regulation tended to be conserved across cell types. Considering splicing maps of cassette exons, we observed that binding of many RBPs had highly similar correlations with either inclusion or exclusion upon knockdown, including alternative splicing regulators HNRNPL and SRSF1 (Fig. 8c-d). We observed that the splicing maps for the same RBP across cell types had significantly higher correlation than random pairings of RBPs, when comparing across all 16 RBPs with both eCLIP and RNA-seq datasets (with sufficient splicing changes) in both K562 and HepG2 (Fig. 8d, Extended Data Fig. 10d).

Cell type-specific regulation of RNAs may also be achieved through differential modulation of RBP levels. To assess which RBPs confer regulation in such a manner, we calculated the expression of each RBP in HepG2 and K562 cells in addition to 31 diverse human tissues from the GTEx Project (Extended Data Fig. 11). Many RBPs had high expression in ENCODE cell lines and across a broad range of human tissues, including

ribosomal proteins (RPL23A, RPS11, RPS24), translation factors (EIF4H, EEF2), and ubiquitously expressed splicing factors (HNRNPC, HNRNPA2B1) among the 10 least tissue-specific RBPs (Fig. 8e). However, several other RBPs had highly tissue-specific expression exhibiting either a pattern of high expression in one or a small number of human tissues (e.g., LIN28B, IGF2BP1/3) or were differentially expressed by orders of magnitude across several human tissues (e.g., IGF2BP2 and APOBEC3C), indicating that the RNA targets and regulatory activity of these RBPs are likely modulated through cell type-specific gene expression programs. Of course, even RBPs with similar mRNA levels across tissues may have different protein levels or activity because of post-translational modifications, which are widespread among RBPs.

Discussion

Our study represents the largest effort to date to systematically study the roles of human RBPs by integrative approaches. The resulting catalog of functional RNA elements substantially expands the repertoire of regulatory components encoded in the human genome. While the impact of DNA binding proteins mostly culminates in effects on gene expression levels, RBP function encompasses a broader range of activities. RBP functions extend outside the nucleus and into the cytoplasm and organelles, and consist of multiple paths by which RNA substrates are altered (splicing, RNA editing/modification, RNA stability, localization, translation), expanding transcriptome and proteome complexity. We demonstrate the effectiveness of combining *in vivo* maps of RNA binding sites using eCLIP with orthogonal approaches, such as *in vitro* evaluation of RNA affinity for the same RBPs, chromatin association by ChIP-seq, and functional assessment of RNA variation by depletion experiments and RNA-seq. At the molecular level, we confirm that *in vivo* and *in vitro* preferences are highly correlated for RBPs interrogated here, and show that CLIP peaks containing motifs reflective of intrinsic RNA affinity are more predictive of regulation. We confirm using unbiased genome-wide analyses that SR and hnRNP proteins have broadly antagonistic effects on alternative splicing. Moreover, we extend genome-wide previous findings that alternative 3' splice site choice results from an "AG" scanning process initiating with branch point recognition, and implicate the upstream 5' splice sites of cassette exons in splicing regulation. We also implicate RNA structures bound by an RBP in processing of mitochondrial transcripts, and elucidate new RNA splicing maps for many RBPs. Furthermore, our data provide the first systematic investigation of chromatin-associated gene regulation and RNA processing at the level of RBP-nucleic acid interactions. At the cellular

level, immunohistochemical analysis with our extensive repository of RBP-specific antibodies place these molecular interactions within subcellular contexts. We confirm localization of many RBPs to nuclear speckles, mitochondria and other compartments, and identify many new proteins resident to these sites, emphasizing the necessity of localization data for interpreting RBP-RNA regulatory networks.

Here, we have surveyed the *in vivo* binding patterns of 126 RBPs, comprising roughly 10% of the human genes predicted to encode proteins that interact directly with RNA. Our observation that additional mapping of new RBPs leads to a greater increase in the functional RNA element coverage than mapping of the same RBPs in additional cell lines argues that expansion of these approaches to additional RBPs will be particularly informative. Additionally, the binding modes *in vivo* are highly preserved across genes expressed similarly in our two cell lines assayed (K562 and HepG2). Nevertheless, mapping of previously characterized RBPs in drastically different cell types (embryonic stem cells, post-mitotic cells such as neurons and muscle cells) and human tissues, with highly distinct transcriptomes, will undoubtedly yield new discoveries. Additionally, RNA processing is highly dynamically regulated during acute or chronic environmental influences such as stress, as new binding sites may arise from both environmental changes in RBP or RNA concentrations as well as from changes in post-translational modifications, binding partners, or subcellular distribution of RBPs. Thus, studying RBP subcellular localization and RBP-RNA substrate regulation under these conditions has potential to reveal new biology.

We expect that the data generated here will provide a useful framework upon which to build analyses of other aspects of RNA regulation, such as microRNA levels, RNA editing and modifications such as pseudouridylation and m6A methylation, translation efficiency, and mRNA half-life measurements. We have yet to integrate *in vivo* RNA structure probing data to evaluate how RBP-mediated RNA processing are influenced by local[41] and long-range RNA structures[42]. As we continue to embark on comprehensively characterizing all functional RNA elements, genome-scale CRISPR/Cas9-inspired genome-editing[43] and RNA modulation[44] technologies will ultimately provide opportunities to study the impact on cellular and organismal phenotypes resulting from disruption of these RNA elements.

Methods

RNA binding protein annotations and domains

RBPs were chosen from a previously described list of 1072 known RBPs, proteins containing RNA binding domains, and proteins characterized as associated with polyadenylated RNA, based on the availability of high quality antibodies[17]. Annotation of RBP function was performed by integration of published literature, with manual inspection of references for less well-established annotations. Annotation of RNA binding domain presence was determined by UniProt Domain Descriptions, and a database of cell-essential genes was obtained from published high-throughput CRISPR screening efforts[45].

eCLIP - experimental methods

Antibodies for eCLIP were pre-screened using a set of defined metrics[17]. A 'biosample' of HepG2 or K562 cells was defined as a batch of cells starting from a single unfrozen stock, passaged for less than 30 days under standard ENCODE reference conditions, and validated for high viability and non-confluent at the time of crosslinking. All cells within a biosample were pooled and UV crosslinked on ice at 400 mJoules/cm2 with 254 nm radiation. The biosample was then split into 20 million cell aliquots for eCLIP experiments.

eCLIP experiments were performed as previously described in a detailed Standard Operating Procedure[18], which is provided as associated documentation with each eCLIP experiment on the ENCODE portal (https://www.encodeproject.org/documents/fa2a3246-6039-46ba-b960-17fe06e7876a/@@download/attachment/CLIP_SOP_v1.0.pdf). Briefly, 20 million crosslinked cells were lysed and sonicated, followed by treatment with RNase I (Thermo Fisher) to fragment RNA. Antibodies were pre-coupled to species-specific (anti-Rabbit IgG or anti-Mouse IgG) Dynabeads (Thermo Fisher), added to lysate, and incubated overnight at 4 deg C. Prior to immunoprecipitation (IP) washes, 2% of sample was removed to serve as the paired input sample. For IP samples, high- and low-salt washes were performed, after which RNA was dephosphorylated with FastAP (Thermo Fisher) and T4 PNK (NEB) at low pH, and a 3' RNA adapter was ligated with T4 RNA Ligase (NEB). 10% of IP and input samples were run on an analytical PAGE Bis-Tris protein gel, transferred to PVDF membrane, blocked in 5% dry milk in TBST, incubated with the same primary antibody used for IP (typically at 1:4000 dilution), washed, incubated with secondary HRP-conjugated species-specific TrueBlot antibody (Rockland), and visualized with standard enhanced chemiluminescence imaging to validate successful IP. 90% of IP and input samples were run on an analytical PAGE Bis-Tris protein gel

and transferred to nitrocellulose membranes, after which the region from the protein size to 75 kDa above protein size was excised from the membrane, treated with Proteinase K (NEB) to release RNA, and concentrated by column purification (Zymo). Input samples were then dephosphorylated with FastAP (Thermo Fisher) and T4 PNK (NEB) at low pH, and a 3' RNA adapter was ligated with T4 RNA Ligase (NEB) to synchronize with IP samples. Reverse transcription was then performed with AffinityScript (Agilent), followed by ExoSAP-IT (Affymetrix) treatment to remove unincorporated primer. RNA was then degraded by alkaline hydrolysis, and a 3' DNA adapter was ligated with T4 RNA Ligase (NEB). qPCR was then used to determine required amplification, followed by PCR with Q5 (NEB) and gel electrophoresis to size-select the final library. Libraries were sequenced on either the HiSeq 2000, 2500, or 4000 platform (Illumina). Each ENCODE eCLIP experiment consisted of IP from two independent biosamples, along with one paired size-matched input (sampled from one of the two IP lysates prior to IP washes).

eCLIP - data processing and peak identification

Data processing, including adapter trimming, repetitive element removal, unique genomic mapping, PCR duplicate removal, and peak calling versus paired size-matched input were performed as previously described in a detailed Standard Operating Procedure[18]. Unless otherwise noted, reproducible and significant peaks were identified by merging peaks identified in each replicate, requiring that the peak meet an irreproducible discovery rate cutoff of 0.01 as well as p-value ≤ 0.001 and fold-enrichment ≥ 8 (using the geometric mean of $\log_2$(fold-enrichment) and $-\log_{10}$(p-value) between the two biological replicates). For submission to the ENCODE portal, eCLIP datasets were required to pass several quality metrics, including minimum read depth, reproducibility, and peak saturation metrics. To quantify binding to snRNA and other multi-copy elements, a separate pipeline was developed to require unique mapping to element families instead of unique genomic positions. All analyses described in this manuscript used mapping to GRCh37 and GENCODE v19 annotations, but mapping to GRCh38 and GENCODE v24 annotations were also deposited at the ENCODE portal.

For analyses using binding considered at the level of regions (e.g. 3'UTR, CDS, or proximal intronic), read density was counted for the indicated region for both IP and paired input, and significance was determined by Fisher Exact test (or Yates' Chi-Square test if all observed and expected values were above 5). Only regions with at least 10 reads in one of IP or input, and where at least 10 reads would be expected in the comparison dataset given the total number of

usable reads, were considered, and significant regions were defined as those with fold-enrichment ≥ 4 and p-value ≤ 0.00001.

To summarize relative enrichment between IP and input, information was defined as the Kullback-Leibler divergence (relative entropy) $p_i * \log_2 (p_i / q_i)$, where p_i is the fraction of total reads in IP that map to a queried element i (peak, gene, or repetitive element), and q_i is the fraction of total reads in input for the same element.

Knockdown followed by RNA-seq (KD/RNAseq)— experimental methods

Individual RBPs were depleted from HepG2 or K562 cells by either RNA interference or CRISPR-mediated gene disruption. RNA interference was performed by transducing cells with lentiviruses expressing shRNAs (TRC collection) targeting an RBP of interest followed by puromycin selection for 5 days. CRISPR-mediated gene disruption was performed by transfecting cells with a plasmid expressing Cas9 and a gRNA targeting the RBP of interest, followed by puromycin selection for 5 days. In each case, depletions were performed in biological duplicate along with a pair of control depletions using a scrambled shRNA or gRNA. RNA was extracted from half of each sample and used to perform qRT-PCR to confirm depletion of the targeted RBP transcript and to prepare RNA-seq libraries using the Illumina Tru-seq stranded mRNA library preparation kit. Protein was extracted from the other half of each sample and used to confirm depletion of the target RBP by Western blotting. Paired-end 100 bp reads were generated from the RNA-seq libraries to an average depth of 63 million reads per replicate.

KD/RNA-seq - data processing

Reads were aligned to both the hg19 assembly of the genome using the Gencode v19 annotation, respectively, using both TopHat2[46] and STAR[47]. Gene expression levels were quantitated using RSEM[48] and Cufflinks[49] and differential expression levels determined using DESeq[50] and Cuffdiff2[51]. Alternative splicing was quantitated using rMATS[52] and alternative poly(A) site use quantitated with MISO[53]. All analyses described in this manuscript used mapping to GRCh37 and GENCODE v19 annotations, but mapping to GRCh38 and GENCODE v24 annotations were also deposited at the ENCODE portal.

RNA Bind-N-Seq (RBNS) - experimental methods

RBNS experiments were performed as indicated in the protocol included on each experiment at the ENCODE portal. Briefly, randomized RNA oligonucleotides (20 or 40 nt) flanked by constant adapter sequences were synthesized and incubated with an SBP-tagged recombinant RBP (consisting minimally of all annotated RNA binding domains) at several concentrations (typically five, ranging from 5-1300 nM). RNA-protein complexes were isolated with streptavidin-conjugated affinity resin and eluted RNA was prepared for deep sequencing, resulting in 15-20 million reads per RBP pulldown concentration with a similar number of input reads sequenced per *in vitro* transcription reaction.

RBNS - data processing

RBNS *k*mer enrichments (*R* values) were calculated as the frequency of each *k*mer in the pulldown library reads divided by its frequency in the input library; enrichments from the pulldown library with the highest individual *k*mer R value were used for each RBP. Mean and SD of *R* values were calculated across all *k*mers for a given *k* to calculate the RBNS Z-score for each *k*mer. RBNS pipeline source code is available at: https://bitbucket.org/pfreese/rbnspipeline.

RBNS motif logos were made using the following iterative procedure for *k*=5: the most enriched 5mer was given a weight equal to its excess enrichment over the input library (=*R*–1), and all occurrences of that 5mer were masked in both the pulldown and input libraries to eliminate subsequent counting of lower-affinity 'shadow' 5mers (e.g., GGGGA shifted by 1 from GGGGG). All enrichments were then recalculated on the masked read sets to obtain the resulting most enriched 5mer and its corresponding weight, with this process continuing until the enrichment Z-score (calculated from the original *R* values) was less than 3. All 5mers determined from this procedure were aligned to minimize mismatches to the most enriched 5mer, with a new motif initiated if the number of mismatches + offsets exceeds 2. The frequencies of each nucleotide in the position weight matrix, as well as the overall percentage of each motif, were determined from the weights of the individual aligned 5mers that went into that motif; empty unaligned positions before or after each aligned 5mer were given pseudocounts of 25% each nucleotide, and outermost positions of the motif logo were trimmed if they had >75% unaligned positions. To improve the robustness of the motif logos, the pulldown and input read sets were each divided in half and the above procedure was performed independently on each half; only 5mers identified in corresponding motif logos from both halves were included in the alignments to make the final motif logo. In Fig. 3a, only the top RBNS motif logo is shown if there were

multiple (all motifs displayed on the ENCODE portal within the "Documents" box of each experiment).

The RBP's top RBNS logos were clustered using Jensen-Shannon divergence (JSD)-based similarities computed by summing the score of the *j* overlapping positions between the motifs of RBP A and RBP B:

$$\sum_{\text{aligned position } i=1..j} I_{A_i} I_{B_i} (1-\sqrt{JSD(A_i,B_i)})$$

where $I_{A,i}$ and $I_{B,i}$ are the information content of motifs A and B at aligned position *i,* and A_i and B_i are the vectors of base frequencies at aligned position i in motifs A and B, respectively (with each A_i and B_i vector summing to 1). This score weights similarity more heavily positions with higher information content and greater numbers of aligned positions. This similarity score was computed for each possible overlap of the two logos (with at least four positions overlapping, i.e., $j \geq 4$), and the top score with its corresponding alignment offset was used. The matrix of these scores was normalized to the maximum score over all RBP pairs and clustered using the linkage function with centroid method in scipy.cluster.hierarchy to obtain the dendrogram shown in Fig. 3a.

Immuno-Fluorescence, Microscopy Imaging and Data Processing

HepG2 cells were seeded in Poly-L-Lysine coated 96-well clear bottom plates (Corning Inc; plate number 3882 half-area microplates), at a concentration of 2,000 cells per well in DMEM + 10% FBS. After 72h in standard growth conditions (i.e. 37°C and 5% CO_2), cells were fixed with 3.7% formaldehyde, permeabilized in PBS + 0.5% Triton X-100 and blocked in PBS + 0.2% Tween-20 + 2% BSA (PBTB), all conducted for 20 min at room temperature. Primary antibodies directed against specific RBPs (all rabbit antibodies) and marker proteins were subsequently applied to the cells at a final concentration of 2 µg/mL in PBTB and incubated overnight at 4°C. The cells were next washed 3 times for 10 min each in PBST and incubated with secondary antibodies (Alexa647 donkey anti-rabbit and Alexa488 donkey anti-mouse, both diluted 1:500 in PBTB) for 90 min at room temperature. After 3 PBTB washes, the cells were counter stained with DAPI for 5 min, washed 3 times in PBS and stored in PBS at 4°C. Subcellular marker antibodies and dilutions used are as follows: rat anti-Alpha Tubulin, MCA78G, 1:200 (Serotec, Bio-Rad); mouse anti-CD63, ab8219, 1:200 (Abcam); mouse anti-Coilin, GTX11822, 1:100 (GeneTex Inc); mouse anti-DCP1a, sc100706, 1:200 (Santa Cruz Biotechnology); mouse anti-Fibrillarin, ab4566, 1:200 dilution (Abcam); mouse anti-GM130, #610822, 1:200 (Becton Dickinson); mouse anti-KDEL, ENZSPA827D, 1:200 (Enzo Life Sciences); mouse anti-Phospho

Tyrosine, #9411S, 1:200 (NEB); mouse anti-PML, sc-966, 1:50 (Santa Cruz Biotechnology); mouse anti-SC35, GTX11826, 1:200 (GeneTex Inc). For staining with Mitotracker (Molecular Probes, M22426), cells were incubated with 100nM of dye in tissue culture media for 45 min at 37°C prior to fixation. For staining with Phalloidin (Sigma, P5282), cells were incubated with 50ug/ml of Phalloidin for 20 min prior DAPI staining.

Imaging was conducted on an ImageXpress Micro high content screening system (Molecular Devices Inc). For each RBP/marker combination, 10-20 high resolution images were acquired in the DAPI, FITC and Cy5 channels, using a 40x objective. Automated laser based auto-focusing and auto-exposure functions were employed for sample imaging, with exposure times ranging from 250-3000ms, 100-500ms and 50-100ms, for RBP, Marker and DAPI channels respectively. Raw unprocessed grayscale images from individual channels were acquired as high resolution TIF files of 726kb each. An in house Matlab script was developed to batch normalize image intensity values and add blue, green or red colors to the respective channels, which were subsequently merged as colour JPEG files. The final images were uploaded on a server accessible through the RBP Image Database website. A MySQL relational database (version 5.1.73) was implemented, along with a MyISAM storage engine, to store the images, data annotations and characteristics. A controlled vocabulary of descriptors was devised to document RBP subcellular localization features.

Image analysis to quantify nuclear/cytoplasmic staining ratios, or to assess the degree of RBP targeting to punctate subcellular structures (e.g. Cajal bodies, nuclear speckles, nuceloli, Golgi, P-bodies), was conducted using ‘Granularity’, ‘Colocalization’ and ‘Multi Wavelength Cell Scoring’ analysis modules from the MetaXpress v3.1 software (Molecular Devices Inc), according to manufacturer recommendations. For localization categories including microtubules, actin, cell cortex, ER, focal adhesions, mitochondria and mitotic apparatus, manual localization grading was conducted by ranking candidate RBPS as strongly or weakly co-localized with respective protein markers. The Circos plot of localization co-occurrance (Fig. 7b) was generated by drawing one line between every pair of categories for each RBP that shared both localization annotations. Nuclear annotations are indicated in purple, cytoplasmic in red, and lines between nuclear and cytoplasmic annotations are indicated in yellow.

ChIP-seq - experimental methods

Chromatin immunoprecipitation was implemented according to ChIP Protocol optimized for RNA-binding proteins (https://www.encodeproject.org/documents/e8a2fef1-580b-45ad-b29c-fffc3d527202/@@download/attachment/ChIP-seq_Protocol_for_RNA-Binding_Proteins_ENCODE_Fu_lab_RuiXiao.pdf). In brief, prior to coupling with RBP antibodies, magnetic beads were equilibrated by washing with ChIP dilution buffer and blocked with glycogen, BSA and tRNA in ChIP dilution buffer. 10-20 million HepG2 and K562 cells were crosslinked in 1% formaldehyde diluted in PBS for 20 minutes and then quenched by adding glycine. Cell nuclei were extracted by resuspending cell pellet with cell lysis buffer with occasional inversion. Nucleus pellets resuspended in nuclear lysis buffer were sonicated with Branson Sonifier cell disruptor. 95% of nuclear lysate diluted in the final concentration of 1% triton X-100, 0.1% sodium deoxycholate and 1X proteinase inhibitor cocktail was subjected to immunoprecipitation with antibody coupled beads and 5% of nuclear lysate was used as input chromatin. Stringent washes were performed before elution. Input and immunoprecipitated chromatin DNAs were recovered by decrosslinking, RNase A digestion, proteinase K treatment, phenol/chloroform extraction and precipitation with ethanol. Library construction was mainly followed the instruction of the Illumina Preparing Samples for ChIP Sequencing. Each library was barcoded for pooled sequencing. DNA Libraries between 200-400 bp were gel purified, quantified with Qubit and subjected to Illumina HiSeq 2000/2500 for sequencing.

ChIP-seq - data processing

Data processing was performed in accordance with ENCODE uniform transcription factor ChIP-seq pipeline (https://www.encodeproject.org/chip-seq/transcription_factor/) and using hg19 as the reference human genome. All datasets containing >10 million usable reads from each replicate, passing IDR, and generating >200 peaks were used for final analysis.

Integrated Analysis

Saturation Analysis

Saturation analysis of eCLIP and KD/RNA-seq data was performed by randomly shuffling the order of datasets 100 times, subsampling 1 through all datasets, and calculating the desired metrics. Gene level saturation analysis of RBP binding was calculated first by taking all unique genes that were bound by an IDR filtered peak in an eCLIP experiment. Then, each eCLIP experiment was iteratively added to the previous experiment, counting only unique genes in any experiment. Saturation analysis of differentially expression genes from KD/RNA-seq was similarly performed, based on differentially expressed genes identified with DESeq2. Genes

were identified as differentially expressed if they had a $|\log_2(\text{fold-change})| > 1$ and an adjusted p-value < 0.05 between knockdown and control. Alternative versions of this analysis used (Fig. 2b) all genes, (Extended Data Fig. 2a) only genes with TPM > 1 in HepG2 and K562 , or (Extended Data Fig. 2b) only genes with TPM > 1 in either HepG2 or K562, using average gene-level expression from two rRNA-depleted RNA-seq experiments in HepG2 (ENCODE accession ENCFF533XPJ, ENCFF321JIT) and K562 (ENCFF286GLL, ENCFF986DBN). The set of differentially expressed and bound genes was determined by taking all genes differentially expressed upon RBP KD that contained at least one IDR-filtered peak in the corresponding eCLIP experiment in the same cell type.

Differentially spliced events were defined as those meeting p-value < 0.05, FDR < 0.1, and $|\Delta\Psi| > 0.05$ from rMATS analysis (described above). The number of unique events was defined as the number of non-overlapping events upon combining all experiments for a given sampling. A differentially spliced event was considered bound if for any RBP in which the event was differentially included upon KD, there was an eCLIP peak for the same RBP in the same cell type between the start of the upstream flanking exon and the end of the downstream flanking exon for cassette exons and mutually exclusive exons, start of the upstream flanking exon and end of the common exon region for A3SS, start of the common exon and end of the common exon region for A5SS, and start of the upstream and stop of the downstream exons for retained introns).

To perform saturation of transcript regions, the highest expressed transcript for each gene was first identified using transcript-level quantifications) from the same rRNA-depleted RNA-seq experiments described above (K562, accession numbers ENCFF424CXV,ENCFF073NHK; HepG2, accession numbers ENCFF205WUQ, ENCFF915JUZ). The following regions were then identified: the entire unspliced transcript (pre-mRNA), all exons (exon), 5’ untranslated regions (5’ UTR), coding sequence (CDS), 3' untranslated regions (3’UTR), all introns (intron), 100nt intronic regions flanking the 5' and 3' splice sites (splice site), proximal intronic regions extending from 100nt to 500nt from the 5' and 3' splice site (prox. intron), and distal intronic regions extending from 500nt and beyond from the 5' and 3' splice sites. Saturation calculations were then performed as described above for all genes (Fig. 2c-d, Extended Data Fig. 2e-g) or only genes with TPM > 1 in both K562 and HepG2 (Extended Data Fig. 2d), and plotted as either the total number of bases covered (Fig. 2c), or the fraction of covered bases divided by the total number of bases in that annotation across all genes (Fig. 2d, Extended Data Fig. 2d). The ratio of bases covered was calculated by dividing the number of bases covered in subsampling of N+1 datasets divided by the number covered in subsampling N datasets.

Analysis of the fold-increase between one and two datasets (Extended Data Fig. 2f) was determined by first taking all 55 RBPs profiled in both HepG2 and K562, and calculating the fold-increase in covered bases by considering 110 comparisons including HepG2 followed by K562 and K562 followed by HepG2. Then, for each of the 110 datasets, 10 random other datasets were chosen from the same cell type, and for each of the 10 the fold-increase in covered bases from adding that dataset to the first was calculated.

To compare the fold-increase between profiling new RBPs in additional cell lines, eCLIP datasets profiling RBFOX2, IGF2BP1, IGF2BP2, and IGF2BP3 in H9 human embryonic stem cells were obtained from the Gene Expression Omnibus (GSE78509)[54], and added as the 181st dataset. These were compared against profiling a new RBP in K562 or HepG2 (calculated by adding each of the 126 profiled RBPs as the 179 (if it was profiled in both cell types) or 180 (if it was profiled in only one cell type) datasets for other RBPs), or a profiled RBP done in second cell type (calculated by sampling 180 datasets, and adding the 181st if it was an RBP already profiled in the other cell type).

Motif comparisons between RBNS and eCLIP

eCLIP 6mer Z-scores in Fig. 3b were calculated as previously described[55]. Briefly, peaks and a shuffled background set of peaks that preserves the region of binding (3'UTR, 5'UTR, CDS, exon, proximal and distal intron) were generated. EMBOSS compseq [http://structure.usc.edu/emboss/compseq.html] was used on these two peak sets and the Z-scores of the difference between real and background 6mer frequencies was calculated.

To produce eCLIP logos in a similar manner for comparison with RBNS logos, an analogous procedure was carried out on the eCLIP peak sequences (only eCLIP peaks at least 2-fold enriched were used): the two halves of the RBNS pulldown read set were replaced with the two eCLIP replicate peak sequence sets (each peak was extended 50 nt upstream of its 5' end as some RBPs have motif enrichments symmetrically around or only upstream of the peak starts), and the input RBNS sequences were replaced by random regions within the same gene as each peak that preserved peak length and transcript region (5' and 3' UTR peaks were chosen randomly within that region; intronic and CDS peaks were shuffled to a position within the same gene that preserved the peak start's distance to the closest intron/exon boundary to match sequence biases resulting from CDS and splicing constraints). The enrichment Z-score threshold for 5mers included in eCLIP logos was 2.8, as this threshold produced eCLIP logos containing the most similar number of 5mers to that of the Z≥3 5mer RBNS logos. Each eCLIP motif logo was filtered to include only 5mers that occurred in both of the corresponding eCLIP

replicate logos. eCLIP motif logos were made separately for all eCLIP peaks, only 3'UTR peaks, only CDS peaks, and only intronic peaks, with the eCLIP logo of those 4 (or 8 if CLIP was performed in both cell types) with highest similarity score to the RBNS logo shown in Fig. 3a, where the similarity score was the same as previously described to cluster RBNS logos (eCLIP logos for all transcript regions shown in Extended Data Fig. 3a). To determine significance of overlap between RBNS and eCLIP, a hypergeometric test was performed with 5mers in all RBNS logos, eCLIP logo 5mers (for peaks in the region with highest similarity score to the RBNS logo), and 5mers in their intersection, relative to the background of all 1,024 5mers; overlap was deemed significant if $P<0.05$. The top 'eCLIP-only' logo in each region was the highest eCLIP logo, if any, comprised of 5mers that had no overlap with any RBNS $Z\geq3$ 5mers (always using at least the top 10 RBNS 5mers if there were fewer than 10 with $Z\geq3$).

All eCLIP/RBNS comparisons were for the same RBP with the following exceptions in which the eCLIP RBP was compared to a closely related RBNS protein: KHDRBS2 eCLIP versus KHDRBS1 RBNS; PABPN1 eCLIP versus PABPN1L RBNS; PTBP1 eCLIP versus PTBP3 RBNS; PUM2 versus PUM1 RBNS; and RBM15 versus RBM15B RBNS.

Splicing regulatory effects of RBNS+ and RBNS– eCLIP peaks

To assess the splicing regulatory effects of RBNS+ and RBNS– eCLIP peaks for Fig. 3c, only rMATS SEs with a Ψ between 0.05 and 0.95 in at least one of the control or KD were considered for each RBP. Each eCLIP peak (extended 50 nt 5' of the peak start) was first checked if it overlapped the SE, and if not then if it overlapped the upstream or downstream flanking 250 nt. To compare the magnitude of splicing changes upon KD for eCLIP+ vs eCLIP– SEs while minimizing the confounding factors of different wildtype host gene expression level and SE Ψ values among these two sets of SEs, a matched set of eCLIP– SEs was created by selecting for each eCLIP+ SE an SE in the same decile of wildtype gene expression and wildtype Ψ for each corresponding SE with an eCLIP peak. A CDF of the ΔΨ changes upon KD was compared for the eCLIP+ versus eCLIP– SEs in each of the 6 SE direction/eCLIP region combinations ([included, excluded SE] x [peak over SE, upstream intron, downstream intron]), with significance $P<0.05$ for a one-sided Wilcoxon rank-sum test that $|\Delta\Psi|_{SE,\ peak} > |\Delta\Psi|_{SE,\ no\ peak}$. If the eCLIP+ vs eCLIP– comparison was significant, the eCLIP peaks were divided into those that did and did not contain the top RBNS 5mer. The ΔΨ values for all RBPs in each of the 6 SE direction/eCLIP regions were combined for comparison in Fig. 3c; see Extended Data. Fig. 4a

for RBPs that were significant in each region (12 included/4 excluded upon KD, upstream intron eCLIP peak; 11 included/2 excluded upon KD, SE eCLIP peak; 7 included/7 excluded upon KD, downstream intron eCLIP peak). To assess eCLIP peaks with or without the top 'eCLIP-only' *k*mer, the top *5*mer from the aforementioned 'eCLIP-only' logo was used from the first region with an eCLIP-only logo among: all peaks; CDS peaks; intron peaks; and 3'UTR peaks (the more highly enriched 5mer if eCLIP was performed in both cell types). The resulting 'eCLIP-only' 5mers for Extended Data Fig. 4b were: CELF1 (CUCUC), EIF4G2 (GUGUG), EWSR1 (CGCGG); FUBP3 (UUGUU); FUS (GUGUG); HNRNPC (GUCGC); HNRNPK (UCCCC); HNRNPL (none); IGF2BP1 (GUGUG); IGF2BP2 (CGCCG); KHDRBS2: (none); KHSRP (none); PABPN1L (CGCGG); PCBP2 (CGGCG); PTBP3 (GAAGA); PUM2 (UUUUU); RBFOX2 (GGGGG); RBM22 (GGUAA); SFPQ (UCCGG); SRSF5 (CGGCG); SRSF9 (CUGGA); TAF15 (AGGGA); TARDBP (GAAGA); TIA1 (CGCCG); TRA2A (GAGGG).

Overlaps between RBP binding and gene expression perturbation upon KD/RNA-seq

To increase sensitivity for gene expression analysis, significant binding was determined at the level of transcript regions (including 5'UTR, CDS, 3'UTR, and introns) instead of using peaks. To identify significant enrichment between binding and expression changes, genes with significantly enriched binding to regions ($p \leq 0.00001$ and $\log_2$(fold-enrichment) ≥ 4, as described above) were overlapped with the set of genes with significantly altered expression in KD/RNA-seq (|log2(fold-change)| > 1 and an adjusted p-value < .05 between knockdown and control from DEseq analysis). Enrichment was calculated separately for knockdown-increased and knockdown-decreased genes, with significance determined by Fisher Exact test (or Yates' Chi-Square test if all observed and expected values were above 5).

For cumulative distribution plots, genes were separated based on their eCLIP fold-enrichment in IP versus input for the indicated transcript region. To filter out non-expressed genes, genes were included only if the region had at least 10 reads in one of IP or input, and where at least 10 reads would be expected in the comparison dataset given the total number of usable reads.

To perform TIA1 motif enrichment analysis, first the fold-enrichment of each 5mer was calculated by comparing the frequency in 3'UTRs of genes increased or decreased upon TIA1 knockdown in K562 or HepG2 with the frequency in a set of control genes upon knockdown (changed genes upon KD: DEseq adjusted P-val < 0.05 and |Fold-Change| > 1.5; control genes: DEseq P-val > 0.5 and |Fold-Change| < 1.1, subsetted to match the starting expression of changing genes upon KD). The top 15 5mers in TIA1 RBNS were then highlighted among in the

ranked ordering of all 1,024 5mers. For positional analysis, a meta-3'UTR was created by normalizing all 3'UTRs to a 100nt window. For each normalized position, the frequency of the top 10 TIA1 RBNS 5mers was calculated for each of the up-regulated, down-regulated, and control gene sets. Significance at each position was determined by $P < 0.05$ in a binomial test comparing the number of up- or down-regulated genes that have one of the top 10 RBNS 5mers at that position under the null frequency that it is equal to the corresponding frequency observed in control genes.

RBP binding correlation with knockdown-perturbed splicing (splicing maps)

RBP binding/splicing maps were generated using eCLIP normalized (reads per million) read densities overlapped with alternatively spliced (AS) regions from rMATS JunctionCountsOnly files from the same cell type. First, the set of differentially alternatively spliced events of the desired type (cassette/skipped exons (SE), alternative 5' splice site (A5SS), or alternative 3' splice site (A3SS) events were identified (Extended Data Fig. 6a), requiring rMATS p-value < 0.05, FDR < 0.1, and $|\Delta\Psi| > 0.05$ in knockdown versus control RNA-seq. To eliminate potential double counting of CLIP densities, overlapping AS events were additionally filtered to choose only the events containing the highest average inclusion junction count (IJC) among all replicates (using the bedtools v2.26 command merge (-o collapse -c 4) and pybedtools 0.7.9).

Next, for each splicing event, read densities were normalized across all regions separately for IP and input, in order to equally weigh each event. Per-position input probability densities were then subtracted from IP probability densities to attain position-level enrichment or depletion, for regions extending 50nt into each exon and 300nt into each intron composing the event, referred to as 'Normalized eCLIP enrichment' (Extended Data Fig. 6b). For shorter exons (<100 nt) and introns (<600nt), densities were only counted until the boundary of the neighboring feature. Skipped exon (SE) maps were plotted using eCLIP densities overlapping the following 4 regions around AS events: 3' end of the upstream exon, 5' end of the cassette, 3' end of the cassette, and 5' end of the downstream exon. Alternative 3' splice site (A3SS) maps were defined with three regions: 3' end of the upstream exon, 5' end of the longer transcript, and the 5' end of the shorter transcript. Alternative 5' splice site (A5SS) maps were defined with three regions: 3' end of the shorter transcript, 3' end of the longer transcript, and the 5' end of the downstream exon.

Plots of eCLIP signal enrichment (referred to as 'splicing maps') were then created by calculating the mean and standard error of the mean over all events after removing the highest (2.5%) and lowest (2.5%) outlying signal at each position, referred to as 'Average eCLIP enrichment' (Extended Data Fig. 6c). Splicing maps were only considered for RBPs with 100 or more altered cassette exon events, or 50 or more alternative 5' or 3' splice site events. As a background reference for cassette exon comparisons, sets of 1,832 (HepG2) and 2,244 (K562) 'native' cassette exons were identified which had $0.05 < \Psi < 0.95$ in at least half of control shRNA RNA-seq datasets for that cell type. Similar sets of 208 (K562) and 164 (HepG2) native alternative 5' splice site and 393 (K562) and 361 (HepG2) native alternative 3' splice site events were identified that had $0.05 < \Psi < 0.95$ in at least half of control shRNA RNA-seq datasets for that cell type. RBP-responsive event eCLIP enrichment was then calculated as eCLIP signal enrichment at RBP-regulated events minus eCLIP signal enrichment at native control events, referred to as 'Enrichment relative to control events' (Extended Data Fig. 6d). Native versus constitutive enrichment (Extended Data Fig. 7c) was similarly calculated as the eCLIP signal enrichment at native exons minus the eCLIP signal enrichment at constitutive exons (defined as exons with zero reads supporting exclusion in any control RNA-seq in that cell type).

Correlation between splicing maps was defined as the Pearson correlation (R) between a vector containing both included-upon knockdown and excluded-upon knockdown RBP-responsive event eCLIP enrichment for each RBP. If an RBP had less than the minimum required number of events (100 for cassette exons or 50 for alternative 5' or 3' splice site events) for either knockdown-included or knockdown-excluded events, the correlation was only calculated using the other event type.

To calculate the number of RBPs bound per exon, the set of spliceosomal RBPs was taken from manual annotation of RBP functions (described above and listed in Supplementary Table 1). The number of reproducible (IDR) peaks at each position relative to splice sites was summed across all RBPs, and divided by the total number of cassette or constitutive exons respectively.

Comparison of DNA- and RNA-binding properties of RBPs

For integrative analyses, DNaseI HS data (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeOpenChromSynth), histone modifications by ChIP-seq from ENCODE/Broad Institute (http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeBroadHistone) and eCLIP-seq data from ENCODE (https://www.encodeproject.org) were downloaded and compared with RBP ChIP-seq data.

To explore the possibility that some RBP chromatin association events might be coupled with their direct RNA binding activities in cells, RNA binding peaks were compared with DNA binding signals as assayed by ChIP-seq to quantify enrichment. Only eCLIP peaks in gene body regions (excluding promoter and terminator regions, defined as the 1kb surrounding regions of TSS and TTS) were considered. The ChIP-seq signals were calculated for each eCLIP peak, together with surrounding regions that are 10 times the length of eCLIP peak on each side. Wilcoxon rank-sum tests were then performed to see whether ChIP-seq signal were enriched at the middle third regions.

To see whether those differentially-expressed genes after RBP knockdown were enriched in RBP binding at chromatin level, equal number of genes with similar expression level either with or without binding were randomly sampled, and the number of differentially-expressed genes after knockdown of the RBP were counted (fold change > 1.5 or < 2/3, adjusted p-value <0.05 by DESeq2), and one-tailed Fisher's exact tests were then performed to test the dependence of RBP binding and differential expression. The above procedure was performed 100 times to give the distribution of the odds ratio. A significant dependence was defined when the null hypothesis was rejected at level of 0.05 for at least 95 times. The correlation between RBP association and genes with regulated alternative splicing events (A3SS, A5SS, RI, MXE and SE events) were investigated similarly.

Analysis of RBP regulatory features in subcellular space

Localization annotations and calculation of nuclear versus cytoplasmic ratio were generated from immunofluorescence imaging as described above. "Nuclear RBPs" were defined as those with nuclear/cytoplasmic ratio ≥ 2, and "Cytoplasmic RBPs" were defined as those with nuclear / cytoplasmic ratio ≤ 0.5. Spliced reads were defined as reads mapping across an annotated GENCODE v19 splice junction (extending at least 10 bases into each exon) and unspliced reads were defined as reads that overlapped an exon-intron junction (extending at least 10 bases into both the exon and intron regions). Significance between groups was determined by Wilcoxon rank sum test.

Prediction of RNA secondary structure was performed using the RNAfold webserver (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi)[56] with default parameters. Shown is the MFE secondary structure prediction.

Preservation of RBP regulation across cell types

To consider binding across cell types, first the highest expressed transcript for each gene was identified using transcript-level quantifications) from the same rRNA-depleted RNA-seq experiments described above (K562, accession numbers ENCFF424CXV,ENCFF073NHK; HepG2, accession numbers ENCFF205WUQ, ENCFF915JUZ) and used as representative for that gene. Next, genes were categorized based on their absolute fold-difference (FD) between K562 and HepG2: unchanged (FD ≤1.2), weakly (1.2 < FD ≤ 2), moderately (2 < FD ≤ 5) or strongly (FD > 5) differential (for each, requiring TPM ≥ 1 in both K562 and HepG2), or cell-type specific genes (TPM < 0.1 in one cell type and TPM ≥ 1 in the other). Peaks were then categorized based upon the expression change of their associated gene. Significance between groups was determined by Kolmogorov-Smirnov test.

Analysis of preservation of binding across cell types was considered in three ways. First, for each peak identified in one cell type, the fold-enrichment for that region in the other cell type was calculated and considered for each gene type (as in Fig. 8a). Two groups of peaks were then identified: those that were ≥ 4-fold enriched in the other cell type, and those that were not enriched in the other cell type. The fraction of peaks associated with a gene class that were either ≥ 4-fold or not enriched were then considered for each gene class separately (Fig. 8b). Second, the set of peaks ≥4-fold enriched (and the set not enriched) was compiled across all genes, and the fraction associated with each gene class were then reported (Extended Data Fig. 10b). Finally, binding preservation was calculated by determining the fraction of IDR peaks identified in one cell type that overlap (requiring at least 1nt overlap) IDR peaks identified in the second cell type (Extended Data Fig. 10c).

Correlation between splicing maps was performed as described above, considering all RBPs profiled by eCLIP and RNA-seq in both K562 and HepG2 that had at least 100 differentially included or excluded cassette exon events in both cell types.

RBP expression in tissues

Tissue specificity was measured as the entropy deviation from a uniform distribution among all tissues as in[1]. For each RBP, the $\log_2$(TPM+1) was calculated for each of the 42 samples (HepG2, K562, and 40 tissues profiled by the GTEx consortium[57]), and the tissue specificity was computed as the difference between the logarithm of the total number of samples (N=42) and the Shannon entropy of the expression values for an RBP:

$$S = H_{max} - H_{obs} = \log_2(N) - \Sigma_{i=1...N}\, p_i \times \log_2(p_i),$$

$$\text{Where } p_i = x_i / \Sigma_{i=1...N}\, x_i$$

for $x_i = \log_2(\mathrm{TPM}_i + 1)$ in sample *i.*

The data used for the analyses were obtained from dbGaP accession number phs000424.v2.p1 in Jan. 2015. TPMs were measured using kallisto[58] on the following samples: Adipose-Subcutaneous: SRR1081567; AdrenalGland: SRR1120913; Artery-Tibial: SRR817094; Bladder: SRR1086236; Brain-Amygdala: SRR1085015; Brain-AnteriorCingulateCortex: SRR814989; Brain-CaudateBasalGanglia: SRR657731; Brain-CerebellarHemisphere: SRR1098519; Brain-Cerebellum: SRR627299; Brain-Cortex: SRR816770; Brain-FrontalCortex: SRR657777; Brain-Hippocampus: SRR614814; Brain-Hypothalamus: SRR661179; Brain-NucleusAccumben: SRR602808; Brain-SpinalCord: SRR613807; Brain-SubstantiaNigra: SRR662138; Breast-MammaryTissue: SRR1084674; Cervix: SRR1096057; Colon: SRR1091524; Esophagus: SRR1085211; FallopianTube: SRR1082520; Heart-LeftVentricle: SRR815517; Kidney-Cortex: SRR809943; Liver: SRR1090556; Lung: SRR1081283; MinorSalivaryGland: SRR1081589; Muscle-Skeletal: SRR820907; Nerve-Tibial: SRR612911; Ovary: SRR1102005; Pancreas: SRR1081259; Pituitary: SRR1077968; Prostate: SRR1099402; Skin: SRR807775; SmallIntestine: SRR1093314; Spleen: SRR1085087; Stomach: SRR814268; Testis: SRR1081449; Thyroid: SRR808886; Uterus: SRR820026; Vagina: SRR1095599.

Acknowledgments

The authors would like to thank Elise Feingold, Mike Pazin, Dan Gilchrist and Grace Xiao for discussions and Cricket Sloan, Jean Davidson, Eurie Hong, and Mike Cherry for assistance with data deposition and distribution to the public. This work was funded by the National Human Genome Research Institute ENCODE Project, contract U54HG007005, to BRG (principal investigator) and X-DF, CBB and GWY (co-principal investigators), as well as a grant from the Fonds de Recherche du Québec-Santé to EL. ELVN is a Merck Fellow of the Damon Runyon Cancer Research Foundation [DRG-2172-13] and is supported by a K99 grant from the NIH [HG009530]. GAP is supported by the NSF graduate research fellowship. XDF and GWY were partially supported by grants from the NIH (HG007005, NS075449).

Author Contributions.

Work on the paper was divided between data production and analysis. The analysts were ELVN, PF, GAP, XW, DD, RX, LPBB, JC, MOD, NJL, SS, BZ, BAY. The data producers were ELVN, XW, SMB, DD, NALC, SO, BS, RX, LZ, CB, KG, CG-B, AH, NJL, HL, TBN, TP, IR, RS, ALL, SA, JB, AS and RW. Substantially larger contributions were made by the joint first authors.

Overall project management was carried out by the senior authors X-DF, EL, CBB, BRG and GWY.

Author Information.

Data sets described here can be obtained from the ENCODE project website at http://www.encodeproject.org via accession numbers in Supplementary Table 2. ELVN and GWY are co-founders of Eclipse BioInnovations Inc. The authors declare no other competing financial interests.

Figure Legends

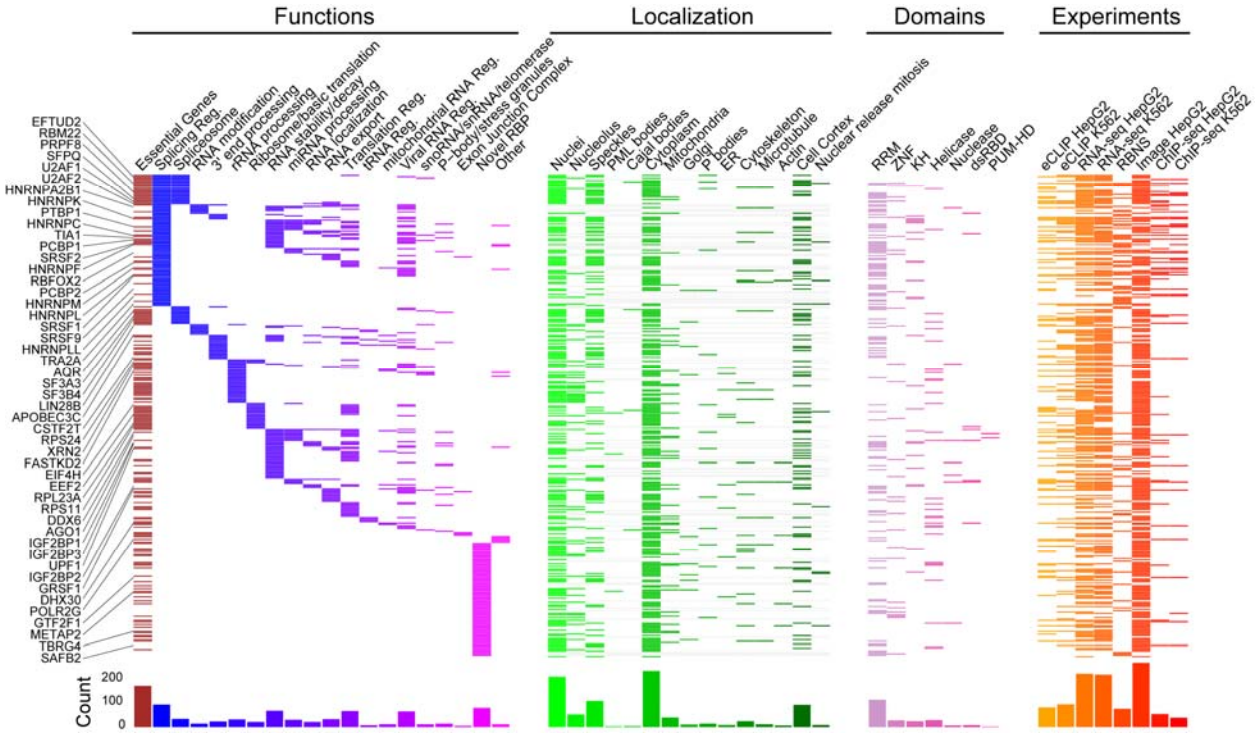


Figure 1 | Overview of experiments and data types.

The 352 RNA binding proteins (RBPs) profiled by at least one ENCODE experiment (orange/red) are shown, with localization by immunofluorescence (green), essential genes from CRISPR screening (maroon), manually annotated RBP functions (blue/purple), and annotated protein domains (pink). Histograms for each category are shown on bottom, and select RBPs highlighted in this study are indicated on left.

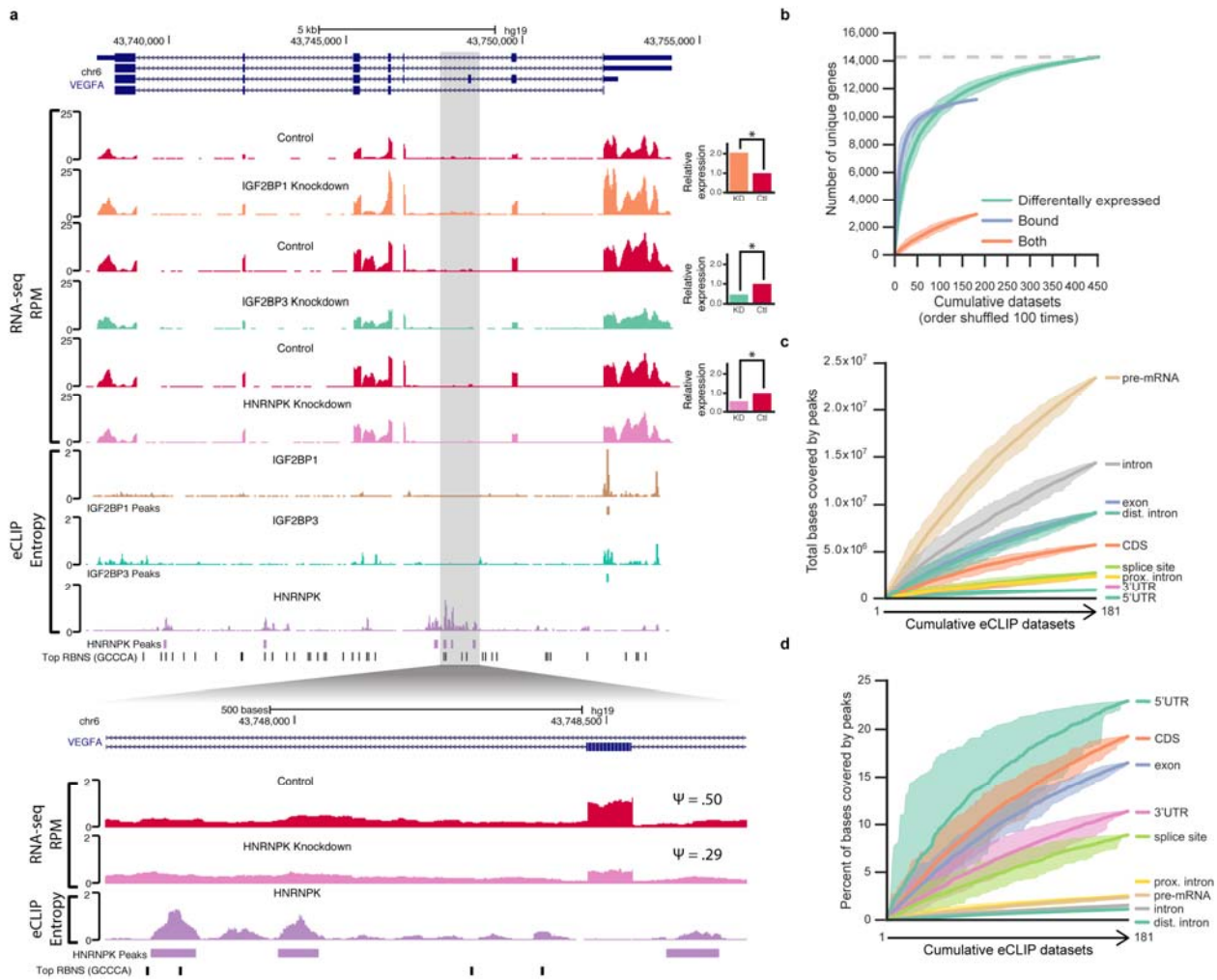


Figure 2 | Integrative analysis of RBP binding and function.

(a) Combinatorial expression and splicing regulation of *VEGFA*. Tracks indicate RNA-seq read density (as reads per million, RPM) and eCLIP relative information (in IP versus input). Asterisk indicates significant expression changes ($|\log_2(\text{fold-change})| > 1$ and p-value < 0.05), and percent spliced in (ψ) is indicated.

(b) Lines indicate the number of genes that are (green) differentially expressed upon RBP knockdown in RNA-seq (FDR < 0.05, p-value < 0.05, and $|\log_2(\text{fold-change})| > 1$), (blue) bound in eCLIP, or (orange) both bound and differentially expressed in the same cell type. The ordering of each data type was randomly shuffled 100 times, with the shaded region indicating the tenth and ninetieth percentiles. Grey line indicates the number of genes expressed with TPM > 1 in either K562 or HepG2.

(c-d) Lines indicate the mean of (c) the cumulative total number of bases covered by peaks, or (d) the cumulative fraction of bases covered by peaks for 100 random orderings of the 181

eCLIP datasets, separated by transcript regions as indicated. Shaded region indicates tenth through ninetieth percentiles.

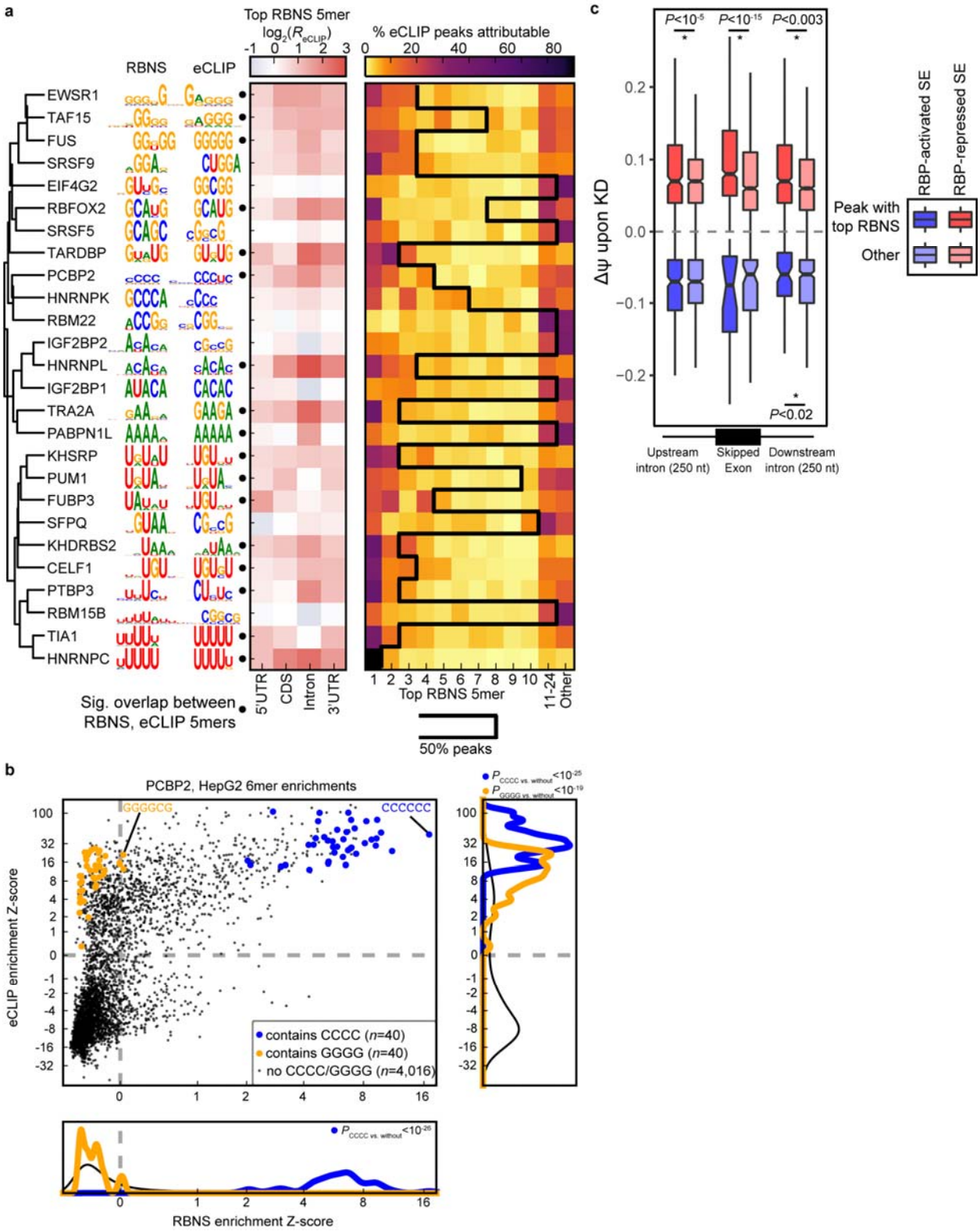


Figure 3 | Sequence-specific binding *in vivo* is determined predominantly by intrinsic RNA affinity of RBPs

(a) Left: Top sequence motif of RBNS versus eCLIP-derived enriched 5mers clustered by similarity of RBNS motifs. Filled circles to the right of the eCLIP logo indicate if the groups of 5mers comprising the RBNS and eCLIP motifs overlap significantly (hypergeometric $P<0.05$). Center: Enrichment of the top RBNS 5mer in eCLIP peaks (R_{eCLIP}) within different genomic regions. Right: The proportion of eCLIP peaks attributed to each of the 10 highest affinity RBNS 5mers, as well as the #11-24 RBNS 5mers combined. The black line indicates the number of top RBNS 5mers required to explain >50% of eCLIP peaks for each RBP (maximum, 24 5mers).
(b) Comparison of PCBP2 *in vivo* vs *in vitro* 6mer enrichments, with 6mers containing CCCC and GGGG highlighted. Significance was determined by Wilcoxon rank-sum test and indicated if $P < 0.05$. x- and y-axes are plotted on an arcsinh scale.
(c) Comparison of the magnitude of splicing change upon RBP knockdown for SEs containing eCLIP peaks with vs. without the top RBNS 5mer, separated by the direction of SE change upon KD and location of the eCLIP peak relative to the SE. Significance was determined by Wilcoxon rank-sum test and indicated if $P < 0.05$.

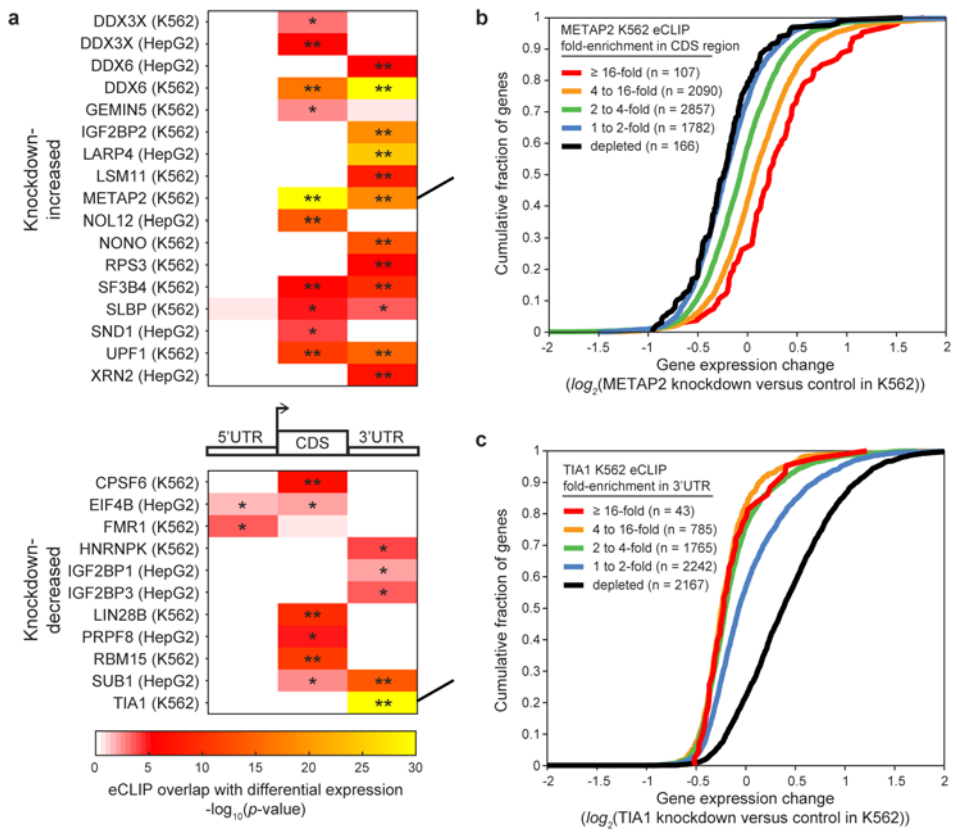


Figure 4 | Association between RBP binding and RNA expression upon knockdown.

(a) Heatmap indicates significance of overlap between genes significantly bound ($p \leq 10^{-5}$ and ≥4-fold enriched in eCLIP versus input) and genes significantly (top) increased or (bottom) decreased ($p < 0.05$ and FDR < 0.05) in RBP knockdown RNA-seq experiments. Significance was determined by Fisher's Exact test or Yates' Chi-Square approximation where appropriate; * indicates $p < 0.05$ and ** indicates $p < 10^{-5}$ after Bonferroni correction. Shown are all overlaps meeting a $p < 0.05$ threshold.

(b-c) Lines indicate cumulative distribution plots of gene expression fold-change (knockdown versus control) for indicated categories of eCLIP binding of (b) METAP2 in K562 and (c) TIA1 in K562.

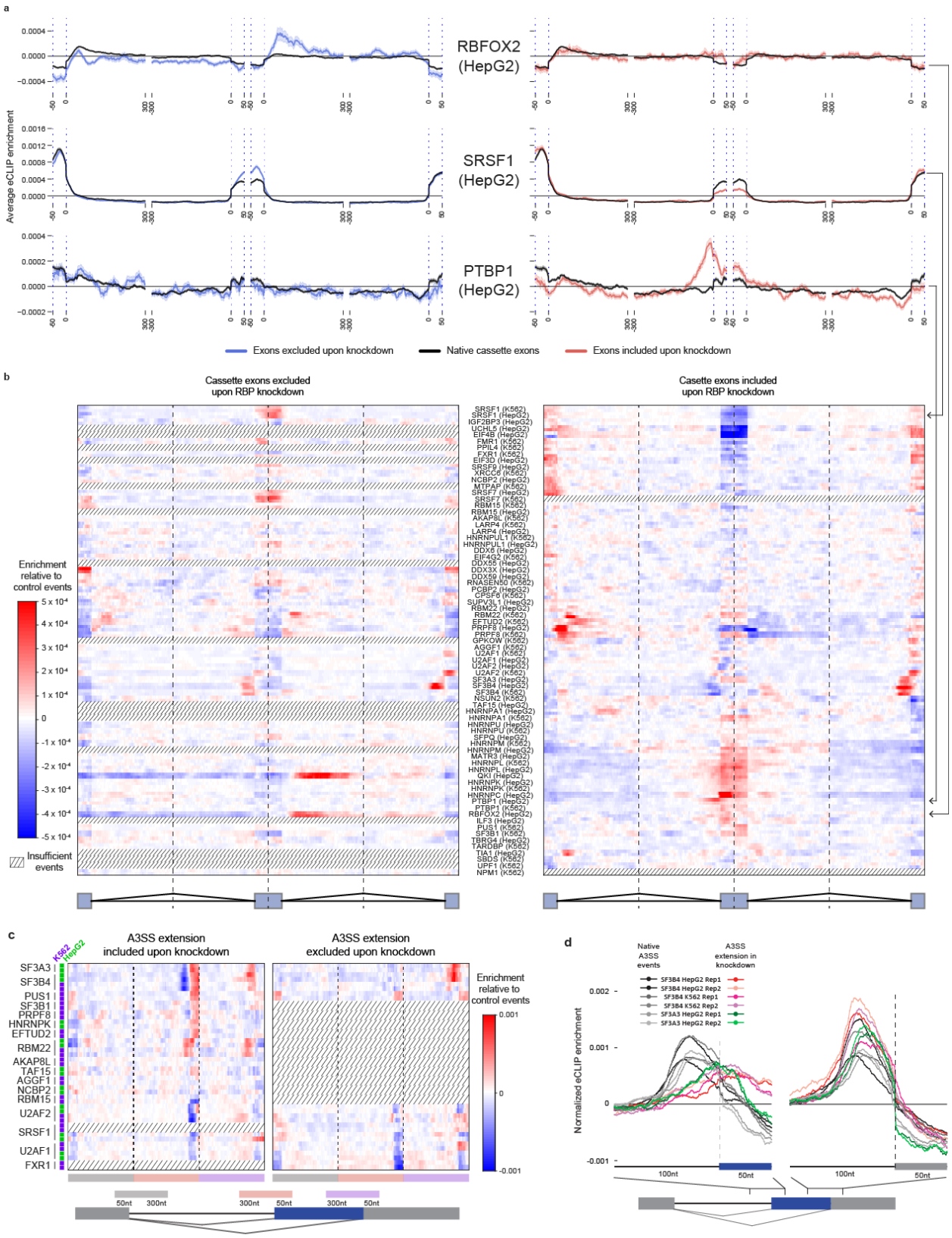

a
RBFOX2 (HepG2)
SRSF1 (HepG2)
PTBP1 (HepG2)
Average eCLIP enrichment
Exons excluded upon knockdown
Native cassette exons
Exons included upon knockdown
b
Cassette exons excluded upon RBP knockdown
Casette exons included upon RBP knockdown
Enrichment relative to control events
Insufficient events
c
A3SS extension included upon knockdown
A3SS extension excluded upon knockdown
K562
HepG2
SF3A3
SF3B4
PUS1
SF3B1
PRPF8
HNRNPK
EFTUD2
RBM22
AKAP8L
TAF15
AGGF1
NCBP2
RBM15
U2AF2
SRSF1
U2AF1
FXR1
Enrichment relative to control events
d
Native A3SS events
A3SS extension in knockdown
SF3B4 HepG2 Rep1
SF3B4 HepG2 Rep2
SF3B4 K562 Rep1
SF3B4 K562 Rep2
SF3A3 HepG2 Rep1
SF3A3 HepG2 Rep2
Normalized eCLIP enrichment
100nt
50nt

Figure 5 | Integration of eCLIP and RNAseq identifies splicing regulatory patterns.

(a) Normalized splicing maps of RBFOX2, SRSF1, and PTBP1 for cassette exons excluded (left) and included (right) upon knockdown, relative to a set of 'native' cassette exons with $0.05 < \Psi < 0.95$ in controls. Lines indicate average eCLIP read density in IP versus input for indicated exon categories, with standard error shown by the shaded area. The displayed region shown extends 50 nt into exons and 300 nt into introns.

(b) Heatmap indicates the difference between normalized eCLIP read density at cassette exons excluded (left) or included (right) upon RBP knockdown versus native cassette exons. Shown are all RBPs with eCLIP and knockdown RNA-seq data, with dashed lines indicating datasets with less than 100 significantly altered events.

(c) As in (b), shown for alternative 3' splice site events. Dashed lines indicate datasets with less than 50 significantly altered events. The displayed regions include the upstream common 5′ splice site (grey box), the extended alternative 3′ splice site (orange box) and the distal alternative 3′ splice site (purple box).

(d) Lines indicate mean normalized eCLIP enrichment in IP versus input for SF3B4 and SF3A3 at (red/purple/green) alternative 3' splice site extensions in RBP knockdown or (black) alternative 3' splice site events in control HepG2 or K562 cells.

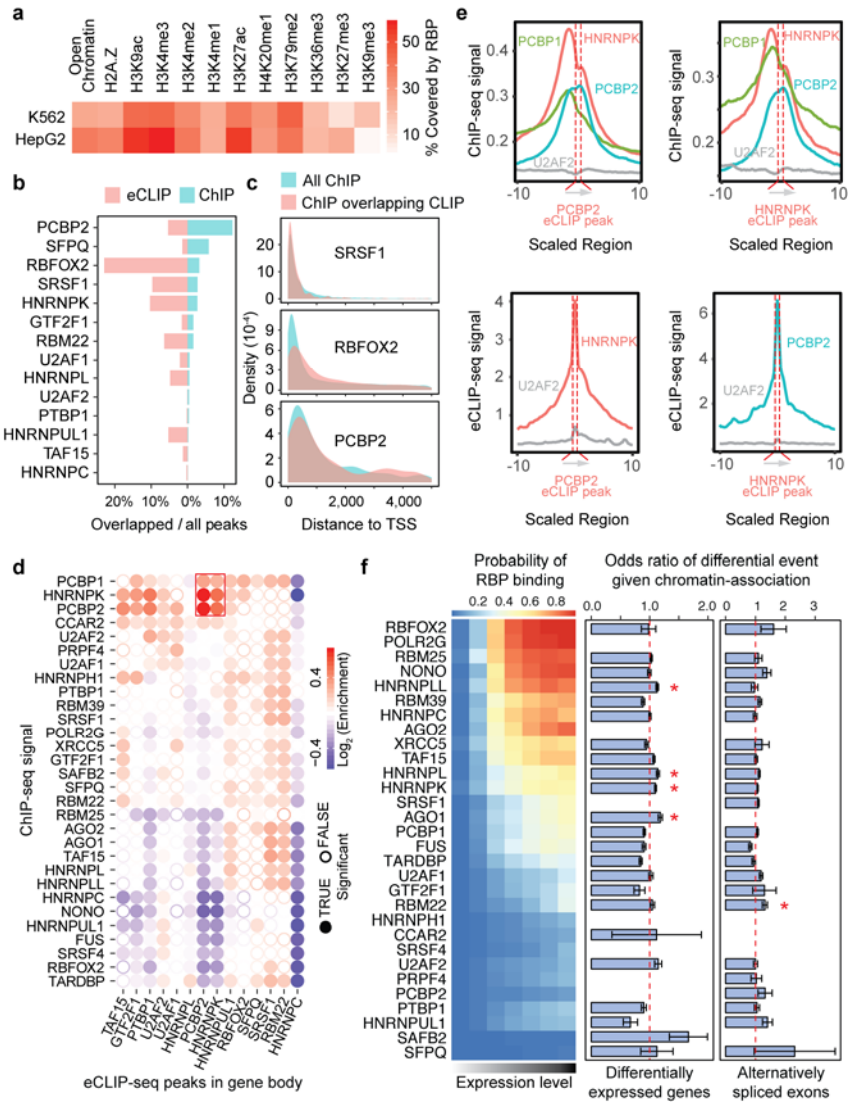


Figure 6 | Chromatin-association of RBPs and overlap with RNA binding

(a) Collective RBP binding activities on specific chromatin regions, including DNase I hypersensitive sites and various histone marks in HepG2 and K562 cells.

(b) Percentage of RBP eCLIP peaks overlapped with corresponding ChIP-seq peaks (pink) or percentage of RBP ChIP-seq peaks overlapped with corresponding eCLIP peaks (green). RBPs are sorted by decreasing level of overlapped ChIP-seq peaks.

(c) Distributions of overall chromatin binding activities (green) and overlapped chromatin and RNA binding activities (pink) of representative RBPs.

(d) Clustering of overlapped chromatin and RNA binding activities of different RBPs in non-promoter regions. Color key indicates the degree of ChIP enrichment at eCLIP peaks relative to surrounding regions with significance at $p<0.001$) or false ($p>0.001$) indicates the significance of the enrichment.

(e) Cross-RBP comparison of chromatin and RNA binding activities. Top: ChIP-seq density of indicated RBPs around PCBP2 or HNRNPK eCLIP peaks. Bottom: eCLIP density of indicated RBPs around PCBP2 or HNRNPK eCLIP peaks.

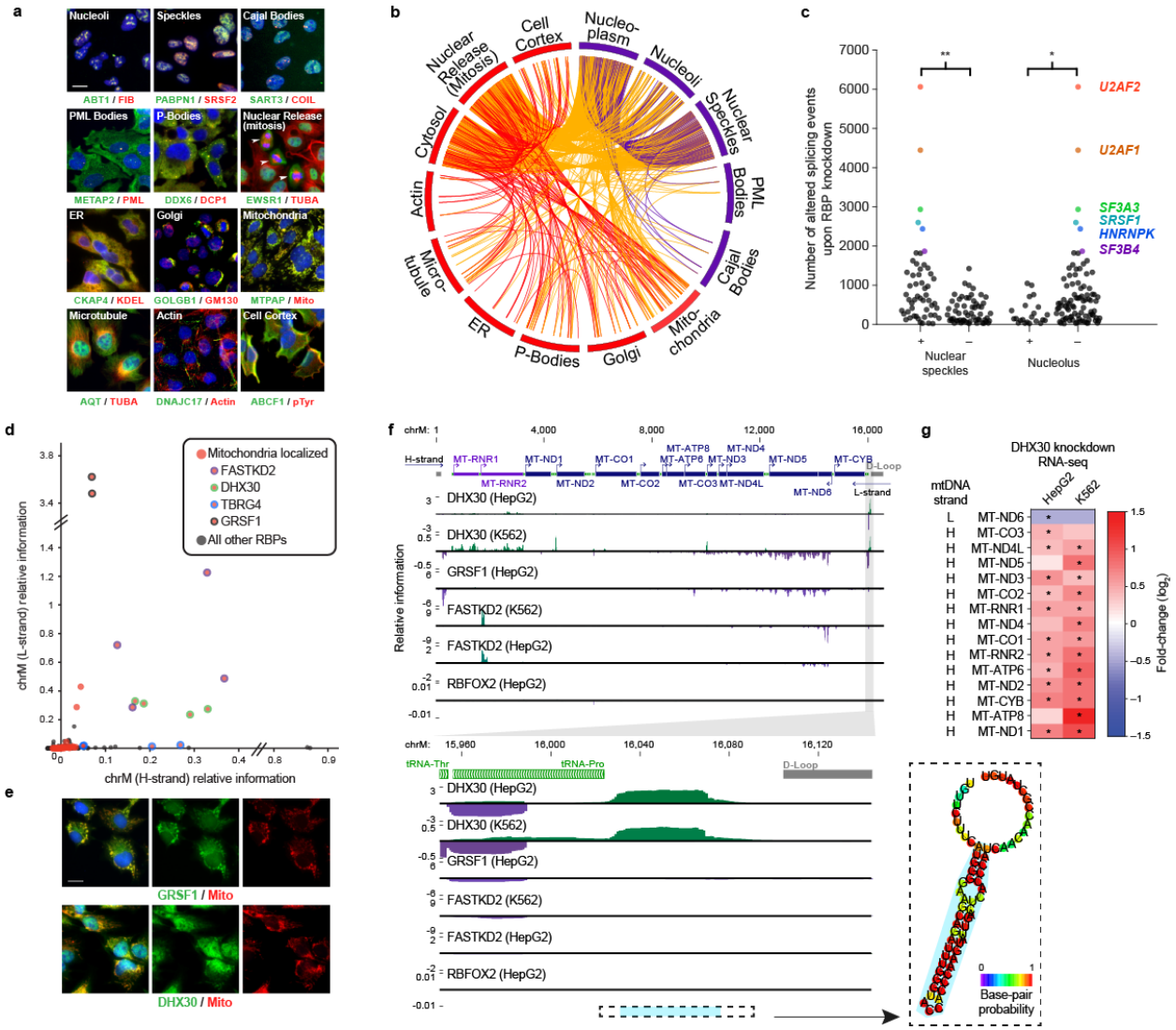


Figure 7 | RBP sub-cellular localization, binding, and regulation

(a) Example RBPs (green) co-localized with twelve interrogated markers (red).

(b) Circos plot with lines indicating co-observed localization patterns (red: within cytoplasm; purple: within nucleus; orange: between cytoplasm/nucleus).

(c) Points indicate the number of differential splicing events observed upon knockdown of each RBP, separated by the presence or absence of localization in nuclear speckles (center) or the nucleolus (right). * indicates $p \leq 0.05$ and ** indicates $p \leq 10^{-4}$ by Wilcoxon rank-sum test.

(d) Points indicate eCLIP relative information content (IP versus input) for mitochondria H-strand (x-axis) or L-strand (y-axis) for RBPs with mitochondrial localization by IF (red, with specific RBPs noted) and all other RBPs (black).

(e) IF images of mitochondrial localization of GRSF and DHX30.

(f) Genome browser tracks indicate binding (as eCLIP relative information content) along (top) the mitochondrial genome or (bottom) a ~300nt region for indicated RBPs. (right) Inset shows RNA secondary structure prediction (RNAfold) for the indicated region in blue.

(g) Heatmap indicates gene expression change upon DHX30 knockdown for all mitochondrial protein-coding and rRNA transcripts. * indicates significant expression changes ($p < 0.05$ and FDR < 0.05 from DEseq analysis).

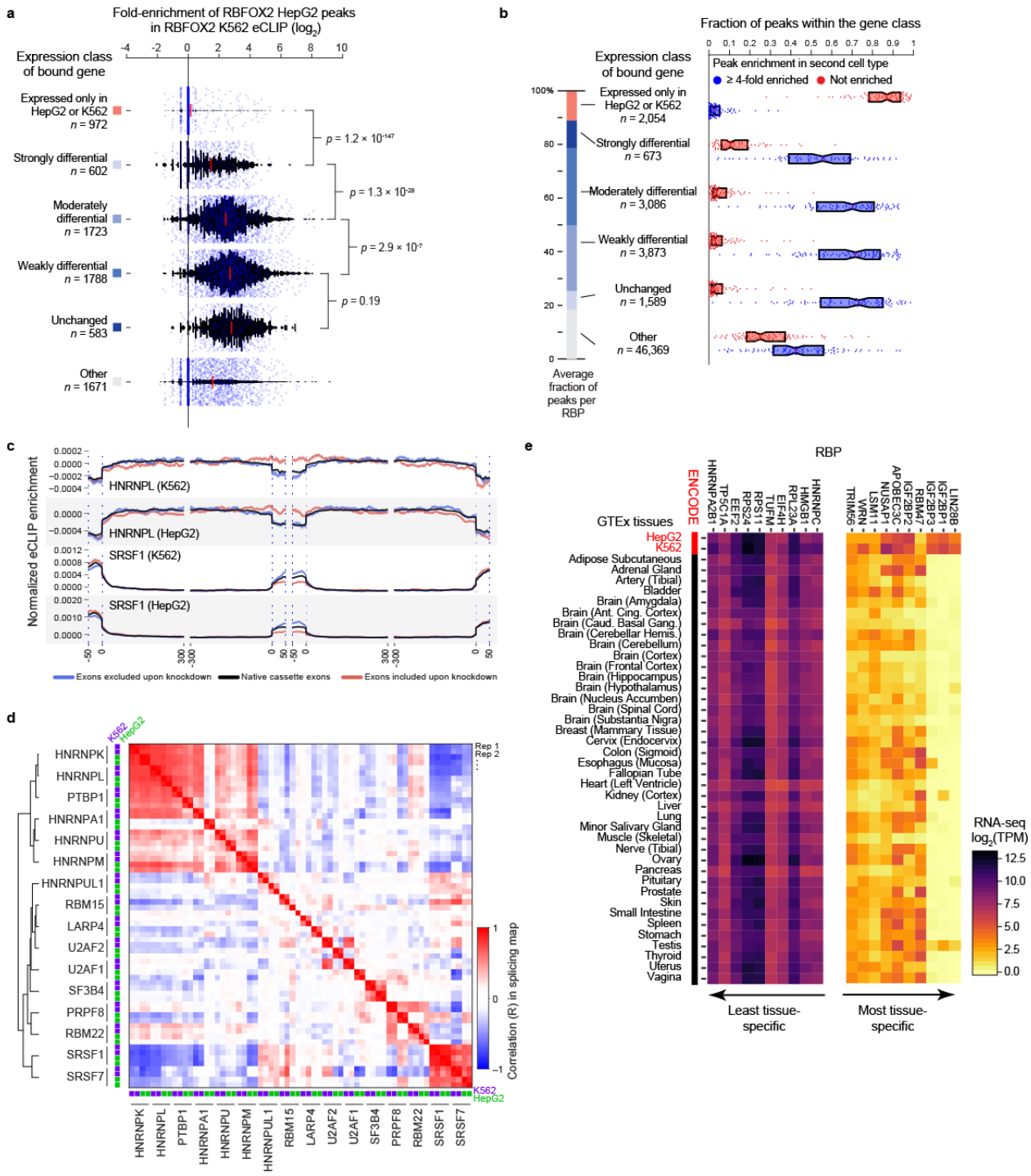


Figure 8 | Preservation of RBP binding and regulation across cell types.

(a) Each point indicates the fold-enrichment in K562 eCLIP of RBFOX2 for a reproducible RBFOX2 eCLIP peak in HepG2, with underlaid black histogram. Peaks are separated based on the relative expression difference of the bound gene between K562 and HepG2: unchanged (fold-difference ≤ 1.2), weakly (1.2 < fold-difference ≤ 2), moderately (2 < fold-difference ≤ 5) or strongly (fold-difference > 5) differential (each of which required expression TPM ≥ 1 in both

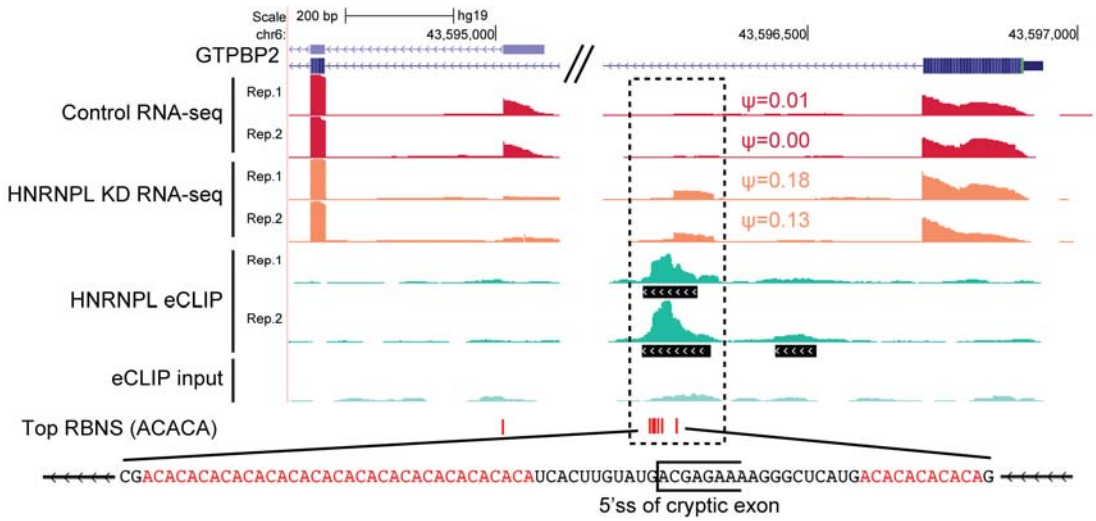
K562 and HepG2), or cell-type specific genes (TPM < 0.1 in one cell type and TPM ≥ 1 in the other). Mean is indicated by red lines, with significance determined by Kolmogorov-Smirnov test.

(b) For each RBP profiled in both K562 and HepG2, points indicate the fraction of peaks in the first cell type associated with a given gene class that are (blue) at least four-fold enriched, or (red) not enriched (fold-enrichment ≤ 0) in the second cell type. Boxes indicate quartiles, with median indicated by red lines. (left) Stacked bar indicates the average fraction of peaks per RBP (see Extended Data Fig. 10a for per-RBP distributions).

(c) Cassette exon splicing maps for HNRNPL and SRSF1 in both K562 and HepG2 cells.

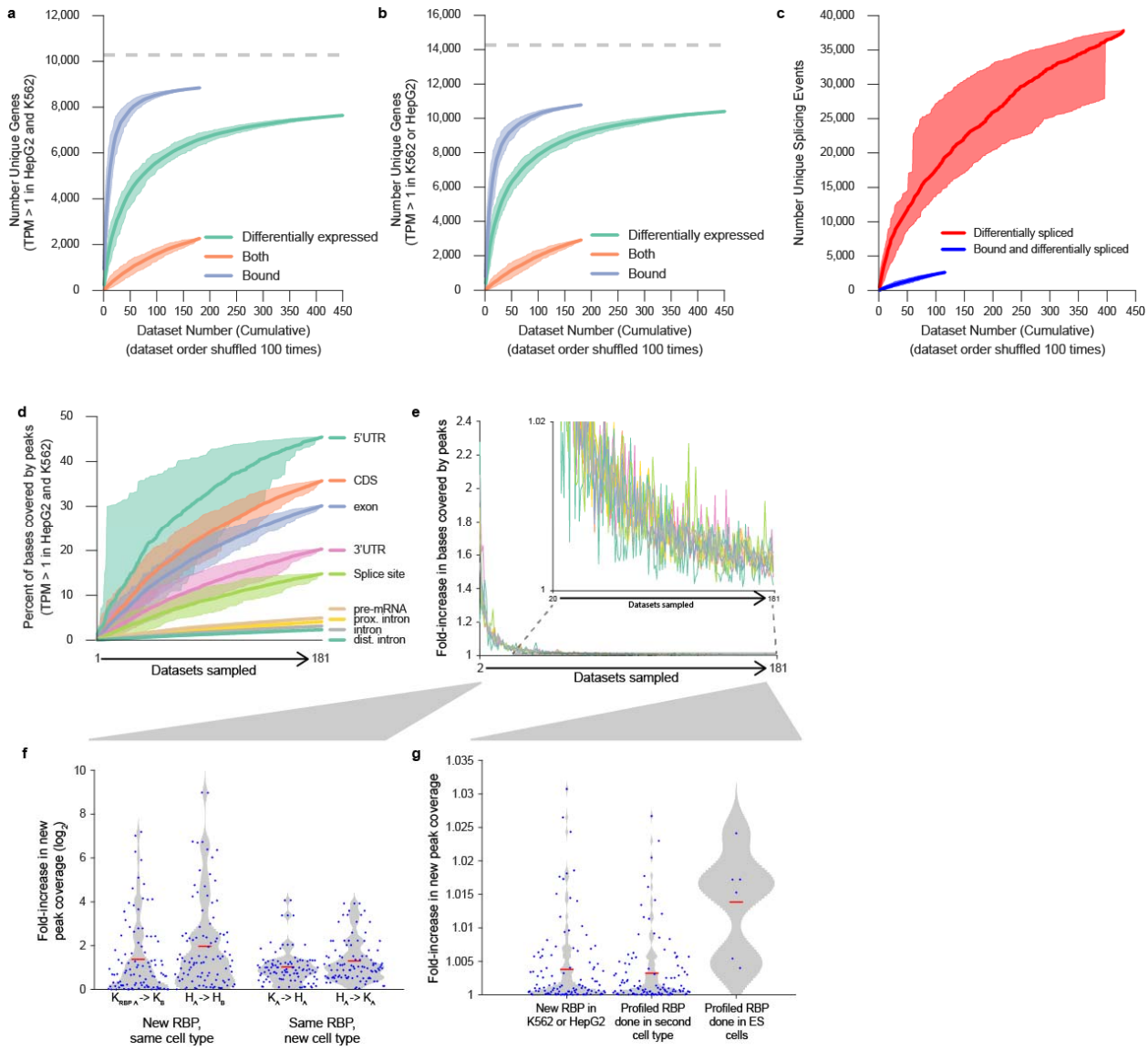
(d) Heatmap indicates correlation (Pearson R) between splicing maps for all RBPs profiled in both K562 and HepG2, hierarchically clustered at the RBP level.

(e) Expression of the 10 RBPs with the highest and lowest tissue-specificity across the two ENCODE cell lines and 40 human tissues.



Extended Data Figure 1 | Integrative analysis of RBP data types in cryptic exon suppression.

Shown is RNA-seq read density (reads per million), eCLIP read density (reads per million), and RBNS motif presence proximal to a 73-nt cryptic exon expected to induce nonsense-mediated decay (NMD) of GTP Binding Protein 2 (a ribosome rescue factor whose loss induces neurodegeneration in certain genetic backgrounds[59]). eCLIP indicates that HNRNPL binds over the cryptic exon 5' splice site in a sequence-specific manner to a region rich in the top RBNS 5mer, ACACA.



Extended Data Figure 2 | Saturation of RBP binding and regulation in the transcriptome

(a-b) Lines indicate the mean of 100 random orderings of each data type for the number of genes that are (green) differentially expressed upon RBP knockdown and RNA-seq (requiring FDR < 0.05, *p*-value < 0.05, and |fold-change| > 1), (blue) bound in eCLIP (overlapped by a IDR-reproducible peak with $p \leq 10^{-3}$ and fold-enrichment ≥ 8 in IP versus input), or (orange) both bound and differentially expressed in the same cell type. Grey dotted line indicates the total number of genes considered, either (a) with TPM > 1 in both HepG2 and K562, or (b) TPM > 1 in either K562 or HepG2. Shaded regions indicate tenth to ninetieth percentile.

(c) Lines indicate the mean of 100 random orderings of datasets for the number of (red) differential splicing events upon RBP knockdown (including cassette exons, alternative 5′ and 3′ splice sites, retained introns, and mutually exclusive exons; requiring FDR < 0.05, *p*-value <

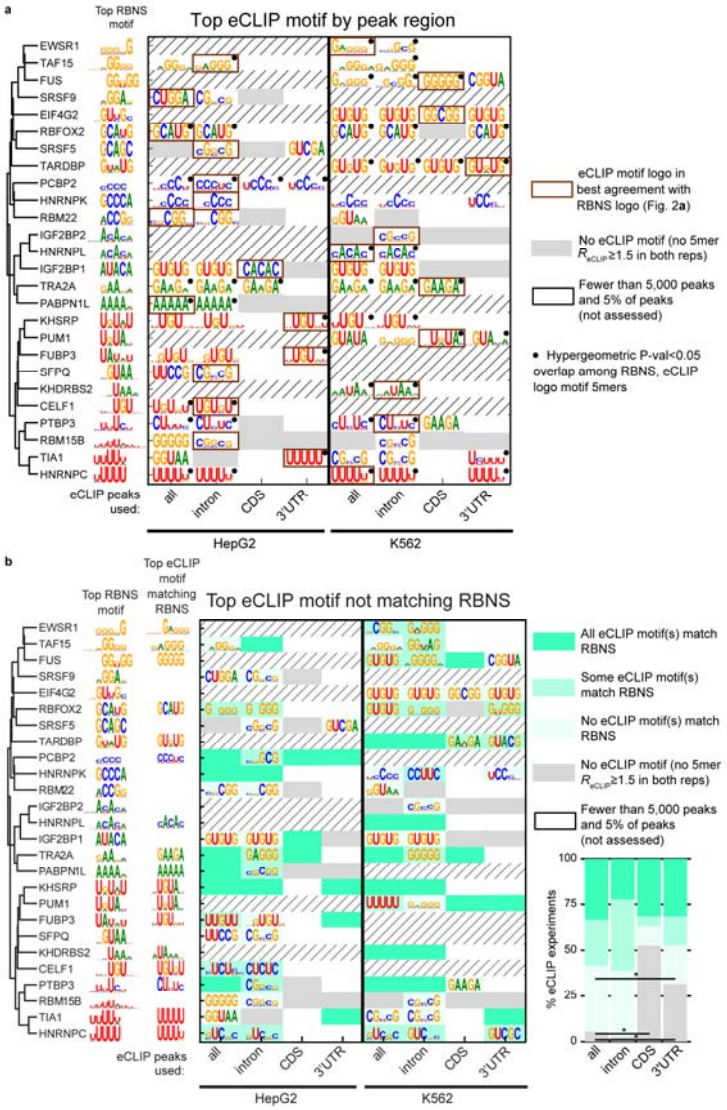
0.05, and ΔΨ > 0.05), and (blue) exons both bound by an RBP and differentially spliced upon RBP knockdown in the same cell type (with binding defined as a peak located anywhere between the upstream intron 5′ splice site and downstream intron 3′ splice site). Shaded regions indicate tenth to ninetieth percentile.

(d) Lines indicate the mean cumulative fraction of bases covered by peaks for 100 random orderings of the 181 eCLIP datasets, separated by transcript regions as indicated, with shaded region indicating tenth and ninetieth percentiles, only considering genes expressed (with TPM > 1) in both K562 and HepG2.

(e) Data and colors as in (d), represented as fold-increase in mean bases covered by peaks from *n* to *n+1* eCLIP datasets.

(f). Points indicate the fold-increase in bases covered by peaks between sampling one or two datasets, separated by whether the second is the same RBP in a new cell type (K_A -> H_A or H_A -> K_A for RBP A profiled in K562 and then HepG2 or HepG2 and then K562 respectively) or a different RBP in the same cell type (K_A -> K_B or H_A -> H_B for RBP A followed by RBP B in either K562 or HepG2 respectively), with kernel smoothed density indicated by the shaded area. Red line indicates mean.

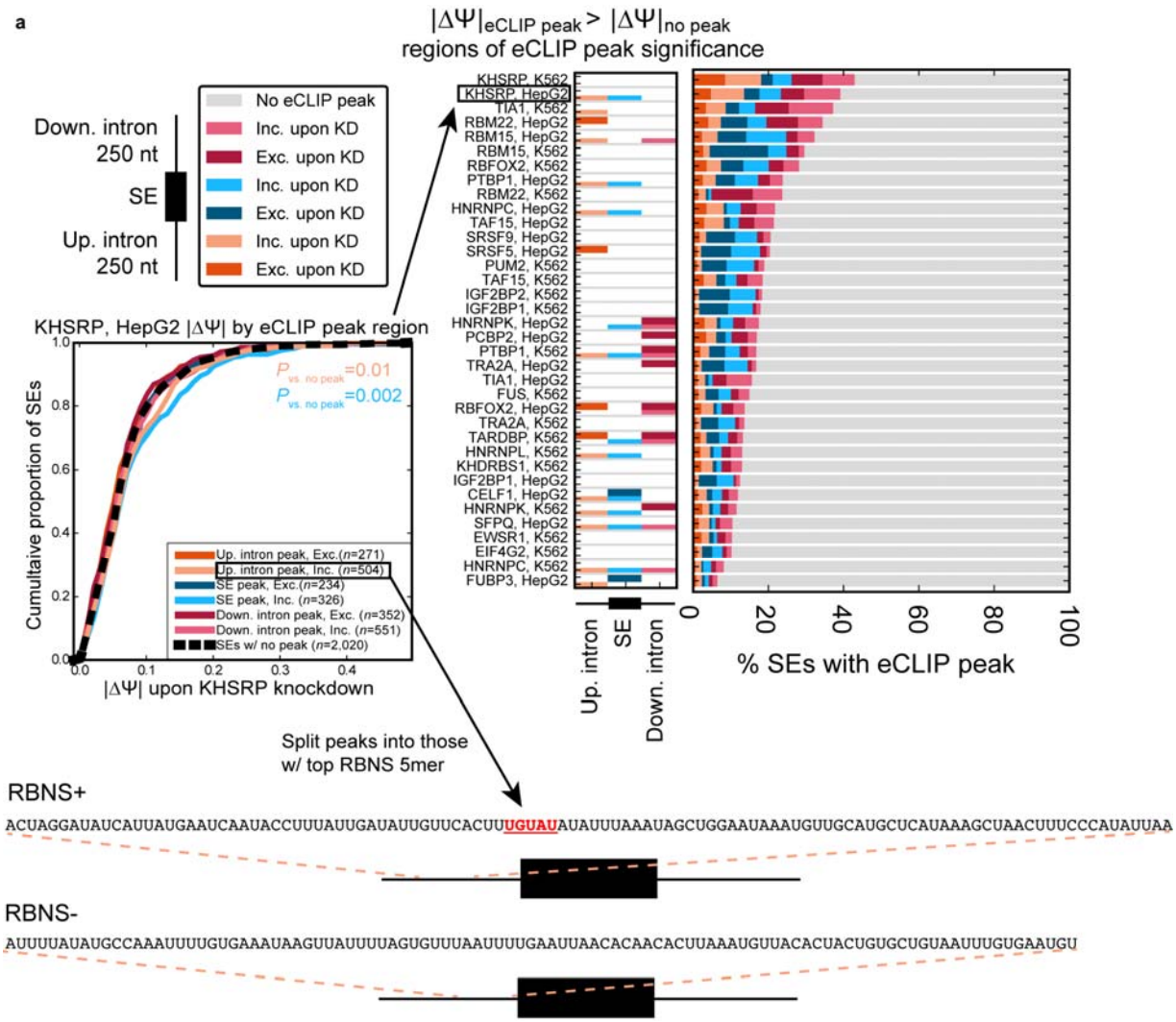
(g) Points indicate the fold-increase in bases covered by peaks between sampling all versus leaving one dataset out, separated by whether the RBP is (left) a newly profiled RBP or (center) a previously profiled RBP profiled in a second cell type (of either K562 or HepG2). (right) The fold-increase observed if an independent eCLIP experiment performed in H1 or H9 human embryonic stem cells is added (including RBFOX2, IGF2BP3, and two replicates each for IGF2BP1 and IGF2BP2). Red line indicates mean.

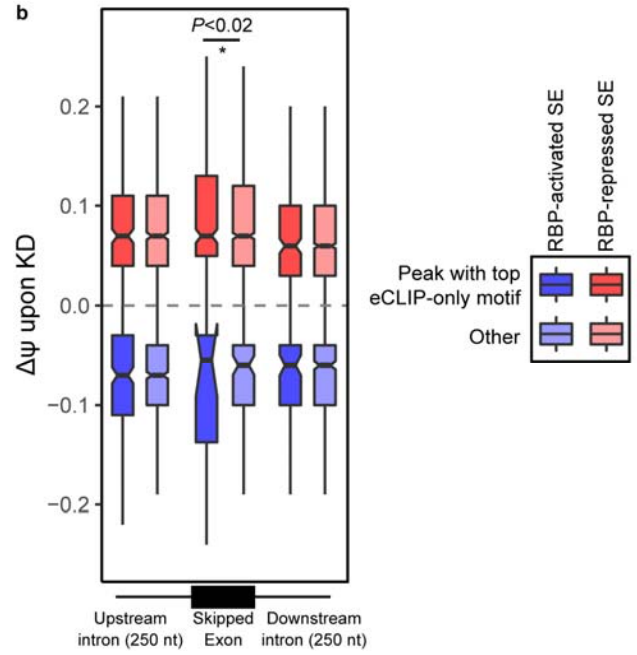

a
Top RBNS motif
Top eCLIP motif by peak region
EWSR1
TAF15
FUS
SRSF9
EIF4G2
RBFOX2
SRSF5
TARDBP
PCBP2
HNRNPK
RBM22
IGF2BP2
HNRNPL
IGF2BP1
TRA2A
PABPN1L
KHSRP
PUM1
FUBP3
SFPQ
KHDRBS2
CELF1
PTBP3
RBM15B
TIA1
HNRNPC
eCLIP peaks used:
all
intron
CDS
3'UTR
HepG2
K562
eCLIP motif logo in best agreement with RBNS logo (Fig. 2a)
No eCLIP motif (no 5mer R_{eCLIP}≥1.5 in both reps)
Fewer than 5,000 peaks and 5% of peaks (not assessed)
• Hypergeometric P-val<0.05 overlap among RBNS, eCLIP logo motif 5mers
b
Top RBNS motif
Top eCLIP motif matching RBNS
Top eCLIP motif not matching RBNS
All eCLIP motif(s) match RBNS
Some eCLIP motif(s) match RBNS
No eCLIP motif(s) match RBNS
No eCLIP motif (no 5mer R_{eCLIP}≥1.5 in both reps)
Fewer than 5,000 peaks and 5% of peaks (not assessed)
% eCLIP experiments
100
75
50
25
0

Extended Data Figure 3 | Comparison of *in vitro* RBNS-derived motifs with *in vivo* eCLIP-derived motifs.

(a) Top motif derived from all eCLIP peaks as well as eCLIP peaks within intronic, CDS, and 3′UTR regions. Motifs were only derived if there were at least 5,000 peaks or 5% of total peaks in that region, averaged over the two eCLIP replicates. Dashed lines indicate eCLIP was not performed in that cell line. Filled circles indicate significant overlap ($P < 0.05$ by hypergeometric test) between RBNS and eCLIP motifs.

(b) The top eCLIP motif that does not match RBNS for the corresponding RBP (if any). The eCLIP motif was considered as matching RBNS if any of its constituent 5mers were among the RBNS $Z \geq 3$ 5mers (always using at least 10 RBNS 5mers if there were fewer with $Z \geq 3$). Dashed lines indicate eCLIP was not performed in that cell line. (right) The percentage of eCLIP experiments aggregated over all RBP/cell types in each category of agreement with RBNS. Horizontal line indicates a significant difference in the proportion of a particular eCLIP/RBNS agreement category between eCLIP analysis of all peaks versus eCLIP analysis of intron, CDS, or 3′UTR peaks ($P < 0.05$ by Fisher's Exact Test).

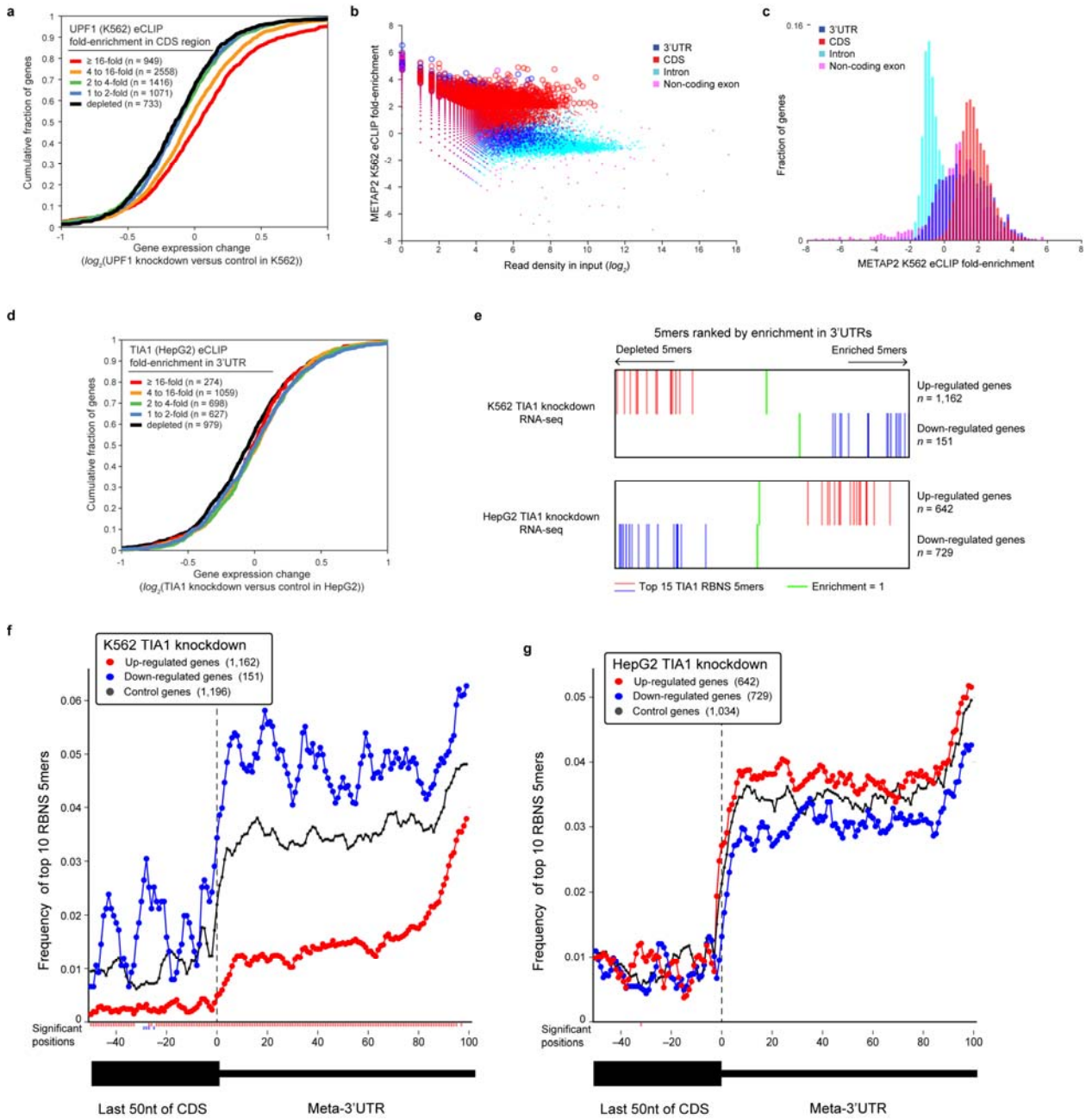

a
|ΔΨ|eCLIP peak > |ΔΨ|no peak
regions of eCLIP peak significance
Down. intron 250 nt
SE
Up. intron 250 nt
No eCLIP peak
Inc. upon KD
Exc. upon KD
Inc. upon KD
Exc. upon KD
Inc. upon KD
Exc. upon KD
KHSRP, HepG2 |ΔΨ| by eCLIP peak region
Cumulative proportion of SEs
|ΔΨ| upon KHSRP knockdown
Up. intron peak, Exc.(n=271)
Up. intron peak, Inc. (n=504)
SE peak, Exc.(n=234)
SE peak, Inc. (n=326)
Down. intron peak, Exc. (n=352)
Down. intron peak, Inc. (n=551)
SEs w/ no peak (n=2,020)
% SEs with eCLIP peak
Up. intron
SE
Down. intron
Split peaks into those w/ top RBNS 5mer
RBNS+
ACUAGGAUAUCAUUAUGAAUCAAUACCUUUAUUGAUAUUGUUCACUUUGUAUAUAUUUAAAUAGCUGGAAUAAAUGUUGCAUGCUCAUAAAGCUAACUUUCCCAUAUUAA
RBNS-
AUUUUAUAUGCCAAAUUUUGUGAAAUAAGUUAUUUUAGUGUUUAAUUUUGAAUUAACACAACACUUAAAUGUUACACUACUGUGCUGUAAUUUGUGAAUGU


b
P<0.02
Δψ upon KD
Upstream intron (250 nt)
Skipped Exon
Downstream intron (250 nt)
RBP-activated SE
RBP-repressed SE
Peak with top eCLIP-only motif
Other

Extended Data Figure 4 | Splicing regulatory activity of RBNS+ and RBNS- eCLIP peaks.

(a) Left: The distribution of ΔΨ changes upon KD in each of the 6 eCLIP+ peak region/SE splicing change types compared to that of eCLIP- SEs for KHSRP in HepG2 (significant if $P<0.05$ by Wilcoxon rank-sum test). Center: Regions of significance for eCLIP+ vs. eCLIP- SEs for each eCLIP experiment. Right: Proportion of SEs in each of the six eCLIP+ types for each eCLIP experiment. Bottom: Classification of eCLIP+ peaks into RBNS+ and RBNS- based on the presence of the top RBNS 5mer, shown here for two of the KHSRP peaks in HepG2.

(b) Same set of RBPs and corresponding eCLIP+ peak region/SE splicing change types as used in Fig. 3c, but separating eCLIP peaks on whether they contain the top 'eCLIP-only' 5mer (based on the motifs from Extended Data Fig. 3b) instead of the top RBNS 5mer.



Extended Data Figure 5 | Association between RBP binding and RNA expression upon knockdown.

(a) Lines indicate cumulative distribution plots of gene expression fold-change (UPF1 knockdown versus control) for indicated categories of UPF1 eCLIP binding in K562 cells.

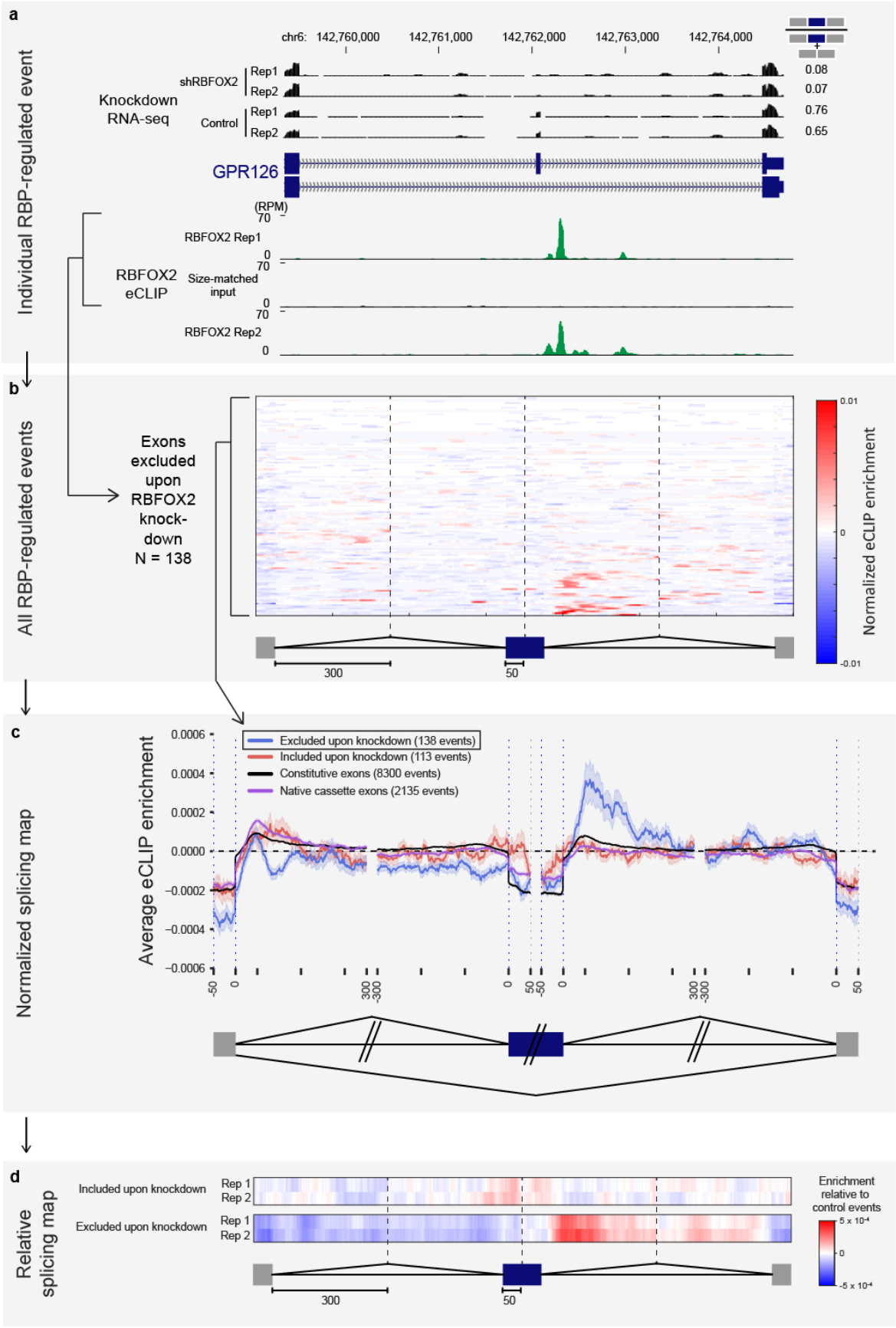
(b-c) METAP2 K562 eCLIP region-level binding at 3'UTR, CDS, intronic, and non-coding exonic regions. (b) Points indicate read density in input (x-axis) versus fold-enrichment in METAP2 eCLIP (y-axis) for indicated transcript regions of all GENCODE v19 genes. Significantly

enriched regions ($p \leq 10^{-5}$ and fold-enrichment ≥ 4) are indicated by open circles. (c) Histogram of METAP2 eCLIP fold-enrichment for the indicated transcript regions.

(d) Cumulative distribution plots of gene expression fold-change (TIA1 HepG2 knockdown versus control) for indicated categories of 3'UTR TIA eCLIP binding.

(e) Enrichment or depletion of the top 15 TIA1 RBNS 5mers in 3'UTRs of genes that are up- and down-regulated upon TIA1 knockdown in K562 and HepG2, relative to their frequency in control gene 3'UTRs (green lines indicate an enrichment of 1 (equal frequency in regulated gene 3'UTRs and control gene 3'UTRs)). All 1,024 5mers are ordered from lowest to highest enrichment from left to right in each row.

(f-g) Position-specific frequency of the top 10 TIA1 RBNS 5mers in the last 50 positions of the CDS and in a meta-3'UTR of (red) up-regulated, (blue) down-regulated, and (black) control genes upon TIA1 knockdown in (f) K562 and (g) HepG2 cells. Positions of motif density significantly different in up- or down-regulated genes relative to control genes are indicated below the x-axis (calculated using a binomial test comparing the number of regulated genes that do versus do not have one of the top 10 RBNS 5mers at that position versus the frequency observed in control genes).


a
Individual RBP-regulated event
chr6: 142,760,000
142,761,000
142,762,000
142,763,000
142,764,000
Knockdown RNA-seq
shRBFOX2
Rep1
Rep2
Control
Rep1
Rep2
0.08
0.07
0.76
0.65
GPR126
(RPM)
70
RBFOX2 Rep1
0
70
RBFOX2 eCLIP
Size-matched input
0
70
RBFOX2 Rep2
0
b
All RBP-regulated events
Exons excluded upon RBFOX2 knock-down N = 138
300
50
0.01
0
-0.01
Normalized eCLIP enrichment
c
Normalized splicing map
Average eCLIP enrichment
0.0006
0.0004
0.0002
0.0000
−0.0002
−0.0004
−0.0006
Excluded upon knockdown (138 events)
Included upon knockdown (113 events)
Constitutive exons (8300 events)
Native cassette exons (2135 events)
-50
0
300
-300
0
50
-50
0
300
-300
0
50
d
Relative splicing map
Included upon knockdown
Rep 1
Rep 2
Excluded upon knockdown
Rep 1
Rep 2
300
50
Enrichment relative to control events
5 x 10⁻⁴
0
-5 x 10⁻⁴

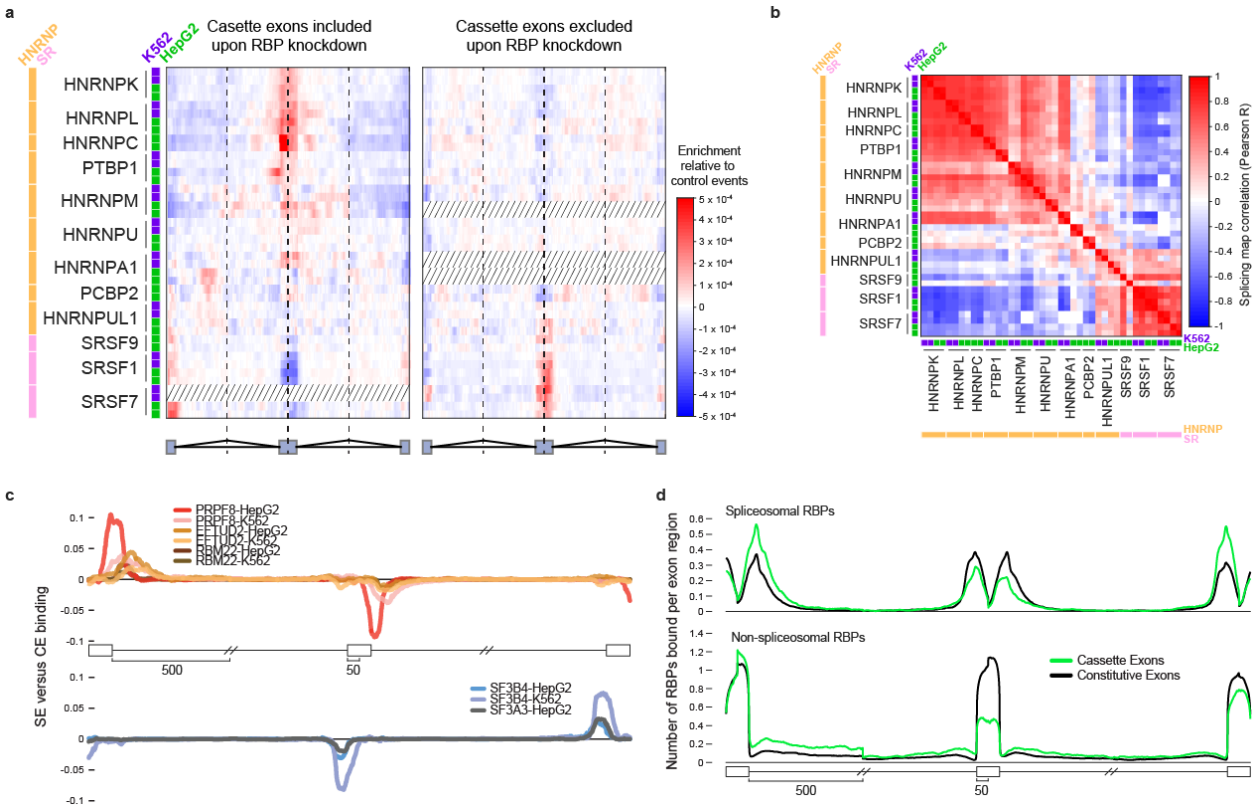
Extended Data Figure 6 | Generation of splicing maps for RBFOX2.

(a) First, individual RBP-regulated splicing events are identified from significant changes in knockdown RNA-seq. Genome browser tracks indicate RNA-seq read density (as reads per million (RPM)) and eCLIP read density (RPM) of RBFOX2 in the same cell type, as well as its paired size-matched input.

(b) Next, each exon is normalized between IP versus input to obtain 'Normalized eCLIP enrichment'. The heatmap indicates normalized eCLIP enrichment for all exons significantly excluded upon RBFOX2 knockdown.

(c) Next, a 'splicing map' is created by calculating the mean and standard error of the mean of normalized eCLIP enrichment for each position across the region, removing the top and bottom 5% outlier values at each position. Lines in splicing map indicate 'Average eCLIP enrichment', defined as the mean normalized eCLIP enrichment for exons (red) included or (blue) excluded upon RBFOX2 knockdown. Also plotted are (purple) a control set of cassette exons (referred to as 'native' cassette exons) in wild-type HepG2 cells and (black) constitutive exons.

(d) A final simplified splicing map vector was calculated by subtracting the normalized eCLIP enrichment of control native cassette exons from that of either included or excluded exons at each position to calculate 'Enrichment relative to control events'.



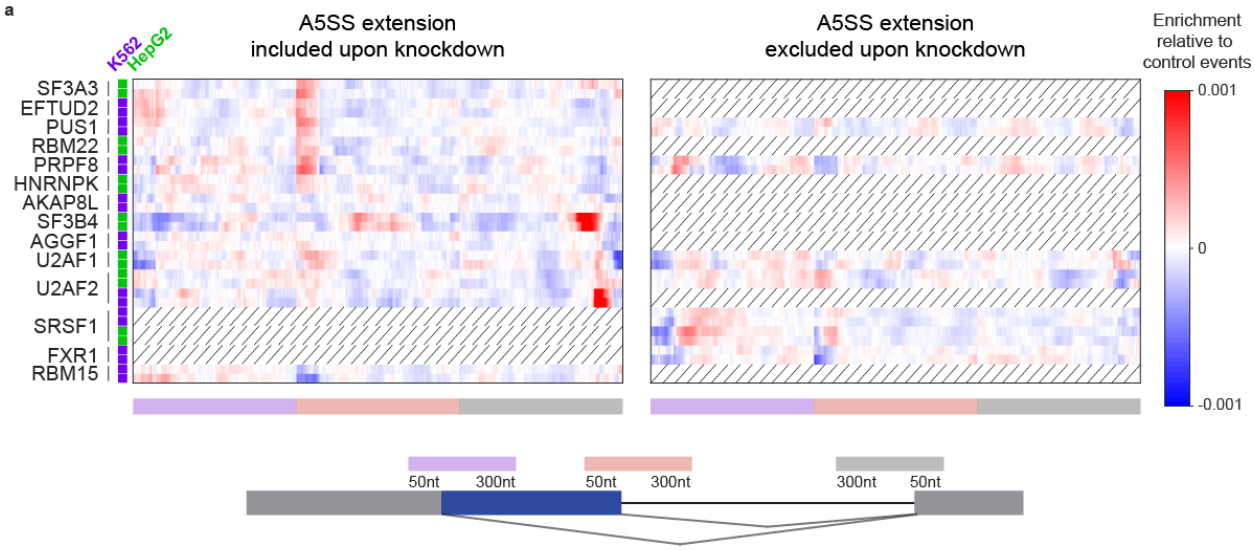
Extended Data Figure 7 | Splicing regulatory patterns of SR, HNRNP, and spliceosomal proteins

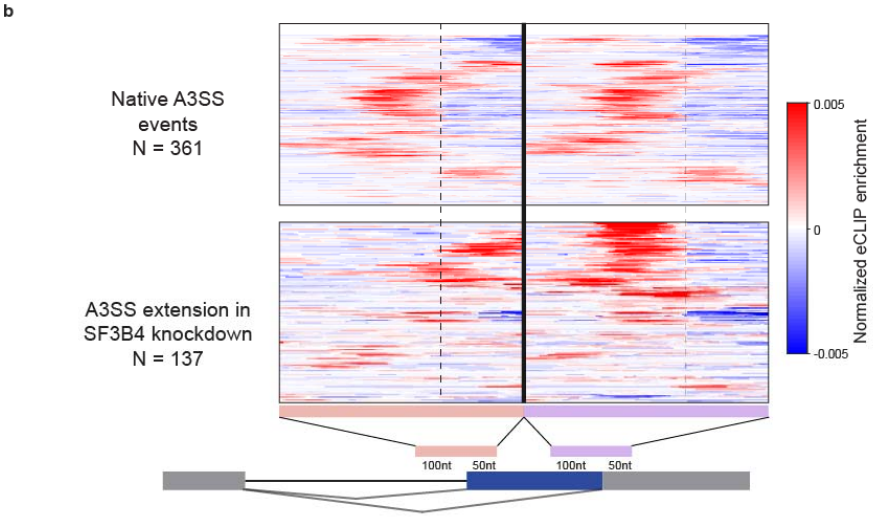
(a) Relative splicing maps for cassette exons included (left) and excluded (right) upon knockdown (as described in Fig. 5b) are shown for all profiled SR and HNRNP proteins. Datasets were hierarchically clustered at the RBP-level, and datasets with less than 100 events are indicated by slashed lines.

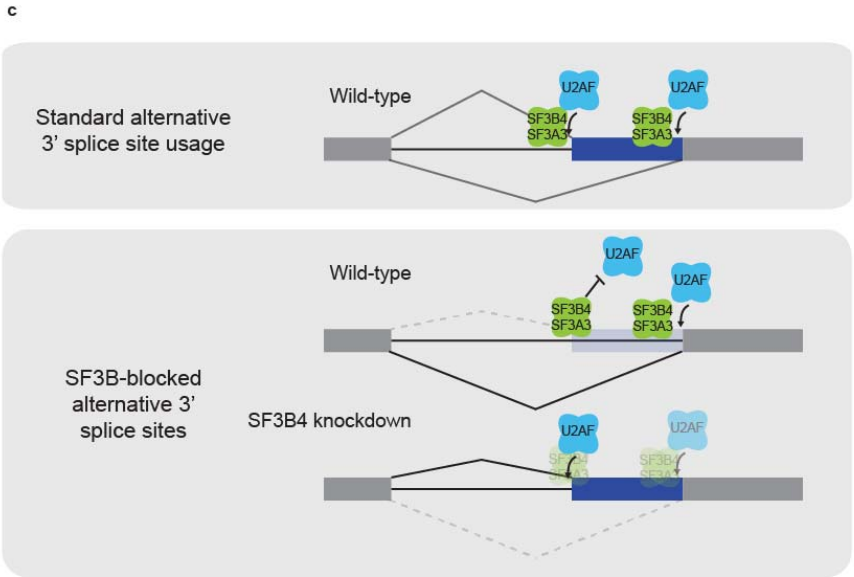
(b) Heatmap indicates the Pearson correlation (R) between splicing maps shown in (a), calculated across both included and excluded exon maps. Datasets are ordered identically to (a).

(c) Average binding patterns for indicated spliceosome-associated RBPs in 50nt exonic and 500nt intronic regions flanking splice sites. Lines indicate the difference in average binding for native cassette exons minus the average at constitutive exons.

(d) Lines indicate the average number of RBPs bound (out of 181 total datasets) in 50nt exonic and 500nt intronic regions flanking splice sites, separated by(top) core spliceosomal RBPs or (bottom) all other RBPs.


a
A5SS extension
included upon knockdown
A5SS extension
excluded upon knockdown
Enrichment
relative to
control events
K562
HepG2
SF3A3
EFTUD2
PUS1
RBM22
PRPF8
HNRNPK
AKAP8L
SF3B4
AGGF1
U2AF1
U2AF2
SRSF1
FXR1
RBM15
0.001
0
-0.001
50nt
300nt
50nt
300nt
300nt
50nt


b
Native A3SS
events
N = 361
A3SS extension in
SF3B4 knockdown
N = 137
0.005
0
-0.005
Normalized eCLIP enrichment
100nt
50nt
100nt
50nt

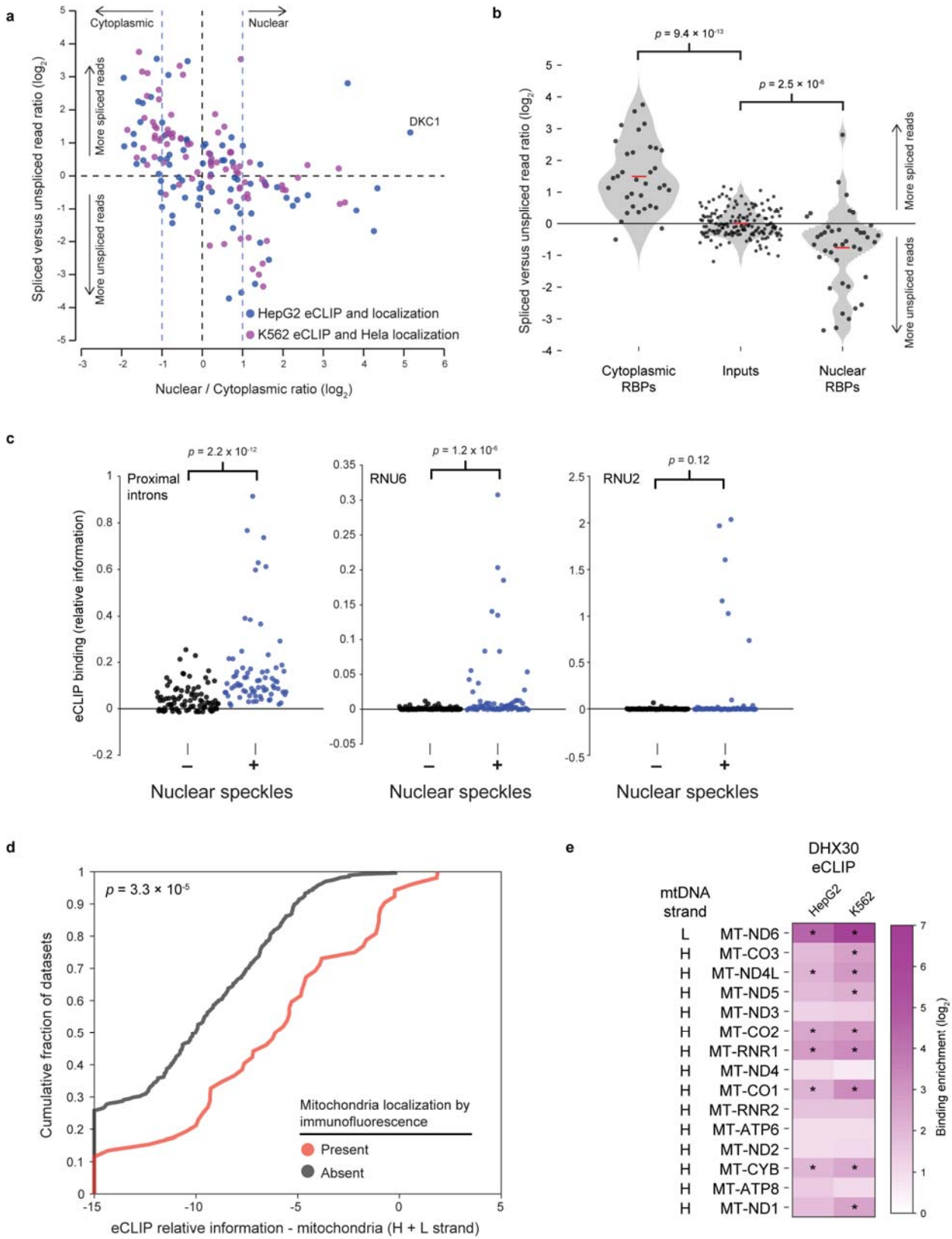

c
Standard alternative
3' splice site usage
Wild-type
U2AF
U2AF
SF3B4
SF3A3
SF3B4
SF3A3
SF3B-blocked
alternative 3'
splice sites
Wild-type
U2AF
U2AF
SF3B4
SF3A3
SF3B4
SF3A3
SF3B4 knockdown
U2AF
U2AF

Extended Data Figure 8 | RNA maps for alternative 5' and 3' splice sites.

a) Heatmap indicates enrichment relative to control events at alternative 5' splice site events, for all RBPs with eCLIP and knockdown RNA-seq data (requiring a minimum of 50 significantly changing events upon knockdown). The region shown extends 50 nt into exons and 300 nt into introns.

(b) Heatmap indicates normalized eCLIP signal for SF3B4 in HepG2 at alternative 3' splice site events either (top) alternatively spliced in wild-type cells or (bottom) events with increased usage of the extended 3' splice site upon SF3B4 knockdown. The region shown extends 50 nt into exons and 100 nt into introns.

(c) Model for SF3B4 and SF3A3 blockage of 3' splice site recognition by U2AF.


a
Cytoplasmic
Nuclear
More spliced reads
More unspliced reads
DKC1
HepG2 eCLIP and localization
K562 eCLIP and Hela localization
Spliced versus unspliced read ratio (log2)
Nuclear / Cytoplasmic ratio (log2)
b
p = 9.4 × 10-13
p = 2.5 × 10-6
Spliced versus unspliced read ratio (log2)
More spliced reads
More unspliced reads
Cytoplasmic RBPs
Inputs
Nuclear RBPs
c
p = 2.2 x 10-12
Proximal introns
p = 1.2 x 10-6
RNU6
p = 0.12
RNU2
eCLIP binding (relative information)
Nuclear speckles
d
p = 3.3 × 10-5
Cumulative fraction of datasets
Mitochondria localization by immunofluorescence
Present
Absent
eCLIP relative information - mitochondria (H + L strand)
e
DHX30 eCLIP
HepG2
K562
mtDNA strand
L MT-ND6
H MT-CO3
H MT-ND4L
H MT-ND5
H MT-ND3
H MT-CO2
H MT-RNR1
H MT-ND4
H MT-CO1
H MT-RNR2
H MT-ATP6
H MT-ND2
H MT-CYB
H MT-ATP8
H MT-ND1
Binding enrichment (log2)

Extended Data Figure 9 | eCLIP binding patterns in subcellular space

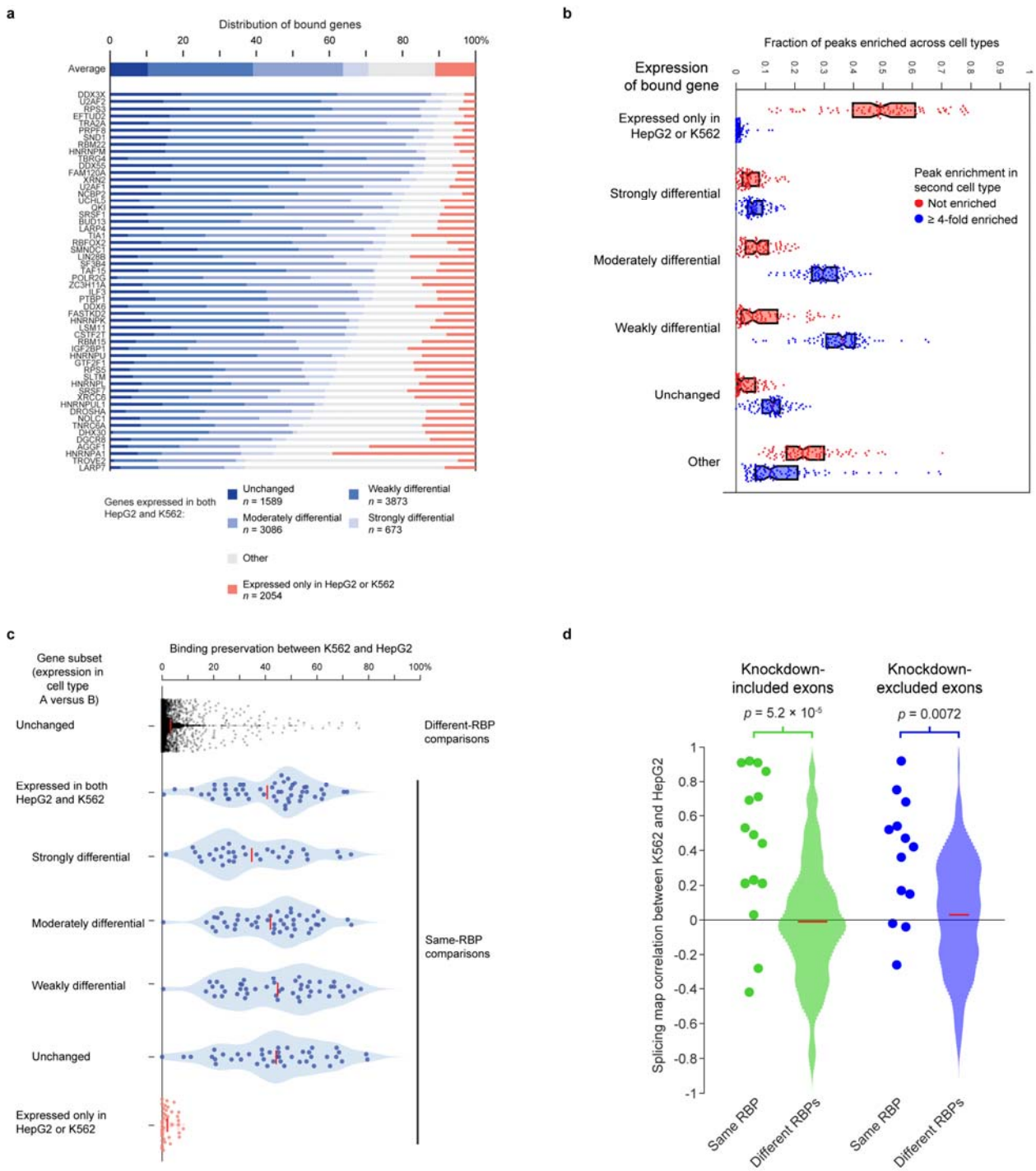
(a) Points indicate nuclear versus cytoplasmic ratio from immunoflourescence (IF) imaging (x-axis) versus ratio of spliced versus unspliced exon junction reads, normalized to paired input (y-axis). RBPs profiled by eCLIP and IF in HepG2 are indicated in blue, and RBPs profiled by eCLIP in K562 (in purple) were paired with IF experiments performed in Hela cells.

(b) Points indicate values as in (a), with RBPs separated into nuclear (nuclear / cytoplasmic ratio > 2) and cytoplasmic (nuclear / cytoplasmic ratio < 0.5). Significance was determined by Kolmogorov-Smirnov test, and red line indicates mean.

(c) Points indicate eCLIP relative information for all profiled RBPs for the indicated snRNA or pre-mRNA region, separated by the observation of co-localization at nuclear speckles. Significance was determined by Wilcoxon rank-sum test.

(d) Cumulative distribution curves indicate total relative information content for the mitochondrial genome for RBPs with mitochondrial localization by IF (red) and all other RBPs (black). Significance was determined by Kolmogorov-Smirnov test.

(e) Heatmap indicates DHX30 eCLIP binding across all exons for all mitochondrial protein-coding and rRNA transcripts.* indicates significant eCLIP binding (fold-enrichment ≥ 4 and $p \leq 0.00001$ in IP versus input).



Extended Data Figure 10 | Preservation of binding across cell types.

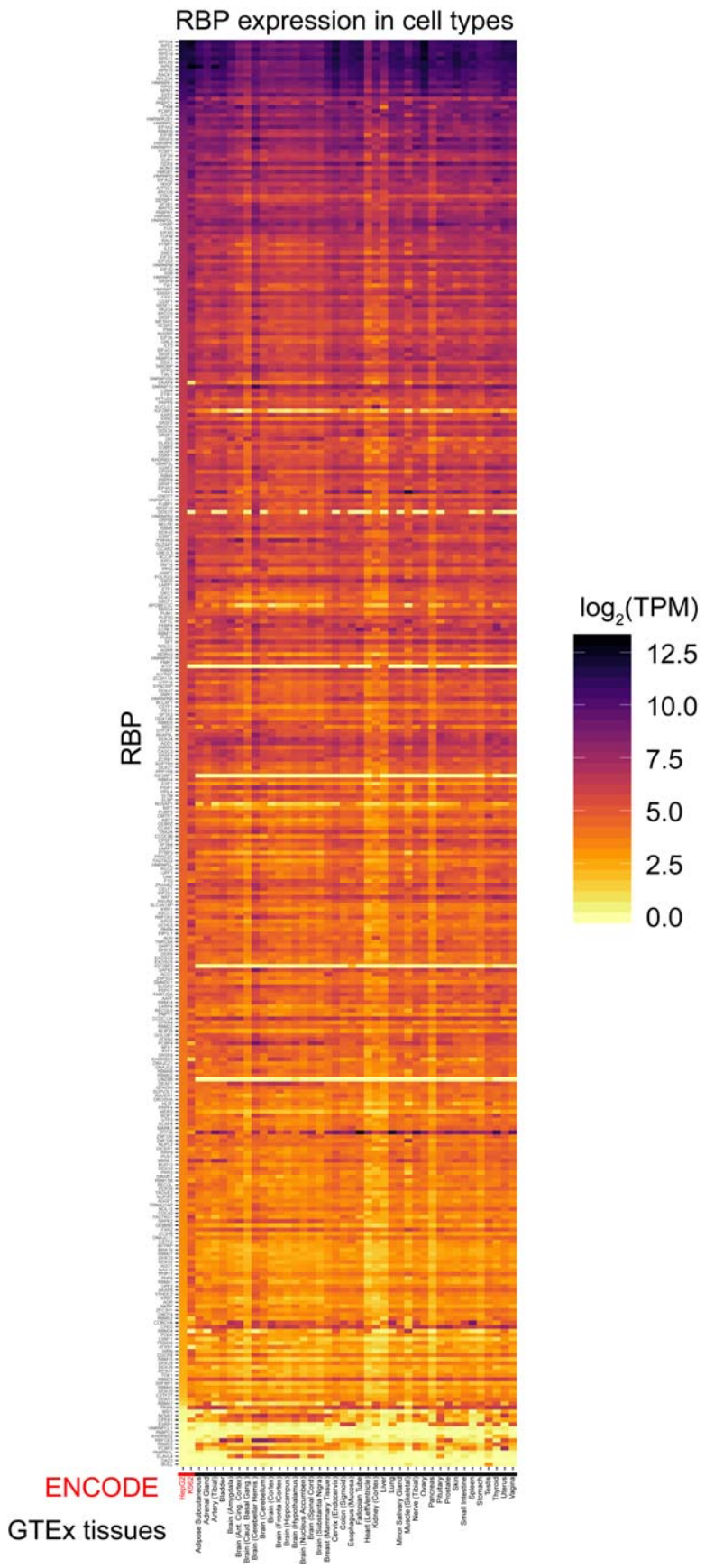
(a) Bars indicate the fraction of peaks observed for each RBP within sets of genes separated by their relative expression change between K562 and HepG2: unchanged (fold-difference ≤ 1.2), weakly (1.2 < fold-difference ≤ 2), moderately (2 < fold-difference ≤ 5) or strongly (fold-difference > 5) differential, or cell-type specific genes (TPM < 0.1 in one cell type and TPM ≥ 1

in the other). *n* indicates the number of genes meeting each criteria. For each RBP, the results shown are for the cell type with fewer total peaks.

(b) Each point represents one eCLIP dataset compared with the same RBP profiled in the second cell type. For the set of peaks from the first cell type that are not enriched (fold-enrichment < 1) in the second cell type, red points indicate the fraction occurring in genes with the indicated expression difference between HepG2 and K562. Blue points similarly indicate the gene distribution of peaks four-fold enriched in the opposite cell type. Boxes indicate quartiles, with median indicated by the central red line.

(c) Points indicate the fraction of overlapping peaks between K562 and HepG2 for RBPs profiled in both cell types (blue or red), or between one RBP in K562 and a second in HepG2 (black), for sets of genes separated by their relative expression change between K562 and HepG2 as in (a).

(d) Plot represents the distribution of Pearson correlations between splicing maps as shown in Figure 8d, separated by whether the comparison is between the same RBP or different RBPs profiled in two different cell types. Different RBPs are shown as smoothed histogram using a Normal kernel, and red line indicates mean. Significance was determined by Kolmogorov-Smirnov test.


RBP expression in cell types
RBP
log2(TPM)
12.5
10.0
7.5
5.0
2.5
0.0
ENCODE
GTEx tissues

Extended Data Figure 11 | Expression of RBPs across tissues and cell types.
Expression of the 352 RBPs (in Transcripts Per Million) investigated in this study in ENCODE cell lines HepG2 and K562 as well as 40 human tissues measured by the GTEx project. RBPs sorted by decreasing expression in HepG2.

Table legends

Supplementary Table 1. Manual annotation of RBP functions.

Supplementary Table 2. ENCODE accession identifiers of datasets used

References

1 Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nature reviews. Genetics* **15**, 829-845, doi:10.1038/nrg3813 (2014).

2 Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters* **582**, 1977-1986, doi:10.1016/j.febslet.2008.03.004 (2008).

3 Licatalosi, D. D. *et al.* HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464-469, doi:10.1038/nature07488 (2008).

4 Kao, D. I., Aldridge, G. M., Weiler, I. J. & Greenough, W. T. Altered mRNA transport, docking, and protein translation in neurons lacking fragile X mental retardation protein. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 15601-15606, doi:10.1073/pnas.1010564107 (2010).

5 King, O. D., Gitler, A. D. & Shorter, J. The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease. *Brain research* **1462**, 61-80, doi:10.1016/j.brainres.2012.01.016 (2012).

6 Lagier-Tourenne, C., Polymenidou, M. & Cleveland, D. W. TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. *Human molecular genetics* **19**, R46-64, doi:10.1093/hmg/ddq137 (2010).

7 Nussbacher, J. K., Batra, R., Lagier-Tourenne, C. & Yeo, G. W. RNA-binding proteins in neurodegeneration: Seq and you shall receive. *Trends in neurosciences* **38**, 226-236, doi:10.1016/j.tins.2015.02.003 (2015).

8 Paronetto, M. P., Achsel, T., Massiello, A., Chalfant, C. E. & Sette, C. The RNA-binding protein Sam68 modulates the alternative splicing of Bcl-x. *The Journal of cell biology* **176**, 929-939, doi:10.1083/jcb.200701005 (2007).

9 Lukong, K. E., Chang, K. W., Khandjian, E. W. & Richard, S. RNA-binding proteins in human genetic disease. *Trends in genetics : TIG* **24**, 416-425, doi:10.1016/j.tig.2008.05.004 (2008).

10 Martini, A. *et al.* Recurrent rearrangement of the Ewing's sarcoma gene, EWSR1, or its homologue, TAF15, with the transcription factor CIZ/NMP4 in acute leukemia. *Cancer research* **62**, 5408-5412 (2002).

11 Sonenberg, N., Morgan, M. A., Testa, D., Colonno, R. J. & Shatkin, A. J. Interaction of a limited set of proteins with different mRNAs and protection of 5'-caps against

pyrophosphatase digestion in initiation complexes. *Nucleic acids research* **7**, 15-29 (1979).

12 Sonenberg, N., Rupprecht, K. M., Hecht, S. M. & Shatkin, A. J. Eukaryotic mRNA cap binding protein: purification by affinity chromatography on sepharose-coupled m7GDP. *Proceedings of the National Academy of Sciences of the United States of America* **76**, 4345-4349 (1979).

13 Baltz, A. G. *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Molecular cell* **46**, 674-690, doi:10.1016/j.molcel.2012.05.021 (2012).

14 Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393-1406, doi:10.1016/j.cell.2012.04.031 (2012).

15 Kwon, S. C. *et al.* The RNA-binding protein repertoire of embryonic stem cells. *Nature structural & molecular biology* **20**, 1122-1130, doi:10.1038/nsmb.2638 (2013).

16 Brannan, K. W. *et al.* SONAR Discovers RNA-Binding Proteins from Analysis of Large-Scale Protein-Protein Interactomes. *Molecular cell* **64**, 282-293, doi:10.1016/j.molcel.2016.09.003 (2016).

17 Sundararaman, B. *et al.* Resources for the Comprehensive Discovery of Functional RNA Elements. *Molecular cell* **61**, 903-913, doi:10.1016/j.molcel.2016.02.012 (2016).

18 Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature methods* **13**, 508-514, doi:10.1038/nmeth.3810 (2016).

19 Lambert, N. *et al.* RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Molecular cell* **54**, 887-900, doi:10.1016/j.molcel.2014.04.016 (2014).

20 Naftelberg, S., Schor, I. E., Ast, G. & Kornblihtt, A. R. Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annual review of biochemistry* **84**, 165-198, doi:10.1146/annurev-biochem-060614-034242 (2015).

21 Ji, X. *et al.* SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. *Cell* **153**, 855-868, doi:10.1016/j.cell.2013.04.028 (2013).

22 Zhang, X. *et al.* An Atomic Structure of the Human Spliceosome. *Cell* **169**, 918-929 e914, doi:10.1016/j.cell.2017.04.033 (2017).

23 Silverman, I. M. *et al.* RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Genome biology* **15**, R3, doi:10.1186/gb-2014-15-1-r3 (2014).

24 Mauger, D. M., Lin, C. & Garcia-Blanco, M. A. hnRNP H and hnRNP F complex with Fox2 to silence fibroblast growth factor receptor 2 exon IIIc. *Molecular and cellular biology* **28**, 5403-5419, doi:10.1128/MCB.00739-08 (2008).

25 Damianov, A. *et al.* Rbfox Proteins Regulate Splicing as Part of a Large Multiprotein Complex LASR. *Cell* **165**, 606-619, doi:10.1016/j.cell.2016.03.040 (2016).

26 Sugimoto, Y. *et al.* Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome biology* **13**, R67, doi:10.1186/gb-2012-13-8-r67 (2012).

27 Hauer, C. *et al.* Improved binding site assignment by high-resolution mapping of RNA-protein interactions using iCLIP. *Nature communications* **6**, 7921, doi:10.1038/ncomms8921 (2015).

28 Yamasaki, S., Stoecklin, G., Kedersha, N., Simarro, M. & Anderson, P. T-cell intracellular antigen-1 (TIA-1)-induced translational silencing promotes the decay of selected mRNAs. *The Journal of biological chemistry* **282**, 30070-30077, doi:10.1074/jbc.M706273200 (2007).

29 Wigington, C. P. *et al.* Post-transcriptional regulation of programmed cell death 4 (PDCD4) mRNA by the RNA-binding proteins human antigen R (HuR) and T-cell intracellular antigen 1 (TIA1). *The Journal of biological chemistry* **290**, 3468-3487, doi:10.1074/jbc.M114.631937 (2015).

30 Ule, J. *et al.* An RNA map predicting Nova-dependent splicing regulation. *Nature* **444**, 580-586, doi:10.1038/nature05304 (2006).

31 Erkelenz, S. *et al.* Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. *Rna* **19**, 96-102, doi:10.1261/rna.037044.112 (2013).

32 Smith, C. W., Chu, T. T. & Nadal-Ginard, B. Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Molecular and cellular biology* **13**, 4939-4952 (1993).

33 Bradley, R. K., Merkin, J., Lambert, N. J. & Burge, C. B. Alternative splicing of RNA triplets is often regulated and accelerates proteome evolution. *PLoS biology* **10**, e1001229, doi:10.1371/journal.pbio.1001229 (2012).

34 Engreitz, J. M., Ollikainen, N. & Guttman, M. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nature reviews. Molecular cell biology* **17**, 756-770, doi:10.1038/nrm.2016.126 (2016).

35 Skalska, L., Beltran-Nebot, M., Ule, J. & Jenner, R. G. Regulatory feedback from nascent RNA to chromatin and transcription. *Nature reviews. Molecular cell biology* **18**, 331-337, doi:10.1038/nrm.2017.12 (2017).

36 Wei, C. *et al.* RBFox2 Binds Nascent RNA to Globally Regulate Polycomb Complex 2 Targeting in Mammalian Genomes. *Molecular cell* **62**, 875-889, doi:10.1016/j.molcel.2016.04.013 (2016).

37 Makeyev, A. V. & Liebhaber, S. A. The poly(C)-binding proteins: a multiplicity of functions and a search for mechanisms. *Rna* **8**, 265-278 (2002).

38 Spector, D. L. & Lamond, A. I. Nuclear speckles. *Cold Spring Harbor perspectives in biology* **3**, doi:10.1101/cshperspect.a000646 (2011).

39 Rackham, O., Mercer, T. R. & Filipovska, A. The human mitochondrial transcriptome and the RNA-binding proteins that regulate its expression. *Wiley interdisciplinary reviews. RNA* **3**, 675-695, doi:10.1002/wrna.1128 (2012).

40 Antonicka, H. & Shoubridge, E. A. Mitochondrial RNA Granules Are Centers for Posttranscriptional RNA Processing and Ribosome Biogenesis. *Cell reports*, doi:10.1016/j.celrep.2015.01.030 (2015).

41 Singh, N. N., Singh, R. N. & Androphy, E. J. Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes. *Nucleic acids research* **35**, 371-389, doi:10.1093/nar/gkl1050 (2007).

42 Lovci, M. T. *et al.* Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nature structural & molecular biology* **20**, 1434-1442, doi:10.1038/nsmb.2699 (2013).

43 Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096, doi:10.1126/science.1258096 (2014).

44 Nelles, D. A. *et al.* Programmable RNA Tracking in Live Cells with CRISPR/Cas9. *Cell* **165**, 488-496, doi:10.1016/j.cell.2016.02.054 (2016).

45 Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science* **350**, 1096-1101, doi:10.1126/science.aac7041 (2015).

46 Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology* **14**, R36, doi:10.1186/gb-2013-14-4-r36 (2013).

47 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

48 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323, doi:10.1186/1471-2105-12-323 (2011).

49 Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols* **7**, 562-578, doi:10.1038/nprot.2012.016 (2012).

50 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome biology* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).

51 Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology* **31**, 46-53, doi:10.1038/nbt.2450 (2013).

52 Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E5593-5601, doi:10.1073/pnas.1419161111 (2014).

53 Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods* **7**, 1009-1015, doi:10.1038/nmeth.1528 (2010).

54 Conway, A. E. *et al.* Enhanced CLIP Uncovers IMP Protein-RNA Targets in Human Pluripotent Stem Cells Important for Cell Adhesion and Survival. *Cell reports* **15**, 666-679, doi:10.1016/j.celrep.2016.03.052 (2016).

55 Kapeli, K. *et al.* Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses. *Nature communications* **7**, 12143, doi:10.1038/ncomms12143 (2016).

56 Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R. & Hofacker, I. L. The Vienna RNA websuite. *Nucleic acids research* **36**, W70-74, doi:10.1093/nar/gkn188 (2008).

57 Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660, doi:10.1126/science.1262110 (2015).

58 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology* **34**, 525-527, doi:10.1038/nbt.3519 (2016).

59 Ishimura, R. *et al.* RNA function. Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration. *Science* **345**, 455-459, doi:10.1126/science.1249749 (2014).