Human cells contain myriad excised linear intron RNAs with links to gene regulation and potential utility as biomarkers

Jun Yao,[1] Shelby Winans,[1] Hengyi Xu,[1] Douglas C. Wu,[1,3] Elizabeth A. Ferrick-Kiddie,[1] Manuel Ares, Jr.,[2] and Alan M. Lambowitz[1,4]

[1]Departments of Molecular Biosciences and Oncology
University of Texas at Austin
Austin TX 78712

[2]Department of Molecular, Cell, and Developmental Biology
University of California, Santa Cruz
Santa Cruz, California 95064

Running Title: Human excised linear intron RNAs

[3]Present address: Invitae Corporation

[4]To whom correspondence should be addressed. E-mail: lambowitz@austin.utexas.edu

ABSTRACT

By using TGIRT-seq, we identified >8,500 short full-length excised linear intron (FLEXI) RNAs in human cells. Subsets of FLEXIs accumulated in a cell-type specific manner, and ~200 corresponded to agotrons or mirtrons or encoded snoRNAs. Analysis of CLIP-seq datasets identified potential interactions between FLEXIs and >100 different RNA-binding proteins (RBPs), 53 of which had binding sites in ≥30 different FLEXIs. In addition to proteins that function in RNA splicing, these 53 RBPs included transcription factors, chromatin remodeling proteins, and cellular growth regulators that impacted FLEXI host gene alternative splicing and/or mRNA levels in knockdown datasets. We computationally identified six groups of RBPs whose binding sites were enriched in different subsets of FLEXIs: AGO1-4 and DICER associated with agotrons and mirtrons; AATF, DKC1, NOLCI, and SMNDC1 associated with snoRNA-encoding FLEXIs; two different combinations of alternative splicing factors found in stress granules; and two novel RBP-intron combinations, one including LARP4 and PABC4, which function together in the cytoplasm to regulate ribosomal protein translation. Our results suggest a model in which proteins involved in transcriptional regulation, alternative splicing, or post-splicing secondary functions bind and stabilize cell-type specific subsets of FLEXIs that perform different biological functions and have potential utility as biomarkers.

Key words: alternative splicing, cancer, diagnostics, miRNA, poly(A), ribosome assembly

INTRODUCTION

Most protein-coding genes in eukaryotes consist of coding regions (exons) separated by non-coding regions (introns), which are removed by RNA splicing to produce mRNAs. RNA splicing is performed by the spliceosome, a large ribonucleoprotein complex that catalyzes transesterification reactions yielding ligated exons and an excised intron lariat RNA, whose 5' end is linked to a branch-point nucleotide near its 3' end by a 2',5'-phosphodiester bond (Wilkinson et al., 2020). After splicing, this bond is typically hydrolyzed by a dedicated debranching enzyme (DBR1) to produce a linear intron RNA, which is rapidly degraded by cellular ribonucleases (Chapman and Boeke, 1991). In a few cases, excised intron RNAs persist after excision, either as branched circular RNAs (lariats whose tails have been removed) or as unbranched linear RNAs, with some found to contribute to cellular or viral regulatory processes (Farrell et al., 1991; Gardner et al., 2012; Kulesza and Shenk, 2006; Morgan et al., 2019; Moss and Steitz, 2013; Pek et al., 2015; Saini et al., 2019; Talhouarne and Gall, 2018; Zhang et al., 2013). The latter include a group of yeast introns that accumulate as unbranched linear RNAs and contribute to cell growth regulation and stress responses, possibly by sequestering spliceosomal proteins in stationary phase or other stress conditions (Morgan et al., 2019; Parenteau et al., 2019). Other examples are mirtrons, structured excised intron RNAs that are debranched by DBR1 and processed by DICER into miRNAs (Berezikov et al., 2007; Okamura et al., 2007; Ruby et al., 2007), and agotrons, structured excised linear intron RNAs that bind AGO2 and can repress target mRNAs in a miRNA-like manner (Hansen et al., 2016).

Recently, we identified 44 different short (≤300 nt) full-length excised linear intron (FLEXI) RNAs in human plasma by using Thermostable Group II Intron Reverse Transcriptase sequencing (TGIRT-seq), which gives continuous end-to-end sequence reads of structured RNAs (Yao et al., 2020). FLEXI RNAs were identified by peak calling as discrete blocks of continuous full-length intron reads that began and ended within 3 nt of the annotated splice sites. Subsets of plasma FLEXIs corresponded to annotated agotrons or pre-miRNAs of annotated mirtrons (denoted mirtron pre-miRNAs) (Yao et al., 2020). Almost all (>95%) of the plasma FLEXIs had

stable predicted secondary structures (≤-30 kcal/mol), and ~50% contained an annotated binding site for one or more RNA-binding proteins (RBPs), which could afford protection from highly active RNases in plasma. These findings suggested that FLEXI RNAs might have utility as stable RNA biomarkers in liquid biopsies (Yao et al., 2020).

Despite the precedents of agotrons and mirtrons, human excised linear intron RNAs have remained largely unexplored. Here, we used TGIRT-seq to search for FLEXIs in different human cells and tissues. We identified >8,500 different FLEXI RNAs, subsets of which accumulated in a cell-type specific manner and have potential utility as biomarkers. The identified FLEXIs included agotrons, mirtron pre-miRNAs, and snoRNA-encoding introns, as well computationally identified subsets that were enriched in binding sites for transcription factors, chromatin remodeling proteins and proteins that regulate cell growth and different steps in gene expression. Our findings suggest novel biological functions for FLEXI RNAs and provide a road map for further analyses of FLEXI-RBP interactions.

RESULTS and DISCUSSION

Identification of FLEXI RNAs in cellular RNA preparations

A search of the human genome (Ensembl GRCh38 Release 93 annotations; http://www.ensembl.org) identified 51,645 short introns (≤300 nt) in 12,020 different genes that could potentially give rise to FLEXI RNAs. To determine which short introns could be detected as FLEXI RNAs in human cells, we carried out TGIRT-seq of rRNA-depleted, unfragmented HEK-293T, K562, and HeLa S3 whole-cell RNAs, as well as Universal Human Reference RNA (UHRR), a commercial mixture of total RNAs from ten human cell lines (Fig. 1). TGIRT-seq is well suited for the detection of excised linear intron RNAs as the high processivity and strand displacement activity of TGIRTs together with reverse transcription at elevated temperatures enable continuous full-length sequence reads of structured small non-coding RNAs (sncRNAs), which are intractable to reverse transcription by the retroviral RTs used in other RNA-seq methods (Katibah et al., 2014; Qin et al., 2016; Shurtleff et al., 2017). In the TGIRT-seq datasets of unfragmented

cellular RNAs obtained for this study, most of the reads corresponded to full-length mature tRNAs, Y RNAs, snoRNAs, and snRNAs (Fig. S1 and S2). Further, because the variety of TGIRT enzyme used in this study has difficulty reading through long stretches of poly(A), reads from mature polyadenylated mRNAs comprised a low percentage of total reads (0.7-5.3%; Fig. S1), leading to the enrichment of reads mapping to excised intron RNAs.

To identify FLEXIs in the TGIRT-seq datasets, we compiled the coordinates of all short introns (≤300 nt) in Ensembl GRCh38 Release 93 annotations into a BED file and searched for the intersections with our data. As done previously for FLEXIs in plasma, we defined FLEXI reads as continuous intron reads that began and ended within 3 nt of the annotated splice sites (Yao et al., 2020). This definition enabled us to include full-length intron reads with small numbers of extra non-templated nucleotides added by cellular processes or during TGIRT-seq library preparations, but excluded long continuous intron reads that may be biologically significant but did not satisfy the strict definition of a FLEXI at either end. The 300-nt size cutoff was based on the size distribution of FLEXIs found in human plasma (Yao et al., 2020) and encompasses 85-92% of the FLEXIs detected in the different RNA preparations, including the most abundant FLEXI RNAs (Fig. S3A). The searches were done by using combined datasets obtained from multiple technical replicate libraries totaling 666 to 768 million mapped paired-end reads for each cellular RNA sample, with subsequent comparisons to biological replicates for four of the cell types (Table S1). In addition to human cellular RNAs, we searched for FLEXIs in remapped datasets of human plasma RNAs from healthy individuals (Yao et al., 2020).

By using this approach, we identified 8,144 different FLEXI RNAs originating from 3,743 different protein-coding genes, lncRNA genes, or pseudogenes (collectively denoted FLEXI host genes). The distribution of these FLEXI RNAs and their host genes between the different cellular RNA samples and plasma is shown in UpSet plots in Fig. 1A. Sixty five of the FLEXIs identified in the four cellular samples and plasma corresponded to annotated agotrons (Hansen et al., 2016) and 114 corresponded to pre-miRNAs of an annotated mirtron (Fig. 1B), both of which were shown previously to be debranched linear intron RNAs (Berezikov et al., 2007; Hansen et al.,

2016). The proportion of FLEXI RNAs corresponding to annotated agotrons or mirtron pre-mRNAs was higher in plasma (15.6%) than in cells (1.5-2.2%; $p < 1x10^{-15}$ by Fisher's exact test, Fig. 1B), possibly reflecting preferential cellular export and/or greater stability of structured RNAs in plasma (Yao et al., 2020). A small number of FLEXI RNAs found in cells but not plasma (43, 0.5% of the total) encoded snoRNAs, all of which were also detected as mature snoRNAs in the same samples (Fig. 1B).

In IGV alignments of reads mapping to a protein-coding gene, FLEXIs stand out from spliced exon and intron-exon junction reads as columns of continuous intron reads that began and ended within 3 nt of the annotated splice sites (Fig. 2A). The IGV alignments for 4I_JUP, an annotated agotron (Hansen et al., 2016), provide an example of an intron for which a high proportion of intron reads (94%) satisfied the conservative definition of a FLEXI in HeLa S3 cells, but had 3' truncations that resulted in lower proportions of the reads (2-3%) being counted as FLEXIs in the other three cellular RNA samples (Fig. 2A, top left). Some FLEXIs showed a propensity for indels at runs of U residues downstream of the putative branch-point nucleotide, but no indication of impediments or misincorporations at the putative branch-point nucleotide expected for a lariat RNA (Fig. 2A and Fig. 3B).

A histogram of the length distribution of all reads that mapped within an intron identified as a FLEXI in any of the four cellular RNA samples showed a peak at 90-100% intron length with only a smattering corresponding to shorter fragments (Fig. 2B). By contrast, reads mapping within other short introns (≤300 nt) annotated in GRCh38 corresponded to heterogeneously sized RNA fragments, as expected for rapid intron turnover after splicing (Fig. 2B). Further, most of the FLEXI reads began and ended directly at the splice sites (87% for the 5'-splice site and 58% for the 3'-splice site), with <10% extending a short distance into a flanking exon (Fig. 2C). Similar FLEXI length distributions were seen in each of the four individual cellular RNA sample types, including for different subsets of FLEXIs described further below, except those containing a binding site for DICER, which showed additional discrete peaks corresponding to shorter intron fragments that may be DICER cleavage products (Fig. S3B and C). RT-qPCR assays for 7

relatively abundant FLEXIs, including an annotated agotron (4I_JUP), confirmed that they were sensitive to digestion by Terminator (Lucigen), a commercial 5' to 3' exonuclease that degrades linear RNAs but not circular or lariat RNAs, which have blocked 5' ends (Fig. S4). Collectively, these findings show that TGIRT-seq can be used to identify large numbers of short full-length excised linear intron RNAs in human cells.

FLEXI RNAs have distinctive structural characteristics

Most of the 8,082 FLEXIs identified in the cellular RNA datasets have sequence characteristics of major U2-type spliceosomal introns (8,004 with GU-AG splice sites and 78 with GC-AG splice sites; Fig. 1C), with only 36 having sequence characteristics of minor U12-type spliceosomal introns (34 with GU-AG and 2 with AU-AC splice sites; Fig. S5A), and 23 having non-canonical splice sites (*e.g.,* AU-AG and AU-AU; Fig. S5B) (Burset et al., 2000; Sheth et al., 2006). The splice-site and branch-point (BP) consensus sequences of FLEXI RNAs matched literature consensus sequences for human introns (Burset et al., 2000; Gao et al., 2008), possibly with a higher propensity for C residues upstream of the branch-point A and 3'-splice site (Fig. 1C for U2-type FLEXIs and Fig. S5A for U12-type FLEXIs) (Gao et al., 2008). Although hydrolytic splicing, which would directly produce an excised linear intron RNA, cannot be excluded, their largely canonical branch-point sequences suggests that most if not all FLEXIs were excised as lariat RNAs and debranched after splicing, as found previously for mirtron pre-miRNAs and agotrons (Hansen et al., 2016; Okamura et al., 2007; Ruby et al., 2007).

We used density plots to compare the cumulative distribution of different properties of the >8,000 FLEXIs detected in the combined TGIRT-seq datasets for human cell lines (K-562, HEK-293T, HeLa S3, and UHHR) and plasma with those of other short introns in the same datasets (Fig. 1D). The plots showed that FLEXI RNAs differed significantly from other short introns in being skewed toward higher GC contents and lower minimum free energies (MFEs; ΔG) for the most stable secondary structure predicted by RNAfold ($p \leq 0.001$ for both by Wilcoxon signed-rank test; Fig. 1D), features that could contribute to their stability as FLEXIs. These

differences were most pronounced for the FLEXIs detected in human plasma, which were a relatively homogeneous subset with peaks at 90-nt length, 70% GC content, and -40 kcal/mole MFE (Fig. 1D). By comparison, the density plots for the FLEXIs detected in cells showed similar peaks but larger shoulders extending to longer lengths, lower GC contents, and less stable predicted secondary structures (≥-25 kcal/mole; Fig. 1D), again suggesting preferential export or greater resistance to plasma nucleases of shorter, more stably structured FLEXI RNAs.

Abundance of FLEXI RNAs

Density plots comparing the abundances (RPM) of FLEXIs in the cellular RNA samples (Fig. 1E) showed that they could be divided into two categories: one corresponding to low abundance FLEXIs (<0.002 RPM), which likely include transient intron RNA turnover intermediates, and the other corresponding to FLEXI RNAs that accumulate to higher abundance in different cell types (peaks between 0.002 and 0.1 RPM with a tail extending to 6.9 RPM) overlapping the lower end of the abundance distribution for snoRNAs (dashed yellow line). In HeLa S3 cells, the peak of more abundant FLEXIs was predominant, and agotrons comprised a higher proportion of more abundant FLEXIs than in the other cell types (Fig. 1E). The relatively abundant FLEXIs included 3,363 detected at ≥0.01 RPM in one or more of the cellular RNA samples.

To estimate copy number per cell values of FLEXI RNAs, we used sncRNAs of known cellular abundance in the TGIRT-seq datasets to produce a linear regression model for the relationship between sequence-bias corrected $\log_{10}$ RPM values and $\log_{10}$ transformed sncRNA copy number per cell values from the literature (Fig. S6A; Table S2). By using this method, we estimated that moderately abundant FLEXIs (0.01-0.99 RPM) were present at 15 to 239 copies per cell, and the most abundant FLEXIs (≥1 RPM) were present at 363 to 1,863 copies per cell (see linear regression plots in Fig. S6A). Based on the linear regression model, FLEXIs previously annotated as agotrons were present at 1 to 1,335 copies per cell; mirtron pre-miRNA FLEXIs at 1 to 225 copies per cell; and snoRNA-encoding FLEXIs at 1 to 265 copies per cell, where 1 is the minimum copy number cell value that could be estimated from the linear regression. Copy

per cell values determined by droplet digital PCR (ddPCR) for a sampling of FLEXIs were generally of the same order of magnitude as the TGIRT-seq quantitation, but varied with the choice of sncRNA standard, with the most abundant FLEXIs estimated to be present at 3,000 to 4,000 copies per cell similar to that of U7 RNA when the latter was used as a standard (Fig. S6B).

Abundant FLEXI RNAs have cell-type specific expression patterns

Scatter plots of FLEXIs detected in pairwise comparisons of different cell lines identified numerous relatively abundant FLEXI RNAs (≥0.01 RPM) that were reproducibly detected and differentially expressed between the two compared cell types (Fig. 3A and Fig. S7). Further, the expression patterns of these relatively abundant FLEXIs were more discriminatory between cell types than those of their corresponding host genes, indicating that FLEXI RNA abundance is controlled independently of host gene mRNA levels (Fig. 3A). IGV alignments showed that factors that may contribute to the more discriminatory expression patterns of FLEXIs included cell-type specific differences in alternative splicing, splicing efficiency, and intron RNA turnover, all of which can result in differences in the relative abundance of different FLEXIs transcribed from the same host gene (Fig. 3B).

The cell-type specific expression patterns of FLEXI RNAs were confirmed by principal component analysis (PCA), t-SNE (Kobak and Berens, 2019) and ZINB-WaVE (Risso et al., 2018), which showed clustering of the FLEXI RNA expression profiles in the cellular RNA datasets obtained for this study (including for MDA-MB-231 cells analyzed below) with those in TGIRT-seq datasets for additional biological and technical replicates for four of the cellular RNA sample types obtained from other studies (Fig. S8A, Table S1). The clustering of cell-type specific expression patterns improved when restricted to more abundant FLEXIs detected at ≥0.01 RPM, but only slightly compared to those for all FLEXIs and did not improve further with higher abundance cutoffs (Fig. S8B). This likely reflects that sample comparisons using the above algorithms are dominated by higher abundance FLEXIs. Collectively, these findings indicate that relatively abundant FLEXI RNAs have cell-type specific expression patterns.

Evolutionary conservation of FLEXIs

PhastCons scores provide a measure of evolutionary conservation, with higher phastCons scores indicating a higher probability that a sequence is conserved and functionally important across a range of organisms (Siepel et al., 2005). Density plots of phastCons scores calculated across 27 primates including humans plus mouse, dog, and armadillo showed that most FLEXIs, including those corresponding to mirtron pre-miRNAs or agotrons, were poorly conserved (peaks at 0.06-0.09 compared to 0.02 for other annotated short introns in GRCh38), but with tails extending to higher phastCons scores. FLEXIs encoding snoRNAs had higher phastCons scores (four at ≥0.5) than did other FLEXIs, likely indicating conservation of the snoRNA component (Fig. 1F, yellow line). The relatively low phastCons scores of most FLEXIs could reflect recent acquisition and/or rapid sequence divergence in the human lineage.

Although most FLEXI RNAs were poorly conserved, 2% (158) of the FLEXIs that were not annotated as mirtrons, agotrons or encoding snoRNAs had phastCons scores ≥0.75, possibly reflecting an evolutionarily conserved sequence-dependent function. Compared to those with lower phastCons scores, these more highly conserved FLEXIs were enriched in introns that were alternatively spliced (69% compared to 35%), shared a 5'- or 3'-splice site with a longer intron (44% compared to 18%), corresponded to a retained intron (47% compared to 25%), or contained an in-frame protein-coding sequence that would be expressed if all or part of the intron was retained within a mRNA (20% compared to 7.8%; Fig. 3C). The proportion of FLEXIs with these characteristics was even higher for the 44 most highly conserved FLEXIs (phastCons scores ≥0.99; Fig. 3C). A FLEXI in the human *EIF1* gene with phastCons score = 1.00 resulted from acquisition of a 3'-splice site in its highly conserved 3' UTR leading to expression of a novel human-specific EIF1 isoform (chr17:41,690,818-41,690,902; Supplemental File) (Kim et al., 2020). These findings suggest that in most if not all cases the sequence conservation of FLEXIs with high phastCons scores reflects evolutionary constraints on different protein isoforms produced by alternative splicing.

FLEXIs contain binding sites for >100 different RBPs with diverse cellular functions

To explore possible biological functions of FLEXIs, we searched for RBPs that have binding sites in FLEXI RNAs in CLIP-seq datasets (ENCODE high confidence eCLIP datasets for K562 and HepG2 cells (Van Nostrand et al., 2016) and AGO1-4 and DICER PAR-CLIP datasets for HEK-293 cells (Hafner et al., 2010; Rybak-Wolf et al., 2014)). More than half of the detected FLEXIs (4,505; 55%) contained a CLIP-seq-identified binding site for one or more of 126 different RBPs, with 53 RBPs (51 in the eCLIP datasets plus DICER and AGO1-4 in the PAR-CLIP datasets) having binding sites in ≥30 different FLEXI RNAs (Fig. 4A, Fig. S9A). These 53 RBPs included spliceosomal proteins and other proteins that function in RNA splicing; DICER, AGO1-4 and other proteins that have miRNA-related functions; and a surprising number of proteins that function in cellular regulatory processes, including transcriptional regulation, chromatin assembly and disassembly, cellular growth regulation, and stress responses (Fig. 4A and Table S3). Although cross-linking does not distinguish whether the RBP was bound to the intron before or after splicing, examination of sequence reads in the ENCODE eCLIP datasets showed that all of the RBPs that have a binding site in ≥30 different FLEXIs were cross-linked to ≥95% full-length intron RNAs corresponding to identified FLEXIs, with 32 showing a ≥2-fold enrichment of such reads compared to the control antibody datasets (Fig. S10). This finding indicates that whether or not the RBP bound to the intron before or after splicing, some fraction of each RBP remained bound to an excised intron RNA and may have protected it from RNase I digestion used for processing the cross-linked RNAs for sequencing in the ENCODE datasets.

RBPs that bind FLEXIs also bind other short and long introns

All 53 proteins in the ENCODE eCLIP and PAR-CLIP datasets that have cross-link identified binding sites in ≥30 different FLEXIs also bind thousands of other short and long introns (Fig. S9B and C). Scatter plots showed that RBPs whose binding sites were significantly over-represented in FLEXIs compared to long introns (significant enrichment defined as ≥1% of total RBP-binding sites; $p \leq 0.05$, Fisher's exact test) included spliceosomal proteins, as well as AGO1-4,

DICER, DDX24, GRWD1 (a histone-binding protein that functions in ribosome assembly and regulates chromatin dynamics), RBM22 (a spliceosomal component that may play a role in regulating cell division), UCHL5 (ubiquitin carboxy-terminal hydrolase), and ZNF622 (a putative transcriptional activator that is a positive regulator of apoptosis; Fig. 4B), with AGO1-4, DICER, and RBM22 also significantly over-represented in FLEXIs compared to other short introns (Fig. 4C). The scatter plots also identified numerous proteins whose binding sites were over-represented in other short or long introns compared to FLEXIs (Fig. 4B and C). The over-representation of spliceosomal protein-binding sites in FLEXI RNAs may reflect that most FLEXIs dissociate more slowly from spliceosomes than do long introns, providing greater opportunity for cross-linking. If so, the abundance of FLEXIs as a whole might be sufficiently high to affect intracellular levels of spliceosomal proteins in response to stimuli that globally affect FLEXI RNA turnover, as found for the collective of yeast linear introns that accumulate under stress conditions (Morgan et al., 2019).

FLEXI RBP knockdowns impact FLEXI host gene mRNA levels and alternative splicing

ENCODE shRNA knockdown datasets for K-562 and/or HepG2 cells were available for all 51 of the RBPs identified in the ENCODE eCLIP datasets as having a binding site in ≥30 different FLEXI RNAs (The ENCODE Project Consortium, 2012; Van Nostrand et al., 2016). By analyzing gene counts in the knockdown datasets, we identified 17 RBPs whose knockdown resulted in significant changes in mRNA expression levels from a group of host genes whose FLEXIs contained a binding site for the knocked down RBP compared to three groups composed of the same number of randomly selected genes ($p \leq 0.05$ by analysis of variance (ANOVA) test; Fig. 5A and Fig. S11; results summarized in grids below the bar graph in Fig. 4A). For GKPOW, LARP4, RPS3, TIAL1, U2AF2 and UCHL5, these changes were significantly biased toward decreased mRNA levels, while for DDX55, RBM22, and ZNF622, the changes were significantly biased toward increased mRNA levels ($p \leq 0.05$, Fisher's exact test based on comparison to all significantly changed genes; Fig. S11). For the remaining RBPs who knockdowns caused significant

changes in FLEXI host gene mRNA levels, the changes were not significantly biased in either direction (Fig. 5A and Fig. S11).

Knockdowns of 27 RBPs resulted in significant changes in alternative splicing (skipped exons (SE) or retained introns (RI)) of FLEXIs that contained a binding site for that RBP, but not FLEXIs that lacked such a binding site nor other short introns that were not detected as FLEXIs (Fig. 5B, Fig. S12, results summarized in Fig. 4A; $p \leq 0.05$, Kolmogorov-Smirnov test). In addition to proteins that function in RNA splicing, these included 17 proteins with no previously known RNA splicing-related function (AATF, DKC1, GRWD1, IGF2BP1, LARP4, LIN28B, LSM11, METAP2, NOLC1, PABPC4, RBM15, RBM22, RPS3, SUB1, UCHL5, XRN2, and YBX3). Knockdown of 9 of these proteins also affected mRNA levels of FLEXI host genes (see grids in Fig. 4A). Collectively, knockdown of 34 of the 51 RBPs identified in the ENCODE eCLIP datasets as cross-linking to ≥30 FLEXIs resulted in significant changes in mRNA expression levels and/or alternative splicing of host genes for bound FLEXIs, indicating functional interactions with introns that give rise to FLEXI RNAs.

Mapping of RBP-binding sites in intron RNAs

To gain insight into how RBP binding might affect the production and function of FLEXI RNAs, we collected CLIP-seq reads containing intron sequences for the 53 RBPs that have a binding site in ≥30 FLEXIs in the ENCODE eCLIP and AGO1-4 and DICER PAR-CLIP datasets and mapped their binding sites in FLEXIs and in other categories of introns (Fig. 5C and Fig. S13, results summarized in Fig. 4A). We found that 37 of these proteins bind at or near the 5'- and 3'-splice sites (SS), while 16 have binding sites more uniformly distributed throughout the body of the intron (I). For those proteins that bind at or near the splice sites, the midpoint of the annotated binding site peak in the FLEXI RNA was either at the splice site or immediately adjacent to it within the intron (SS-I) or flanking exon (SS-E). Notably, the RBPs whose binding sites were at or near the 5'- and 3'-splice sites included not only splicing factors, but also many proteins whose primary function is not RNA splicing, including AGO1-4, BCLAF1 (BCL2-2-associated

transcription factor), DDX3X, DDX24, DDX55, FXR2, G3BP1, GEMIN5, GPKOW, GRWD1, IGF2BP, LARP4, LIN28B, LSM11, METAP2, NCBP2, PABPC4, PABPN1, PUM1, RPS3, SND1, SRSF1, SUB1 (Transcriptional Regulator), UCHL5, YBX3, ZNF622, and ZNF800). The RBPs that have binding sites more uniformly distributed throughout the FLEXI included AATF (Apoptosis Antagonizing Transcription Factor), DICER, DKC1 (Dyskerin), EFTUD2, KHSRP, NOLC1, PCBP1, PCBP2, PRPF4, RBFOX2, RBM5, RBM15, RBM22, SMNDC1, TIA1, and XRN2, a nuclear 5' to 3' exonuclease that degrades excised linear intron RNAs (Fig. 5C and Fig. S13). In some cases, RBPs whose binding sites were localized near the 5'- and 3'-splice sites in long introns had binding sites more centrally localized in FLEXIs (*e.g.,* RBM5; Fig. 5C; RBM15, TIA1, TIAL1; Fig. S13). The RBP-binding site locations suggest possible mechanisms by which individual RBPs may affect the production and/or stability of FLEXI RNAs.

Identification of RBP-binding sites enriched in agotrons, mirtrons, snoRNA-encoding FLEXIs, and evolutionarily conserved FLEXIs

Although FLEXIs as a whole are enriched in binding sites for spliceosomal proteins (Fig. 4B), it remained possible that subsets of FLEXI RNAs might be enriched in binding sites for RBPs that perform specialized biological functions. Such FLEXIs seemed likely to include agotrons and mirtrons, which bind AGO1-4 and DICER, proteins that recognize branch-point occluding secondary structures whose formation may require full or partial dissociation of the spliceosome (Berezikov et al., 2007; Hansen et al., 2016). Consistent with this prediction, scatter plots confirmed that FLEXIs annotated as agotrons (n=61) or mirtrons (n=110) in the cellular RNA samples were significantly enriched in binding sites for AGO1-4 and DICER relative to those for spliceosomal or other proteins (significant enrichment defined as ≥2% of total RBP-binding sites; $p \leq 0.05$ by Fisher's exact test; Fig. 6A and B, left panel). Density plots showed that the cellular FLEXI RNAs annotated as agotrons were a smaller, more homogeneously sized GC-rich subset of FLEXIs than were pre-miRNAs of annotated mirtrons (Fig. 6A and B, right panel).

Similar analyses showed that the FLEXIs encoding annotated snoRNAs were significantly enriched in bindings sites for DKC1 (Dyskerin) and NOLC1 (Nucleolar and Coiled Body Phosphoprotein 1, alias Nopp140), which have snoRNA-related functions (Alawi and Lin, 2010; Wang et al., 2002), along with several other proteins, including SMNDC1 (Survival of Motor Neuron Domain Containing 1, alias SPF30), which is better known for binding snRNAs, and AATF (Apoptosis Antagonizing Transcription Factor) (Fig. 6C). The latter protein was found previously to bind both DKC1 transcripts and snoRNAs, potentially linking the ribosome biosynthesis function of snoRNAs to the regulation of cell proliferation, apoptosis, and p53 activity (Kaiser et al., 2019) (see legend Fig. 6C for differences in RBP-binding sites in FLEXIs encoding H/ACA-box and CD-box snoRNAs).

Finally, the 158 highly conserved FLEXIs (phastCons scores ≥ 0.75), which included a high proportion of alternatively spliced introns (see above), were significantly enriched in binding sites for splicing factor U2AF2 and splicing regulator TIAL1 as well as the transcriptional regulators BCLAF1 and ZNF622, potentially linking alternative splicing of these introns to transcriptional and other cell regulatory pathways (Fig. 6D; Table S3), a role previously suggested for BCLAF1 (Savage et al., 2014; Vohhodina et al., 2017). Collectively, the above findings show that specific subsets of FLEXI RNAs are enriched in binding sites for RBPs that preform specialized biological functions relative to those for spliceosomal proteins and led us to wonder whether such enrichment might be a distinguishing feature of RBPs that bind or remain bound to excised intron RNAs after complete or partial dissociation from the spliceosome.

Computational identification of RBPs whose binding sites are co-enriched in different subsets of FLEXI RNAs

Encouraged by the above findings, we attempted to mathematically identify other RBPs whose binding sites were co-enriched in different subsets of FLEXIs RNAs without *a priori* assumptions about the function of the protein or bound FLEXIs. To do so, we first extracted subsets of FLEXIs with a CLIP-seq identified binding site for 47 of the 53 RBPs that bind ≥30 different

FLEXIs in the combined CLIP-seq datasets, excluding six core spliceosomal proteins (AQR, BUD13, EFTUD2, PPIG, PRPF8, and SF3B4), whose abundant binding sites would otherwise complicate the analysis. We then identified which of the 126 RBPs that bind FLEXIs were co-enriched in each of the 47 subsets by comparing the frequencies of different RBP-binding sites in each subset to those in all FLEXs (significant enrichment defined as abundance ≥2%, adjusted p-value ≤0.05). RBPs whose binding sites were significantly over- or under-represented in the tested subset and those that did not meet these criteria were key coded, and hierarchal clustering was performed using the Gower's distance between the 47 subsets (Gower, 1971). The results were displayed as a dendrogram heat map for FLEXIs in each of the four cellular RNA datasets (HEK-293T, K-562, HeLa S3, and UHHR; Fig. S14).

We thus identified six clusters of RBPs whose binding sites were co-enriched in different subsets of FLEXI RNAs (Fig. 7A). Each RBP cluster consists of two or more proteins whose binding sites were significantly enriched in the subset of FLEXIs comprising the cluster in all four of the cellular RNA datasets plus additional proteins that met the abundance and significance criteria for association with the cluster in some cell types but not others. Scatter plots confirmed that binding sites for the identified proteins were significantly enriched relative to those for most or all of the six core spliceosomal proteins in the subsets of FLEXIs comprising the clusters (Fig. 6E-J). Independently generated scatter plots for the subset of FLEXIs with a binding site for each individual RBP in the cluster confirmed the co-enrichment of binding sites for other RBPs in the same cluster (Fig. S15).

Clusters V and VI (Fig. 7A) recapitulated and extended the findings above for agotron/mirtron and snoRNA-encoding FLEXIs. Cluster VI consists of AGO1-4 and DICER, which bind overlapping subsets of 479 GC-rich FLEXIs (Fig. 6J). Notably, only 23 of the 250 FLEXIs that have a binding site for AGO1-4 in the AGO1-4 PAR-CLIP dataset corresponded to an annotated agotron (Hansen et al., 2016), and only 44 of the 308 FLEXIs that have a binding site for DICER in the DICER PAR-CLIP dataset corresponded to the pre-miRNA of an annotated mirtron (Wen et al., 2015). The other FLEXIs bound by these proteins could be unannotated

agotrons or mirtrons, could be cleaved by DICER to yield unannotated short regulatory RNAs, or could reflect binding of AGO1-4 and DICER to structured introns that could evolve into agotrons or mirtrons, a possibility suggested previously for mirtrons (Ruby et al., 2007).

Cluster V consists of four proteins identified above as binding snoRNA-encoding FLEXIs (DKC1, NOLC1, AATF, and SMDC1), with the first three satisfying the abundance and significance criteria for association with this cluster in all four of the cellular RNA samples and SMDC1 in only two of the four cellular RNA samples (Fig. 7A). Collectively, these four proteins have binding sites in 160 FLEXIs, only 43 of which ENCODE annotated snoRNAs. The FLEXIs bound by these proteins are longer and have lower GC% content than other FLEXIs, with many having stable predicted secondary structures that might include features recognized by the snoRNA- and snRNA-binding proteins comprising the clusters (Fig. 6I).

Cluster III and IV each consist of two splicing-related proteins (splicing regulators TIA1 and TIAL1 for cluster III and splicing factors U2AF1 an U2AF2 for cluster IV), whose binding sites were significantly enriched relative to those for core spliceosomal proteins in the FLEXIs comprising the cluster in all four cell types, with splicing regulator KHSRP satisfying the abundance and significance criteria for association with this cluster only in HEK-293T cells (Fig. 7A). The RBPs in clusters III and IV have binding sites in 100 and 427 FLEXIs, respectively, which collectively have significantly lower GC content than other FLEXIs (Fig. 6G and H). Four of these proteins (TIA1, TIAL1, U2AF1, and U2AF2) were identified above as impacting FLEXI host gene mRNAs expression levels and/or alternative splicing of bound FLEXIs in knockdown datasets, and three (TIAL1, U2AF1, and U2AF2) were found to have binding sites significantly enriched in FLEXIs with high phastCons scores, many of which are alternatively spliced to generate different protein isoforms. Four of the five proteins in these clusters (TIA1, TIAL1, U2AF1, and KHSRP; Fig. 7A) have been found associated with stress granules (Nunes et al., 2019), possibly reflecting a post-splicing destination for FLEXIs bound by these proteins.

Clusters I and II were the largest in terms of numbers of proteins and most interesting in revealing novel RBP-intron associations. Cluster I is comprised of two RBPs (LARP4 and

PABPC4), which satisfy the abundance and significance criteria for association with the cluster in all four cellular RNA samples, and six other proteins: (Transcription Regulator SUB1, DDX3X, RPS3, NCBP2, DDX55, and Methionine Aminopeptidase 2 (METAP2)), which satisfy these criteria in HEK-293T cells but not in some of the other cell types (Fig. 7A). These eight proteins have binding sites in 386 FLEXIs, with all eight binding at or near the 5'- and 3'-splice sites and knock down of seven impacting expression of FLEXI hosts genes and/or alternative splicing of bound FLEXI RNAs (Fig. 5, Fig. 7A, and Figs. S11-S13). LARP4 and PABP were found previously to function together to regulate the stability and translation of ribosomal protein mRNAs (Mattijssen et al., 2017), and the remaining six proteins also have mRNA synthesis or translation-related functions and/or could link these functions to other cellular regulatory processes (Table S3). All of the RBPs in this cluster except for SUB1 are present in the cytoplasm, consistent with a translation-related function (Fig. 7A).

Finally, cluster II is comprised of five proteins (BCLAF1, UCHL5, ZNF622, TRA2A, ZNF800) that satisfy the abundance and significance criteria for association with this cluster in all of the cellular RNA sample types, two other proteins (GRWD1 and PUM1) that satisfy these criteria in three of the four cellular RNA sample types, and two (DDX24 and FXR2) that satisfy these criteria in only one of the cellular RNA sample types (Fig. 7A). Collectively, these nine proteins have binding sites in 976 FLEXIs, with all nine binding near the 5'- and 3'-splice sites. Three of these proteins (UCHL5, ZNF622, and GRWD1) impacted mRNA expression levels of FLEXI host genes or alternative splicing at bound FLEXIs in knockdown datasets, and two others (BCLAF1 and ZNF622) have binding sites enriched in FLEXIs with high phastCons scores, many of which are alternatively spliced introns (Fig. 6D). Eight of these proteins are present in the cytoplasm, with three (UCHL5, ZNF622, and TRA2A) distinguished from other RBPs in the clusters as being associated with the cytoskeleton, microtubules, actin, or the cell cortex.

Simulations using ten different subsets of 2,000 randomly selected other (*i.e.*, non-FLEXI) short or long introns identified many of the same protein clusters, but in most cases less clearly defined than those identified for FLEXIs and with additional protein clusters not seen for

FLEXIs (Fig. S16A). Cluster IV (U2AF1/U2AF2) was found in all simulations for other short and long intron, while cluster III (TIA1/TIAL1) was found in all simulations for non-FLEXI short introns, but only 40% of the simulations for long introns. Clusters I and II were found in ≥90% of the simulations for other short and long introns, but with only two proteins found consistently in each cluster (LARP4 and PABPC4 for cluster I and UCHL5 and ZNF622 for cluster 2). Three of the cluster V proteins (AATF/DKC1/NOLC1) were found together in 90% of simulations for other short introns but in only 50% of simulations for long introns, while cluster I (AGO1-4/DICER) did not appear in any simulations for non-FLEXI short introns and in only 40% of the simulations for long introns. Four additional protein clusters (PCBP1/PCBP2, METAP2/YBX3, LSM11/PABPN1, RBFOX2/PHRP4) were found in ≥60% of simulations for other short introns but less frequently or not at all in FLEXIs or long introns. These findings suggest both common and different protein associations for the three classes of introns. The clearer definition of the six clusters identified for FLEXIs may reflect that the proteins in these clusters bind together in the same short, relatively stable excised intron RNAs rather than to fragments of more rapidly turning over introns.

FLEXI RNAs associated with RBP clusters originate from host genes with related biological functions

To determine if the host genes encoding FLEXIs in the six identified clusters might be functionally related, we analyzed gene ontology (GO) terms of host genes bound by the same RBPs using DAVID bioinformatic tools and displayed the results as a heat map for each of the RBPs (Fig. 7B). Overall, the heat map for GO terms related to host gene function was remarkably congruent with the identified RBP clusters. The host genes encoding FLEXIs bound by AGO1-4 and DICER (Cluster VI) encode proteins involved in translation and focal adhesion as well as structural components of the ribosome and proteins found in nuclear speckles, while those encoding FLEXIs in cluster II, III, and IV encode proteins involved in RNA splicing, gene expression and mRNA processing, with the host genes for FLEXIs bound by some proteins cluster II also

involved in cell-cell adhesion. The host genes encoding FLEXIs bound by the snoRNA-related proteins in cluster V have functions related to ribosomes, translation, and rRNA processing, with the host genes for FLEXIs bound SMNDC1 enriched in additional GO terms. Notably, the GO terms for the host genes encoding FLEXIs in clusters I, which have binding sites LARP4 and PABPC4, which function together in regulating the stability and translation of ribosomal protein mRNAs, largely overlap those for the host genes for the snoRNA-related FLEXIs in cluster V and were similarly enriched in genes encoding ribosomal proteins (12% and 10.6% compared to 1.4% in all FLEXIs; $p<0.001$), suggesting that cluster I FLEXIs may function in a parallel pathway for regulating ribosome biogenesis. Collectively, these findings indicate that the host genes encoding FLEXI RNAs that bind the same subsets of RBPs are functionally related and could be coordinately regulated to produce different subsets FLEXI RNPs.

FLEXI as potential cancer biomarkers

The cell type-specific expression patterns of FLEXI RNAs suggested that they might be useful as RNA biomarkers. As an initial test of this possibility, we compared FLEXI RNAs in matched tumor and neighboring healthy tissue from two breast cancer patients (patients A and B; PR^+, ER^+, $HER2^-$ and PR unknown, ER^-, $HER2^-$, respectively) and two breast cancer cell lines (MDA-MB-231 and MCF7). UpSet plots for the relatively abundant FLEXIs (≥0.01 RPM) showed hundreds of differences in FLEXI RNAs and FLEXI host genes between the cancer and healthy samples (Fig. 8A). Further, scatter plots comparing the FLEXI RNAs detected in the matched healthy and tumor samples from patients A and B identified multiple candidate FLEXI RNA biomarkers (5 and 43 in patients A and B, respectively) that were detected at ≥0.01 RPM in all replicate libraries from the cancer patient tissue, but not detected in the matched healthy tissue (red dots and protein names in red in Fig. 8B). As for the previous comparisons between different cellular RNA samples (Fig. 3A), the scatter plots showed a wider spectrum of differences than did those for mRNAs from the same host genes (Fig. 8B). Notably, FLEXI RNAs were more abundant in

the highly metastatic MDA-MB-231 cells than in MCF-7 cells, with density plot profiles suggesting upregulation of snoRNA-encoding FLEXIs and agotrons in MDA-MB-231 (Fig. S17).

FLEXI host genes in all four cancer samples were significantly enriched ($p \leq 0.05$) for hallmark gene sets (Liberzon et al., 2015) that are dysregulated in many cancers, including glycolysis, G2M checkpoint, UV response up, and PI3K/AKT/MTOR signaling; Fig. 8C, pathway names in red). Gene sets that were significantly enriched in one or more of the cancer samples but not in the healthy controls included mitotic spindle, MYC targets V1 and V2, estrogen response early and late, androgen response, oxidative phosphorylation, mTORC1 signaling, apical junction, and cholesterol homeostasis (Fig. 8C, pathway names in orange).

Only a small number of the FLEXIs identified as potential biomarkers in the UpSet and scatter plots corresponded to annotated oncogenes (asterisks in Fig. 8A and B). Additional UpSet plots for FLEXI RNAs originating from known oncogenes identified 169 that were up upregulated and 81 that were downregulated ≥2-fold in one or more of the cancer samples compared to the healthy controls (Fig. S18A and B). Similar patterns of up- and down regulation in the cancer samples were seen for tumor suppressor genes (Zhao et al., 2016) (Fig. S18C and D). These findings reflect that FLEXI RNA abundance is dictated not only by transcription but also by alternative splicing and intron RNA turnover. They do not preclude oncogene or tumor suppressor gene FLEXIs from serving as biomarkers for a specific cancer so long as they are reproducibly up or downregulated in that cancer. However, most of the FLEXIs that best discriminated between the cancer and healthy samples arose from genes that were strongly up or downregulated in response to oncogenesis but not oncogenes that drive this process.

Summary, Perspectives, and Limitations

Here, we used TGIRT-seq of rRNA-depleted, unfragmented RNA preparations to identify thousands of FLEXI RNAs in human cells and tissues. In total, including the initial cellular and plasma RNA samples and subsequent cancer samples, we identified 8,687 different FLEXIs expressed from 3,923 host genes, representing ~17% of the 51,645 short introns (≤300 nt)

annotated in Ensembl GRCh38 Release 93 annotations. That FLEXIs are linear RNAs was inferred from the characteristics of TGIRT-seq reads (Fig. 2) and confirmed by digestion with a 5' to 3' exonuclease (Fig. S4). Most FLEXIs have high GC content and stable predicted secondary structures, which may contribute to their stability as FLEXIs (Fig. 1D). In addition to low abundance FLEXIs, which likely include transient intron RNA turnover intermediates, we identified subsets of FLEXIs in each cell type that accumulated to relatively high abundance, including many with abundances of hundreds to thousands of copies per cell (Fig. 1E and Fig. S6). These relatively abundant FLEXIs had reproducible cell-type specific expression patterns (Fig. 3, Fig. S7, and Fig. S8), suggesting that FLEXIs may have utility as RNA biomarkers in cells as well as plasma.

By searching CLIP-seq datasets, we identified 126 different RBPs that have annotated binding sites in FLEXI RNAs, including 53 RBPs that have binding sites in ≥30 different FLEXIs on which we focused further analysis (Fig. 4A). These 53 RPBs included spliceosomal proteins and other proteins that function in RNA splicing, AGO1-4 and DICER and other proteins that have a miRNA-related functions, and a surprising number of proteins that function in cell regulatory processes, including transcription factors (AATF, BCLAF1, and SUB1, a chromatin remodeling protein (GRWD1), proteins that function in translational regulation (LARP4, METAP2, PABC4), and cellular growth regulators (DDX3 and IGF2B1; Fig. 4A; Table S3). Although CLIP-seq does not distinguish whether an RBP bound to an intron before or after splicing, all of the RBPs identified as having binding sites in ≥30 FLEXI RNAs in the ENCODE eCLIP datasets were associated with ≥95% full-length reads corresponding to identified FLEXIs, despite the use of RNase I in processing the cross-linked RNAs for sequencing (Fig. 4 and Fig. S10). The ability RBPs to protect short excised intron RNAs from nucleases that function in intron turnover presumably contributes to their accumulation as FLEXIs.

Addressing the biological significance of FLEXI-RBP interactions by analysis of ENCODE knockdown datasets, we found that knockdown of 34 of the RBPs that have binding sites in ≥30 different FLEXIs resulted in significant changes in alternative splicing of bound FLEXIs and/or

mRNA levels of FLEXIs that contained a binding site for that RBP relative to control groups (Fig. 4A, Fig. 5A and B, Fig. S11, and Fig. S12). Surprisingly 17 of the 27 RBPs whose knock-down impacted alternative splicing of bound FLEXIs had no previously known RNA splicing-related function, including the cell regulatory proteins AATF, RPS3, SUB1, GRWD1, IGF2B, LARP4, PABC4, and XRN2 a nuclear 5' to 3' exonuclease that degrades linear RNAs, (see Results for complete list). The involvement of transcription, translation, and cell growth regulatory proteins in RNA splicing provides a means of linking splicing to other cellular regulatory pathways, a role suggested previously for BCLAF1 (Savage et al., 2014; Vohhodina et al., 2017) and for binding of AATF to snoRNAs (Kaiser et al., 2019).

By mapping RBP-binding sites in CLIP-seq datasets, we found that 37 of the 53 RBPs that bind ≥30 FLEXI RNAs bind at or near the 5'- and 3'-splice sites, while 16 have binding sites more uniformly distributed throughout the body of the intron (Fig. 5C and Fig. S13). The binding of RBPs to the 5'- and 3'-splice sites in a pre-mRNA could be facilitated for FLEXIs that fold into a stable secondary structure that brings the 5'- and 3'-splice sites into proximity, as suggested for the binding of the spliceosome to the 5'- and 3'-spice sites of human short introns (Abebrese et al., 2017). Alternatively, some RBPs may bind to the exposed 5' and 3' ends of the FLEXI RNA after complete or partial spliceosome dissociation. In either case, RBPs bound simultaneously to the 5'- and 3'-ends of an excised intron would effectively circularize the intron RNA, protecting it from exonucleases that function in intron RNA turnover and potentially contributing to its accumulation as a FLEXI.

Compared to other short and long introns, FLEXIs as a whole were enriched in binding sites for core spliceosomal proteins (Fig. 4B and C), possibly reflecting that most FLEXIs dissociate more slowly from the spliceosome than other introns, providing greater opportunity for cross-linking. However, we also found that subsets of FLEXIs with known or hypothesized biological functions were enriched in binding sites for RBPs related to those functions, possibly reflecting that these RBPs remained bound to the excised intron RNA after dissociation from the spliceosome. These subsets of FLEXIs included agotrons and mirtrons, which were enriched in binding

sites for AGO1-4 and DICER (Fig. 6A and B); snoRNA-encoding FLEXIs, which were enriched in proteins with snoRNA-related functions (Fig. 6C); and evolutionarily conserved FLEXIs, which were enriched in binding sites for transcription factors and splicing regulators that may control the production of different protein isoforms from FLEXI host genes (Fig. 6D).

Building on these findings, we used a computational method with no *a priori* assumptions about RBP or FLEXI function to identify six clusters of proteins whose binding sites were co-enriched relative to those for spliceosomal and other proteins in different subsets of FLEXI RNAs (Fig. 7 and Fig. S14). This computational method recapitulated the findings that agotrons and mirtrons are enriched in binding sites for AGO1-4 and DICER and that snoRNA-encoding FLEXIs are enriched in binding sites for snoRNA related proteins, but also identified many additional FLEXIs bound by these proteins that were not previously annotated as agotrons, mirtrons, or snoRNA-encoding introns. The computational method also identified two sets of splicing factors and regulators that were co-enriched in different subsets of FLEXIs, and two groups of RBPs that may relate to novel functions of FLEXI RNAs (Fig. 7A). Particularly suggestive of the latter was cluster I, which included two proteins, LARP4 and PABC4, which were known to function together to regulate the stability and translation of ribosomal protein mRNAs (Mattijssen et al., 2017), along with additional proteins that have translation-related functions (NCBP2, METAP2, and RPS3), or that could link these functions to transcriptional regulation of other cell regulatory pathways (SUB1, RPS3, DDX3X, and DDX55). Further, the host genes encoding the FLEXIs in cluster I have GO terms similar to those for the snoRNA-related FLEXIs in cluster V (Fig. 7B) suggesting that the FLEXIs and RBPs of cluster I may function similarly to snoRNAs in a feedback loop regulating ribosome assembly or function (Fig. 7B). Along the same lines, cluster II included three proteins (UCHL5, ZNF622, and TRA2A) that were distinguished from other proteins associated with FLEXI/RBP clusters as being localized to ER, cytoskeleton, and cell cortex, possibly reflecting a cytoskeleton or membrane-associated function, and clusters III and IV consisted of five splicing factors or regulators, four of which localize to stress granules (TIA1, TIAL1, U2AF1, and KHSRP) (Nunes et al., 2019), possibly a post-

splicing destination for the bound FLEXIs (Fig. 7A). Collectively, these findings suggest that some FLEXIs may function as scaffold to co-localize or co-regulate multiple proteins with related biological functions.

Our computational identification of intron-RBP associations is undoubtedly incomplete, as it depends upon assigning a relatively arbitrary significance criterion for association of an RBP with the cluster and would miss proteins that are not efficiently cross-linked to FLEXI RNAs. Minimally, however, our findings demonstrate that different subsets of RBPs bind specifically to distinct subsets of FLEXIs and set the stage for more detailed analysis of these RBP-intron associations.

Finally, the relatively abundant FLEXIs that have cell-type specific expression patterns may have utility as cellular RNA biomarkers. In addition to being linked to the transcription of thousands of protein-coding and lncRNA genes, FLEXI RNA levels also reflect differences in alternative splicing and intron RNA stability, providing higher resolution of cellular differences than mRNAs transcribed from the same gene. Because many FLEXIs are predicted to fold into stable secondary structures, their initial identification may be best done by TGIRT-seq, which yields full-length, end-to-end sequence reads of structured RNAs, possibly utilizing size-selected small RNA preparations in which FLEXIs are likely to be enriched. Once identified, the validation of candidate FLEXI biomarkers and their routine monitoring would best be done by methods such as RT-qPCR, microarrays, other hybridization-based assays, or targeted RNA-seq that can give rapid quantitative read outs in targeted RNA panels for different diseases.

Data access

TGIRT-seq datasets have been deposited in the Sequence Read Archive (SRA) under accession numbers PRJNA648481 and PRJNA640428. A gene counts table, dataset metadata file, FLEXI metadata file, RBP annotation file, and scripts used for data processing and plotting have been deposited in GitHub: https://github.com/reykeryao/FLEXI.

COMPETING INTEREST STATEMENT

Thermostable group II intron reverse transcriptase (TGIRT) enzymes and methods for their use are the subject of patents and patent applications that have been licensed by the University of Texas and East Tennessee State University to InGex, LLC. A.M.L., some former and present members of the Lambowitz laboratory, and the University of Texas are minority equity holders in InGex and receive royalty payments from the sale of TGIRT-enzymes and kits and from the sublicensing of intellectual property by InGex to other companies. A.M.L., J.Y., H.X. and D.C.W. are inventors on a patent application filed by the University of Texas at Austin for the use of full-length excised intron RNAs and intron RNA fragments as biomarkers. S.W., E.F.K. and M.A. have no competing interests.

ACKNOWLEDGMENTS

We thank Marta Mastroianni and Ryan Nottingham (University of Texas at Austin) for cell culture and RNA preparations, Blerta Xhemalce (University of Texas at Austin) for a fresh stock of MDA-MB-231 cells, Camilo Molina Romero and Kavyaa Choudhary (University of Texas at Austin) for help with validating amplicons for terminator and digital PCR analysis, and Jim Manley (Columbia), Richard Maraia (NIH/NICHD), Carla Van Den Berg (University of Texas at Austin), and Sandra Wolin (NIH/NCI) for comments on the manuscript. The authors acknowledge the Texas Advanced Computing Center (TACC; http://www.tacc.utexas.edu) at the University of Texas at Austin for providing high performance computing resources that have contributed to the research results reported in this paper (URL: http://www.tacc.utexas.edu). This work was supported by NIH grants R01 GM37949 and R35 GM136216 to AML. MA was supported by NIH grant R01 GM040478.

Figure 1. Characteristics of FLEXI RNAs in human cells and plasma. (*A*) UpSet plots of FLEXI RNAs and their host genes identified in unfragmented RNA preparations from the indicated samples. Different FLEXIs from the same host gene were aggregated into one entry for that gene. *(B*) Numbers and percentages of identified FLEXIs that correspond to annotated agotrons, pre-miRNA of annotated mirtrons, or encode annotated snoRNAs in each of the samples. "Total" indicates the total number of different FLEXIs in each category in this collection of samples, and GRCh38 indicates the number of annotated short introns (≤300 nt) in each category in the GRCh38 human genome reference sequence. "Agotron and Mirtron" indicates the number of FLEXIs annotated as both an agotron and mirtron, and "Agotron or Mirtron" indicates the number and percentages of FLEXIs annotated as either an agotron or mirtron. The number and percentages of FLEXIs that contain an embedded snoRNAs and the number of those snoRNAs that are small Cajal body-specific snoRNAs (scaRNAs) are also indicated. *(C*) 5'- and 3'-splice site (5'SS and 3'SS, respectively) and branch-point (BP) consensus sequences of FLEXIs corresponding to GC-AG and GU-AG U2-type spliceosomal introns in a merged dataset for the four cellular RNA samples compared to a literature consensus for human U2-type spliceosomal introns (Gao et al., 2008; Sheth et al., 2006) and to consensus sequences for other Ensemble GRCh38-annotated U2-type short (≤300 nt) and long (>300 nt) introns. The number of introns for each consensus sequence is indicated to the right. (*D*) Density distribution plots of length, GC content, and minimum free energy (MFE) for the most stable RNA secondary structure predicted by RNAfold for FLEXIs detected in a merged dataset for the four cellular RNA samples (black) or human plasma (red) compared to other GRCh38-annotated short introns (≤300 nt, dashed gray line). (*E*) Density distribution plots of the abundance of different categories of FLEXIs in different cellular RNA samples color coded as indicated below the plots. The abundance distribution of annotated mature snoRNAs (dashed yellow line), and the abundance of different sncRNAs (gray, arrows) in the same samples are shown for comparison. Red dotted lines separate low abundance FLEXIs (<0.002 RPM) from higher abundance FLEXIs (≥0.002 RPM). (*F*) Density distribution plots of phastCons scores of different categories of FLEXIs detected in the cellular

and plasma RNA samples compared to other annotated short introns (≤300 nt). PhastCons scores were the average score across all intron bases calculated from multiple sequences alignment of 27 primates, including humans plus mouse, dog, and armadillo.

Figure 2. FLEXI RNAs are detected by TGIRT-seq as continuous full-length, end-to-end reads of excised intron RNAs. *(A)* IGV screenshots showing read alignments for FLEXI RNAs in different cellular RNA samples. Gene names are at the top with the arrow below indicating the 5' to 3' orientation of the encoded RNA followed by tracts showing gene annotations (exons, thick bars; introns, thin lines) with read alignments shown below. Reads are color coded by cell type as shown in the Figure. The length of the FLEXI that is the focus of each panel is shown above the intron annotation, and the number of reads assigned to that FLEXI (continuous reads that begin and end within 3 nt of the annotated 5'- and 3'-splice sites) compared to total reads for that intron are shown within the read alignments. Reads were down sampled to a maximum of 100 for display in IGV. Mismatched nucleotides in boxes at the 5' end of RNA reads are due to non-templated nucleotide addition (NTA) to the 3' end of cDNAs during TGIRT-seq library preparation. *(B)* Histogram of the length distribution of reads of any length mapping within introns identified as FLEXI RNAs in a merged dataset for the four cellular RNA samples (red) compared to those mapping within other Ensemble GRCh38-annotated short introns (≤300 nt) in the same samples (blue). Relative length was calculated from the read spans of TGIRT-seq reads normalized to the length of the corresponding intron. Introns encoding embedded snoRNAs or scaRNAs were removed for the comparisons to avoid interference from mature snoRNA reads. *(C)* Three-dimensional bar chart showing the percentage of FLEXI RNA reads in a merged dataset for the cellular RNA samples ending at different positions around the intron-exon junctions.

Figure 3. FLEXI RNA profiles are cell-type specific due to differences in host gene transcription, alternative splicing, and intron RNA turnover. (*A*) Scatter plots comparing $\log_2$-transformed RPM of FLEXI RNA reads (top row) and all reads that mapped to FLEXI host

genes (bottom row) in pairwise comparisons of cellular RNA samples. Relatively abundant FLEXIs detected at ≥0.01 RPM in all replicates of one or both of the compared cell types and their host genes are marked in red and blue for the samples compared on the y-axis and x-axis, respectively. *r* is the Pearson correlation coefficient *(B*) IGV screenshots showing examples of cell-type specific differences in FLEXI RNA abundance due to differences in alternative splicing, splicing efficiency and/or differential stability of FLEXI RNAs encoded by the same host gene. Gene maps for different RNA isoforms generated by alternative splicing of FLEXI RNAs are shown at the bottom with highlighted introns arbitrarily named A to F from the 5' to 3' end of the gene. Reads for different cellular RNA samples are color coded as indicated at the top right. Reads mapping to exons or non-FLEXI introns were omitted for clarity. Mismatched nucleotides in boxes at the 5' end of RNA reads are due to non-templated nucleotide addition (NTA) to the 3' end of cDNAs during TGIRT-seq library preparation. Reads were down sampled to a maximum of 100 for display in IGV. (*C*) Characteristics of highly conserved FLEXIs in a merged dataset for the four cellular RNA samples. FLEXIs were divided into three groups based on phastCons scores for comparisons (<0.75, ≥0.75, and ≥0.99). The number of FLEXIs in each group is shown to the left, and the numbers and percentages of those FLEXIs in different categories are shown to the right. Alt. Spl. indicates FLEXIs that are alternatively spliced to generate different protein isoforms based on Ensemble GRCh38-annotation. Shared SS indicates FLEXIs that share a 5'- or 3'-splice site with a longer intron (>300 nt). RI indicates FLEXIs that are retained in some annotated alternatively spliced transcripts. In frame RI indicates retained FLEXIs that have reading frames in frame with that of the flanking exons.

Figure 4. FLEXI interactions with RNA-binding proteins. (*A*) Bar graph showing the number of different cellular FLEXI RNAs that have a CLIP-seq identified binding site for the 53 RBPs that have binding sites in ≥30 different FLEXI RNAs in a merged dataset for the four cellular RNA samples. An extended bar graph for all 126 RBPs that have a CLIP-seq identified binding site in a FLEXI RNA is shown in Fig. S9A. The grids below the bar graph show RBP function

and cellular localization information from (Van Nostrand et al., 2020) supplemented by the mammalian RNA granule and stress granule protein databases (Nunes et al., 2019; Youn et al., 2019) and the UniProt database (The UniProt Consortium, 2018), followed by RBP-binding site and protein function information obtained in this study from analysis of CLIP-seq and knock-down datasets: RNA-binding sites: I, within the intron; SS-I, at or adjacent to the 5'- and/or 3'-splice site within the intron; SS-E, adjacent to splice site in flanking exons (Fig. 5C and Fig. S13). mRNA levels in RBP knockdown datasets: light blue and red, significant bias toward higher ($\log_2$FC>0) or lower ($\log_2$FC<0) mRNA levels, respectively, compared to all significantly changed genes; gray, significant change in FLEXI host gene mRNA expression levels, but not significantly biased in either direction (Fig. 5A and Fig. S11). Alternative Splicing (Alt. Spl.) in RBP knockdown datasets: RI, significant increase in retained introns; SE, significant increase in skipped exons; Fig. 5B and Fig. S12). Lower GC%: RBPs whose binding sites were significantly enriched in FLEXIs having lower GC content than other FLEXIs (Fig. S15). phastCons scores ≥0.75: RBPs whose binding sites were significantly enriched in evolutionarily conserved FLEXIs with phastCons ≥0.75 (Fig. 6D). (*B* and *C*) Scatter plots comparing the abundance (percentage) of binding sites for different RBPs in FLEXI RNAs with those in longer introns (>300 nt) or other short introns (≤300 nt) annotated in GRCh38. RBPs whose binding sites were ≥1% of the total RBP-binding sites and significantly different between the compared groups ($p \leq 0.05$ calculated by Fisher's exact test) are indicated by the name of the protein color coded by protein function as indicated at the top of Figure. For each panel, the box in the left-hand scatter plot delineates a region that is expanded in the right-hand scatter plot. The circle in the right-hand scatter plot of panel C delineates 12 additional RBPs whose binding sites meet the abundance and significance criteria for enrichment in long introns compared to FLEXIs but were too numerous to name in the Figure (FAM120A, GTF2F1, HNRNPC, HNRNPK, HNRNPL, HNRNPM, ILF3, MATR3, NONO, PCBP2, SUGP2, and TARDBP).

Figure 5. Impact of FLEXI interactions with RNA-binding proteins on FLEXI host gene mRNA abundance and alternative splicing. (*A*). Number of FLEXI host genes with significantly changed mRNA levels, including both under- or overexpressed genes, in RBP knockdown versus control datasets. RNA-seq datasets from RBP knockdown (shRNA or CRISPR) and control datasets for K-562 and HepG2 cells were obtained from ENCODE for all RBPs that bind ≥30 different FLEXIs. Changes in mRNA abundance for host genes with FLEXIs that contain a binding site for the knocked down RBP compared to three randomly chosen genes sets with the same number of genes were calculated from the knockdown datasets by using DESeq2, and the significance of the changes was determined by Analysis of Variance (ANOVA): **** $p<0.0001$, *** $p<0.001$, ** $p<0.01$, * $p\leq0.05$. Volcano plots showing the magnitude and direction of changes for all RBPs whose knockdowns caused significant changes in FLEXI host gene mRNA levels are shown in Fig. S11. (*B*) Alternative splicing changes involving FLEXI introns in RBP knockdowns versus control datasets calculated using rMATS (Shen et al., 2014). Density plots show the inclusion level difference for retained introns (RI) or skipped exons (SE) for FLEXIs that contain a binding site for the indicated RBP (denoted bound FLEXIs, red), FLEXIs that do not contain a binding site for the indicated RBP (denoted unbound FLEXIs, blue), and other short introns (black). In the RI plots, shifts to the right or left represent an increase or decrease in retained introns, respectively, and in the SE plots, shifts to the left or right indicate an increase or decrease in skipped exons, respectively. Statistical significance was calculated by Kolmogorov-Smirnov test. Plots for all RBP knockdowns that caused a significant change ($p\leq0.05$) involving a bound FLEXI are shown in Fig. S12. (*C*) Density plots comparing the location of ENCODE eCLIP identified RBP-binding sites within FLEXIs to those in other short or long introns. RBP-binding sites overlapping with introns were identified by intersecting annotated RBP-binding sites with IDR analysis from the eCLIP datasets with the intron coordinates using BedTools, and plotted as the mid-point of the annotated RBP-binding sites normalize as a percentage of intron length, with 0% and 100% corresponding to the 5' and 3' end of the intron, respectively. In each panel, binding-site distributions are shown for all FLEXIs (red), all other short introns (≤300 nt;

blue), and all longer introns (>300 nt; black) bound by the indicated RBP. Peak positions in the density plots for FLEXI RNAs are indicated by red dashed. Density plots of the mid-points of RBP-binding sites for all 51 ENCODE eCLIP RBPs that bind ≥30 different FLEXIs are shown in Fig. S13.

Figure 6. Enrichment of RBP-binding sites for different RNA-binding proteins in structurally distinct subsets of FLEXI RNAs. Each panel shows scatter plots identifying RBPs whose binding sites are enriched in different subsets of FLEXI RNAs (left) and density plots comparing the characteristics of those FLEXIs to all other FLEXIs (right) in a merged dataset for the cellular RNA samples. (*A*) FLEXIs annotated as agotrons; (*B*) FLEXIs corresponding to pre-miRNAs of annotated mirtrons; (*C*) FLEXIs encoding an annotated snoRNA. Further analysis showed that that AATF cross-linked only to FLEXIs encoding C/D-box snoRNAs, while DKC1, which is a core H/ACA-box snoRNA binding protein (Massenet et al., 2017), cross-linked to FLEXIs encoding both H/ACA- and C/D-box snoRNAs. These cross-linking patterns raise the possibly that binding of DKC1 to FLEXIs encoding C/D-box snoRNAs is mediated by its interaction with AATF. (*D*) conserved FLEXIs with phastCons score ≥0.75; (*E-J*) FLEXIs bound by different bioinformatically identified clusters of RBPs whose binding sites were enriched in different subsets of FLEXI RNAs (see Fig. 7). The number of different FLEXIs comprising each subset is indicated in parentheses at the top. RBP-binding site annotations were from the ENCODE 150 RBP eCLIP dataset with irreproducible discovery rate (IDR) and AGO1-4 and DICER PAR-CLIP datasets. In the scatter plots, RBP-binding sites whose abundance was ≥2% of all RBP-binding sites and significantly different between the compared subset and all other FLEXIs (adjusted p≤0.05; calculated by Fisher's exact test and adjusted by the Benjamini-Hochberg procedure) are labeled with the name of the RBP color coded by protein function as indicated at the bottom of the Figure. The density distribution plots compare the length, GC content, and MFE for the most stable secondary structure predicted by RNAfold for the FLEXI RNAs comprising the subset (red) to those of all other detected FLEXIs (black). p-values are shown at the top left

of those density plots in which the distribution for the subset of FLEXIs differed significantly from other FLEXIs ($p<0.01$ and FDR≤0.05 as determined by 1,000 Monte-Carlo simulation). Independently generated scatter plots and density plots for each individual RBP that binds ≥30 different FLEXI RNAs are shown in Fig. S15.

Figure 7. Identification of clusters of RBPs whose binding sites were co-enriched in subsets of FLEXI RNAs. (*A*) Hierarchical clustering of FLEXI RNAs based on patterns of over- and under-represented RBP-binding sites. Hierarchical clustering was done as described in Results and Methods and in Fig. S14. Clusters I to VI are delineated at the top of the Figure, and the names of the RBPs comprising the cluster are shown at the bottom of the Figure. The proteins that met the abundance and significance criteria for association with the clusters in HEK-293T, HeLa S3, K-562, and UHHR are indicated by filled boxes at the top. The grids below summarize RBP-binding site and protein function information obtained in this study from the analysis of public CLIP-seq and knockdown datasets followed by protein function and cellular localization information from (Van Nostrand et al., 2020) supplemented by the mammalian RNA granule and stress granule protein databases (Nunes et al., 2019; Youn et al., 2019) and the UniProt database (The UniProt Consortium, 2018) (see legend Fig. 4A for details). (*B*) A heatmap of GO terms enriched in host genes of FLEXI RNAs containing binding sites for different RBPs. GO enrichment analysis was performed with DAVID bioinformatics tools, using all FLEXI host genes as the background. Similar results were obtained using all ENSEMBL-annotated genes as background (not shown).

Figure 8. FLEXI RNAs in breast cancer tumors and cell lines. (*A*) UpSet plots of FLEXI RNAs detected at ≥0.01 RPM and their host genes in unfragmented RNA preparations from matched cancer/healthy breast tissues from patients A (PR^+, ER^+, $HER2^-$) and B (PR unknown, ER^-, $HER2^-$) and breast cancer cell lines MDA-MB-231 and MCF7. Different FLEXI RNAs from the same host gene were aggregated into one entry for that gene. FLEXI RNAs and FLEXI

host genes are listed below some sample groups in descending order of RPM with the range of RPM indicated below each column. (*B*) Scatter plots comparing $\log_2$ transformed RPM of FLEXI RNA reads in unfragmented RNA preparations (left plot in each pair) and all reads mapping to FLEXI host genes (right plot in each pair) in chemically fragmented RNA preparations from cancer and healthy breast tissues from patients A and B. FLEXI RNAs present at ≥0.01 RPM and detected in all replicate libraries from the cancer tissue but not in matched healthy tissue are indicated in red with up to the 10 most abundant FLEXIs satisfying these criteria listed to the right. An additional 33 FLEXIs meeting these criteria in patient B are listed in the Supplemental file. In panels A and B, FLEXIs originating from annotated oncogenes and oncogenes that encode FLEXI RNAs are indicated with an asterisk. (*C*) GO enrichment analysis of genes encoding detected FLEXI RNAs in 50 hallmark gene sets (MSigDB) in combined patient A + B healthy tissue and cancer sample datasets. Names of pathways significantly enriched ($p \leq 0.05$) in all or at least one cancer sample are in red and orange, respectively.

Figure S1. TGIRT-seq of rRNA-depleted unfragmented cellular RNAs. (*A*) Stacked bar graphs showing the percentage of reads mapping to different categories of annotated genomic features in combined datasets for the indicated samples. Genomic features follow Ensembl GRCh38 Release 93 annotations. rRNA includes cellular and mitochondrial (Mt) rRNAs, and protein coding includes all transcripts of protein-coding genes from both the nuclear and Mt genomes. (*B*) Stacked bar graphs showing the percentage of bases mapping to different regions of the sense strand of protein-coding genes. CDS, coding sequences; intergenic, regions upstream or downstream of transcription start and stop sites annotated in RefSeq; intron, intronic regions; UTR, 5'- or 3'-untranslated regions. Healthy and Cancer indicate tumor and neighboring healthy tissues from breast cancer patients A and B (see Fig. 8).

Figure S2. Integrative Genomics Viewer (IGV) screenshots showing coverage tracts and read alignments for sncRNAs detected in TGIRT-seq datasets of unfragmented cellular

RNAs. The name of the sncRNA is shown at the top with its length indicated in parentheses and the arrow indicating the 5' to 3' orientation of the RNA. Coverage tracts (gray) and read alignments for combined technical replicates of each cellular RNA sample type color coded as indicated in the Figure are shown below. Reads were down sampled to a maximum of 100 for display in IGV. Gray in the coverage tracts indicates bases in the read that match the reference base, and other colors indicate bases in the read that do not match the reference base (red, thymidine; green, adenosine; blue, cytidine; and brown, guanosine). Misincorporation at known sites of tRNA post-transcriptionally modified bases are highlighted in the alignments: m^1A58: 1-methyl-adenosine at position 58; I: inosine. CCA indicates the post-transcriptionally added 3' CCA sequence of tRNAs mapped against a reference set of mature tRNA sequences.

Figure S3. Characteristics of FLEXI RNAs. (*A*) Density plots showing the distribution of length of FLEXIs (≤300 nt; red) and long FLEXIs (>300 nt; blue) in combined datasets for each of the four cellular RNA samples. (*B*) Density plots showing the length distribution of all reads of any length mapping within different categories of FLEXI RNAs in combined datasets for each of the four cellular RNA samples compared to those mapping to other Ensemble GRCh38-annotated short introns (≤300 nt) in the same combined datasets. Separate plots color coded by category as shown in the Figure are overlaid for FLEXI RNAs containing annotated binding sites for AGO1-4, DICER, five core spliceosome proteins (AQR, BUD13, EFTUD2, PRPF8 and SF3B4), other annotated RBPs, FLEXIs without (w/o) annotated RBP-binding sites, and other GRCh38-annotated short introns (≤300 nt). Relative length was calculated from the read spans of TGIRT-seq reads normalized to the length of the corresponding intron. Introns encoding embedded snoRNAs were removed prior to calculating the distributions to avoid interference from mature snoRNA reads. *(C)* Three-dimensional bar chart showing the percentage of FLEXI RNA reads ending at different positions around the intron-exon junctions in each of combined datasets for each of the four cellular RNA samples.

Figure S4. FLEXI RNAs are susceptible to 5' to 3' exonuclease digestion. Unfragmented HEK-293T and UHHR RNAs (500 ng) or three synthetic oligonucleotide controls corresponding to FLEXI 3I_RAN (100 ng; linear with or without a 5' phosphate or circular generated by T4 RNA Ligase I (New England Biolabs)) were treated with Terminator exonuclease (Lucigen) or mock treated. Reverse transcription with Maxima H reverse transcriptase (ThermoFisher) was then performed using gene-specific primers near the 3' end of the RNA for 7 different FLEXIs, 3 sncRNAs, and the 3I_RAN FLEXI oligonucleotide controls, the latter using the same primer used for 3I_RAN FLEXI in unfragmented cellular RNA (Table S4). Quantitative PCR (qPCR) was done using SYBR Green with oligonucleotide primers listed in Table S4. Data are reported for three replicates as a percent of FLEXI RNA remaining after exonuclease treatment relative to a mock-treated RNA sample, run in parallel. Fold change was calculated using the delta Ct method. U6, which has a 5' γ-monomethyl phosphate cap (Singh and Reddy, 1989), was insensitive to the exonuclease treatment, as expected. Linear forms of the oligonucleotide control were sensitive to exonuclease treatment whereas the circular form was less sensitive. Sensitivity of the linear oligonucleotide with a 5'-phosphate could reflect incomplete phosphorylation of the synthetic oligonucleotide or reported off target activity in the standard terminator buffer, Buffer A, which is the recommended buffer in the manufacturer's instructions for achieving maximal efficiency of Terminator digestion. The sensitivity of one replicate of the circular oligonucleotide to Terminator treatment might reflect incomplete circularization of the RNA used for that replicate. PCR specificity was confirmed by gel extracting UHHR qPCR amplicons followed by TOPO-TA cloning and Sanger sequencing.

Figure S5. FLEXI RNAs corresponding to U12-type introns or with non-GU-AG 5'- and 3'-splice sites. (*A*) 5'- and 3'-splice sites (5'SS and 3'SS, respectively) and branch-point (BP) consensus sequences of GU-AG and AU-AC U12-type FLEXIs detected in combined datasets for each of the four cellular RNA samples (HEK-293T, HeLa S3, K-562, and UHHR) compared to a literature consensus sequence for human U12-type introns (top) (Burset et al., 2000; Gao et

al., 2008) and consensus sequences for other Ensemble GRCh38-annotated U12-type short (≤300 nt) and long (>300 nt) introns in the same combined datasets (below). The number of introns for each consensus sequence is indicated to the right of the sequence. (*B*) IGV screenshots showing examples of coverage tracts and read alignments for FLEXI RNAs having U12-type or non-canonical 5'- and 3'-splice sites. Dinucleotides at the 5'- and 3'-ends of the intron are indicated at the upper left. The host gene name and genomic coordinates of the FLEXI are shown at the top with the arrow below indicating the 5' to 3' orientation of the encoded RNA followed by tracts showing the genomic sequence and gene annotations for different transcript isoforms (exons, thick bars; introns, thin lines). Read alignments for FLEXI RNAs are shown below the tracts and are color coded by sample type as indicated in the key at the upper right of the panel. Reads were down sampled to a maximum of 100 for display in IGV.

Figure S6. Estimates of FLEXI RNA abundance using a linear regression model based on sncRNA standards in TGIRT-seq datasets and droplet digital PCR. (*A*) Linear regression of molecule per cell values based on RPM values of sncRNAs of known cellular abundance in TGIRT-seq datasets. The scatter plots show the relationship between $\log_{10}$ transformed copy number per cell values for sncRNAs reported in the literature (Tycowski et al., 2006) and $\log_{10}$ transformed RPM of sncRNAs in combined datasets for each of the four cellular RNA samples before (top panels) or after (bottom) bias-correction of TGIRT-seq reads using a random forest regression model (Xu et al., 2019). The linear regression was plotted as a light blue line with 95% confidence intervals shown as dashed blue lines. sncRNA used in the linear regression were major spliceosomal snRNAs U1, U2, U4, U5, and U6; minor spliceosomal snRNAs U4ATAC and U6ATAC; U7, 7SL and 7SK RNA; MRP and RNase P RNA (Table S2) (Tycowski et al., 2006). Pearson (*r*) correlation coefficients are shown in the upper left of each panel. A test of the linear regression model (lrm) for three sncRNAs (U7, SNORD14B, and SNORD44) used as ddPCR standards below gave copy number per cell estimates consistent with literature values (Tycowski et al., 2006), but with copy number per cell values varying between different cell

types (see Table at the bottom of panel A). (*B*) Abundance estimates for FLEXI RNAs based on the TGIRT-seq linear regression model and droplet digital PCR (ddPCR). The left-hand columns show RPM and copy number per cell values of FLEXIs estimated using the linear regression model in cellular RNA samples. The right-hand columns show copy number per cell values based on ddPCR abundance relative to three different sncRNA standards run in parallel (U7, SNORD14B, and SNORD44). For ddPCR, reverse transcription was performed on DNase-treated, size-selected (≤200 nt) RNA preparations using Maxima H reverse transcriptase with a gene-specific primer at the 3' end of the putative FLEXI or sncRNA. ddPCR was then performed using primer sets within the FLEXI sequence or spanning the 5'exon-FLEXI junction, with 5'exon-FLEXI junction signals acting as a control to detect amplification of any RNAs with retained FLEXI introns, which was subtracted from the FLEXI signal when applicable. Copy number per cell values for FLEXIs were calculated based on copy number per cell values for U7, SNORD14B, and SNORD44 in each cellular RNA sample, calculated using the linear regression model without or with TGIRT-seq bias correction (see Table at the bottom of panel A). The ddPCR values relative to U7 RNA, which has the closest cellular abundance to FLEXIs, showed general agreement with the copy number per cell values estimated from the linear regression model without TGIRT-seq bias correction, while the other sncRNA controls, predicted to be in much higher abundance than FLEXIs, gave FLEXI quantifications closer to the bias-corrected linear regression model. Primer amplification specificity was verified by gel extraction of a representative set of matched qPCR amplicons, which were then Sanger sequenced directly or after TOPO-TA cloning (ThermoFisher). The combined TGIRT-seq and ddPCR approach gives greater confidence in the identity and abundance of FLEXIs. Although ddPCR offers a vast improvement in precision compared to qPCR, it is still subject to PCR artifacts and primer-design constraints. Using multiple sncRNA controls concurrently allows for an estimate range for FLEXI abundance, which can minimize the impact of inaccuracies in quantifying any individual sncRNA.

Figure S7. Scatter plots comparing log_2-transformed RPM of relatively abundant (≥0.01 RPM) FLEXI RNAs in biological replicates (Bio 1 and Bio 2) for four different cellular RNA samples. Bio 1 is a merged dataset for all technical replicates for the indicated cellular RNA samples obtained for this study, and Bio 2 is a merged dataset for all technical replicates of different RNA preparations from the same cell types obtained for other studies (including a different lot number for UHHR; see Table S1). FLEXIs that were detected in all Bio 1 technical replicates are shown as red dots. *r* is the Pearson correlation coefficient.

Figure S8. Clustering analysis of detected FLEXI RNAs. (*A*) and (*B*) PCA, PCA-initialized t-SNE, and ZINB-WaVE analysis for FLEXI RNAs detected at ≥1 read and ≥0.01 RPM, respectively, in datasets for biological and technical replicates of rRNA-depleted unfragmented cellular RNAs (Table S1). t-SNE (Kobak and Berens, 2019) and ZINB-WaVE (Risso et al., 2018) are algorithms widely used for the analysis of single cell RNA-seq datasets with zero-inflated counts. Top rows are clustering of technical replicates from difference cell types used in this study. Middle and bottom rows show clustering of datasets obtained for this study (circle, biological replicate 1; Bio 1) and biological replicate datasets for rRNA-depleted unfragmented HEK-293T, K-562, MDA-MB-231, and UHRR RNAs obtained for other studies (triangle, biological replicate 2; Bio 2) (see Table S1), before and after batch effect correction by ZINB-WaVE, respectively.

Figure S9. Bar graphs showing the number of FLEXIs and other classes of introns in a merged dataset for the four cellular samples that have a CLIP-seq identified RBP-binding site for the indicated RBP. (*A*) FLEXI RNAs. (*B*) Other Ensembl GRCh38-annotated short introns (≤300 nt); (*C*) Ensembl GRCh38-annotated long introns (>300 nt). Bars are color coded by RBP function as shown at the top. Asterisks in panels B and C indicate the 53 proteins identified as binding 30 or more different FLEXIs.

Figure S10. Enrichment of FLEXI reads in ENCODE eCLIP datasets. eCLIP data were obtained from ENCODE for all RBPs that bind ≥30 FLEXIs, and ≥95% full-length intron reads were identified by intersecting the mapped eCLIP reads with FLEXI coordinates. Data are shown as an enrichment of ≥95% full-length intron reads for each RBP, calculated as a ratio of reads in the eCLIP versus the control antibody dataset (**** $p<0.0001$, * $p<0.05$, Fisher's least significant difference test). The dashed line indicates two-fold enrichment. Names of proteins are color coded by protein function as indicated in the Figure.

Figure S11. Volcano plots showing gene expression changes in FLEXI host genes in RBP knockdown versus control datasets. The plots show $-\log_{10}$ transformed adjusted p-values versus $\log_2$ transformed fold changes for ENSEMBL-annotated genes in knockdown versus control datasets for the indicated RBPs in K-562 and HepG2 cells (GENCODE). Host genes of FLEXI RNAs are shown as red dots. Other genes with or without significant expression changes (adjusted p-value ≤0.05) are shown as black or gray dots, respectively. RBP knockdown that had a significant bias towards increased or decreased mRNA levels in the group of host genes encoding FLEXIs with a binding site for the RBP are indicated by up (light blue) or down (red) arrows next to the RBP name (p-value ≤0.05 determined by Fisher's exact test comparing the ratio of significantly up ($\log_2FC>0$) or down ($\log_2FC<0$) regulated host genes containing a FLEXI with a binding site for the RBP to that in all significantly changed FLEXI host genes). RBP knockdowns that resulted in significant changes in mRNA levels from host genes containing a FLEXI with a binding site for the knocked down RNA but no significant directional bias are indicated by a bidirectional arrow next to the RBP name. Datasets that were not available for an RBP in one of the two cell lines were left as a blank space.

Figure S12. Splicing changes at or adjacent to introns that give rise to FLEXI RNAs in RBP knockdown versus control datasets. Splicing changes were calculated using rMATS (Shen et al., 2014). Density plots show inclusion of retained introns (RI) or skipped exons (SE)

for bound FLEXIs (red), unbound FLEXIs (blue), and other short introns (black) for RBP knockdowns that caused a significant change ($p \leq 0.05$) in splicing at or adjacent to bound FLEXIs compared to the other categories of introns. In the RI plots, shifts to the right or left indicate an increase or decrease in retained introns, respectively, and in the SE plots, shifts to the left or right indicate an increase or decrease in skipped exons, respectively. Statistical significance was calculated by Kolmogorov-Smirnov test. Names of RBP are color coded by protein function: red, RNA splicing-related function; orange, miRNA-related function; black, other function. Datasets that were not available for a specific RBP in one of the two cell lines were left as a blank space.

Figure S13. Density plots showing the location of eCLIP-identified RBP-binding sites within FLEXIs and other classes of introns. (*A*) RBPs with an RNA splicing-related function. (*B*) RBPs with either a miRNA-related function or miRNA- and splicing-related functions. (*C*) RBPs that bind snoRNAs. (*D*) RBPs with other functions. RBP-binding sites overlapping with introns were identified by intersecting annotated RBP-binding sites with IDR analysis from the eCLIP datasets with the intron coordinates using BedTools and plotted as the mid-point of the annotated RBP-binding sites normalized as a percentage of intron length, with 0% and 100% corresponding to the 5' and 3' end of the intron, respectively. For each RBP, binding-site distributions are shown for FLEXIs (red), other short introns (≤300 nt; blue), and long introns (>300 nt; black). Vertical red dashed lines indicate the position of peaks in the density plots for FLEXI RNAs. Names of RBP are color coded by protein function: RNA splicing-related function, red; miRNA-related function, orange; both RNA splicing- and miRNA-related functions, blue; other functions, black. Blank spaces were left in cases where datasets were not available for an RBP in one of the two cell lines used to obtain the eCLIP datasets.

Figure S14. Hierarchical clustering of RBPs based on patterns of over- and under-represented RBP-binding sites in FLEXI RNAs. Subsets of FLEXIs that have a CLIP-seq identified binding site for each of the 47 RBPs that have binding sites in ≥30 different FLEXI RNAs in

merged TGIRT-seq datasets were extracted for each of the cellular RNA samples. Contingency tables comparing the frequency of RBP-binding sites in each of the 47 subsets in each of the cellular RNA samples were then generated for all 126 RBPs identified as having a binding site in a FLEXI RNA (124 RBPs from ENCODE eCLIP dataset plus AGO1-4 and DICER from PAR-CLIP dataset). Multiple binding sites for the same RBP in the same FLEXI were counted as one binding site, and p-values were calculated by Fisher's exact test and adjusted by the Benjamini-Hochberg procedure. RBP-binding sites were then key-coded as not significantly different or as significantly over- or under-represented (≥2% abundance, $p \leq 0.05$) in the tested subset of FLEXIs compared to all FLEXIs, and matrices of the Gower's distance between the 47 subsets in each of the cellular RNA samples were calculated using the key-coded information and used as input for hierarchical clustering by the complete linkage clustering method (Defays, 1977). The results were displayed as a heat map to identify subsets of FLEXIs showing similar patterns of significantly over- and under-represented RBP-binding sites in each of the cellular RNA samples, with the color scale based on $-\log_{10}$ transformed adjusted p-values. Significantly co-enriched RBPs are indicated by an X in the heatmap box, and clusters of co-enriched RBPs are delineated in larger boxes. Proteins names are color coded by function as indicated in the Figure.

Figure S15. Scatter plots and density plots comparing RBP-binding sites and characteristics of subsets of FLEXIs bound by each of 47 RBPs that have binding sites in ≥30 different FLEXI RNAs. Plots are shown for each protein in Clusters I to VI, followed by the remainder of the 47 RBPs that have binding sites in ≥30 different FLEXI RNAs. In the scatter plots, RBP-binding sites whose relative abundance was significantly over-represented or under-represented in the tested subset of FLEXI compared to other FLEXIs (≥2% abundance, $p \leq 0.05$ calculated by Fisher's exact test and adjusted by the Benjamini-Hochberg procedure) are labeled by the name of the protein color coded by protein function as indicated in the Figure. The density distribution plots compare the length, GC content, and MFE for the most stable secondary structure predicted by RNAfold for the FLEXI RNAs comprising the subset of FLEXIs containing a binding site for

the (red) compared to all other FLEXIs (black). The number of FLEXIs comprising each subset is indicated in parentheses next to the name of the RBP. p-values are shown at the top left of those density plots in which the distribution for the subset of FLEXIs differed significantly from other FLEXIs ($p<0.01$ and FDR$\leq$0.05 as determined by 1,000 Monte-Carlo simulation).

Figure S16. Identification of clusters of RBPs whose binding sites are enriched in randomly selected subsets of other short and long introns. (*A*) Hierarchical clustering was carried out as in Fig. 7 for ten sets of 2,000 randomly selected other (non-FLEXI) short introns and long introns (Sim1-10). The clusters of RBPs enriched in subsets of FLEXI RNAs in HEK-293T, HeLa S3, K-562, and UHHR are shown at the top for comparison. (*B*) A heatmap of GO terms enriched in host genes of FLEXI RNAs containing binding sites for different RBPs. GO enrichment analysis was performed using DAVID bioinformatics tools, using all FLEXI host genes as the background. Similar results were obtained using all ENSEMBL-annotated genes as background (not shown).

Figure S17. Density distribution plots of the abundance of different categories of FLEXI RNAs in MDA-MB-231 and MCF-7 cells. FLEXI RNA abundance (RPM) was calculated from combined datasets for technical replicates of each sample. Different categories of FLEXI RNAs are color coded as indicated in the Figure. The abundance distribution of annotated mature snoRNAs (dashed yellow line), and the abundance of different sncRNAs (arrows) in the same samples are shown for comparison.

Figure S18. Oncogene FLEXI RNAs in breast cancer tumors and cell lines. (*A* and *B*) UpSet plots of upregulated (A) and downregulated (B) oncogene FLEXI RNAs in rRNA-depleted unfragmented RNA preparations from matched cancer/healthy breast tissues from patients A and B and breast cancer cell lines MDA-MB-231 and MCF7. (*C* and *D*) UpSet plots of upregulated (C) and downregulated (D) tumor suppressor gene (TSG) FLEXI RNAs in the same unfragmented

RNA preparations. Up and down regulated FLEXI RNAs were defined as those with an RPM-fold change ≥2. The most abundant oncogene or tumor suppressor gene FLEXI RNAs (up to a limit of 10) are listed below some sample groups, with the range of RPM values indicated in parentheses at the bottom. Different FLEXIs originating from the *FASN* oncogene (highlighted in red) showed notably different patterns of up- or down-regulation in the same cancer samples.

Methods

DNA and RNA oligonucleotides

DNA and RNA oligonucleotides were purchased from Integrated DNA Technologies (IDT) in RNase-free, HPLC-purified form. The DNA and RNA oligonucleotides used for TGIRT-seq are shown in Table S4. R2R DNA oligonucleotides with 3' A, C, G, and T residues were hand-mixed in equimolar amounts prior to annealing to the R2 RNA oligonucleotide to generate RNA/DNA heteroduplexes with a 1-nt 3'-DNA overhang for RNA-seq adapter addition by TGIRT template switching.

The oligonucleotides used Terminator exonuclease and ddPCR experiments are listed in Table S5 and include reverse transcriptase primers, PCR primers, and the 3I_RAN FLEXI synthetic oligonucleotide controls with and without a 5' phosphate. The sequence specificity of all primer sets (Table S5) was confirmed by gel purifying a representative subset of qPCR amplicons and either directly Sanger sequencing or by TOPO-TA cloning (ThermoFisher) followed by Sanger sequencing. A circular form of the 3I_RAN FLEXI synthetic oligonucleotide control was generated by incubating the 5'phosphorylated-3I_RAN FLEXI oligonucleotide (500 ng) with T4 RNA Ligase I (10 U; New England Biolabs) for 2 h at 25 °C and 2 min at 95 °C. The circularized oligonucleotide was cleaned up with an RNA Clean & Concentrator kit (Zymo Research), using 8 sample volumes of ethanol added during purification (denoted 8X ethanol method) to maximize the recovery of small RNAs concurrently with the longer RNAs.

RNA preparations

Universal Human Reference RNA (UHRR) was purchased from Agilent (product name: Quantitative PCR Human Reference Total RNA), and HeLa S3 and MCF-7 RNAs were purchased from Thermo Fisher. RNAs from matched frozen healthy/tumor tissues of breast cancer patients were purchased from Origene (500 ng; Patient A: PR^+, ER^+, $HER2^-$, CR562524/CR543839; Patient B: PR unknown, ER^-, $HER2^-$, CR560540/CR532030).

K-562, HEK-293T/17, and MDA-MB-231 RNAs were isolated from cultured cells by using a mirVana miRNA Isolation Kit (Thermo Fisher). K-562 cells (ATCC CTL-243) were maintained in Iscove's Modified Dulbecco's Medium (IMDM) + 4 mM L-glutamine and 25 mM HEPES; Thermo Fisher) supplemented with 10% Fetal Bovine Serum (FBS; Gemini Bio-Products), and approximately 2 x 10^6 cells were used for RNA extraction. HEK-293T/17 cells (ATCC CRL-11268) were maintained in Dulbecco's Modified Eagle Medium (DMEM) + 4.5 g/L D-glucose, 4 mM L-glutamine, and 1 mM sodium pyruvate, supplemented with 10% FBS, and approximately 4 x 10^6 cells were used for RNA extraction. MDA-MB-231 cells (ATCC HTB-26) were maintained in DMEM + 4.5 g/L D-glucose and 4 mM L-glutamine (Thermo Fisher) supplemented with 10% FBS and 1X PSQ (Penicillin, Streptomycin, and Glutamine: Thermo Fisher), and approximately 4 x 10^6 cells were used for RNA extraction. All cells were maintained at 37 °C in a humidified 5% CO_2 atmosphere.

For RNA isolation, cells were harvested by centrifugation (after trypsinization for HEK-293T/17 and MDA-MB-231 cells) at 300 x g for 10 min at 4 °C and washed twice by centrifugation with cold Dulbecco's Phosphate Buffered Saline (Thermo Fisher). The indicated number of cells (see above) was then resuspended in 600 µL of mirVana Lysis Buffer and RNA was isolated according to the kit manufacturer's protocol with elution in a final volume of 100 µL.

Terminator exonuclease assay

Unfragmented HEK-293T and UHHR total RNAs (500 ng) or synthetic 3I_RAN FLEXI oligonucleotide controls (100 ng; oligonucleotide with or without a 5'phosphate or circularized

oligonucleotide) were treated with 1 U Terminator exonuclease (Lucigen) or mock treated in the manufacturer's Buffer A for 1 h at 37 °C followed by 1 h at 30 °C. The terminator reactions were cleaned up with the RNA Clean & Concentrator kit (8X ethanol method; Zymo Research). The manufacturer's instructions note that the standard buffer (Buffer A) is the preferred reaction buffer for maximal exonuclease activity, but that non-5'-monophosphate-specific activity, including activity with RNAs having a 5' hydroxyl activity, might also result. Reverse transcription was performed by using a gene specific primer (1 μM final concentration) located near the 3' end of the FLEXI intron with 100 U Maxima H reverse transcriptase (ThermoFisher) in the manufacturer's buffer and incubated at 25 °C for 10 min, 50 °C for 30 min, 65 °C for 30 min and 85 °C for 5 min. qPCR reactions were carried out with 1 μL cDNA and PowerUp SYBR Green master mix (Life Technologies) using a ViiA 7 Real Time PCR system (ThermoFisher). Delta CT values were calculated to obtain a fold change of RNA levels in the treated compared to the mock treated sample. The primer sequences and 3I_RAN FLEXI synthetic oligonucleotide control sequences are shown in Table S5.

Droplet digital PCR

Unfragmented HeLa S3, UHRR or HEK-293T RNAs (5 μg) were DNased with 2 U TURBO DNase (Invitrogen) and purified using an RNA Clean & Concentrator kit (Zymo Research) following the manufacturer's small-size selection method, to separate RNAs ≤200 nt. The input of the size-selected RNAs for reverse transcription reactions were scaled to have a similar SNORD14B control amplification when compared with total RNA inputs (HEK-293T and HeLa S3, 5 ng; UHHR, 1 ng). The RNAs were reverse transcribed by using 1 μM final concentration of a gene specific primer (Table S5) located at the 3' end of the FLEXI intron with 100 U Maxima H Reverse Transcriptase (ThermoFisher) in the manufacturer's buffer and incubated at 25 °C for 10 min, 50 °C for 30 min, 65 °C for 30 min and 85 °C for 5 min. Matched control samples containing the small size-selected RNA, but omitting the RT (RT-) were performed for every reaction. Droplet digital PCR (ddPCR) was performed on 1 μL cDNA using QX200 ddPCR

EvaGreen Supermix (Biorad) according to manufacturer's protocol with the QX200 Droplet Digital PCR system (BioRad). Primer sets (Table S5) were designed to encompass either the FLEXI intron sequence or to span the upstream 5'exon-FLEXI junction in order to quantify the FLEXI RNA compared to any precursor or mature mRNA with retained FLEXI introns. The same gene-specific-generated cDNA was used as the PCR template for quantifying both the FLEXI and 5' exon-FLEXI junctions. Due to the short size of the FLEXI sequences, primer design was constrained to the best available primer location that maximized length and sequence specificity. As a result of the primer constraints, some primers amplified a smaller region of the FLEXI sequence than the TGIRT-seq criteria (*i.e.,* 5' and 3' ends within 3 nt of the annotated splice sites), which could contribute to differences between ddPCR and TGIRT-seq quantification. 5'-exon-FLEXI junction primer pairs used a different reverse primer than the FLEXI, that was inset closer to the 5'exon-FLEXI junction, so that the size of the amplicon would be more similar to the size of the compared FLEXI amplicon. As a consequence of these constraints, some primer pairs amplified background non-specific products in addition to verified products. For each reaction replicate, any RT- signal was subtracted, and three background-subtracted reaction replicates were averaged for each primer pair. The majority of the 5'exon-FLEXI junction reactions either gave no product or amplified non-specific products. However, a few 5'exon-FLEXI junction reactions produced validated products sequenced by direct Sanger sequencing and/or TOPO-TA cloning and Sanger sequencing of gel extracted, matched qPCR amplicons (HEK-293T, RPS2; HeLa S3, ACTB; UHRR, POLG and ACTB). These sequence-validated 5'exon-FLEXI junction ddPCR signals were subtracted from the FLEXI ddPCR signals to obtain the final ddPCR estimate for abundance of the FLEXI. FLEXI copies per reaction were then normalized to the controls performed in parallel: U7 (chosen for similar abundance to FLEXIs), SNORD14B (Dodd et al., 2013), and SNORD44 (Sharbati et al., 2012). Finally, the normalized FLEXI copies per reaction were converted to copy number per cell estimates by multiplying by respective control values generated by the linear regression model without or with TGIRT-seq bias correction.

TGIRT-seq

To remove residual DNA, UHRR and HeLa S3 RNAs (1 μg) and patients A and B healthy and cancer tissue RNAs (500 ng) were incubated with 20 U exonuclease I (Lucigen) and 2 U Baseline-ZERO DNase (Lucigen) in Baseline-ZERO DNase Buffer for 30 min at 37 °C, and K562, MDA-MB-231 and HEK-293T cell RNAs (5 μg) were incubated with 2 U TURBO DNase for 30 min at 37 °C (ThermoFisher). After DNA digestion, RNA was cleaned up with an RNA Clean & Concentrator kit (8X ethanol method; Zymo Research). The eluted RNAs were ribodepleted by using the rRNA removal section of a TruSeq Stranded Total RNA Library Prep Human/Mouse/Rat kit (Illumina), with the supernatant from the magnetic-bead separation cleaned-up by using a Zymo RNA Clean & Concentrator kit with 8X ethanol. After checking RNA concentration and length by using an Agilent 2100 Bioanalyzer with a 6000 RNA Pico chip, RNAs were aliquoted into ~20 ng portions and stored at -80 °C until use.

For the preparation of samples containing chemically fragmented long RNAs, RNA preparations were treated with exonuclease I and Baseline-Zero DNase to remove residual DNA and ribodepleted, as described above. The supernatant from the magnetic-bead separation after ribodepletion was then cleaned-up with a Zymo RNA Clean & Concentrator kit using the manufacturer's two-fraction protocol, which separates RNAs into long and short RNA fractions (200-nt cut-off). The long RNAs were then fragmented to 70-100 nt by using an NEBNext Magnesium RNA Fragmentation Module (94 °C for 7 min; New England Biolabs). After clean-up by using a Zymo RNA Clean & Concentrator kit (8X ethanol protocol), the fragmented long RNAs were combined with the unfragmented short RNAs and treated with T4 polynucleotide kinase (Epicentre) to remove 3' phosphates (Xu et al., 2019), followed by clean-up using a Zymo RNA Clean & Concentrator kit (8X ethanol protocol). After confirming the RNA fragment size range and RNA concentration by using an Agilent 2100 Bioanalyzer with a 6000 RNA Pico chip, the RNA was aliquoted into 4 ng portions for storage in -80 °C.

TGIRT-seq libraries were prepared as described (Xu et al., 2019) using 20-50 ng of ribodepleted unfragmented cellular RNA or 4-10 ng of ribodepleted chemically fragmented

cellular RNA. Template-switching and reverse transcription reactions were done with 1 μM TGIRT-III (InGex) and 100 nM pre-annealed R2 RNA/R2R DNA starter duplex in 20 μl of reaction medium containing 450 mM NaCl, 5 mM $MgCl_2$, 20 mM Tris-HCl, pH 7.5 and 5 mM DTT. Reactions were set up with all components except dNTPs, pre-incubated for 30 min at room temperature, a step that increases the efficiency of RNA-seq adapter addition by TGIRT template switching, and initiated by adding dNTPs (final concentrations 1 mM each of dATP, dCTP, dGTP, and dTTP). The reactions were incubated for 15 min at 60 °C and then terminated by adding 1 μl 5 M NaOH to degrade RNA and heating at 95 °C for 5 min followed by neutralization with 1 μl 5 M HCl and one round of MinElute column clean-up (Qiagen). The R1R DNA adapter was adenylated by using a 5' DNA Adenylation kit (New England Biolabs) and then ligated to the 3' end of the cDNA by using thermostable 5' App DNA/RNA Ligase (New England Biolabs) for 2 h at 65 °C. The ligated products were purified by using a MinElute Reaction Cleanup Kit and amplified by PCR with Phusion High-Fidelity DNA polymerase (Thermo Fisher Scientific): denaturation at 98 °C for 5 sec followed by 12 cycles of 98 °C 5 sec, 60 °C 10 sec, 72 °C 15 sec and then held at 4 °C until removed. The PCR products were cleaned up by using Agencourt AMPure XP beads (1.4X volume; Beckman Coulter) and sequenced on an Illumina NextSeq 500 to obtain 2 x 75 nt paired-end reads or on an Illumina NovaSeq 6000 to obtain 2 x 150 nt paired-end reads at the Genome Sequence and Analysis Facility of the University of Texas at Austin.

TGIRT-seq datasets of commercial human plasma pooled from multiple healthy individuals were obtained in a previous study (Yao et al., 2020) and deposited in the National Center for Biotechnology Information Sequence Read Archive under accession number PRJNA640428. TGIRT-seq datasets for biological replicates of unfragmented cellular RNAs from HEK-293T were previously deposited in the National Center for Biotechnology Information Sequence Read Archive under accession number PRJNA389469 (samples WC and YC). TGIRT-seq datasets for biological replicates of unfragmented cellular RNAs from K-562, MDA-MB-231 and UHRR were previously deposited in the National Center for Biotechnology Information Sequence Read

Archive under accession number PRJNA722757 (K-562 cells, samples UFC1-7; MDA-MB-231, samples MDA1-5; and UHRR, samples UHRR0-9).

Sequencing data processing

Illumina TruSeq adapters and PCR primer sequences were trimmed from the reads with Cutadapt v2.8 (sequencing quality score cut-off at 20; p-value <0.01) (Martin, 2011) and reads <15-nt after trimming were discarded. To minimize mismapping, we used a sequential mapping strategy. First, reads were mapped to the human mitochondrial genome (Ensembl GRCh38 Release 93) and the *Escherichia coli* genome (GeneBank: NC_000913) using HISAT2 v2.1.0 (Kim et al., 2019) with customized settings (-k 10 --rfg 1,3 --rdg 1,3 --mp 4,2 --no-mixed --no-discordant --no-spliced-alignment) to filter out reads derived from mitochondrial and *E. coli* RNAs (denoted Pass 1). Unmapped read from Pass1 were then mapped to a collection of customized references sequences for human sncRNAs (miRNA, tRNA, Y RNA, Vault RNA, 7SL RNA, 7SK RNA) and rRNAs (2.2-kb 5S rRNA repeats from the 5S rRNA cluster on chromosome 1 (1q42, GeneBank: X12811) and 43-kb 45S rRNA containing 5.8S, 18S and 28S rRNAs from clusters on chromosomes 13,14,15, 21, and 22 (GeneBank: U13369), using HISAT2 with the following settings -k 20 --rdg 1,3 --rfg 1,3 --mp 2,1 --no-mixed --no-discordant --no-spliced-alignment --norc (denoted Pass 2). Unmapped reads from Pass 2 were then mapped to the human genome reference sequence (Ensembl GRCh38 Release 93) using HISAT2 with settings optimized for non-spliced mapping (-k 10 --rdg 1,3 --rfg 1,3 --mp 4,2 --no-mixed --no-discordant --no-spliced-alignment) (denoted Pass 3) followed by splice aware mapping (-k 10 --rdg 1,3 --rfg 1,3 --mp 4,2 --no-mixed --no-discordant --dta) (denoted Pass 4). Finally, the remaining unmapped reads were mapped to Ensembl GRCh38 Release 93 by Bowtie 2 v2.2.5 (Langmead and Salzberg, 2012) using local alignment (-k 10 --rdg 1,3 --rfg 1,3 --mp 4 --ma 1 --no-mixed --no-discordant --very-sensitive-local) to improve mapping rates for reads containing post-transcriptionally added 5' or 3' nucleotides (poly(A) or poly(U)), short untrimmed adapter sequences, and non-templated nucleotides added to the 3' end of the cDNAs by TGIRT-III during TGIRT-seq library preparation

(denoted Pass 5). For reads that map to multiple genomic loci with the same mapping score in passes 3 to 5, the alignment with the shortest distance between the two paired ends (*i.e.*, the shortest read span) was selected. In the case of ties (*i.e.*, reads with the same mapping score and read span), reads mapping to a chromosome were selected over reads mapping to scaffold sequences, and in other cases, the read was assigned randomly to one of the tied choices. The filtered multiply mapped reads were then combined with the uniquely mapped reads from Passes 3-5 by using SAMtools v1.10 (Li et al., 2009).

To generate counts for different genomic features, the mapped reads were intersected with Ensembl GRCh38 Release 93 gene annotations plus the *RNY5* gene and its 10 pseudogenes, which are not annotated in this release. Coverage of each feature was calculated by BEDTools v2.29.2 (Quinlan, 2014). To avoid miscounting reads with embedded sncRNAs that were not filtered out in Pass 2 (*e.g.,* snoRNAs), reads were first intersected with sncRNA annotations and the remaining reads were then intersected with the annotations for protein-coding genes RNAs, lincRNAs, antisense RNAs, and other lncRNAs to get the read count for each annotated feature. Coverage plots and read alignments were created by using Integrative Genomics Viewer v2.6.2 (IGV) (Robinson et al., 2011). Genes with >100 mapped reads were down sampled to 100 mapped reads in IGV for visualization.

FLEXI identification and characterization

To identify short introns that could give rise to FLEXI RNAs, intron annotations were extracted from Ensemble GRCh38 Release 93 gene annotations using a customized script and filtered to remove introns >300 nt as well as duplicate intron annotations from different mRNA isoforms. To identify FLEXI RNAs, mapped reads were intersected with the short intron annotations using BEDTools, and read pairs (Read 1 and Read 2) ending at or within 3 nucleotides of annotated 5'- and 3'-splice sites were identified as corresponding to FLEXI RNAs. Exon-junction reads were defined as those in which one of the paired-end read mapped within the intron and the other within a flanking exon.

UpSet plots of FLEXI RNAs from different sample types were plotted by using the ComplexHeatmap package v2.2.0 in R (Gu et al., 2016). For UpSet plots of FLEXI host genes, FLEXI RNAs were aggregated by Ensemble ID, and different FLEXI RNAs from the same gene were combined into one entry. Density distribution plots and scatter plots were plotted by using R.

FLEXI RNAs corresponding to annotated mirtrons or agotrons were identified by intersecting FLEXI RNA coordinates with those of annotated mirtrons (Wen et al., 2015) and agotrons (Hansen et al., 2016). FLEXI RNAs containing embedded snoRNAs were identified by intersecting the FLEXI RNA coordinates with those of annotated snoRNA and scaRNA from Ensembl GRCh38 annotations.

5'- and 3'-splice sites (SS) and branch-point (BP) consensus sequences of human U2- and U12-type spliceosomal introns were obtained from previous publications (Gao et al., 2008; Sheth et al., 2006). FLEXI RNAs corresponding to U12-type introns were identified by searching for: (i) FLEXI RNAs with AU-AC ends and (ii) the 5'-splice site consensus sequence of U12-type introns with GU-AG ends (Sheth et al., 2006) using FIMO (Grant et al., 2011) with the following settings: FIMO --text --norc <GU_AG_U12_5SS motif file> <sequence file>. Splice-site consensus sequences of FLEXI RNAs were calculated from nucleotides frequencies at the 5' and 3' ends of the introns. The U2-type intron branch-point consensus sequence of intron RNAs was identified by searching for U2-type branch-point consensus using FIMO (Grant et al., 2011). The U12-type intron branch-point consensus was determined by searching for motifs enriched within 40 nt of the 3' end of the introns using MEME (Bailey et al., 2009) with settings: meme <sequence file> -rna -oc <output folder> -mod anr -nmotifs 100 -minw 6 -minsites 100 -markov order 1 -evt 0.05. The branch-point consensus sequence of U12-type FLEXI RNAs (2 with AU-AC ends and 34 with GU-AG matching the 5' sequence of GU-AG U12-type introns) was determined by manual sequence alignment and calculation of nucleotide frequencies. Motif logos were plotted from the nucleotide frequency tables of each motif using scripts from MEME suite (Bailey et al., 2009).

Clustering analysis

PCA analysis of cell-type specific FLEXI RNA profiles in replicate cellular RNA datasets was plotted using R, and PCA initialized t-SNE and ZINB-WaVE analyses of these datasets were plotted using the Rtsne (https://github.com/jkrijthe/Rtsne) and zinbwave (https://github.com/drisso/zinbwave) packages in R (Risso et al., 2018). Batch effect correction was done by using ZINB-WaVE to include the batch information as a sample-level covariate. The normalized counts outputted by ZINB-WaVE were used for PCA analysis, and the *W* matrix outputted by ZINB-WaVE was used for visualization directly or by t-SNE.

RBP-binding site analysis

FLEXI RNAs containing RBP-binding sites were identified by intersecting the FLEXI RNA co-ordinates with ENCODE annotated binding sites of 150 RBPs from the eCLIP-seq with IDR analysis (Van Nostrand et al., 2016) and crosslink-centered regions (CCRs) identified from the DICER PAR-CLIP (Rybak-Wolf et al., 2014) and AGO1-4 PAR-CLIP (Hafner et al., 2010) datasets by using BEDTools. Multiple binding sites for the same RBP in an intron were collapsed and counted as one site for that RBP. Functional annotations, localization patterns, and predicted RNA-binding domains of RBPs in the ENCODE eCLIP dataset were based on Table S2 of (Van Nostrand et al., 2020). RBPs found in stress granules were as annotated in the RNA Granule and Mammalian Stress Granules Proteome (MSGP) databases (Nunes et al., 2019; Youn et al., 2019). The functional annotations, localization patterns, and RNA-binding domains of AGO1-4 and DICER were from the UniProt database (The UniProt Consortium, 2018). Near full length FLEXI or other short intron reads were identified in ENCODE eCLIP datasets by intersecting mapped eCLIP reads with FLEXI or other short intron coordinates using BedTools.

Gene expression changes in RBP knockdown datasets were calculated from changes in mRNA abundance measured by using DESeq2 after GC correction with Salmon (Patro et al., 2017) and CQN (Hansen et al., 2012). The number of significant gene expression changes (DESeq2-calculated adjusted p-value<0.05) for FLEXI host genes was compared to three gene

sets containing the same number of randomly selected genes, and ANOVA was used to identify knockdowns that resulted in significantly more gene expression changes among FLEXI host genes than expected by random chance. To identify FLEXI host genes whose changes in mRNA levels were significantly biased in one direction, the numbers of upregulated and downregulated host genes containing a FLEXI with a binding site for the knocked-down RBP (defined as $\log_2$-transformed fold change>0, adjusted p-value<0.05) were compared to those for all significantly changed genes by Fisher's exact test, and those with p-value <0.05 for difference in one or the other direction were considered significantly biased in that direction. Volcano plots were generating by plotting $\log_2$ transformed fold changes in the mRNA levels between the knockdown and control datasets versus $-\log_{10}$ adjusted p-values.

Changes in alternative splicing were calculated by using rMATS (Shen et al., 2014). Coordinates of adjacent introns or of retained introns from splicing change files (MATS output) were intersected with coordinates of bound FLEXIs (*i.e.*, FLEXIs that contain a binding site for a RBP), unbound FLEXIs (*i.e.*, FLEXIs that do not contain a binding site for the same RBP), and other short introns, and the results were displayed as a density plot (Seaborn distplot) showing the inclusion level difference for the same intron between the knockdown and control datatsets.

To map RBP-binding sites within introns, RBP-binding sites from ENCODE eCLIP datasets were intersected with FLEXI, other short intron, or long intron coordinates (BedTools). Any binding site overlapping with any portion of the intron was used. The location of the binding site with respect to normalized intron length was calculated from the distance between the mid-point of the annotated binding site and the beginning or end of the intron. Data were displayed as a density plot of the mid-point of the binding sites normalized by intron length.

For hierarchical clustering analysis of the RBP-binding site co-enrichment, we first extracted subsets of FLEXIs that have a binding site for each of 47 RBPs that have binding sites in ≥30 different FLEXI RNAs from datasets for each of the cellular RNA samples. We then identified over- and under-represented RBPs in each of the 47 subsets for each cellular RNA sample by generating a contingency table comparing the frequency of binding sites for the 126 RBPs

that have an identified binding site in a FLEXI RNA (Fig. S9) in the tested subset to its frequency in all FLEXIs. Multiple binding sites for the same RBP in the same FLEXI were counted as one binding site. p-values were calculated using Fisher's exact test and then adjusted by the Benjamini-Hochberg procedure and $\log_{10}$ transformed. RBP binding sites were then key-coded as not significantly different or as significantly over- or under-represented (≥2% abundance, $p \leq 0.05$) in the tested subset of FLEXIs compared to all FLEXIs, and matrices of the Gower's distance (Gower, 1971) were calculated using the key-coded information from the above analyses and used as input for hierarchical clustering by the complete linkage clustering method (Defays, 1977). The results were displayed as a heat map plotted using the R package cluster and ComplexHeatmap, respectively.

GO enrichment analysis of host genes encoding FLEXIs bound by different RBPs was performed by using DAVID bioinformatics tools (Huang et al., 2009) with all FLEXI host genes as the background. Hierarchical clustering was performed based on p-values for GO term enrichment of FLEXI host genes bound by the same RBPs using the Seaborn ClusterMap package in Python (https://zenodo.org).

REFERENCES

Abebrese, E.L., Ali, S.H., Arnold, Z.R., Andrews, V.M., Armstrong, K., Burns, L., Crowder, H.R., Day, R.T., Jr., Hsu, D.G., Jarrell, K., et al. (2017). Identification of human short introns. PLOS ONE *12*, e0175393. 10.1371/journal.pone.0175393.

Alawi, F., and Lin, P. (2010). Loss of dyskerin reduces the accumulation of a subset of H/ACA snoRNA-derived miRNA. Cell Cycle *9*, 2467-2469. 10.4161/cc.9.12.11922.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202-W208. 10.1093/nar/gkp335.

Berezikov, E., Chung, W.-J., Willis, J., Cuppen, E., and Lai, E.C. (2007). Mammalian mirtron genes. Mol. Cell *28*, 328-336. 10.1016/j.molcel.2007.09.028.

Burset, M., Seledtsov, I.A., and Solovyev, V.V. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. *28*, 4364-4375. 10.1093/nar/28.21.4364.

Chapman, K.B., and Boeke, J.D. (1991). Isolation and characterization of the gene encoding yeast debranching enzyme. Cell *65*, 483-492. 10.1016/0092-8674(91)90466-C.

Defays, D. (1977). An efficient algorithm for a complete link method. The Computer Journal *20*, 364-366. 10.1093/comjnl/20.4.364.

Dodd, D.W., Gagnon, K.T., and Corey, D.R. (2013). Digital quantitation of potential therapeutic target RNAs. Nucleic Acid Ther *23*, 188-194. 10.1089/nat.2013.0427.

Farrell, M.J., Dobson, A.T., and Feldman, L.T. (1991). Herpes simplex virus latency-associated transcript is a stable intron. Proc. Natl. Acad. Sci. U.S.A. *88*, 790-794. 10.1073/pnas.88.3.790.

Gao, K., Masuda, A., Matsuura, T., and Ohno, K. (2008). Human branch point consensus sequence is yUnAy. Nucleic Acids Res. *36*, 2257-2267. 10.1093/nar/gkn073.

Gardner, E.J., Nizami, Z.F., Talbot, C.C., and Gall, J.G. (2012). Stable intronic sequence RNA (sisRNA), a new class of noncoding RNA from the oocyte nucleus of *Xenopus tropicalis*. Genes Dev. *26*, 2550-2559. 10.1101/gad.202184.112.

Gower, J.C. (1971). A General Coefficient of Similarity and Some of Its Properties. Biometrics *27*, 857. 10.2307/2528823.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017-1018. 10.1093/bioinformatics/btr064.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847-2849. 10.1093/bioinformatics/btw313.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.-C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129-141. 10.1016/j.cell.2010.03.009.

Hansen, K.D., Irizarry, R.A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics *13*, 204-216. 10.1093/biostatistics/kxr054.

Hansen, T.B., Venø, M.T., Jensen, T.I., Schaefer, A., Damgaard, C.K., and Kjems, J. (2016). Argonaute-associated short introns are a novel class of gene regulators. Nat Commun *7*, 11538. 10.1038/ncomms11538.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44-57. 10.1038/nprot.2008.211.

Kaiser, R.W.J., Ignarski, M., Van Nostrand, E.L., Frese, C.K., Jain, M., Cukoski, S., Heinen, H., Schaechter, M., Seufert, L., Bunte, K., et al. (2019). A protein-RNA interaction atlas of the ribosome biogenesis factor AATF. Sci. Rep. *9*, 11071. 10.1038/s41598-019-47552-3.

Katibah, G.E., Qin, Y., Sidote, D.J., Yao, J., Lambowitz, A.M., and Collins, K. (2014). Broad and adaptable RNA structure recognition by the human interferon-induced tetratricopeptide repeat protein IFIT5. Proc. Natl. Acad. Sci. U.S.A. *111*, 12025–12030. 10.1073/pnas.1412842111.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907-915. 10.1038/s41587-019-0201-4.

Kim, P., Yang, M., Yiya, K., Zhao, W., and Zhou, X. (2020). ExonSkipDB: functional annotation of exon skipping event in human. Nucleic Acids Res. *48*, D896-D907. 10.1093/nar/gkz917.

Kobak, D., and Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. Nat Commun *10*, 5416. 10.1038/s41467-019-13056-x.

Kulesza, C.A., and Shenk, T. (2006). Murine cytomegalovirus encodes a stable intron that facilitates persistent replication in the mouse. Proc. Natl. Acad. Sci. U.S.A. *103*, 18302-18307. 10.1073/pnas.0608718103.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. 10.1038/nmeth.1923.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079. 10.1093/bioinformatics/btp352.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417-425. 10.1016/j.cels.2015.12.004.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, pp. 10-12.

Massenet, S., Bertrand, E., and Verheggen, C. (2017). Assembly and trafficking of box C/D and H/ACA snoRNPs. RNA Biol *14*, 680-692. 10.1080/15476286.2016.1243646.

Mattijssen, S., Arimbasseri, A.G., Iben, J.R., Gaidamakov, S., Lee, J., Hafner, M., and Maraia, R.J. (2017). LARP4 mRNA codon-tRNA match contributes to LARP4 activity for ribosomal protein mRNA poly(A) tail length protection. eLife *6*, e28889. 10.7554/eLife.28889.

Morgan, J.T., Fink, G.R., and Bartel, D.P. (2019). Excised linear introns regulate growth in yeast. Nature *565*, 606-611. 10.1038/s41586-018-0828-1.

Moss, W.N., and Steitz, J.A. (2013). Genome-wide analyses of Epstein-Barr virus reveal conserved RNA structures and a novel stable intronic sequence RNA. BMC Genomics *14*, 543. 10.1186/1471-2164-14-543.

Nunes, C., Mestre, I., Marcelo, A., Koppenol, R., Matos, C.A., and Nóbrega, C. (2019). MSGP: the first database of the protein components of the mammalian stress granules. Database *2019*, baz031. 10.1093/database/baz031.

Okamura, K., Hagen, J.W., Duan, H., Tyler, D.M., and Lai, E.C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. Cell *130*, 89-100. 10.1016/j.cell.2007.06.028.

Parenteau, J., Maignon, L., Berthoumieux, M., Catala, M., Gagnon, V., and Abou Elela, S. (2019). Introns are mediators of cell response to starvation. Nature *565*, 612-617. 10.1038/s41586-018-0859-7.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods *14*, 417-419. 10.1038/nmeth.4197.

Pek, J.W., Osman, I., Tay, M.L.-I., and Zheng, R.T. (2015). Stable intronic sequence RNAs have possible regulatory roles in *Drosophila melanogaster*. J. Cell Biol. *211*, 243-251. 10.1083/jcb.201507065.

Qin, Y., Yao, J., Wu, D.C., Nottingham, R.M., Mohr, S., Hunicke-Smith, S., and Lambowitz, A.M. (2016). High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. RNA *22*, 111–128. 10.1261/rna.054809.115.

Quinlan, A.R. (2014). BEDTools: the swiss-army tool for genome feature analysis. Curr. Protoc. Bioinformatics *47*, 11.12.11-34. 10.1002/0471250953.bi1112s47.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2018). A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun *9*, 284. 10.1038/s41467-017-02554-5.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer.

Ruby, J.G., Jan, C.H., and Bartel, D.P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83-86. 10.1038/nature05983.

Rybak-Wolf, A., Jens, M., Murakawa, Y., Herzog, M., Landthaler, M., and Rajewsky, N. (2014). A variety of Dicer substrates in human and *C. elegans*. Cell *159*, 1153-1167. 10.1016/j.cell.2014.10.040.

Saini, H., Bicknell, A.A., Eddy, S.R., and Moore, M.J. (2019). Free circular introns with an unusual branchpoint in neuronal projections. eLife *8*, e47809. 10.7554/eLife.47809.

Savage, K.I., Gorski, J.J., Barros, E.M., Irwin, G.W., Manti, L., Powell, A.J., Pellagatti, A., Lukashchuk, N., McCance, D.J., McCluggage, W.G., et al. (2014). Identification of a BRCA1-mRNA splicing complex required for efficient DNA repair and maintenance of genomic stability. Mol Cell *54*, 445-459. 10.1016/j.molcel.2014.03.021.

Sharbati, S., Sharbati, J., Hoeke, L., Bohmer, M., and Einspanier, R. (2012). Quantification and accurate normalisation of small RNAs through new custom RT-qPCR arrays demonstrates Salmonella-induced microRNAs in human monocytes. BMC Genomics *13*, 23. 10.1186/1471-2164-13-23.

Shen, S., Park, J.W., Lu, Z.-x., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proceedings of the National Academy of Sciences *111*, E5593-E5601. 10.1073/pnas.1419161111.

Sheth, N., Roca, X., Hastings, M.L., Roeder, T., Krainer, A.R., and Sachidanandam, R. (2006). Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Res. *34*, 3955-3967. 10.1093/nar/gkl556.

Shurtleff, M.J., Yao, J., Qin, Y., Nottingham, R.M., Temoche-Diaz, M.M., Schekman, R., and Lambowitz, A.M. (2017). Broad role for YBX1 in defining the small noncoding RNA composition of exosomes. Proc. Natl. Acad. Sci. U.S.A. *114*, E8987-E8995. 10.1073/pnas.1712108114.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Research *15*, 1034-1050. 10.1101/gr.3715005.

Singh, R., and Reddy, R. (1989). Gamma-monomethyl phosphate: a cap structure in spliceosomal U6 small nuclear RNA. Proc Natl Acad Sci U S A *86*, 8280-8283. 10.1073/pnas.86.21.8280.

Talhouarne, G.J.S., and Gall, J.G. (2018). Lariat intronic RNAs in the cytoplasm of vertebrate cells. Proc. Natl. Acad. Sci. U.S.A. *115*, E7970-E7977. 10.1073/pnas.1808816115.

The ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74. 10.1038/nature11247.

The UniProt Consortium (2018). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506-D515. 10.1093/nar/gky1049.

Tycowski, K.T., Kolev, N.G., Conrad, N.K., Fok, V., and Steitz, J.A. (2006). The ever-growing world of small nuclear ribonucleoproteins. In The RNA World, Third Edition, R.F. Gesteland, T.R. Cech, and J.F. Atkins, eds. (Cold Spring Harbor Laboratory Press, NY.), pp. 327-368.

Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods *13*, 508-514. 10.1038/nmeth.3810.

Van Nostrand, E.L., Pratt, G.A., Yee, B.A., Wheeler, E.C., Blue, S.M., Mueller, J., Park, S.S., Garcia, K.E., Gelboin-Burkhart, C., Nguyen, T.B., et al. (2020). Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. Genome Biol. *21*, 90. 10.1186/s13059-020-01982-9.

Vohhodina, J., Barros, E.M., Savage, A.L., Liberante, F.G., Manti, L., Bankhead, P., Cosgrove, N., Madden, A.F., Harkin, D.P., and Savage, K.I. (2017). The RNA processing factors THRAP3 and BCLAF1 promote the DNA damage response through selective mRNA splicing and nuclear export. Nucleic Acids Res *45*, 12816-12833. 10.1093/nar/gkx1046.

Wang, C., Query, C.C., and Meier, U.T. (2002). Immunopurified small nucleolar ribonucleoprotein particles pseudouridylate rRNA independently of their association with phosphorylated Nopp140. Mol Cell Biol *22*, 8457-8466. 10.1128/mcb.22.24.8457-8466.2002.

Wen, J., Ladewig, E., Shenker, S., Mohammed, J., and Lai, E.C. (2015). Analysis of nearly one thousand mammalian mirtrons reveals novel features of Dicer substrates. PLoS Computat. Biol. *11*, e1004441. 10.1371/journal.pcbi.1004441.

Wilkinson, M.E., Charenton, C., and Nagai, K. (2020). RNA splicing by the spliceosome. Annu. Rev. Biochem. *89*, 359-388. 10.1146/annurev-biochem-091719-064225.

Xu, H., Yao, J., Wu, D.C., and Lambowitz, A.M. (2019). Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. Sci. Rep. *9*, 7953. 10.1038/s41598-019-44457-z.

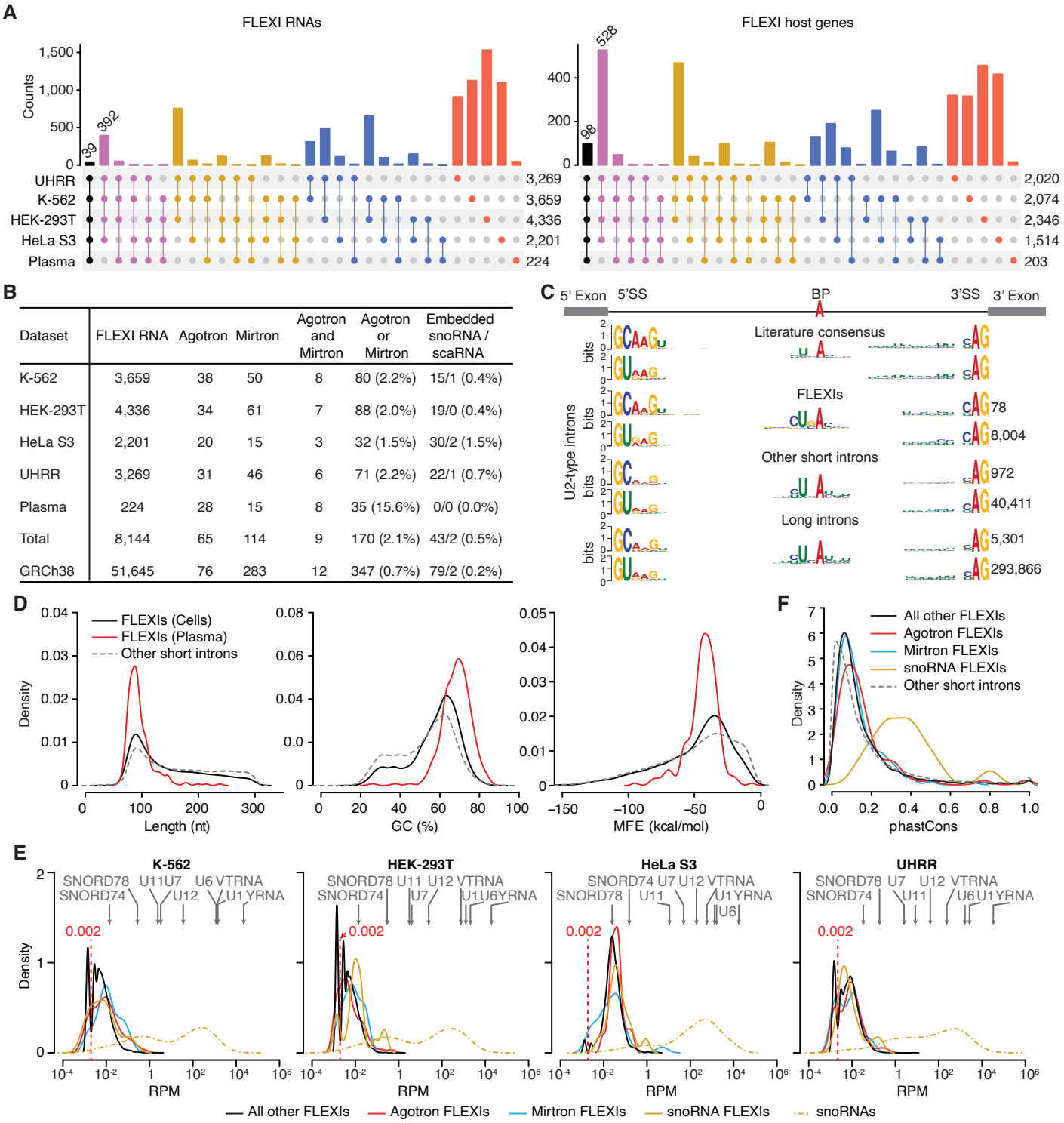
Yao, J., Wu, D.C., Nottingham, R.M., and Lambowitz, A.M. (2020). Identification of protein-protected mRNA fragments and structured excised intron RNAs in human plasma by TGIRT-seq peak calling. eLife *9*, e60743. 10.7554/eLife.60743.

Youn, J.-Y., Dyakov, B.J.A., Zhang, J., Knight, J.D.R., Vernon, R.M., Forman-Kay, J.D., and Gingras, A.-C. (2019). Properties of stress granule and P-body proteomes. Mol. Cell *76*, 286-294. https://doi.org/10.1016/j.molcel.2019.09.014.

Zhang, Y., Zhang, X.-O., Chen, T., Xiang, J.-F., Yin, Q.-F., Xing, Y.-H., Zhu, S., Yang, L., and Chen, L.-L. (2013). Circular intronic long noncoding RNAs. Mol. Cell *51*, 792-806. 10.1016/j.molcel.2013.08.017.

Zhao, M., Kim, P., Mitra, R., Zhao, J., and Zhao, Z. (2016). TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Res. *44*, D1023-D1031. 10.1093/nar/gkv1268.

Fig. 1



B

Dataset	FLEXI RNA	Agotron	Mirtron	Agotron and Mirtron	Agotron or Mirtron	Embedded snoRNA / scaRNA
K-562	3,659	38	50	8	80 (2.2%)	15/1 (0.4%)
HEK-293T	4,336	34	61	7	88 (2.0%)	19/0 (0.4%)
HeLa S3	2,201	20	15	3	32 (1.5%)	30/2 (1.5%)
UHRR	3,269	31	46	6	71 (2.2%)	22/1 (0.7%)
Plasma	224	28	15	8	35 (15.6%)	0/0 (0.0%)
Total	8,144	65	114	9	170 (2.1%)	43/2 (0.5%)
GRCh38	51,645	76	283	12	347 (0.7%)	79/2 (0.2%)

Fig. 2

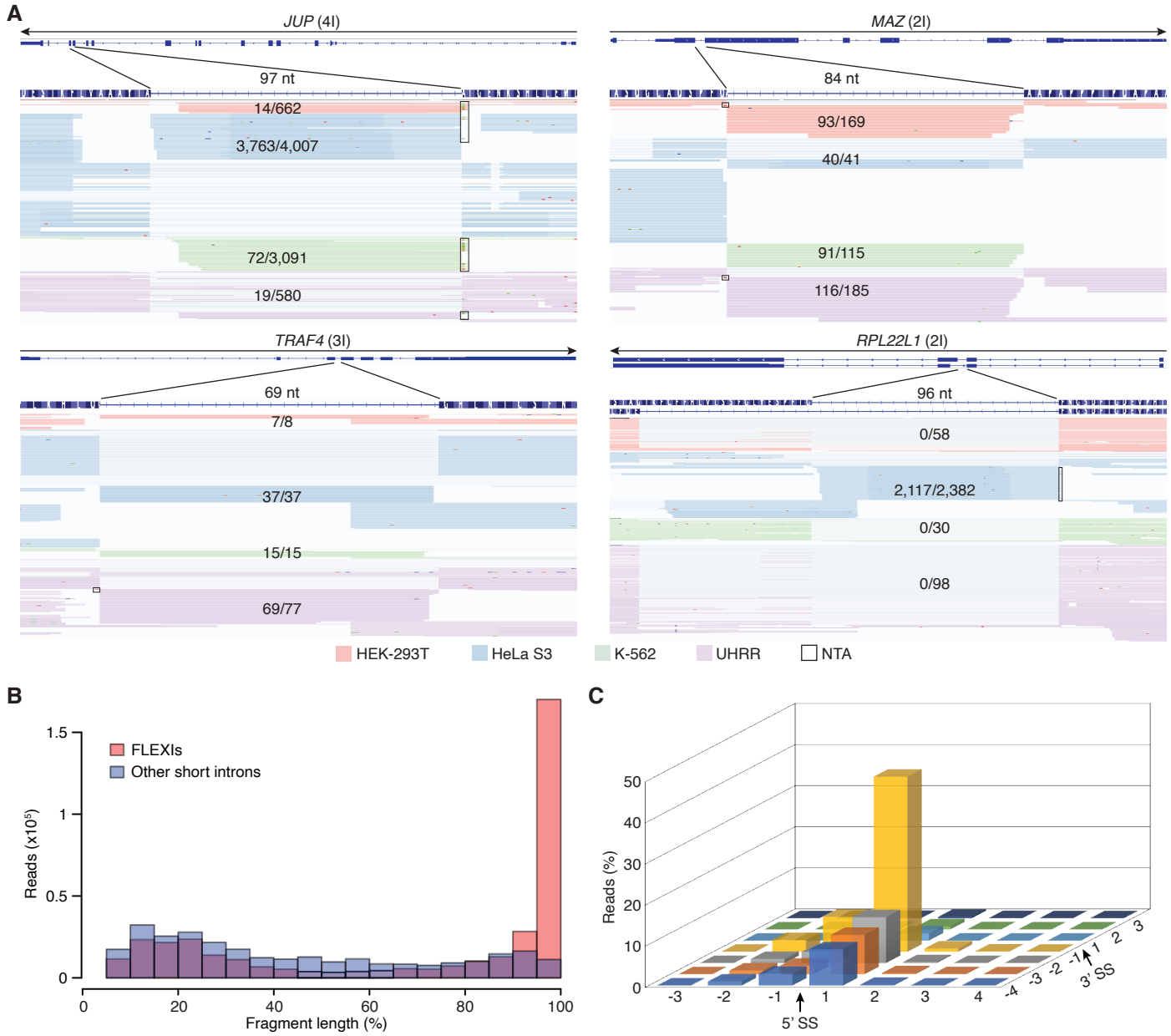
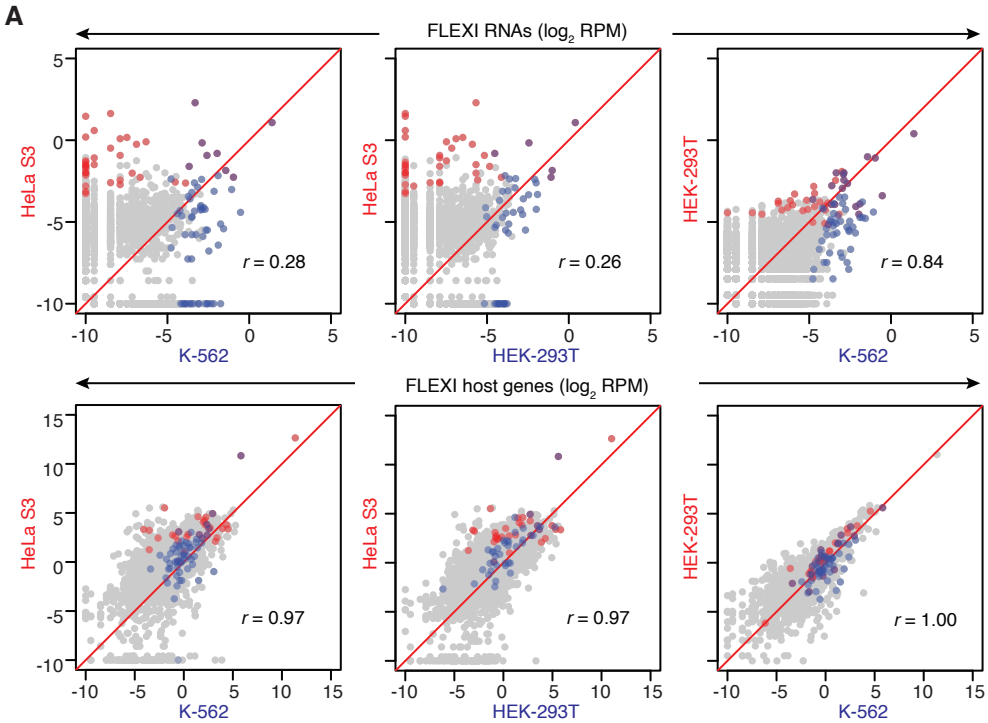
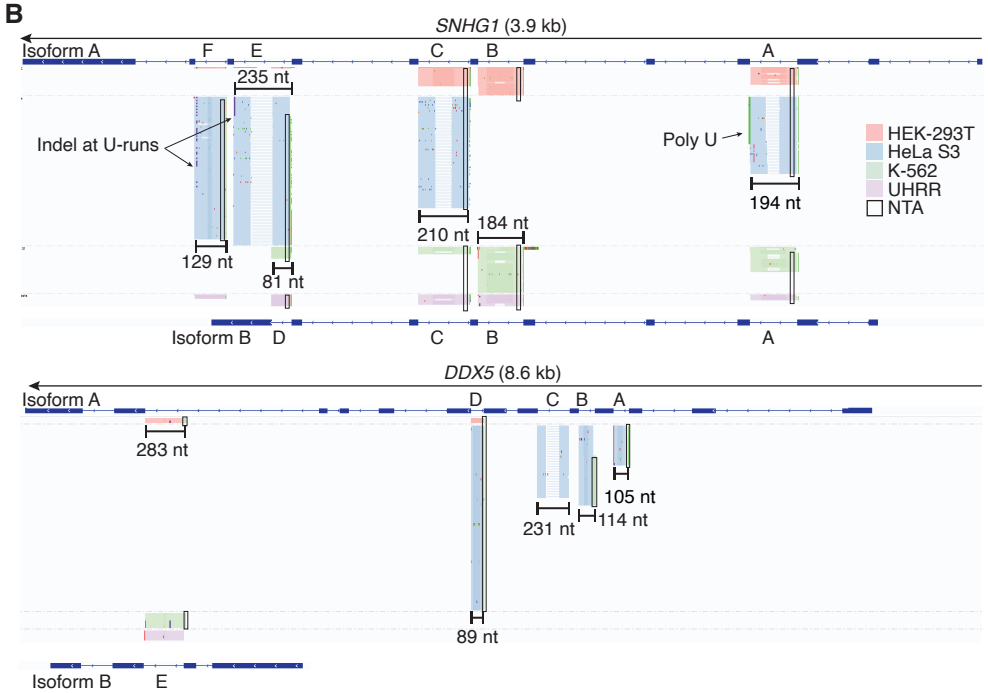

A
JUP (4I)
97 nt
14/662
3,763/4,007
72/3,091
19/580
MAZ (2I)
84 nt
93/169
40/41
91/115
116/185
TRAF4 (3I)
69 nt
7/8
37/37
15/15
69/77
RPL22L1 (2I)
96 nt
0/58
2,117/2,382
0/30
0/98
HEK-293T
HeLa S3
K-562
UHRR
NTA
B
FLEXIs
Other short introns
Reads (x10⁵)
Fragment length (%)
C
Reads (%)
5' SS
3' SS

Fig. 3





C

FLEXIs	phastCons	Protein coding	Alt. Spl.	Shared SS	RI	In frame RI
7,940	<0.75	7,673 (97%)	2,788 (35%)	1,450 (18%)	1,984 (25%)	616 (7.8%)
158	≥0.75	154 (97%)	109 (69%)	70 (44%)	75 (47%)	32 (20%)
44	≥0.99	44 (100%)	34 (77%)	31 (70%)	29 (66%)	13 (30%)

Fig. 4

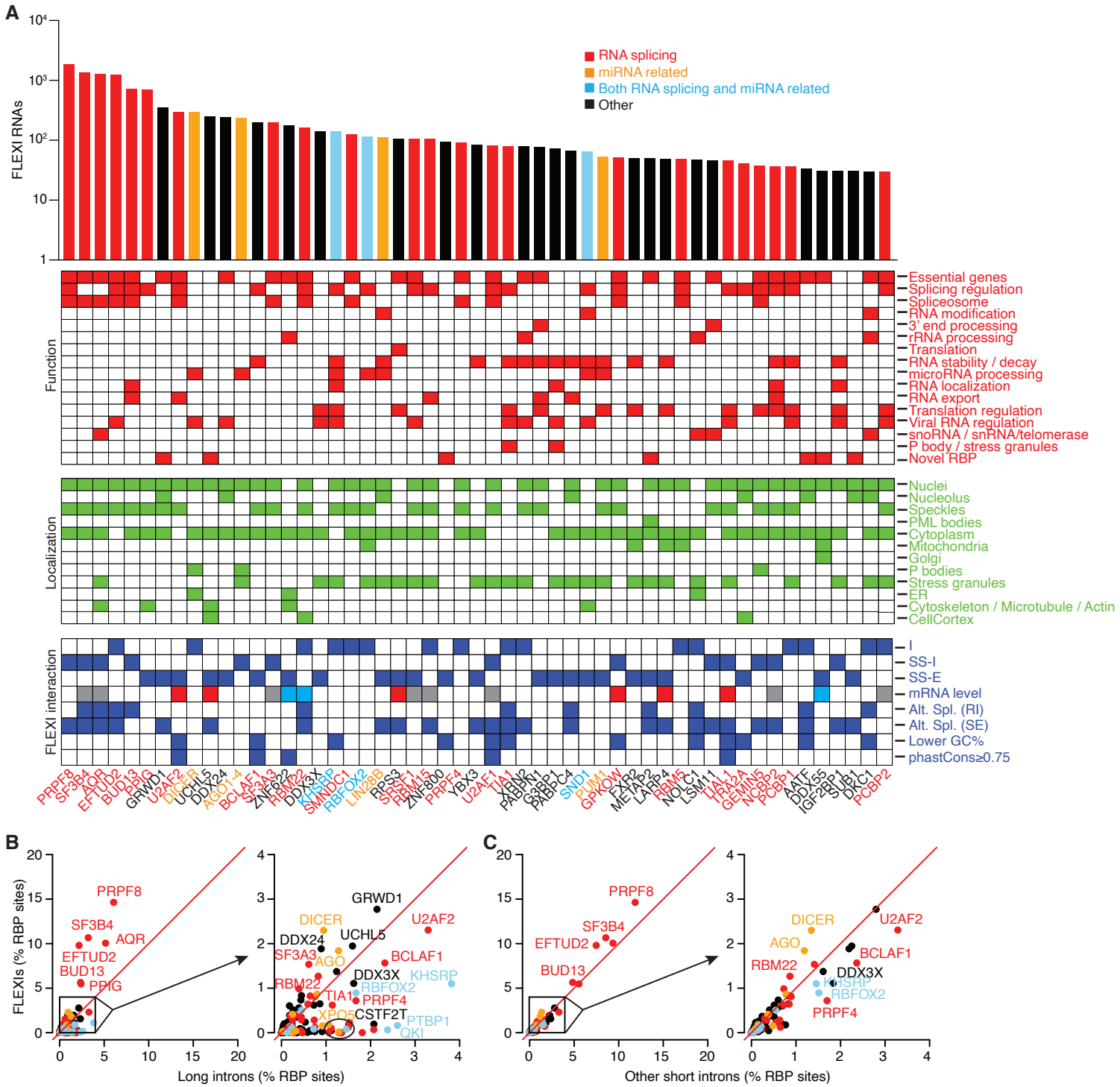

A
FLEXI RNAs
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other
Function
Essential genes
Splicing regulation
Spliceosome
RNA modification
3' end processing
rRNA processing
Translation
RNA stability / decay
microRNA processing
RNA localization
RNA export
Translation regulation
Viral RNA regulation
snoRNA / snRNA/telomerase
P body / stress granules
Novel RBP
Localization
Nuclei
Nucleolus
Speckles
PML bodies
Cytoplasm
Mitochondria
Golgi
P bodies
Stress granules
ER
Cytoskeleton / Microtubule / Actin
CellCortex
FLEXI interaction
I
SS-I
SS-E
mRNA level
Alt. Spl. (RI)
Alt. Spl. (SE)
Lower GC%
phastCons≥0.75
PRPF8 SF3B4 AQR EFTUD2 BUD13 PPIG GRWD1 U2AF2 DICER UCHL5 DDX24 AGO1-4 BCLAF1 SF3A3 ZNF622 RBM22 DDX3X KHSRP SMNDC1 RBFOX2 LIN28B RPS3 SRSF1 RBM15 ZNF800 PRPF4 YBX3 U2AF1 TIA1 XRN2 PABPN1 G3BP1 PABPC4 SND1 PUM1 GPKOW FXR2 METAP2 LARP4 RBM5 NOLC1 LSM11 UAL1 TRA2A GEMIN5 NCBP2 PCBP1 AATF DDX55 IGF2BP1 SUB1 DKC1 PCBP2
B
FLEXIs (% RBP sites)
Long introns (% RBP sites)
C
Other short introns (% RBP sites)

Fig. 5

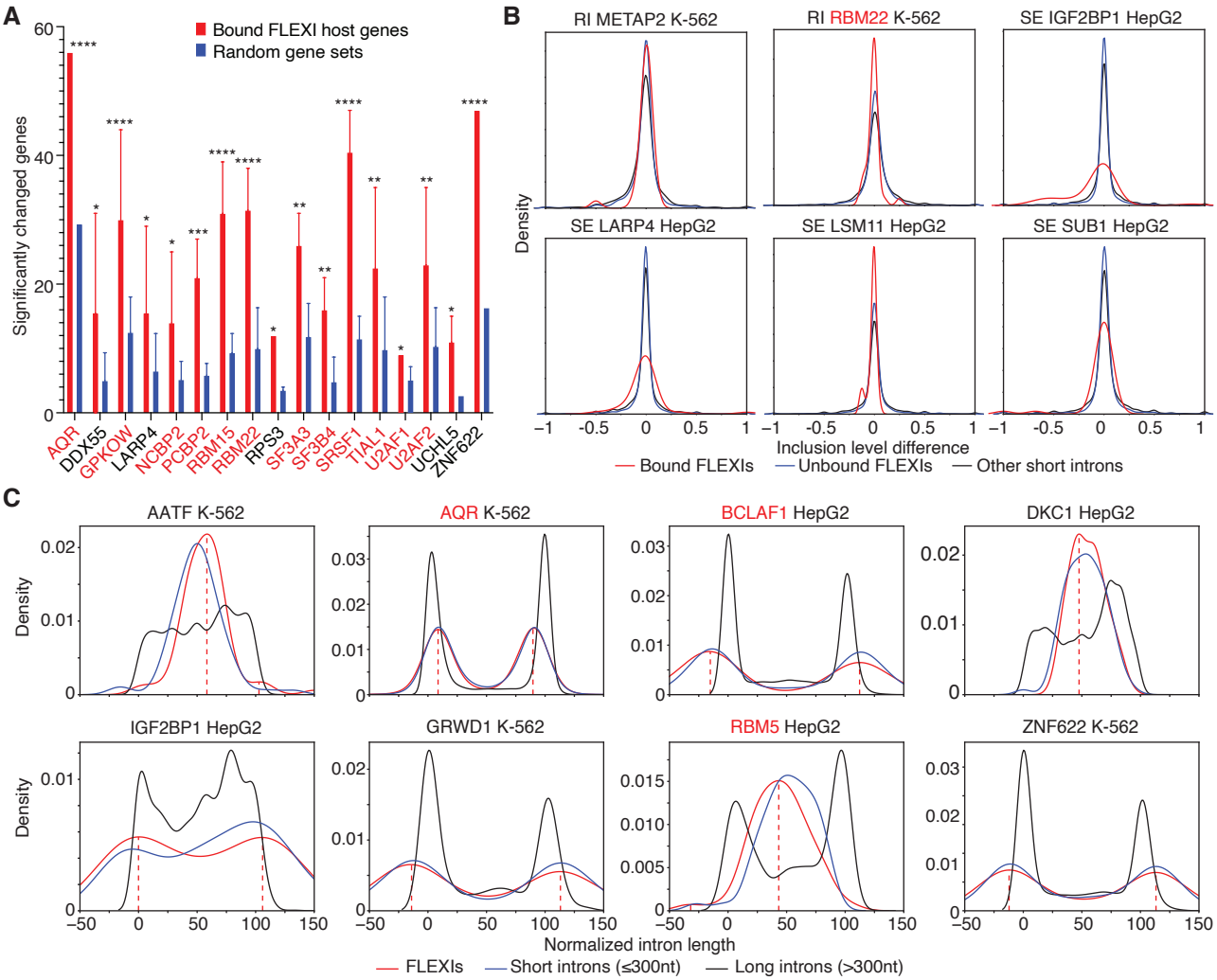

A
Bound FLEXI host genes
Random gene sets
Significantly changed genes
AQR
DDX55
GPKOW
LARP4
NCBP2
PCBP2
RBM15
RBM22
RPS3
SF3A3
SF3B4
SRSF1
TIAL1
U2AF1
U2AF2
UCHL5
ZNF622
B
RI METAP2 K-562
RI RBM22 K-562
SE IGF2BP1 HepG2
SE LARP4 HepG2
SE LSM11 HepG2
SE SUB1 HepG2
Density
Inclusion level difference
Bound FLEXIs
Unbound FLEXIs
Other short introns
C
AATF K-562
AQR K-562
BCLAF1 HepG2
DKC1 HepG2
IGF2BP1 HepG2
GRWD1 K-562
RBM5 HepG2
ZNF622 K-562
Density
Normalized intron length
FLEXIs
Short introns (≤300nt)
Long introns (>300nt)

Fig. 6

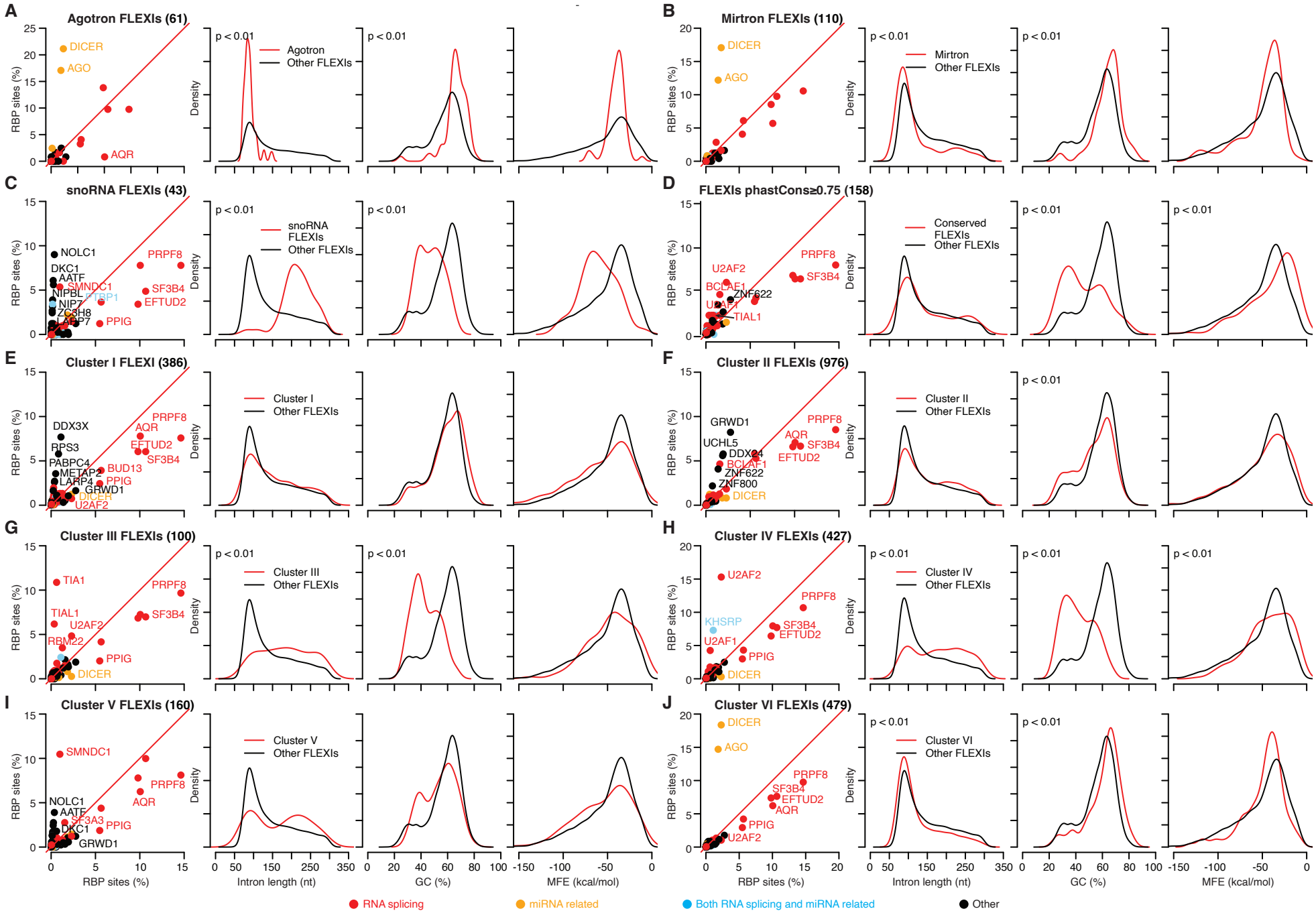

A
Agotron FLEXIs (61)
DICER
AGO
AQR
p < 0.01
Agotron
Other FLEXIs
B
Mirtron FLEXIs (110)
Mirtron
Other FLEXIs
C
snoRNA FLEXIs (43)
NOLC1
DKC1
AATF
SMNDC1
NIPBL
NIP7
ZC3H8
LARP7
PRPF8
SF3B4
EFTUD2
PPIG
snoRNA FLEXIs
Other FLEXIs
D
FLEXIs phastCons≥0.75 (158)
U2AF2
BCLAF1
ZNF622
TIAL1
Conserved FLEXIs
Other FLEXIs
E
Cluster I FLEXI (386)
DDX3X
RPS3
PABPC4
METAP2
LARP4
GRWD1
DICER
U2AF2
BUD13
Cluster I
Other FLEXIs
F
Cluster II FLEXIs (976)
GRWD1
UCHL5
DDX24
ZNF800
Cluster II
Other FLEXIs
G
Cluster III FLEXIs (100)
TIA1
TIAL1
RBM22
Cluster III
Other FLEXIs
H
Cluster IV FLEXIs (427)
KHSRP
U2AF1
Cluster IV
Other FLEXIs
I
Cluster V FLEXIs (160)
SF3A3
Cluster V
Other FLEXIs
J
Cluster VI FLEXIs (479)
Cluster VI
Other FLEXIs
RBP sites (%)
Density
Intron length (nt)
GC (%)
MFE (kcal/mol)
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other

Fig. 7

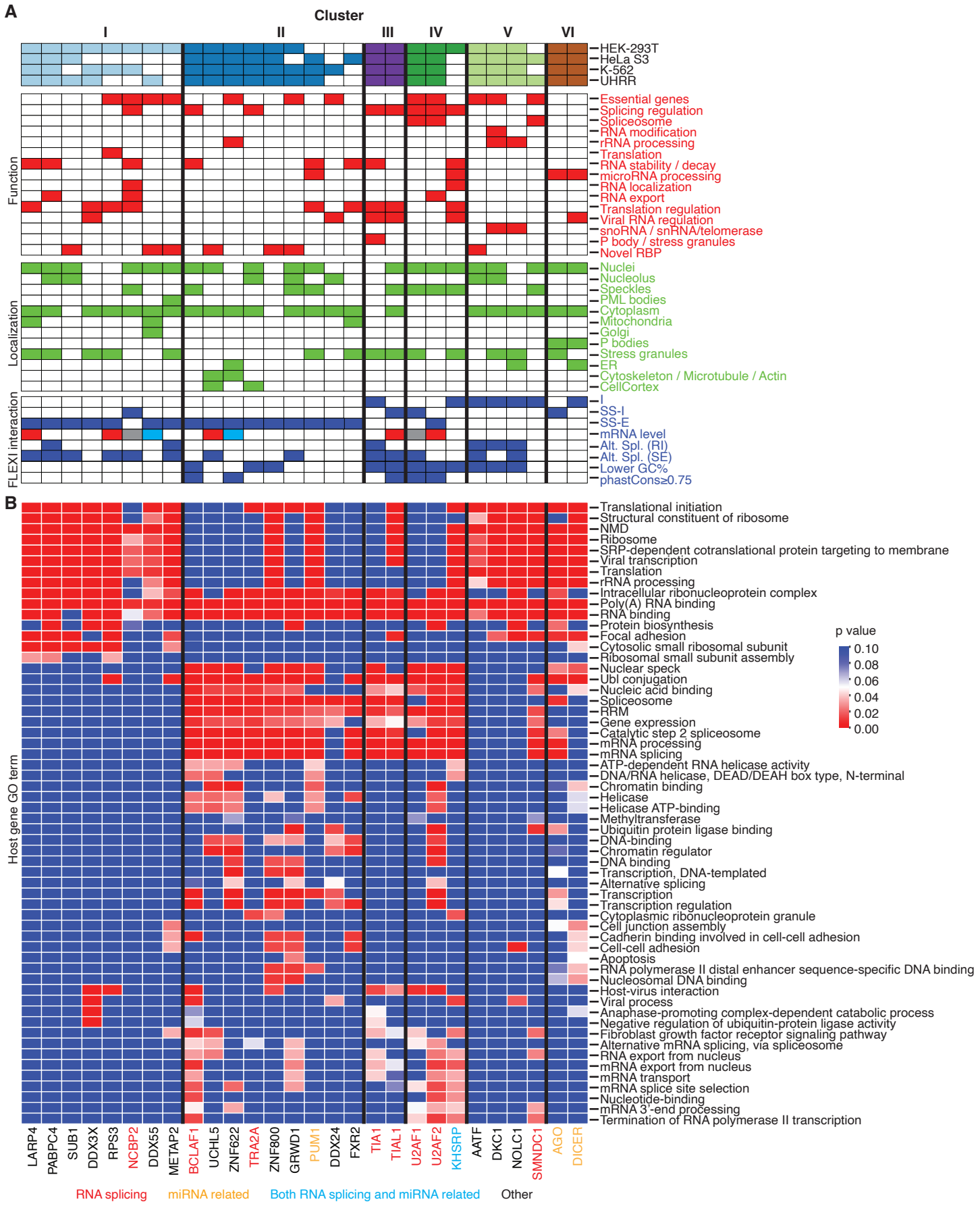

A
Cluster
I
II
III
IV
V
VI
HEK-293T
HeLa S3
K-562
UHRR
Function
Essential genes
Splicing regulation
Spliceosome
RNA modification
rRNA processing
Translation
RNA stability / decay
microRNA processing
RNA localization
RNA export
Translation regulation
Viral RNA regulation
snoRNA / snRNA/telomerase
P body / stress granules
Novel RBP
Localization
Nuclei
Nucleolus
Speckles
PML bodies
Cytoplasm
Mitochondria
Golgi
P bodies
Stress granules
ER
Cytoskeleton / Microtubule / Actin
CellCortex
FLEXI interaction
I
SS-I
SS-E
mRNA level
Alt. Spl. (RI)
Alt. Spl. (SE)
Lower GC%
phastCons≥0.75
B
Host gene GO term
Translational initiation
Structural constituent of ribosome
NMD
Ribosome
SRP-dependent cotranslational protein targeting to membrane
Viral transcription
Translation
rRNA processing
Intracellular ribonucleoprotein complex
Poly(A) RNA binding
RNA binding
Protein biosynthesis
Focal adhesion
Cytosolic small ribosomal subunit
Ribosomal small subunit assembly
Nuclear speck
Ubl conjugation
Nucleic acid binding
Spliceosome
RRM
Gene expression
Catalytic step 2 spliceosome
mRNA processing
mRNA splicing
ATP-dependent RNA helicase activity
DNA/RNA helicase, DEAD/DEAH box type, N-terminal
Chromatin binding
Helicase
Helicase ATP-binding
Methyltransferase
Ubiquitin protein ligase binding
DNA-binding
Chromatin regulator
DNA binding
Transcription, DNA-templated
Alternative splicing
Transcription
Transcription regulation
Cytoplasmic ribonucleoprotein granule
Cell junction assembly
Cadherin binding involved in cell-cell adhesion
Cell-cell adhesion
Apoptosis
RNA polymerase II distal enhancer sequence-specific DNA binding
Nucleosomal DNA binding
Host-virus interaction
Viral process
Anaphase-promoting complex-dependent catabolic process
Negative regulation of ubiquitin-protein ligase activity
Fibroblast growth factor receptor signaling pathway
Alternative mRNA splicing, via spliceosome
RNA export from nucleus
mRNA export from nucleus
mRNA transport
mRNA splice site selection
Nucleotide-binding
mRNA 3'-end processing
Termination of RNA polymerase II transcription
p value
0.10
0.08
0.06
0.04
0.02
0.00
LARP4
PABPC4
SUB1
DDX3X
RPS3
NCBP2
DDX55
METAP2
BCLAF1
UCHL5
ZNF622
TRA2A
ZNF800
GRWD1
PUM1
DDX24
FXR2
TIA1
TIAL1
U2AF1
U2AF2
KHSRP
AATF
DKC1
NOLC1
SMNDC1
AGO
DICER
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other

Fig. 8

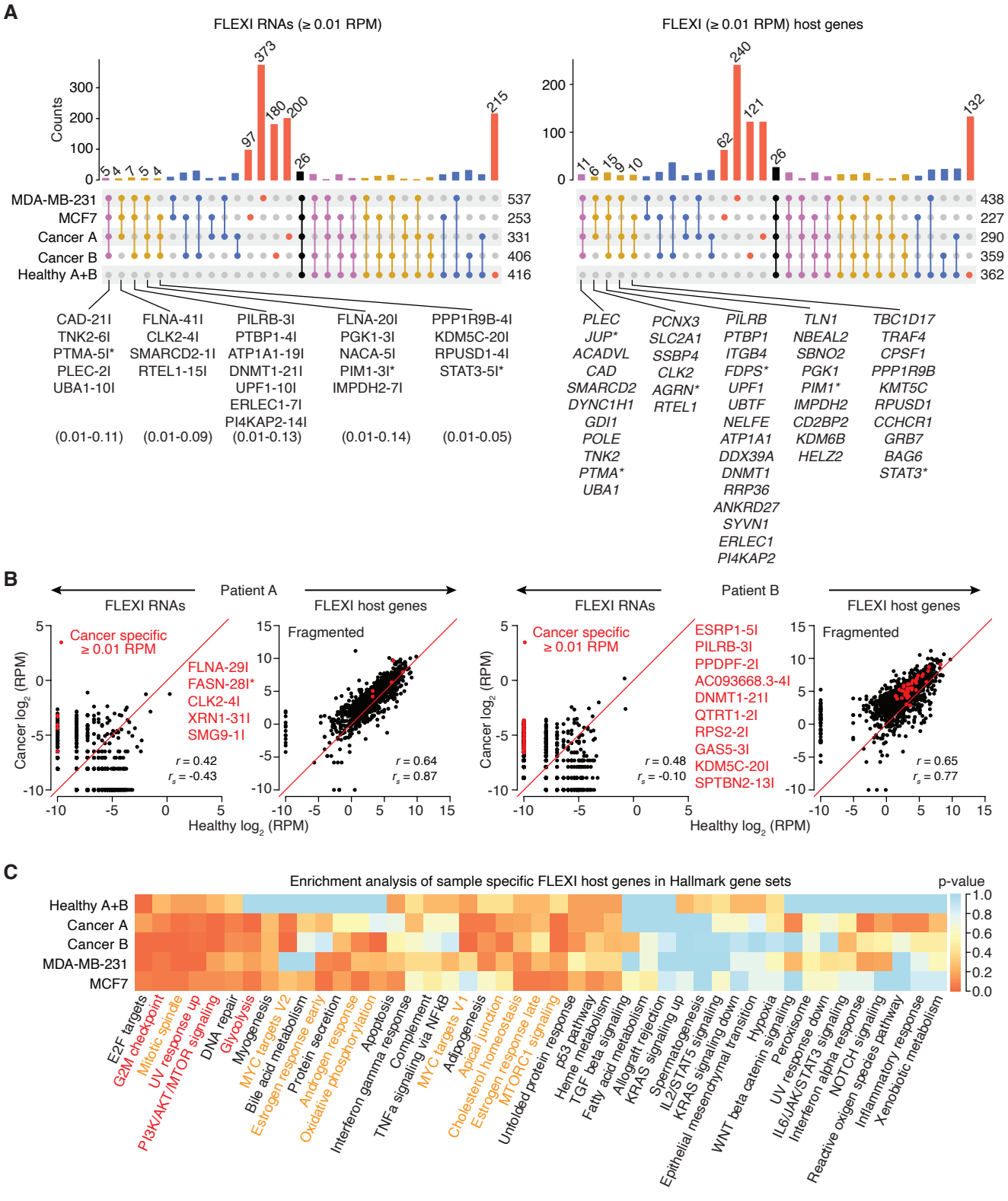

A
FLEXI RNAs (≥ 0.01 RPM)
FLEXI (≥ 0.01 RPM) host genes
Counts
MDA-MB-231
MCF7
Cancer A
Cancer B
Healthy A+B
537
253
331
406
416
438
227
290
359
362
CAD-21I TNK2-6I PTMA-5I* PLEC-2I UBA1-10I
FLNA-41I CLK2-4I SMARCD2-1I RTEL1-15I
PILRB-3I PTBP1-4I ATP1A1-19I DNMT1-21I UPF1-10I ERLEC1-7I PI4KAP2-14I
FLNA-20I PGK1-3I NACA-5I PIM1-3I* IMPDH2-7I
PPP1R9B-4I KDM5C-20I RPUSD1-4I STAT3-5I*
(0.01-0.11) (0.01-0.09) (0.01-0.13) (0.01-0.14) (0.01-0.05)
PLEC JUP* ACADVL CAD SMARCD2 DYNC1H1 GDI1 POLE TNK2 PTMA* UBA1
PCNX3 SLC2A1 SSBP4 CLK2 AGRN* RTEL1
PILRB PTBP1 ITGB4 FDPS* UPF1 UBTF NELFE ATP1A1 DDX39A DNMT1 RRP36 ANKRD27 SYVN1 ERLEC1 PI4KAP2
TLN1 NBEAL2 SBNO2 PGK1 PIM1* IMPDH2 CD2BP2 KDM6B HELZ2
TBC1D17 TRAF4 CPSF1 PPP1R9B KMT5C RPUSD1 CCHCR1 GRB7 BAG6 STAT3*
B
Patient A
FLEXI RNAs
FLEXI host genes
Cancer specific ≥ 0.01 RPM
FLNA-29I FASN-28I* CLK2-4I XRN1-31I SMG9-1I
Fragmented
r = 0.42 r_s = -0.43
r = 0.64 r_s = 0.87
Cancer log2 (RPM)
Healthy log2 (RPM)
Patient B
ESRP1-5I PILRB-3I PPDPF-2I AC093668.3-4I DNMT1-21I QTRT1-2I RPS2-2I GAS5-3I KDM5C-20I SPTBN2-13I
r = 0.48 r_s = -0.10
r = 0.65 r_s = 0.77
C
Enrichment analysis of sample specific FLEXI host genes in Hallmark gene sets
p-value
Healthy A+B
Cancer A
Cancer B
MDA-MB-231
MCF7
E2F targets
G2M checkpoint
Mitotic spindle
UV response up
PI3K/AKT/MTOR signaling
DNA repair
Glycolysis
Myogenesis
MYC targets V2
Estrogen response early
Androgen response
Oxidative phosphorylation
Bile acid metabolism
Protein secretion
Apoptosis
Interferon gamma response
Complement
TNFa signaling via NFKB
MYC targets V1
Adipogenesis
Apical junction
Cholesterol homeostasis
Estrogen response late
MTORC1 signaling
Unfolded protein response
p53 pathway
Heme metabolism
TGF beta signaling
Fatty acid metabolism
Allograft rejection
KRAS signaling up
Spermatogenesis
IL2/STAT5 signaling
KRAS signaling down
Epithelial mesenchymal transition
Hypoxia
WNT beta catenin signaling
Peroxisome
UV response down
IL6/JAK/STAT3 signaling
Interferon alpha response
NOTCH signaling
Reactive oxigen species pathway
Inflammatory response
Xenobiotic metabolism

Fig. S1

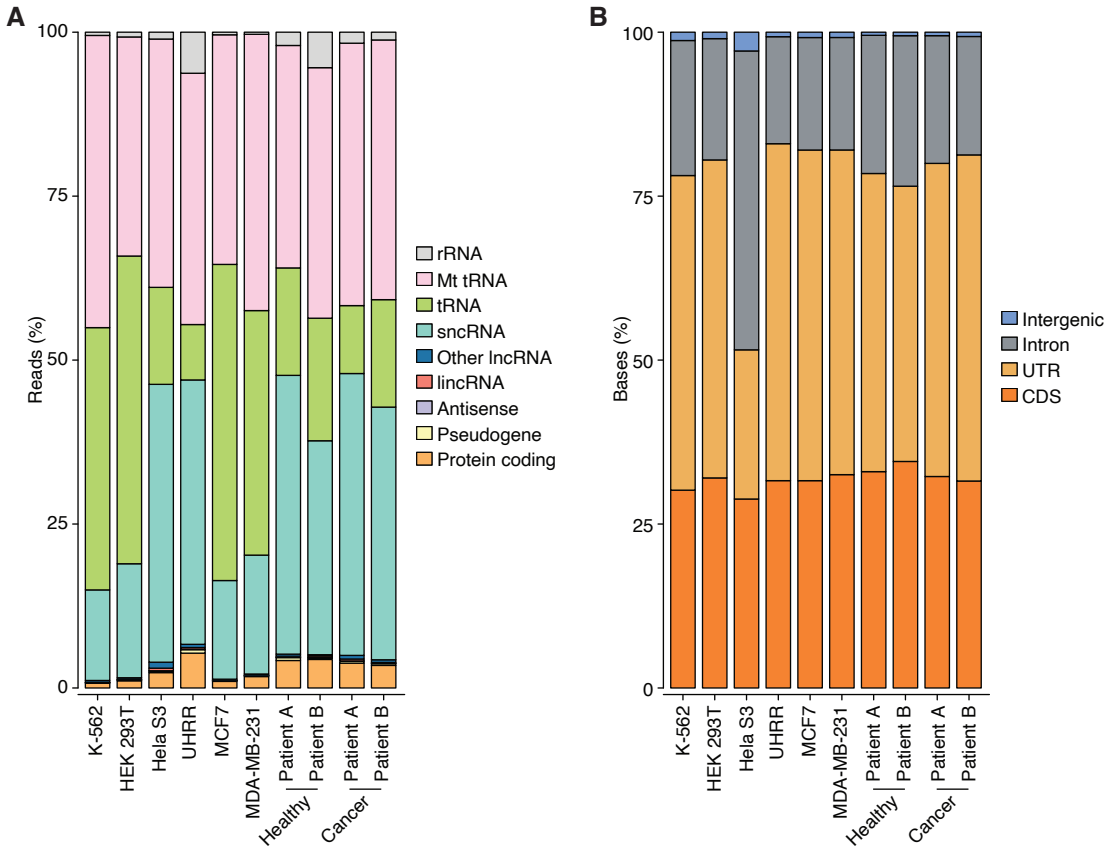

A
100
75
50
25
0
Reads (%)
rRNA
Mt tRNA
tRNA
sncRNA
Other lncRNA
lincRNA
Antisense
Pseudogene
Protein coding
K-562
HEK 293T
Hela S3
UHRR
MCF7
MDA-MB-231
Patient A
Patient B
Patient A
Patient B
Healthy
Cancer
B
100
75
50
25
0
Bases (%)
Intergenic
Intron
UTR
CDS
K-562
HEK 293T
Hela S3
UHRR
MCF7
MDA-MB-231
Patient A
Patient B
Patient A
Patient B
Healthy
Cancer

Fig. S2

snoRA24 (131 nt)
snoRD74 (72 nt)
HEK-293T
HeLa S3
K-562
UHRR
snoRD78 (65 nt)
snoRD104 (70 nt)
RNY1 (113 nt)
RNY5 (84 nt)
VTRNA1-1 (99 nt)
VTRNA2-1 (101 nt)
RNU5B-1 (117 nt)
RNU6ATAC (125 nt)
RN7SK (337 nt)
RN7SL1 (302 nt)
tRNA-GlyCCC (74 nt)
tRNA-ValAAC (76 nt)
m¹A58
CCA
I
m¹A58
CCA

Fig. S3

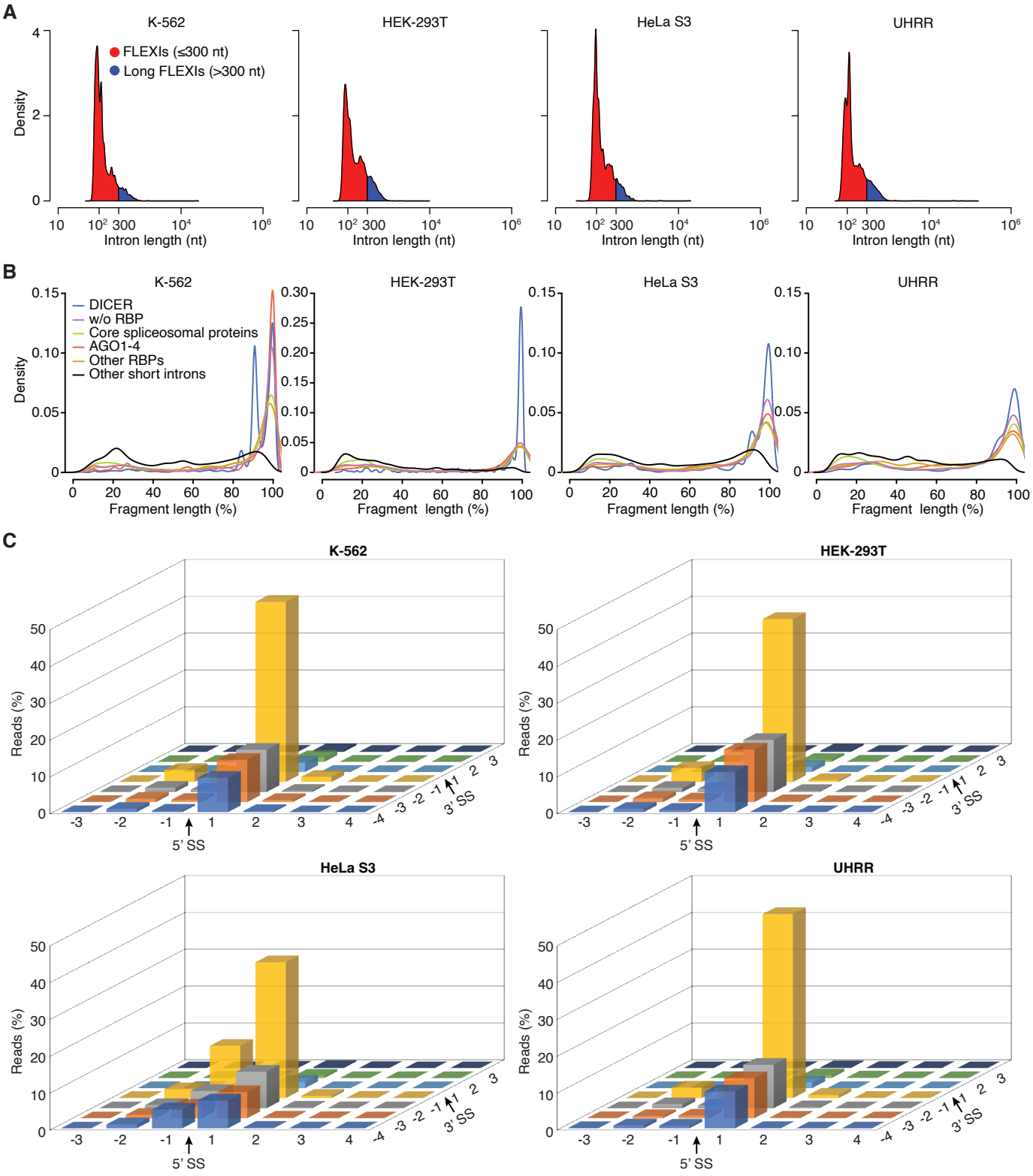

A
K-562
HEK-293T
HeLa S3
UHRR
FLEXIs (≤300 nt)
Long FLEXIs (>300 nt)
Density
Intron length (nt)
B
DICER
w/o RBP
Core spliceosomal proteins
AGO1-4
Other RBPs
Other short introns
Fragment length (%)
C
Reads (%)
5' SS
3' SS

Fig. S4

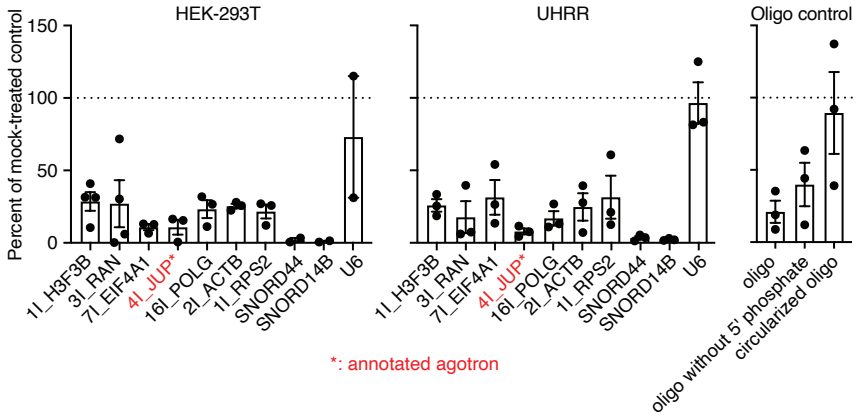

HEK-293T
UHRR
Oligo control
Percent of mock-treated control
150
100
50
0
1I_H3F3B
3I_RAN
7I_EIF4A1
4I_JUP*
16I_POLG
2I_ACTB
1I_RPS2
SNORD44
SNORD14B
U6
oligo
oligo without 5' phosphate
circularized oligo
*: annotated agotron

Fig. S5

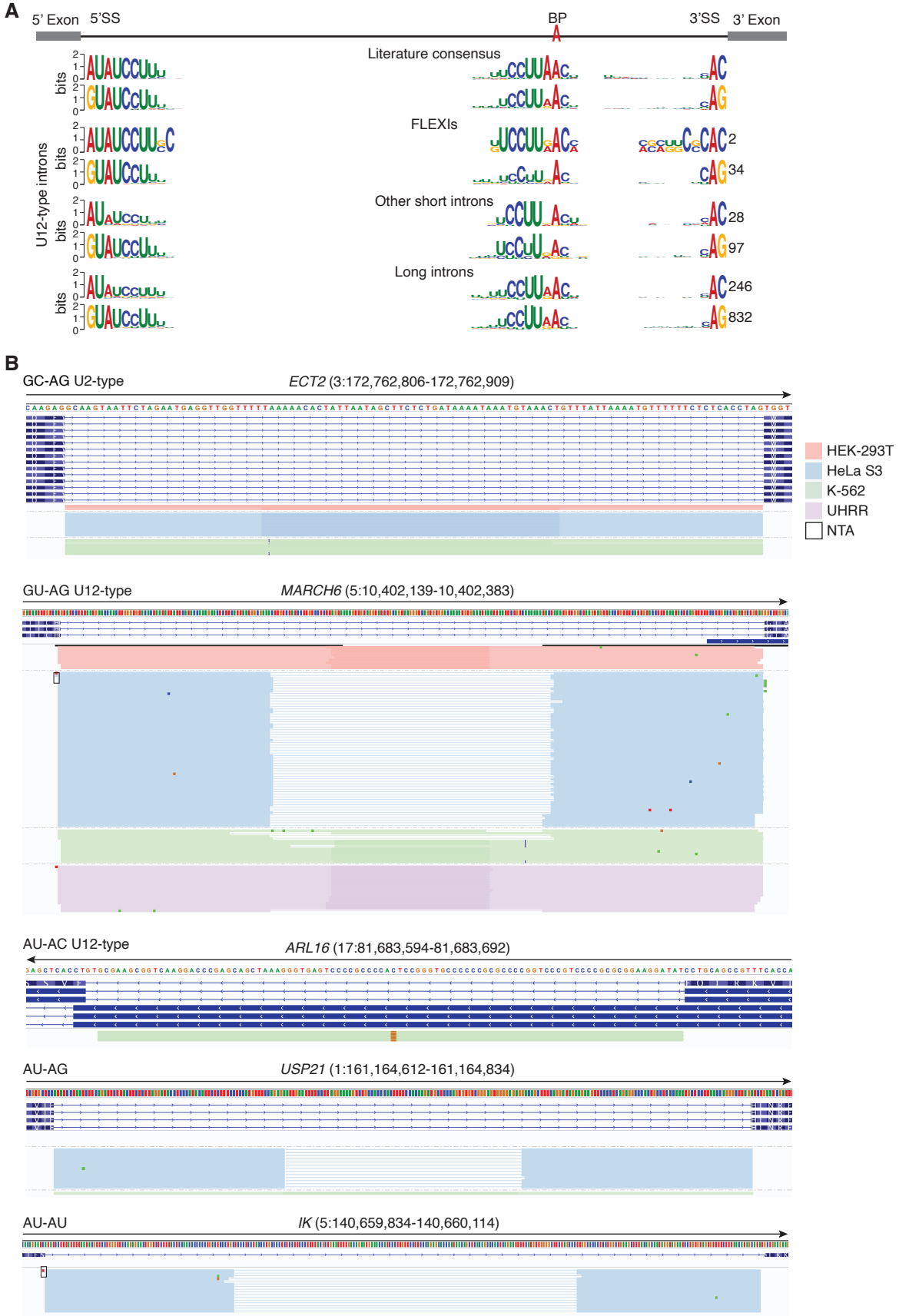
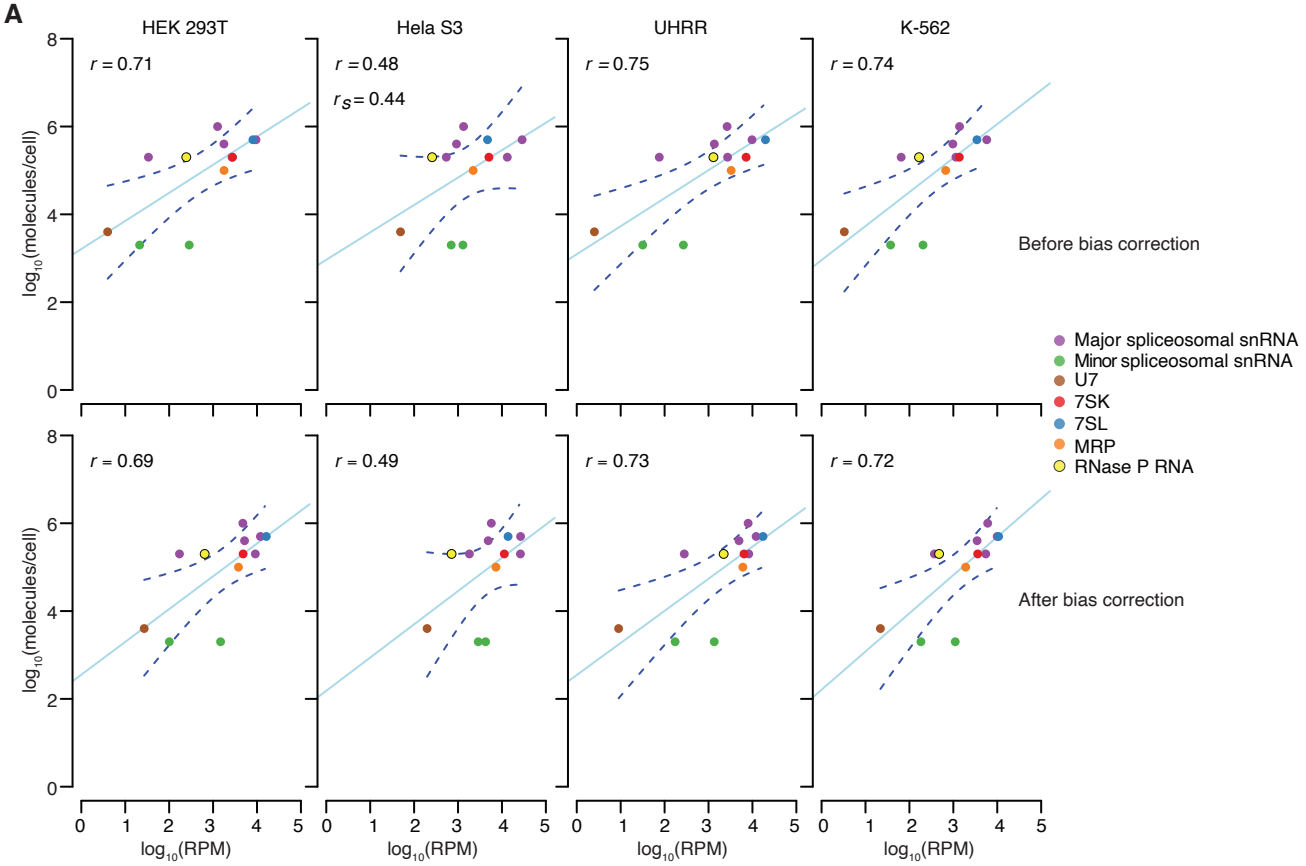

A
5' Exon
5'SS
BP
3'SS
3' Exon
Literature consensus
FLEXIs
Other short introns
Long introns
U12-type introns
bits
2
34
28
97
246
832
B
GC-AG U2-type
ECT2 (3:172,762,806-172,762,909)
HEK-293T
HeLa S3
K-562
UHRR
NTA
GU-AG U12-type
MARCH6 (5:10,402,139-10,402,383)
AU-AC U12-type
ARL16 (17:81,683,594-81,683,692)
AU-AG
USP21 (1:161,164,612-161,164,834)
AU-AU
IK (5:140,659,834-140,660,114)

Fig. S6

A



Copies/cell	Literature value	TGIRT-seq (lrm, before bias correction)			TGIRT-seq (lrm, after bias correction)		
		HEK-293T	HeLa S3	UHRR	HEK-293T	HeLa S3	UHRR
U7	4,000	3,918	10,459	2,225	4,171	8,355	1,736
SNORD14B	10,000	19,026	69,469	15,746	18,833	53,881	14,851
SNORD44	10,000	18,739	57,072	27.924	14,441	53,140	21,991

B

FLEXIs	TGIRT-seq				ddPCR (U7)		ddPCR (SNORD14B)		ddPCR (SNORD44)	
	RPM		Copies/cell		Copies/cell		Copies/cell		Copies/cell	
	Original	Bias corrected	Original	Bias corrected	Original	Bias corrected	Original	Bias corrected	Original	Bias corrected
HEK 293T										
1I_H3F3B	0.03	0.04	180	30	15	16	1	1	1	1
3I_RAN	0.08	0.12	325	74	254	270	21	21	21	16
7I_EIF4A1	0.49	1.36	1,033	448	1,296	1,379	106	105	109	84
4I_JUP*	0.02	0.03	131	23	729	776	60	59	61	47
16I_POLG	0.24	0.26	644	129	25	26	2	2	2	2
2I_ACTB	1.32	2.01	1,933	598	3,393	3,612	278	275	285	220
1I_RPS2	0.47	0.54	1,001	225	–	–	–	–	–	–
Hela S3										
1I_H3F3B	–	–	–	–	73	58	9	7	6	6
3I_RAN	0.09	0.11	203	29	401	320	50	39	33	31
7I_EIF4A1	0.28	0.41	413	78	1,686	1,347	210	163	140	130
4I_JUP*	4.90	17.53	2,483	1,335	1,403	1,121	175	136	117	109
16I_POLG	0.19	0.38	327	73	–	–	–	–	–	–
2I_ACTB	2.11	3.71	1,468	412	907	725	113	88	75	70
1I_RPS2	0.21	0.31	346	62	4,310	3,443	538	417	358	334
UHRR										
1I_H3F3B	0.05	0.05	185	38	180	140	76	71	12	9
3I_RAN	0.06	0.03	195	26	186	145	78	74	12	10
7I_EIF4A1	0.29	0.41	565	181	230	180	97	92	15	12
4I_JUP*	0.03	0.06	127	43	996	777	420	396	67	53
16I_POLG	0.25	0.24	504	122	–	–	–	–	–	–
2I_ACTB	6.90	9.88	4,232	1,863	1,768	1,379	746	704	118	93
1I_RPS2	0.18	0.17	411	94	3,833	2,990	1,618	1,526	257	202

*: annotated agotron

Fig. S7

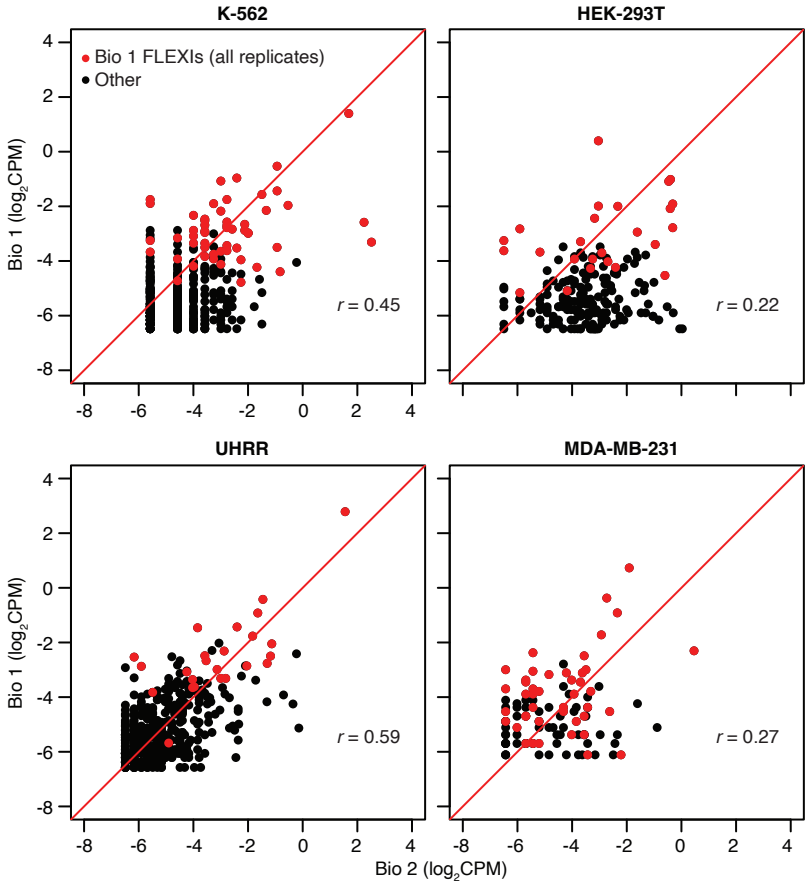

K-562
HEK-293T
UHRR
MDA-MB-231
Bio 1 FLEXIs (all replicates)
Other
r = 0.45
r = 0.22
r = 0.59
r = 0.27
Bio 1 (log2CPM)
Bio 2 (log2CPM)
4
2
0
-2
-4
-6
-8

Fig. S8

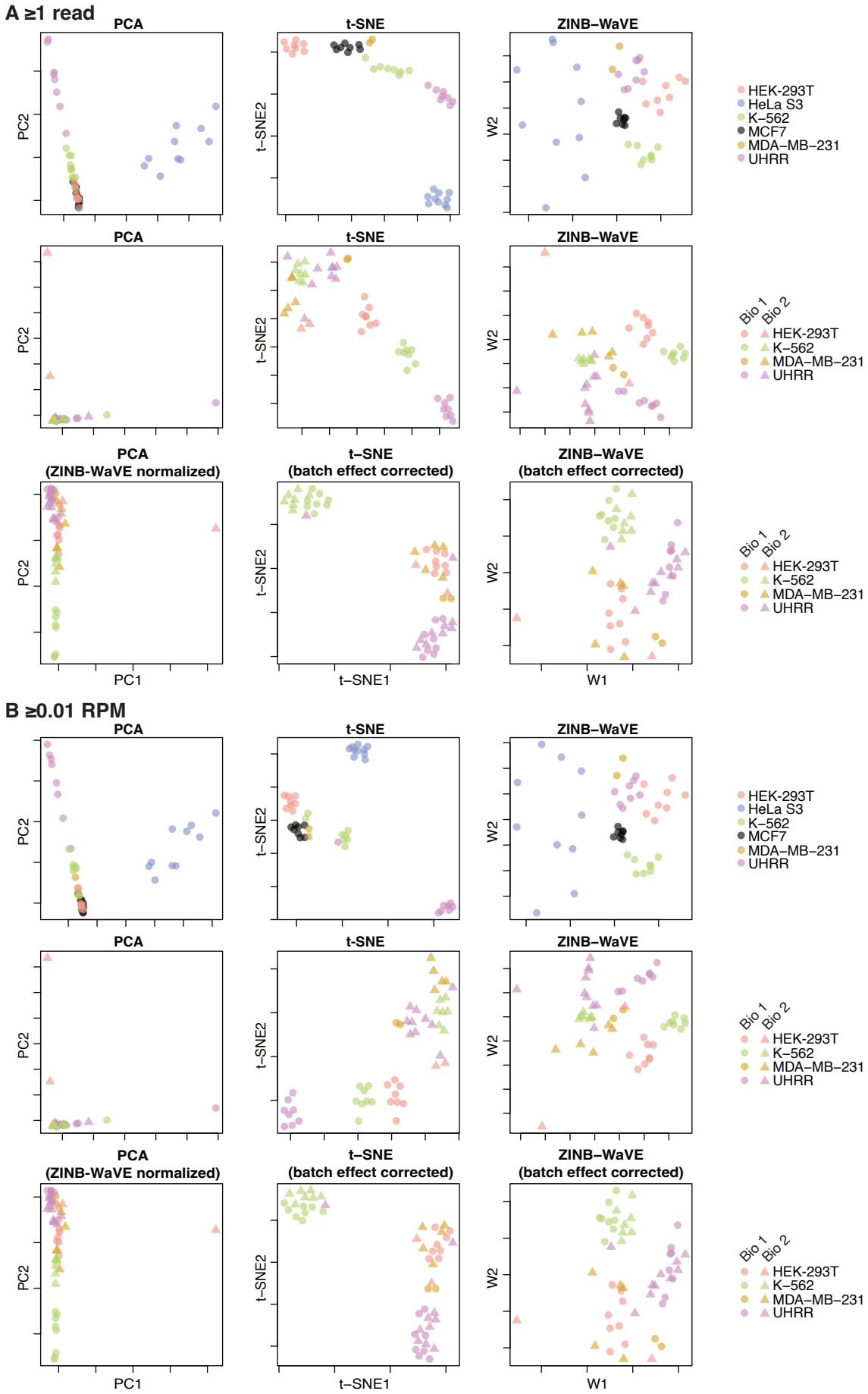

A ≥1 read
PCA
t-SNE
ZINB–WaVE
PC2
t–SNE2
W2
HEK-293T
HeLa S3
K–562
MCF7
MDA–MB–231
UHRR
Bio 1
Bio 2
PCA
(ZINB-WaVE normalized)
t–SNE
(batch effect corrected)
ZINB–WaVE
(batch effect corrected)
PC1
t–SNE1
W1
B ≥0.01 RPM

Fig. S9

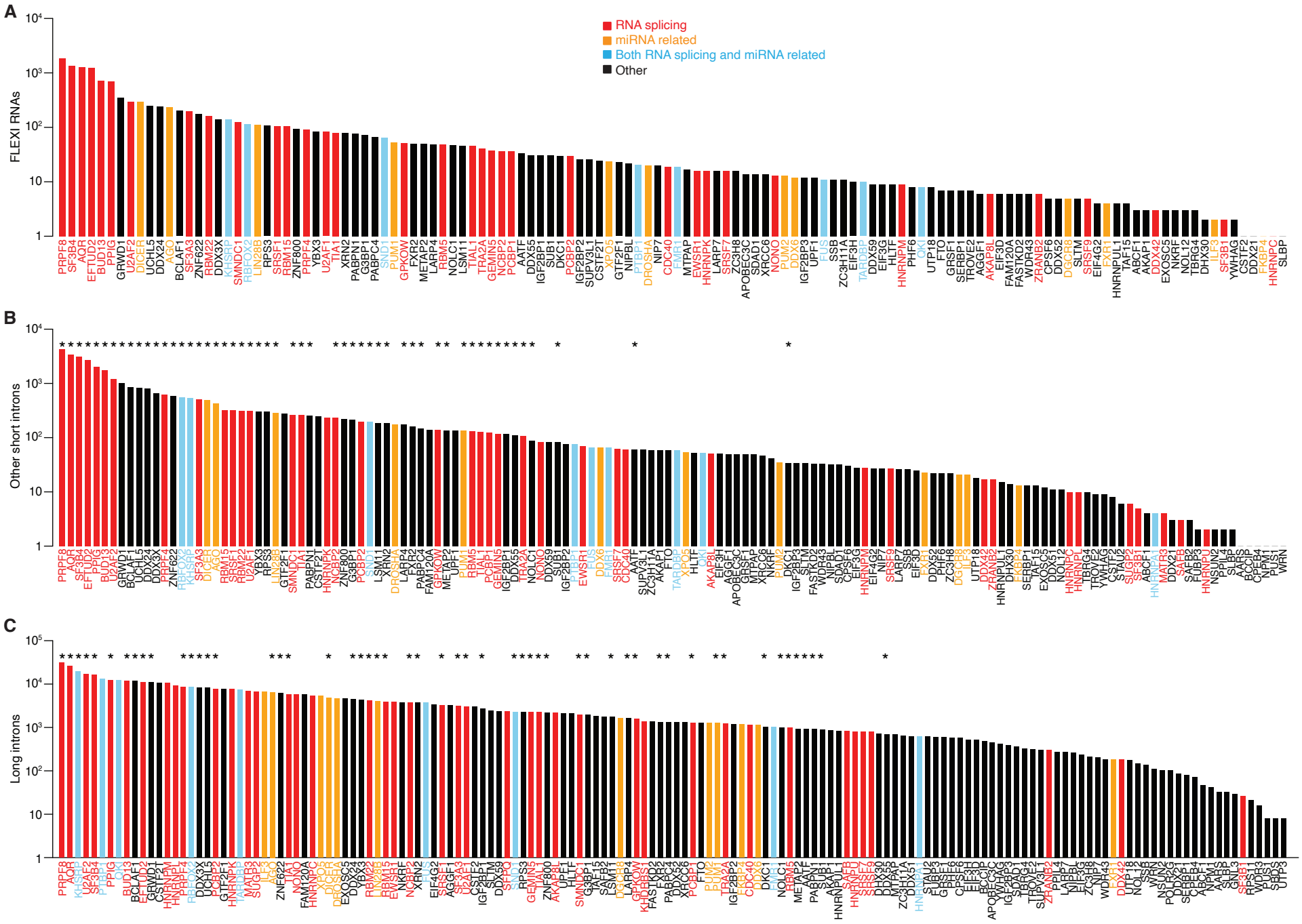

A
FLEXI RNAs
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other
B
Other short introns
C
Long introns

Fig. S10

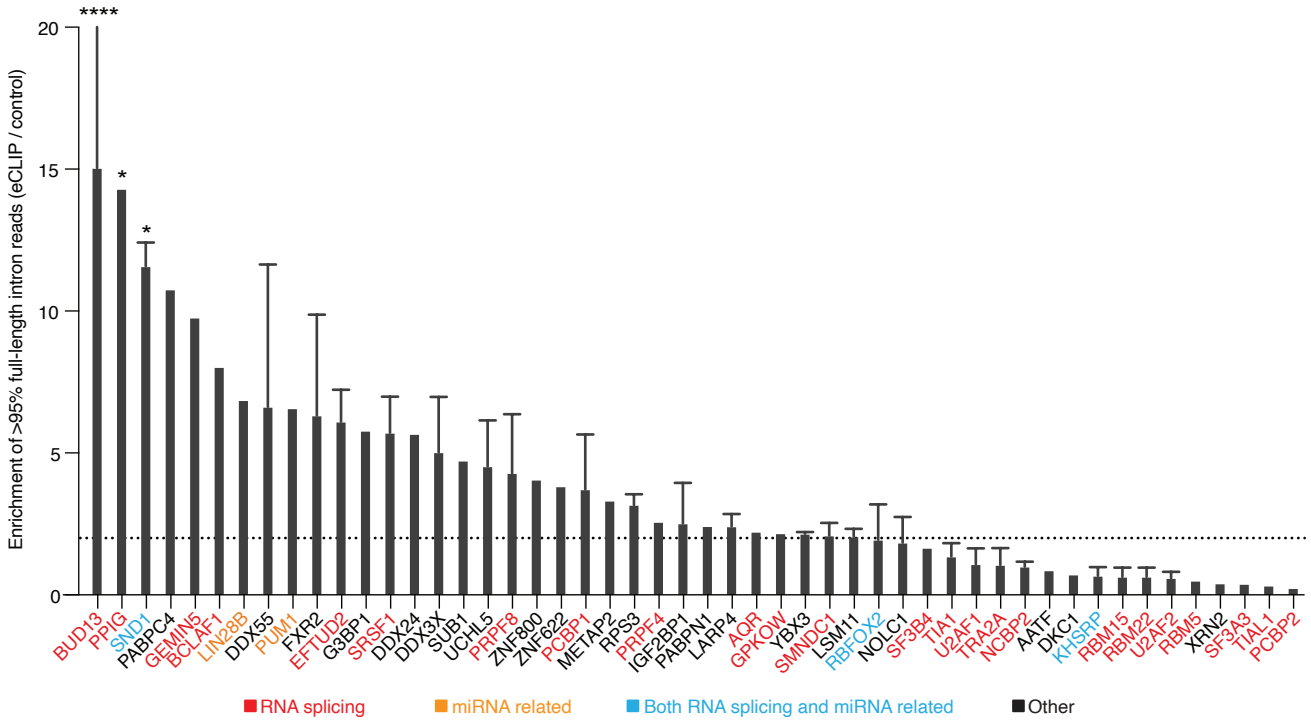

Enrichment of >95% full-length intron reads (eCLIP / control)
20
15
10
5
0
****
*
*
BUD13
PPIG
SND1
PABPC4
GEMIN5
BCLAF1
LIN28B
DDX55
PUM1
FXR2
EFTUD2
G3BP1
SRSF1
DDX24
DDX3X
SUB1
UCHL5
PRPF8
ZNF800
ZNF622
PCBP1
METAP2
RPS3
PRPF4
IGF2BP1
PABPN1
LARP4
AQR
GPKOW
YBX3
SMNDC1
LSM11
RBFOX2
NOLC1
SF3B4
TIA1
U2AF1
TRA2A
NCBP2
AATF
DKC1
KHSRP
RBM15
RBM22
U2AF2
RBM5
XRN2
SF3A3
TIAL1
PCBP2
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other

Fig. S11

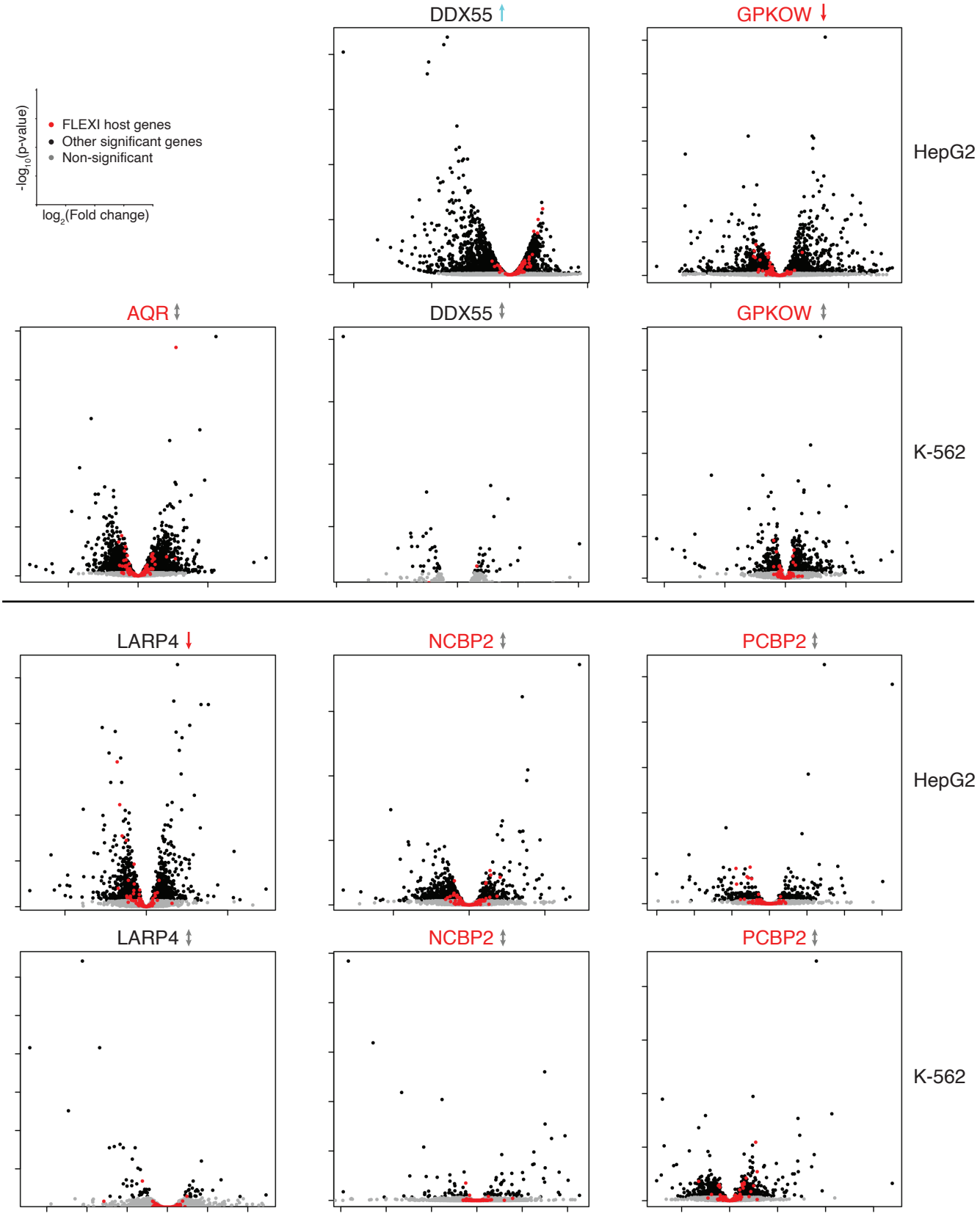

DDX55
GPKOW
HepG2
FLEXI host genes
Other significant genes
Non-significant
-log10(p-value)
log2(Fold change)
AQR
DDX55
GPKOW
K-562
LARP4
NCBP2
PCBP2
HepG2
LARP4
NCBP2
PCBP2
K-562

Fig. S11

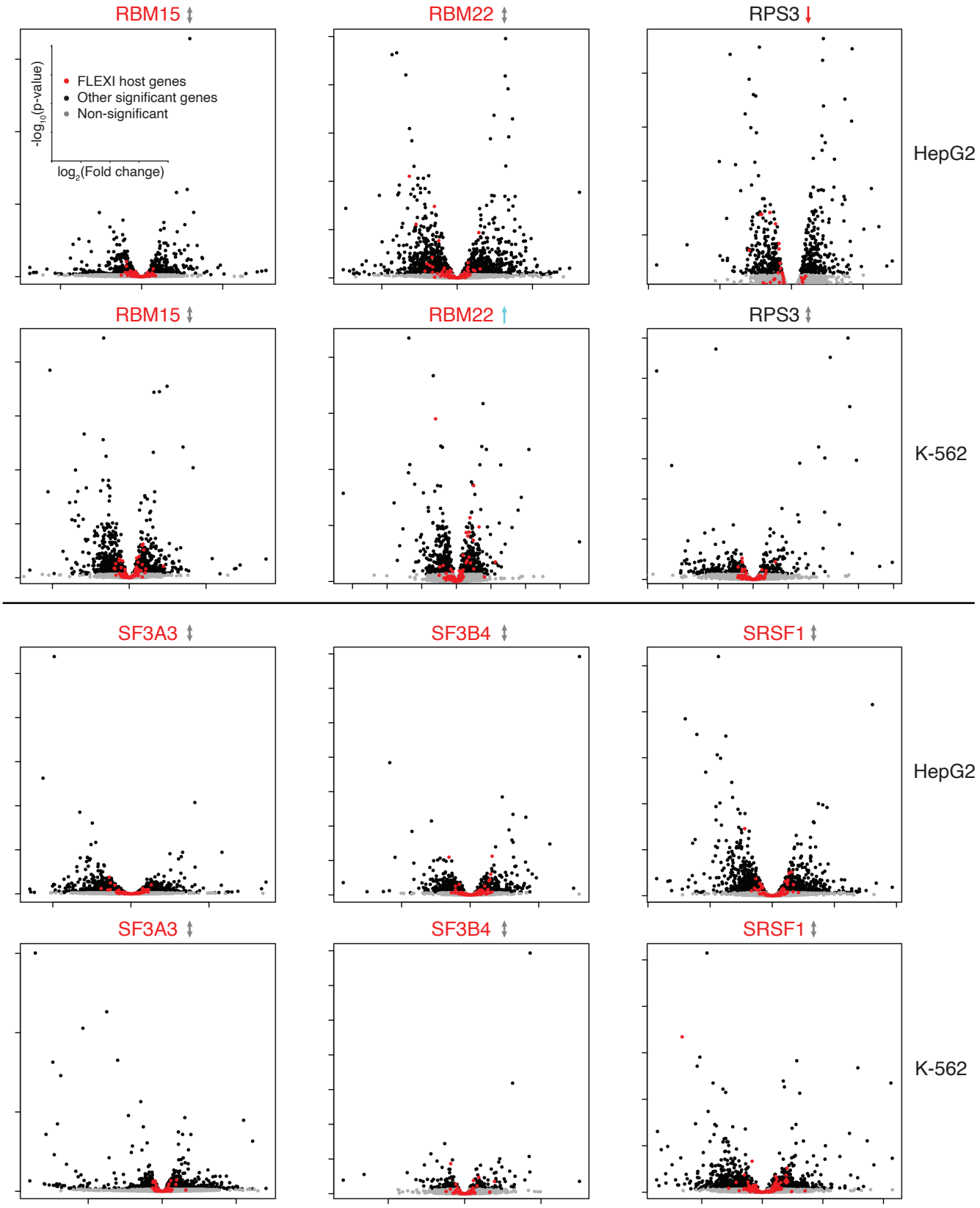

RBM15
RBM22
RPS3
FLEXI host genes
Other significant genes
Non-significant
-log10(p-value)
log2(Fold change)
HepG2
RBM15
RBM22
RPS3
K-562
SF3A3
SF3B4
SRSF1
HepG2
SF3A3
SF3B4
SRSF1
K-562

Fig. S11

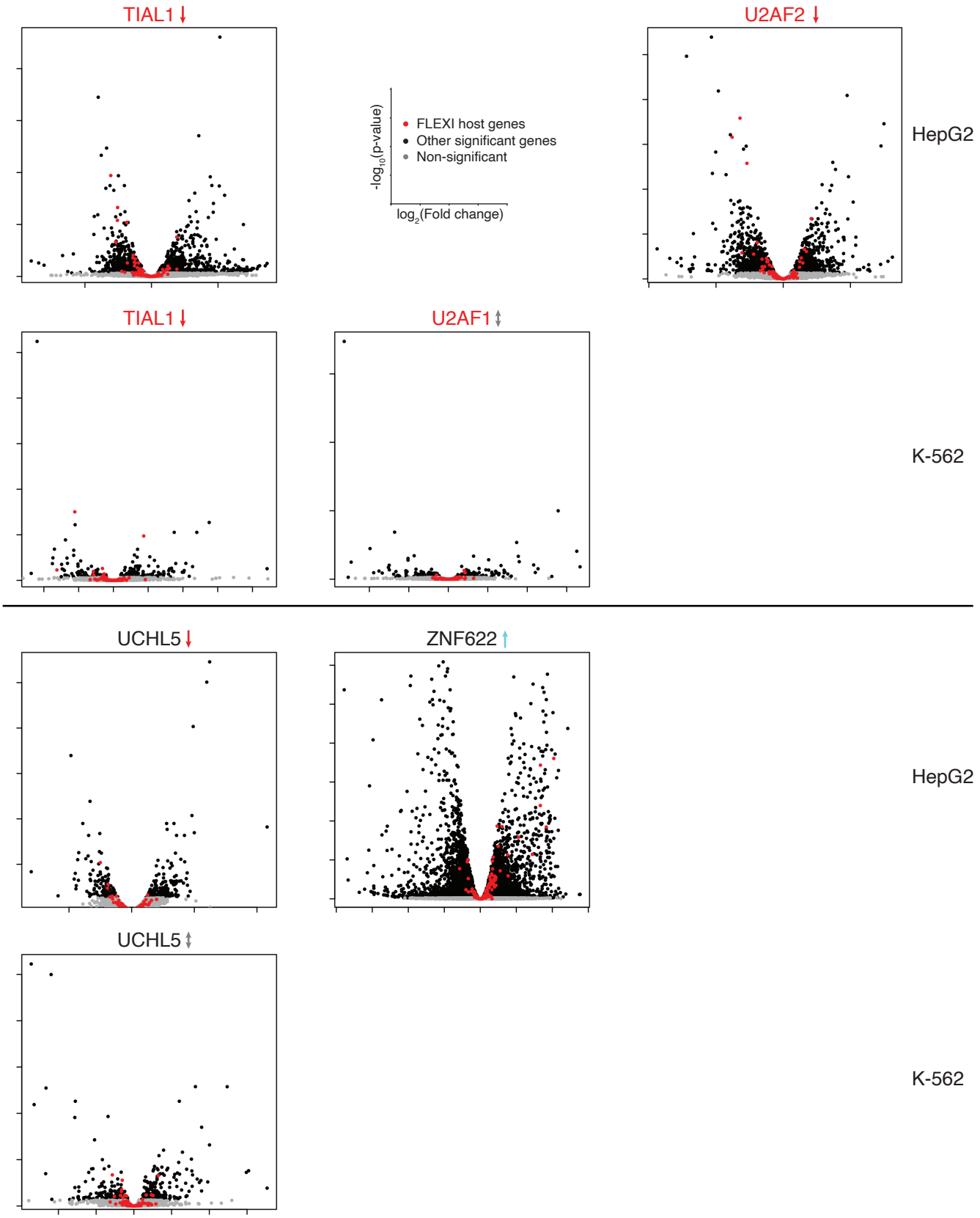

TIAL1
U2AF2
HepG2
-log10(p-value)
log2(Fold change)
FLEXI host genes
Other significant genes
Non-significant
TIAL1
U2AF1
K-562
UCHL5
ZNF622
HepG2
UCHL5
K-562

Fig. S12

A. Retained Intron

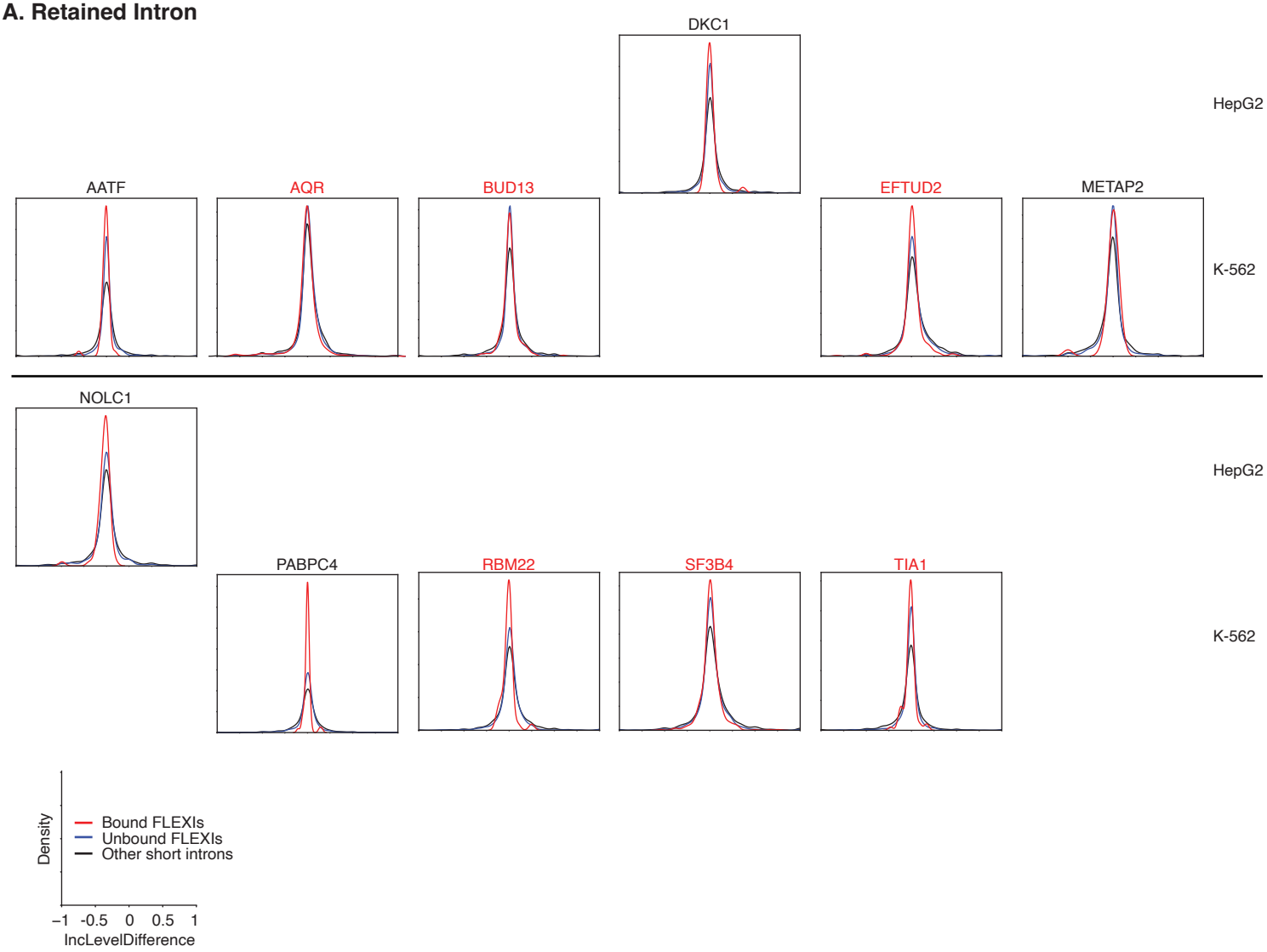


Fig. S12

B. Skipped Exon

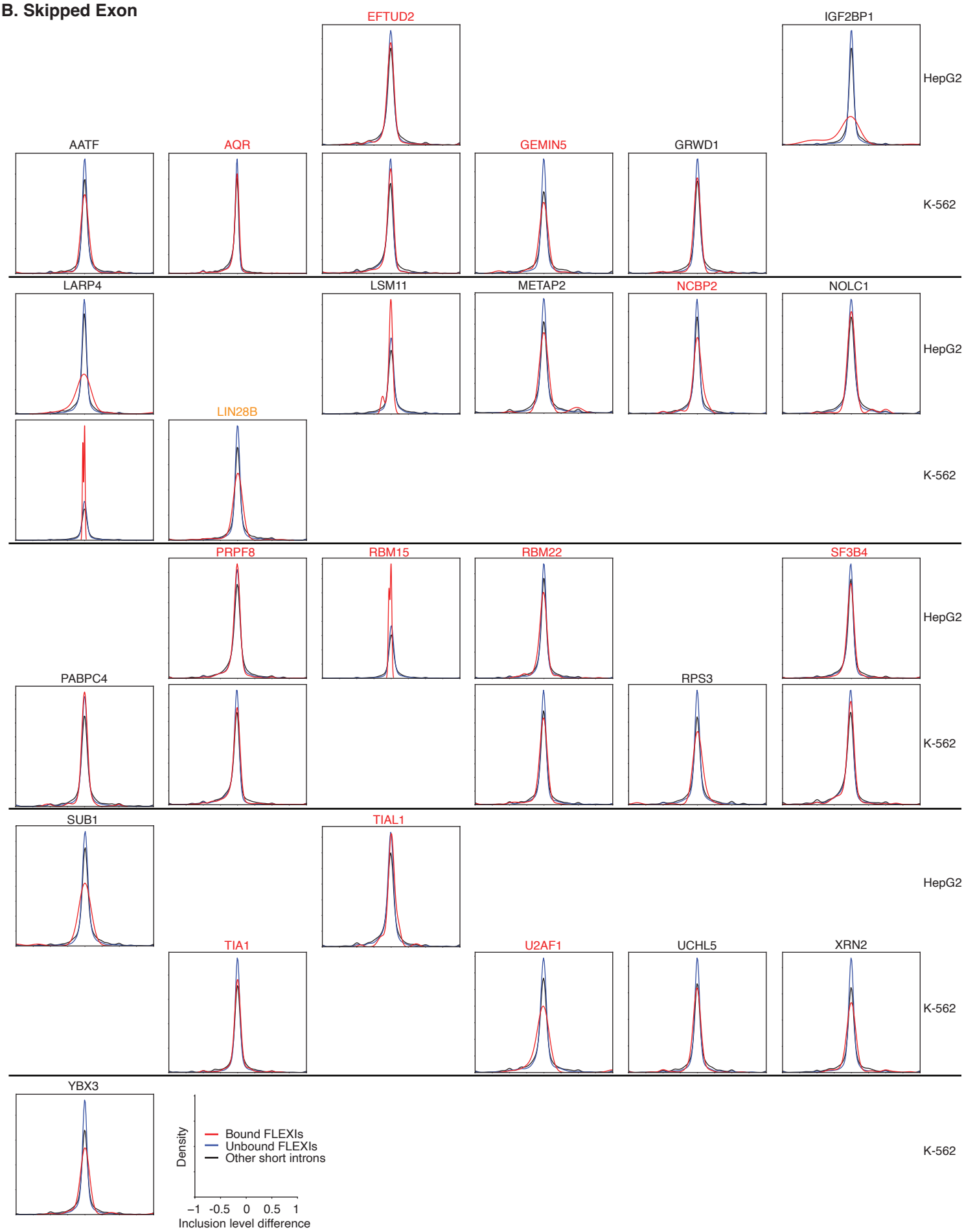


Fig. S13

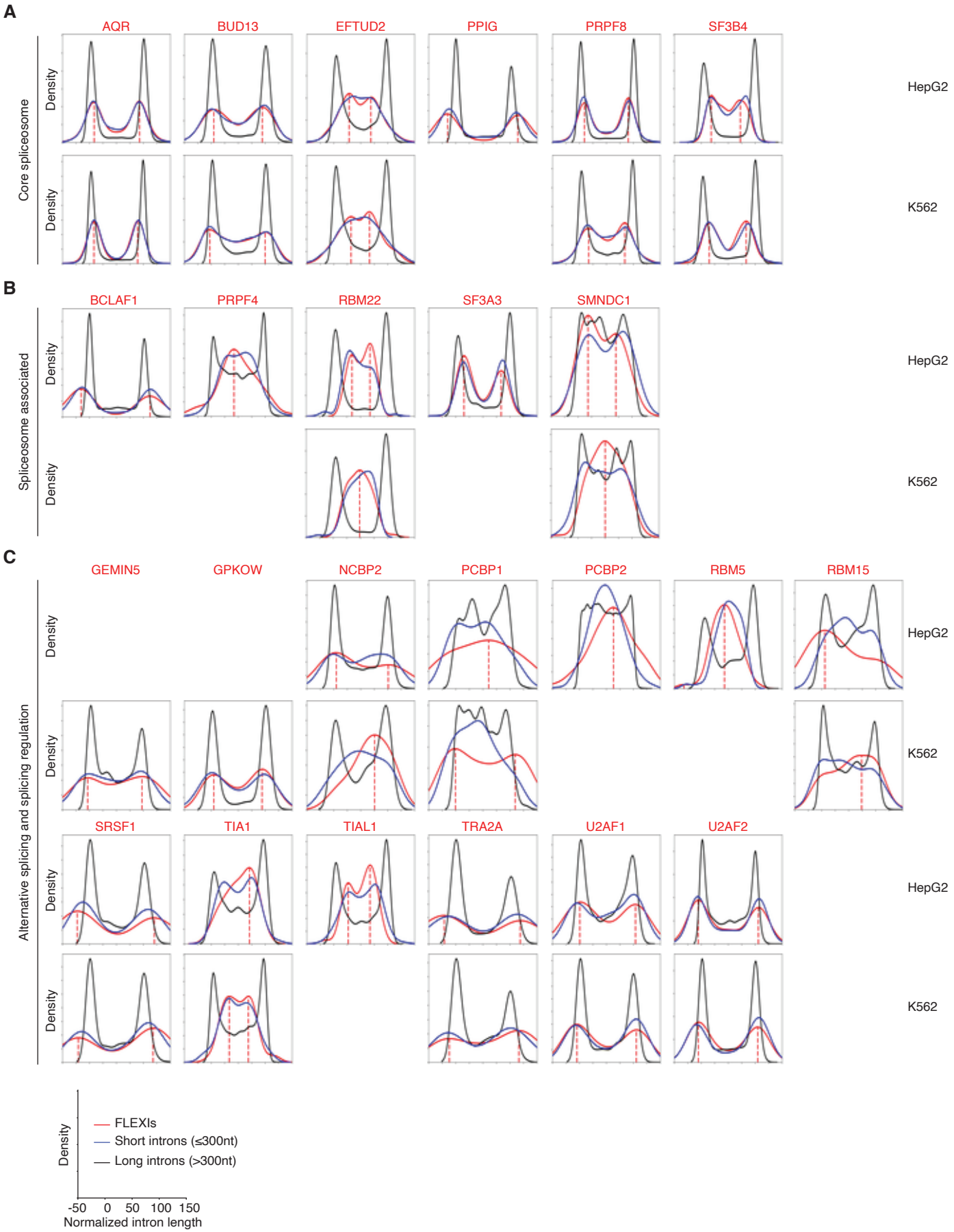

A
AQR
BUD13
EFTUD2
PPIG
PRPF8
SF3B4
Core spliceosome
Density
HepG2
K562
B
BCLAF1
PRPF4
RBM22
SF3A3
SMNDC1
Spliceosome associated
C
GEMIN5
GPKOW
NCBP2
PCBP1
PCBP2
RBM5
RBM15
SRSF1
TIA1
TIAL1
TRA2A
U2AF1
U2AF2
Alternative splicing and splicing regulation
FLEXIs
Short introns (≤300nt)
Long introns (>300nt)
-50 0 50 100 150
Normalized intron length

Fig. S13

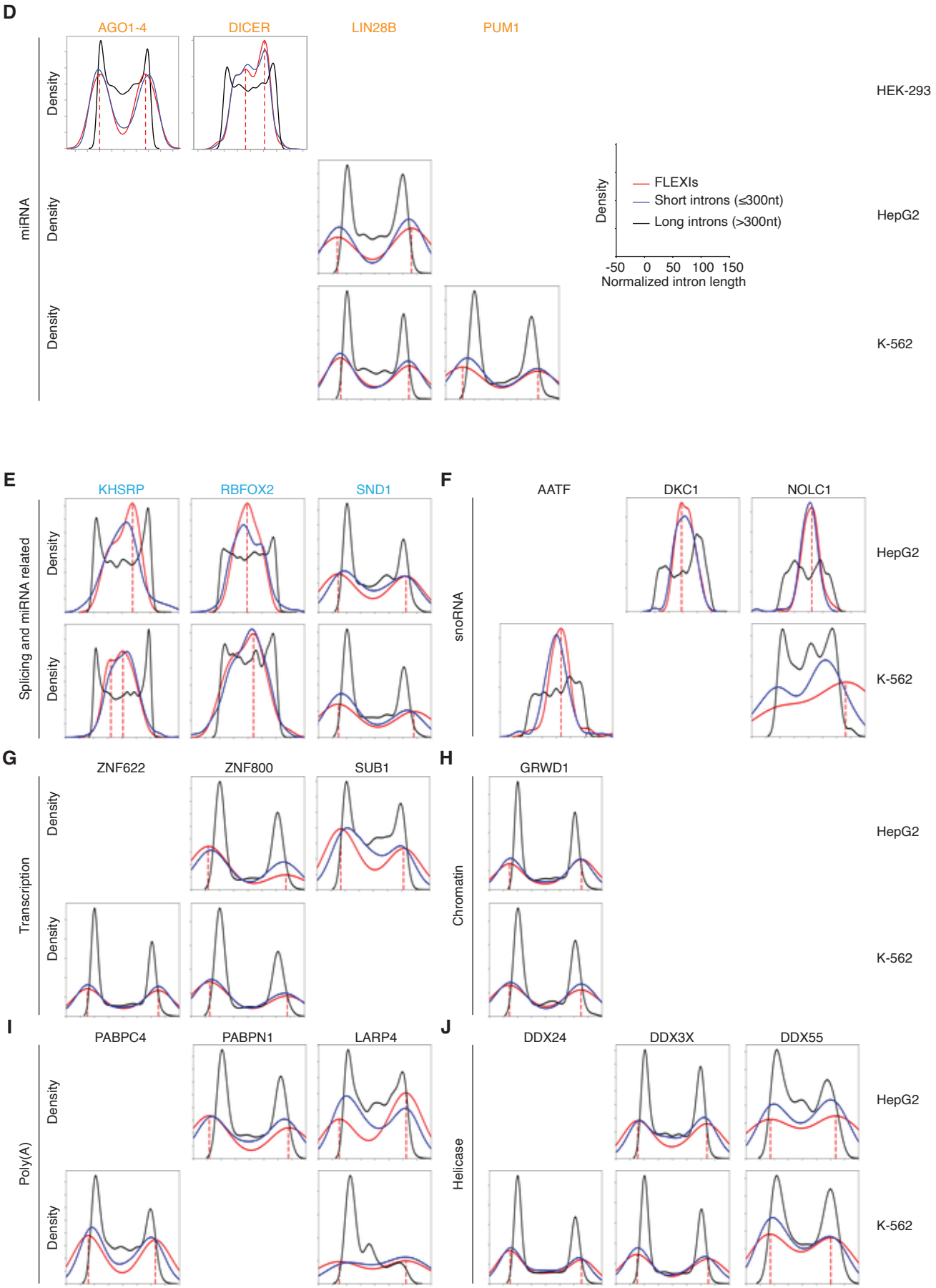

D
AGO1-4
DICER
LIN28B
PUM1
miRNA
Density
HEK-293
HepG2
K-562
FLEXIs
Short introns (≤300nt)
Long introns (>300nt)
-50 0 50 100 150
Normalized intron length
E
KHSRP
RBFOX2
SND1
Splicing and miRNA related
F
AATF
DKC1
NOLC1
snoRNA
G
ZNF622
ZNF800
SUB1
Transcription
H
GRWD1
Chromatin
I
PABPC4
PABPN1
LARP4
Poly(A)
J
DDX24
DDX3X
DDX55
Helicase

Fig. S13

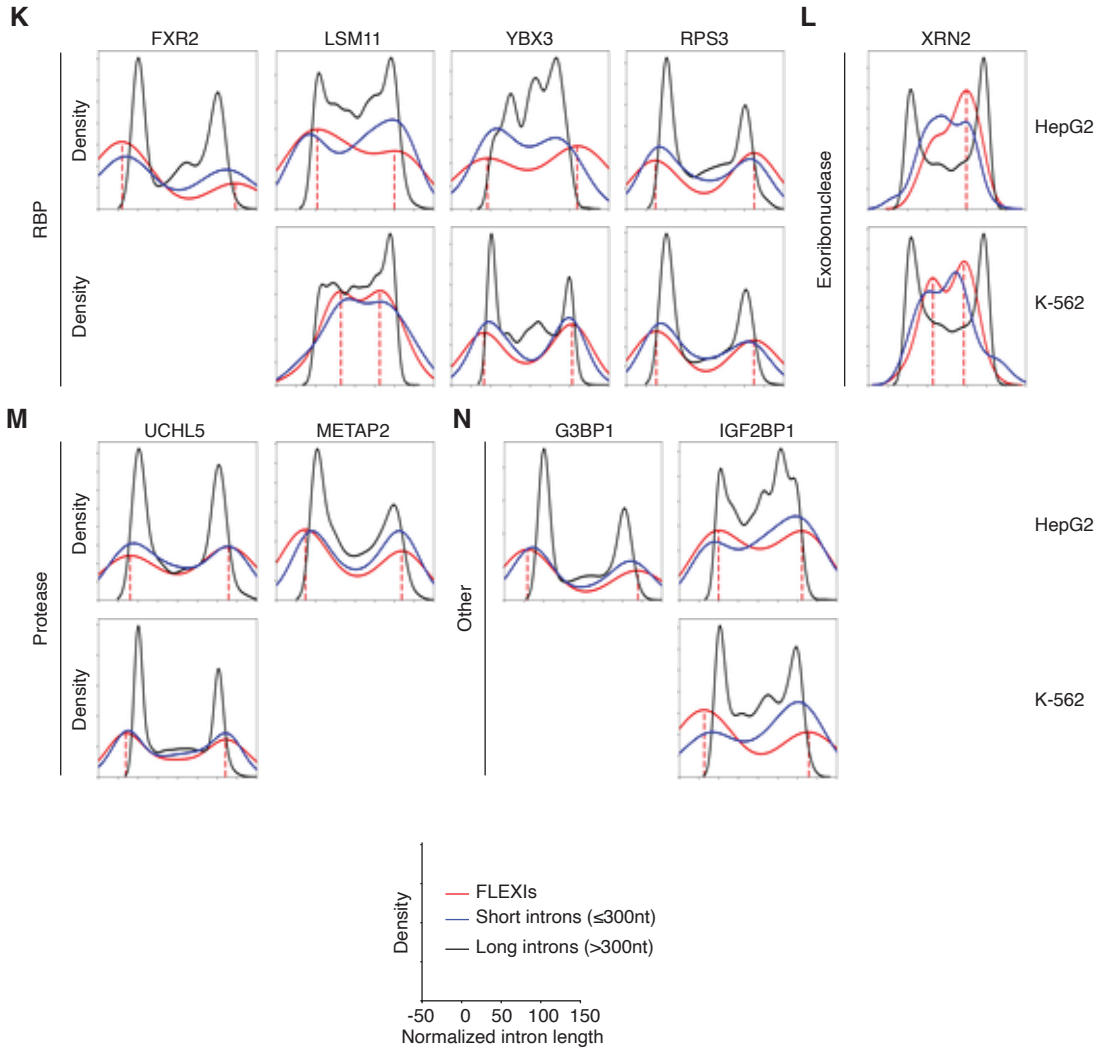

K
FXR2
LSM11
YBX3
RPS3
RBP
Density
Density
L
XRN2
Exoribonuclease
HepG2
K-562
M
UCHL5
METAP2
Protease
Density
Density
N
G3BP1
IGF2BP1
Other
HepG2
K-562
Density
FLEXIs
Short introns (≤300nt)
Long introns (>300nt)
-50
0
50
100
150
Normalized intron length

Fig. S14

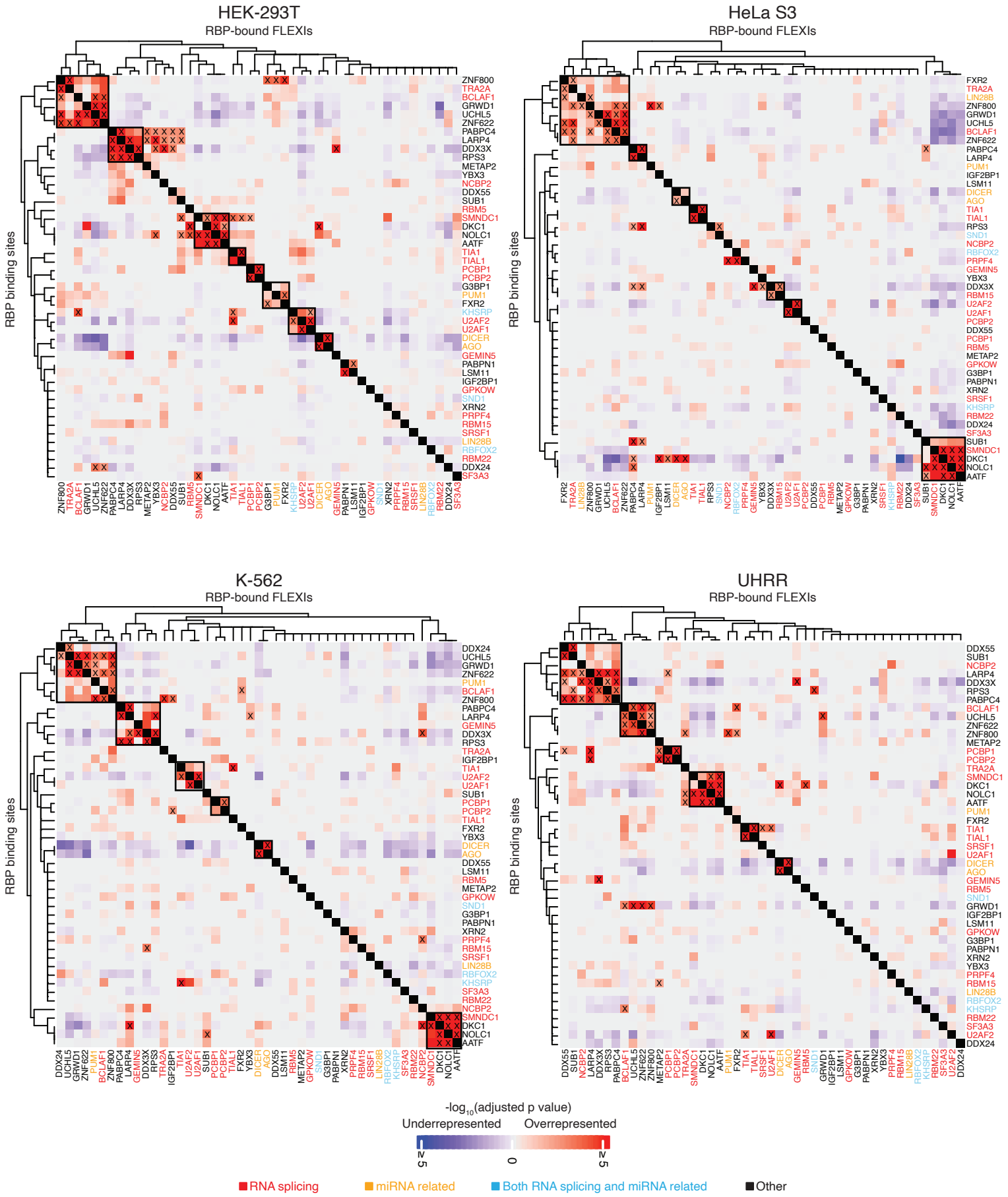

HEK-293T
RBP-bound FLEXIs
HeLa S3
RBP-bound FLEXIs
K-562
RBP-bound FLEXIs
UHRR
RBP-bound FLEXIs
RBP binding sites
-log10(adjusted p value)
Underrepresented
Overrepresented
5
0
5
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other

Fig. S15

A. Cluster I

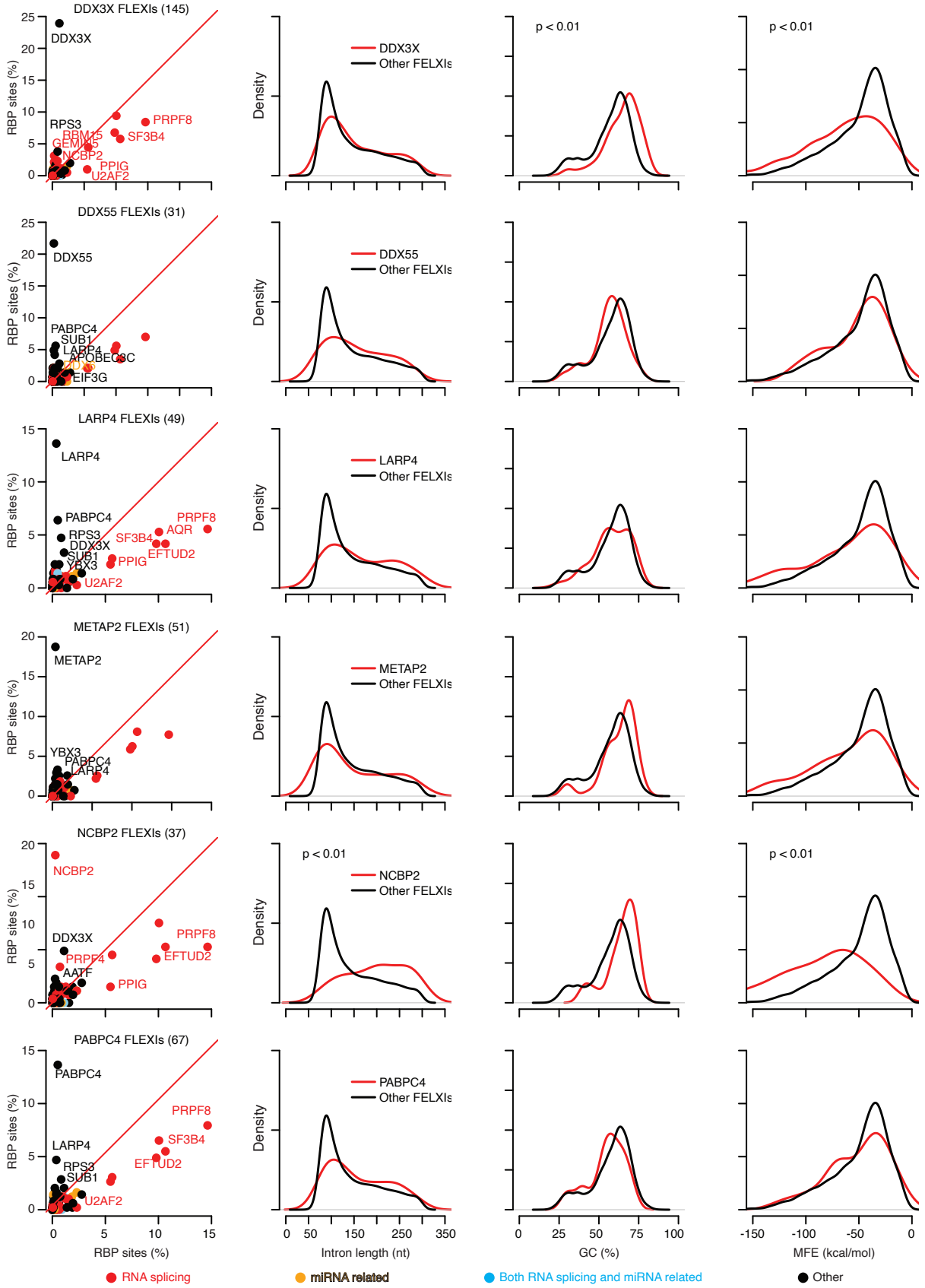
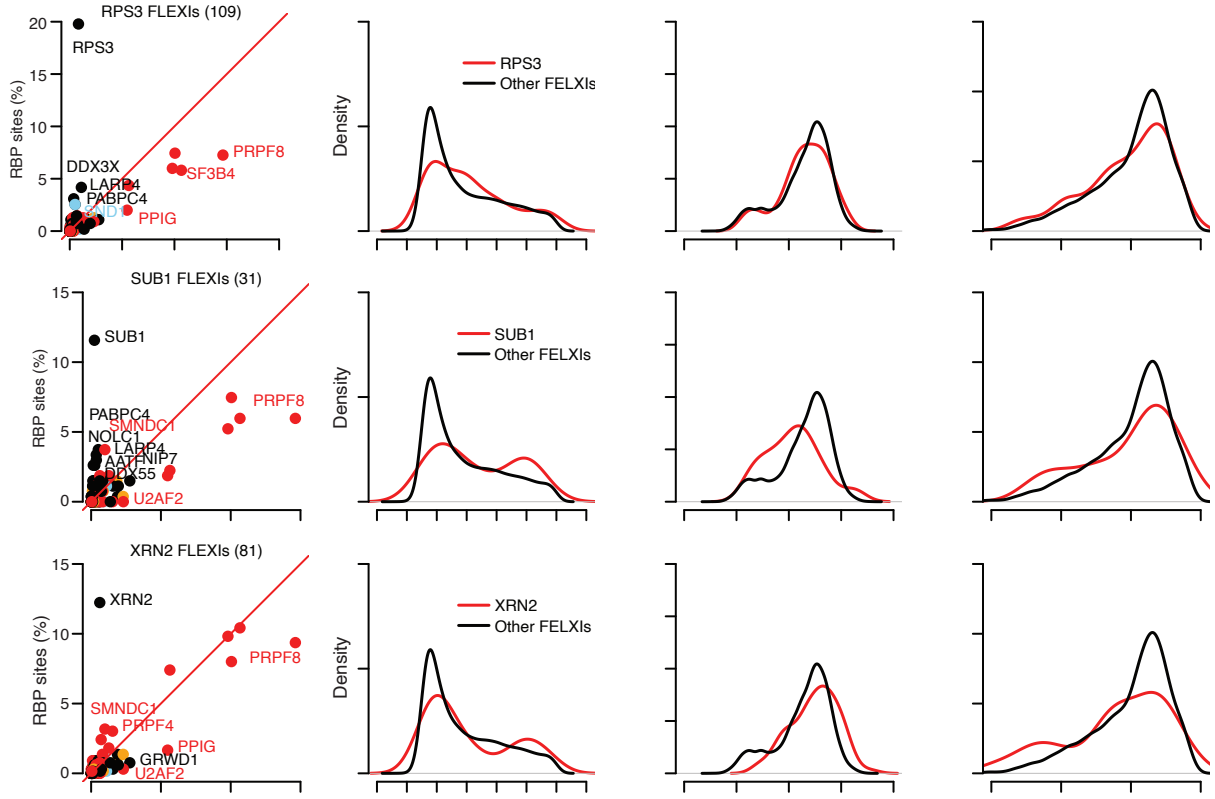


Fig. S15

A. Cluster I



B. Cluster II

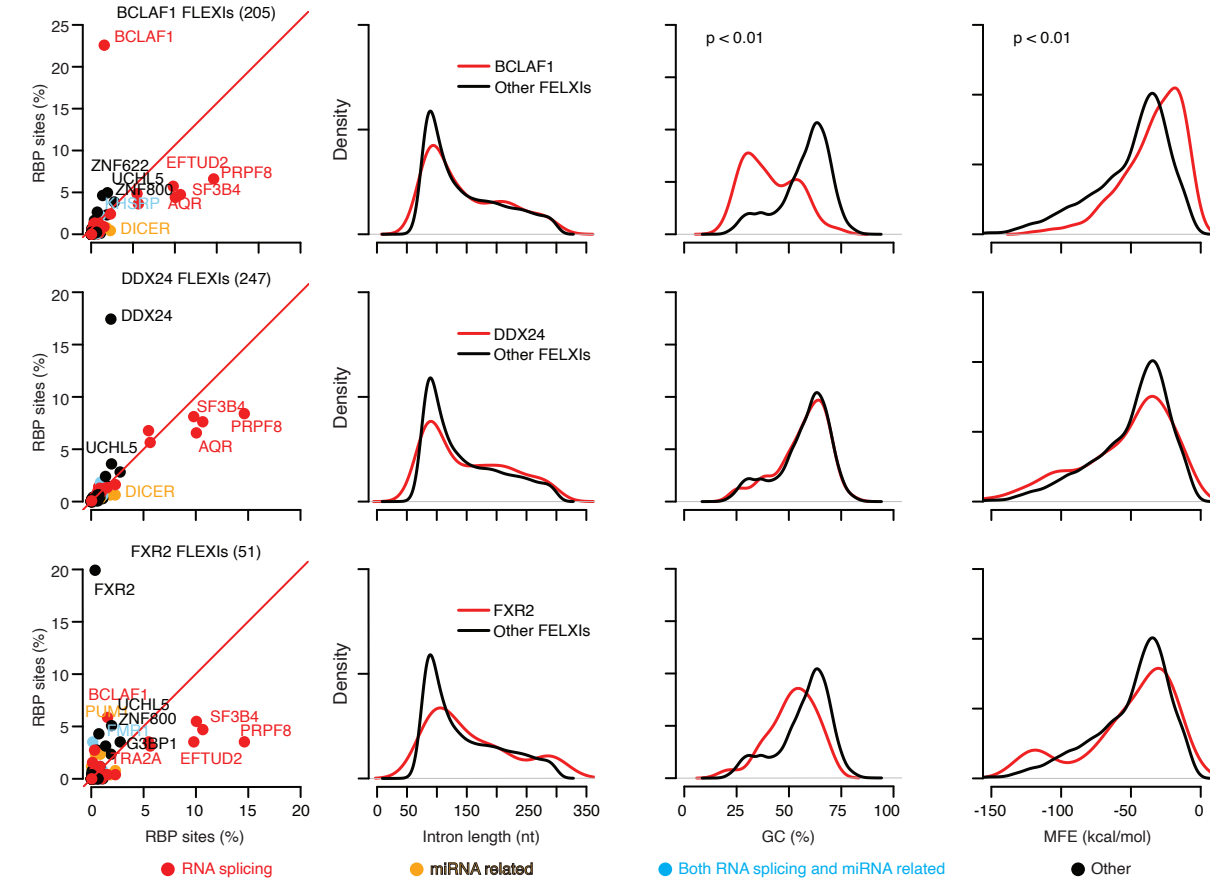


Fig. S15

B. Cluster II

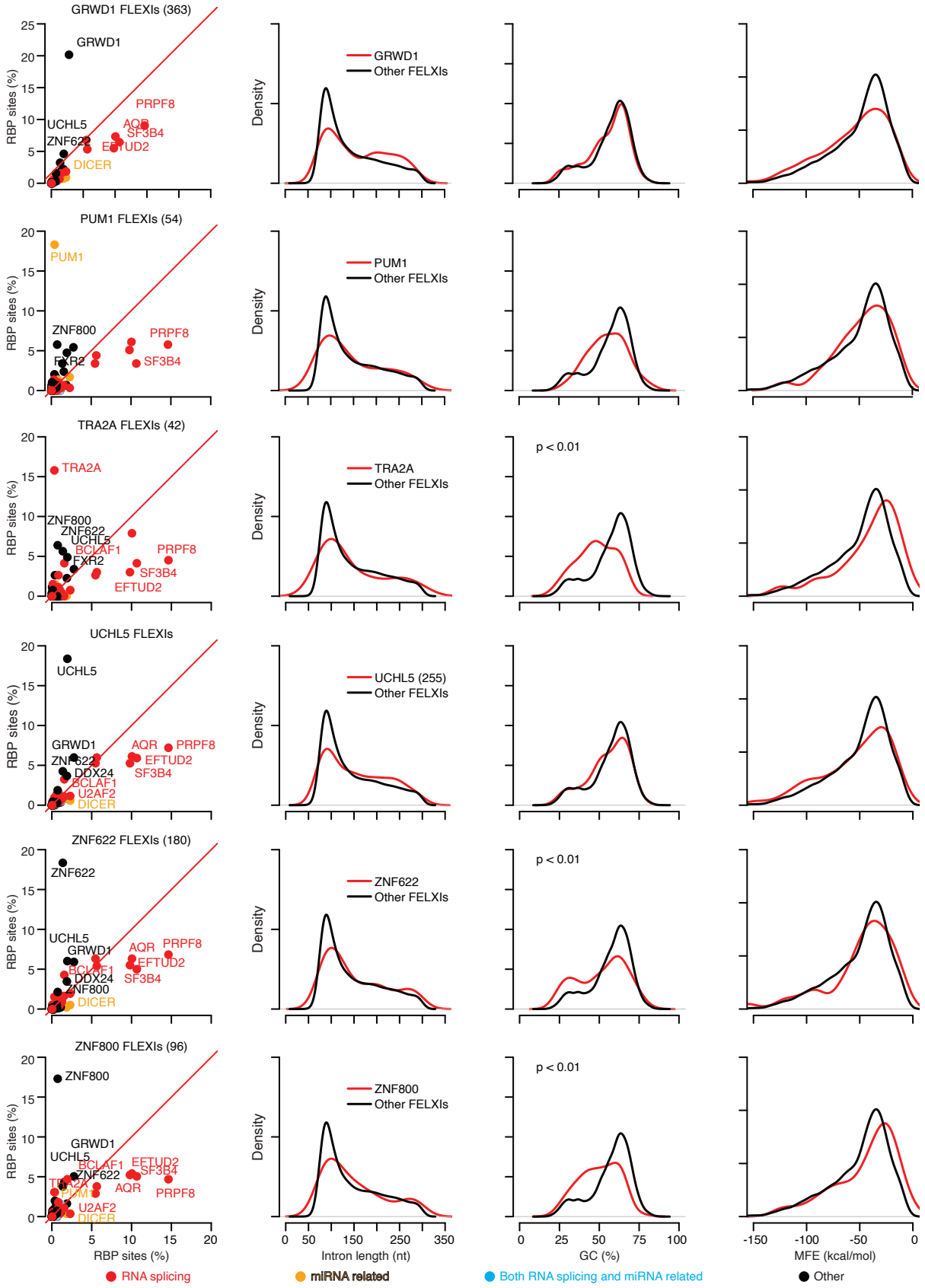
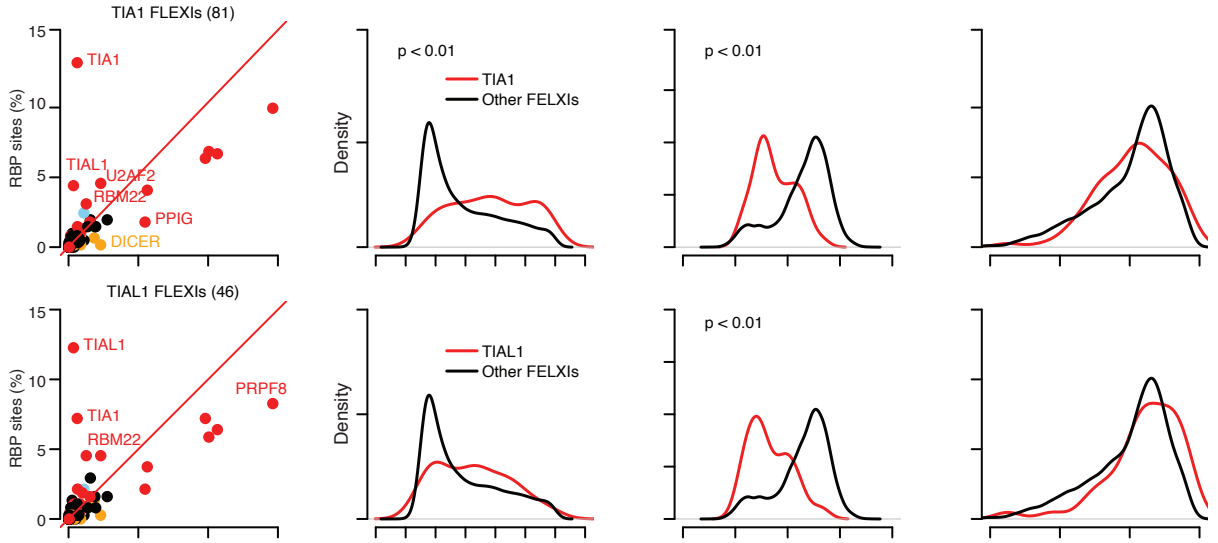
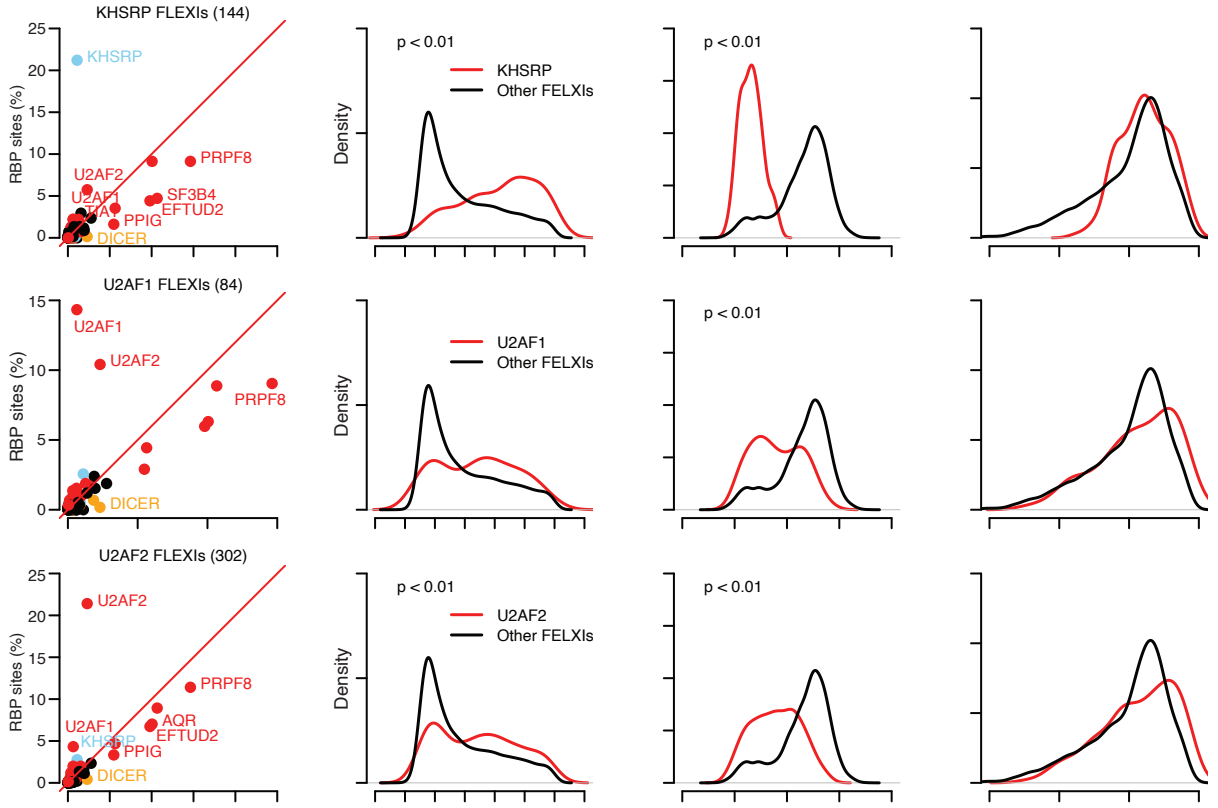


Fig. S15

C. Cluster III



D. Cluster IV



E. Cluster V

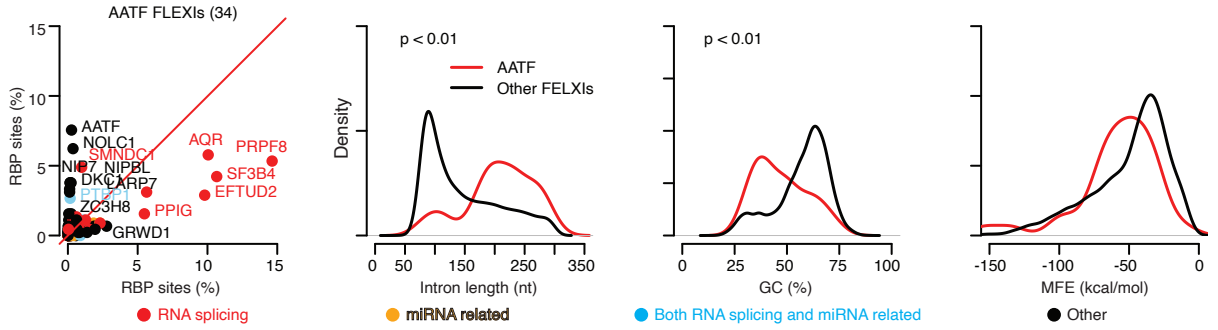
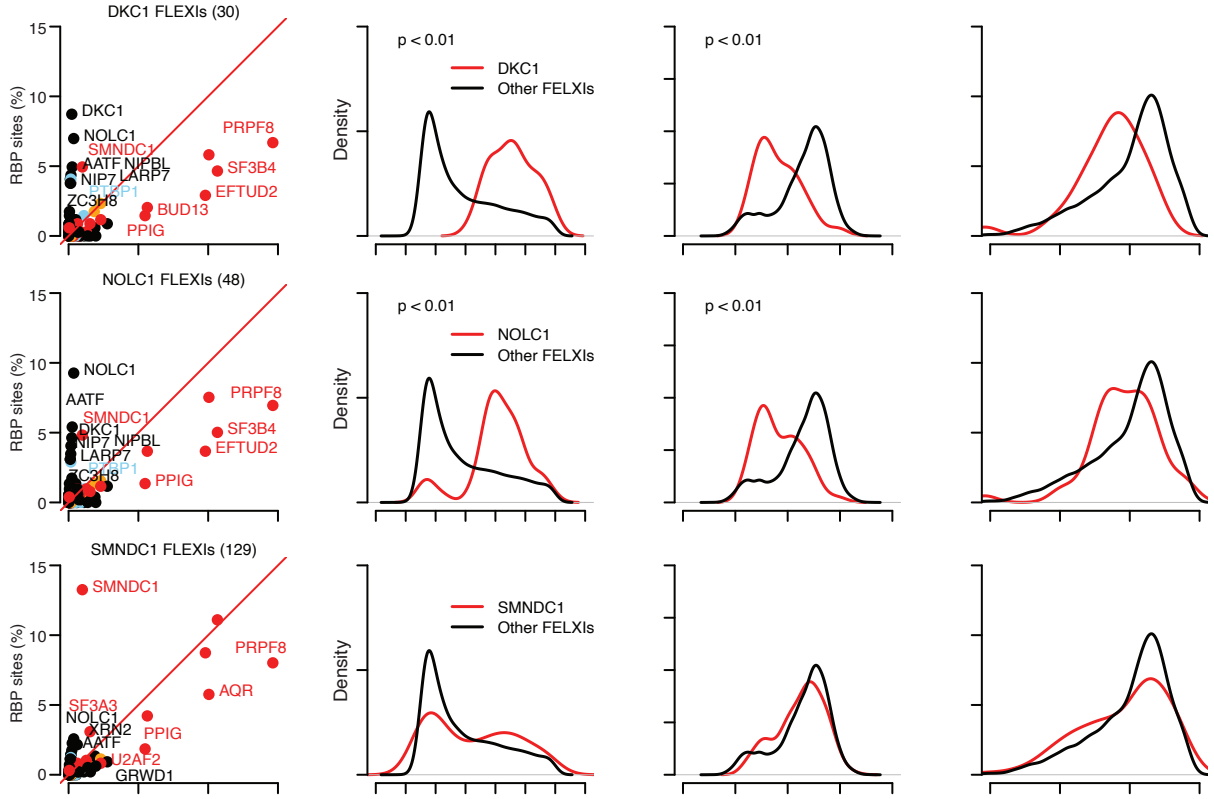


Fig. S15

E. Cluster V



F. Cluster VI

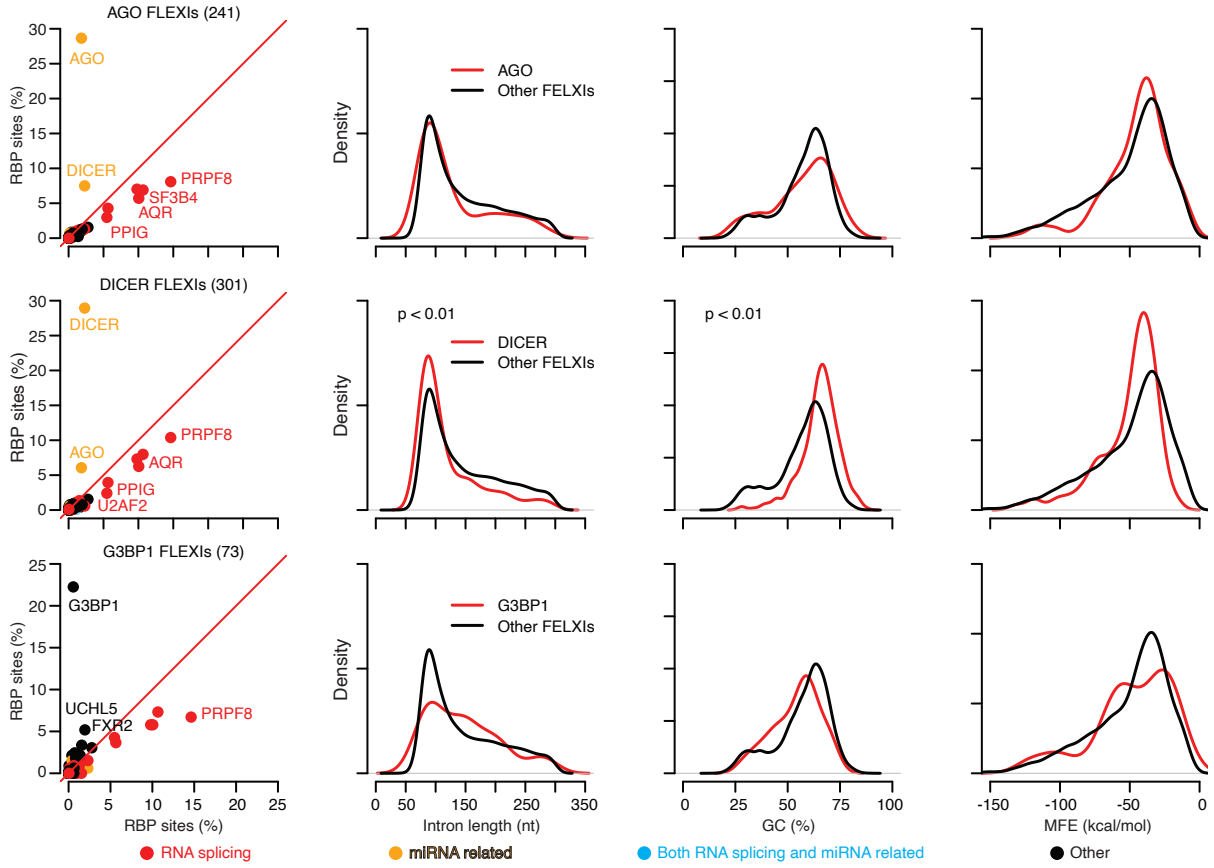


Fig. S15

G. Core spliceosome RBPs

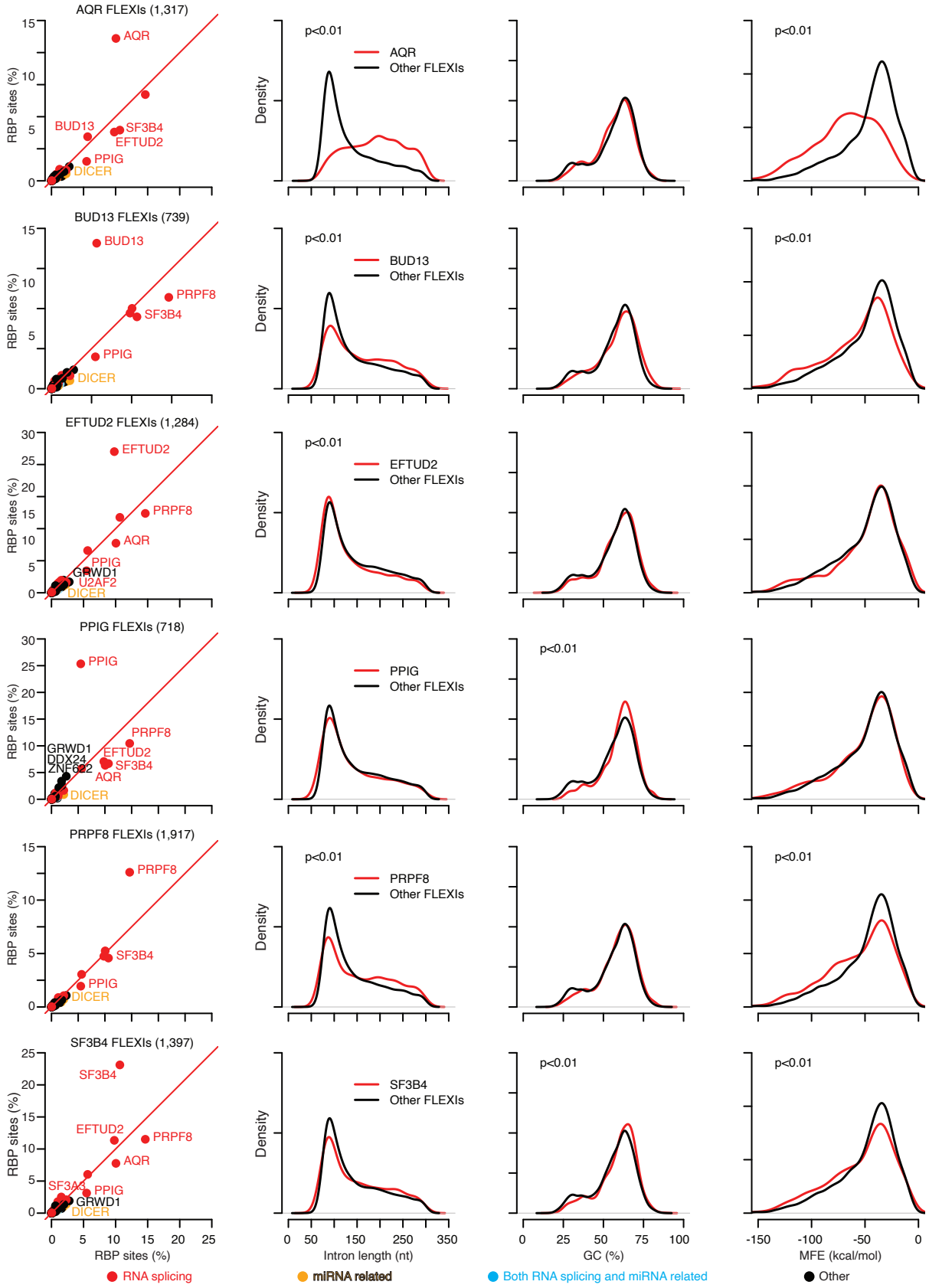


Fig. S15

H. Other

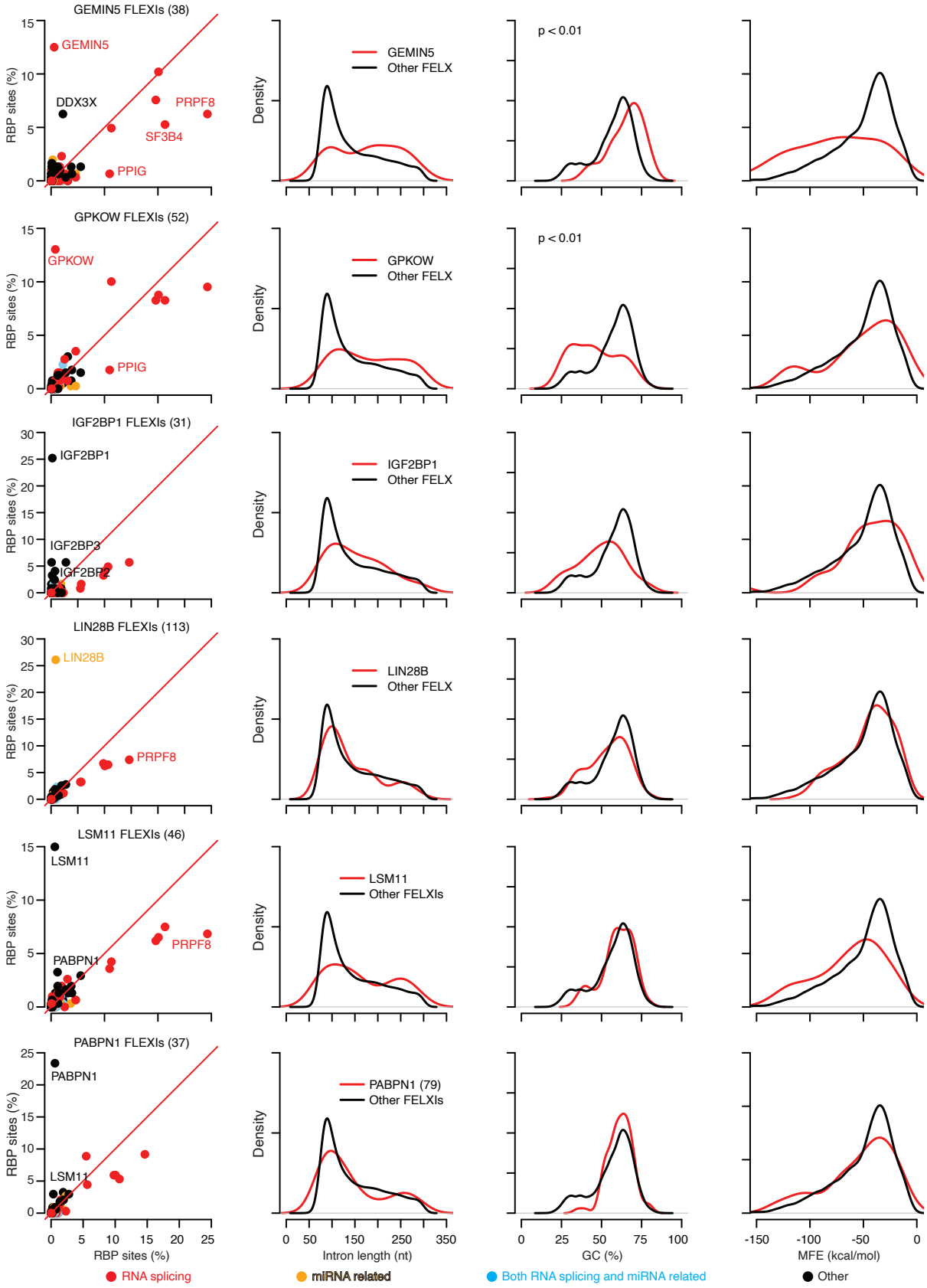


Fig. S15

H. Other

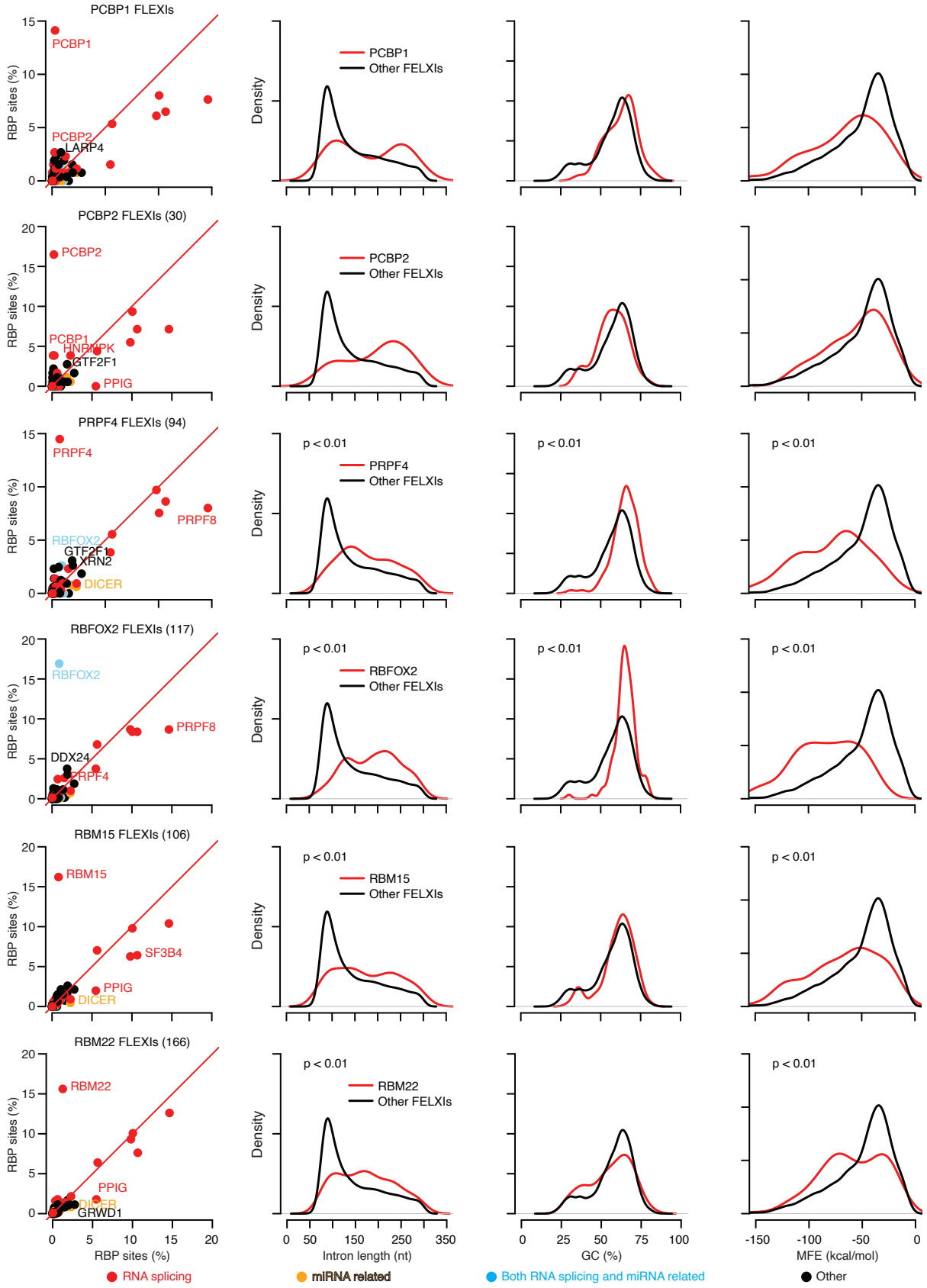


Fig. S15

H. Other

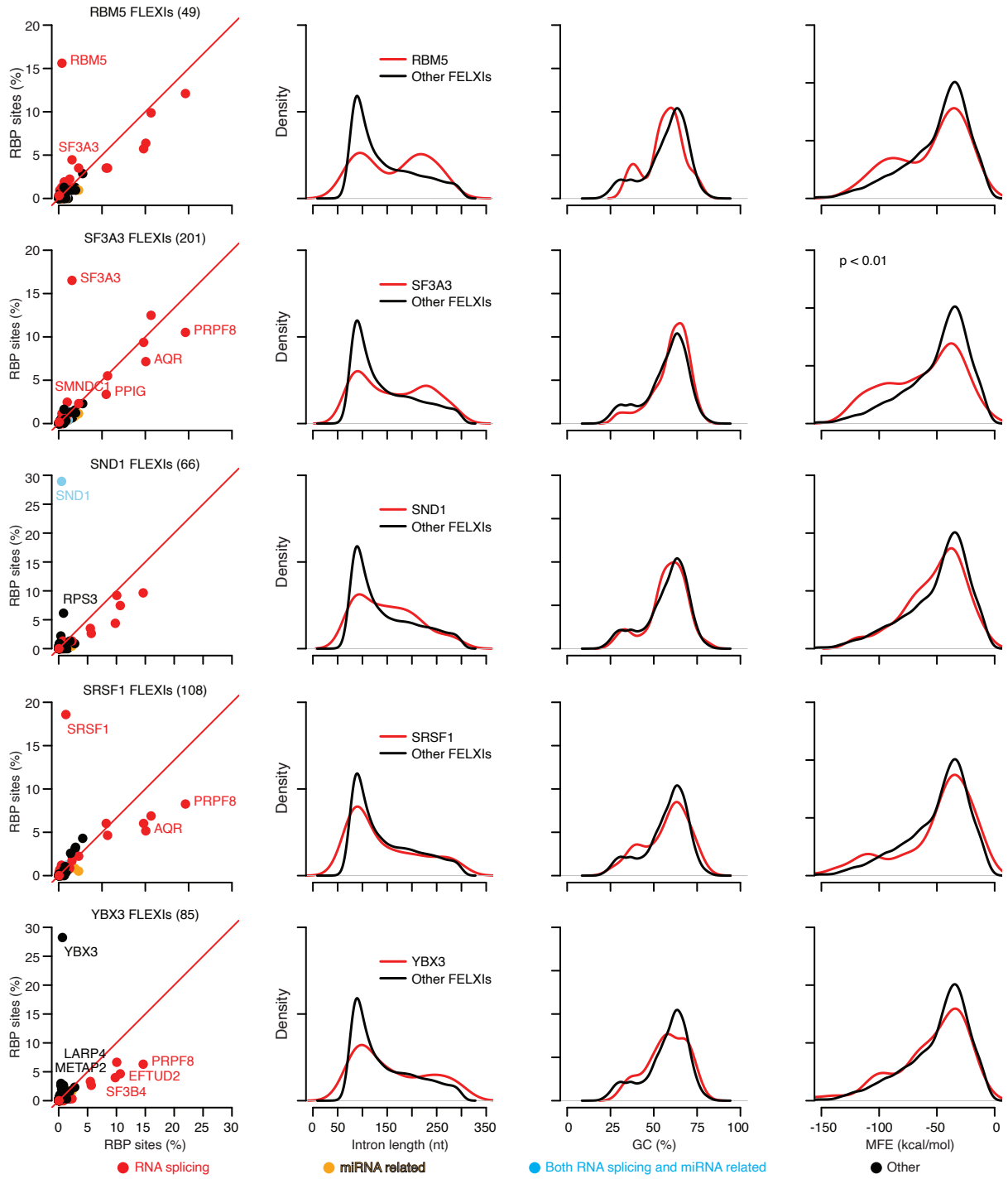


Fig. S16

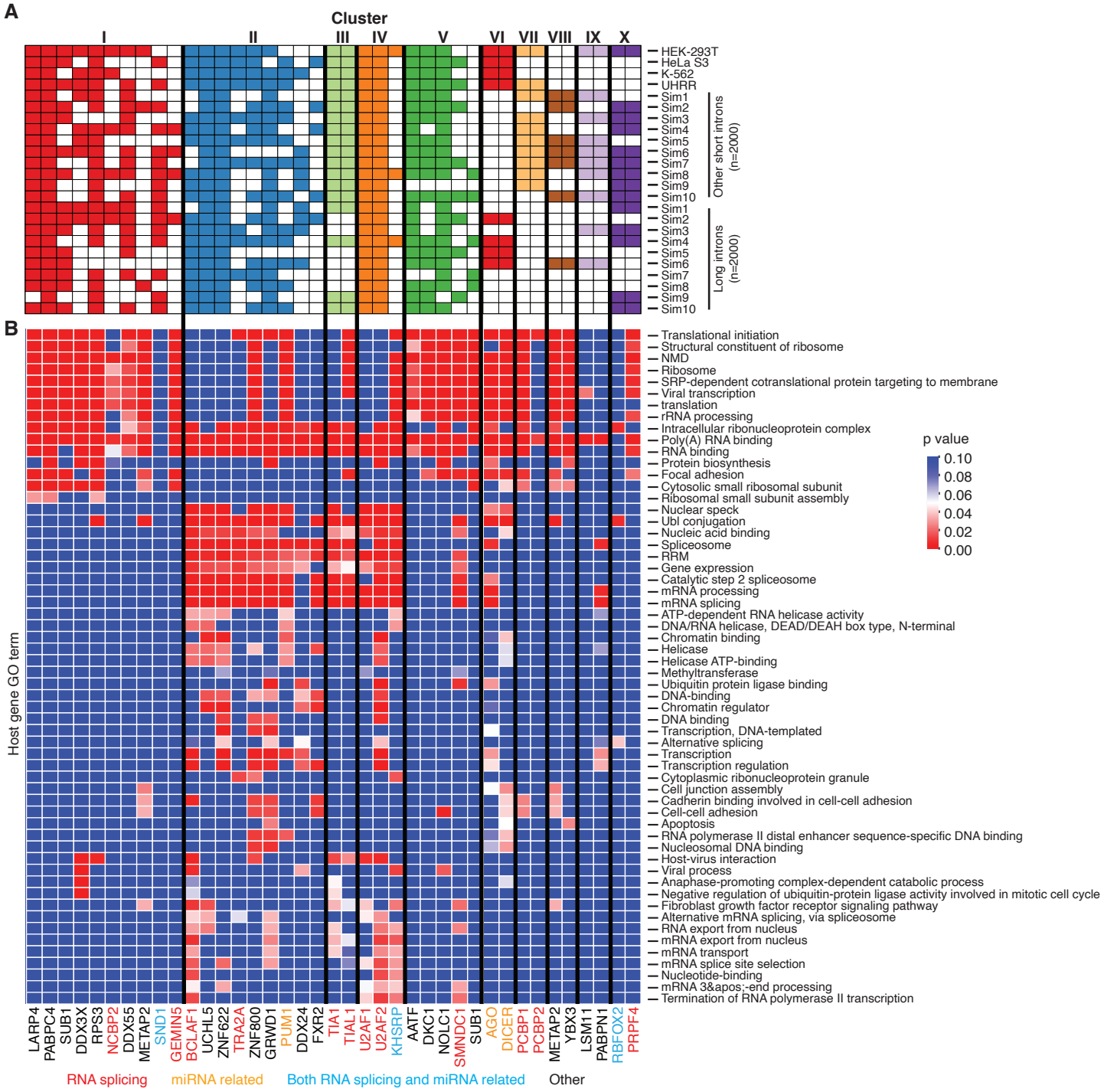

A
Cluster
I
II
III
IV
V
VI
VII
VIII
IX
X
HEK-293T
HeLa S3
K-562
UHRR
Sim1
Sim2
Sim3
Sim4
Sim5
Sim6
Sim7
Sim8
Sim9
Sim10
Other short introns (n=2000)
Long introns (n=2000)
B
Host gene GO term
Translational initiation
Structural constituent of ribosome
NMD
Ribosome
SRP-dependent cotranslational protein targeting to membrane
Viral transcription
translation
rRNA processing
Intracellular ribonucleoprotein complex
Poly(A) RNA binding
RNA binding
Protein biosynthesis
Focal adhesion
Cytosolic small ribosomal subunit
Ribosomal small subunit assembly
Nuclear speck
Ubl conjugation
Nucleic acid binding
Spliceosome
RRM
Gene expression
Catalytic step 2 spliceosome
mRNA processing
mRNA splicing
ATP-dependent RNA helicase activity
DNA/RNA helicase, DEAD/DEAH box type, N-terminal
Chromatin binding
Helicase
Helicase ATP-binding
Methyltransferase
Ubiquitin protein ligase binding
DNA-binding
Chromatin regulator
DNA binding
Transcription, DNA-templated
Alternative splicing
Transcription
Transcription regulation
Cytoplasmic ribonucleoprotein granule
Cell junction assembly
Cadherin binding involved in cell-cell adhesion
Cell-cell adhesion
Apoptosis
RNA polymerase II distal enhancer sequence-specific DNA binding
Nucleosomal DNA binding
Host-virus interaction
Viral process
Anaphase-promoting complex-dependent catabolic process
Negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
Fibroblast growth factor receptor signaling pathway
Alternative mRNA splicing, via spliceosome
RNA export from nucleus
mRNA export from nucleus
mRNA transport
mRNA splice site selection
Nucleotide-binding
mRNA 3'-end processing
Termination of RNA polymerase II transcription
p value
0.10
0.08
0.06
0.04
0.02
0.00
LARP4
PABPC4
SUB1
DDX3X
RPS3
NCBP2
DDX55
METAP2
SND1
GEMIN5
BCLAF1
UCHL5
ZNF622
TRA2A
ZNF800
GRWD1
PUM1
DDX24
FXR2
TIA1
TIAL1
U2AF1
U2AF2
KHSRP
AATF
DKC1
NOLC1
SMNDC1
SUB1
AGO
DICER
PCBP1
PCBP2
METAP2
YBX3
LSM11
PABPN1
RBFOX2
PRPF4
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other

Fig. S17

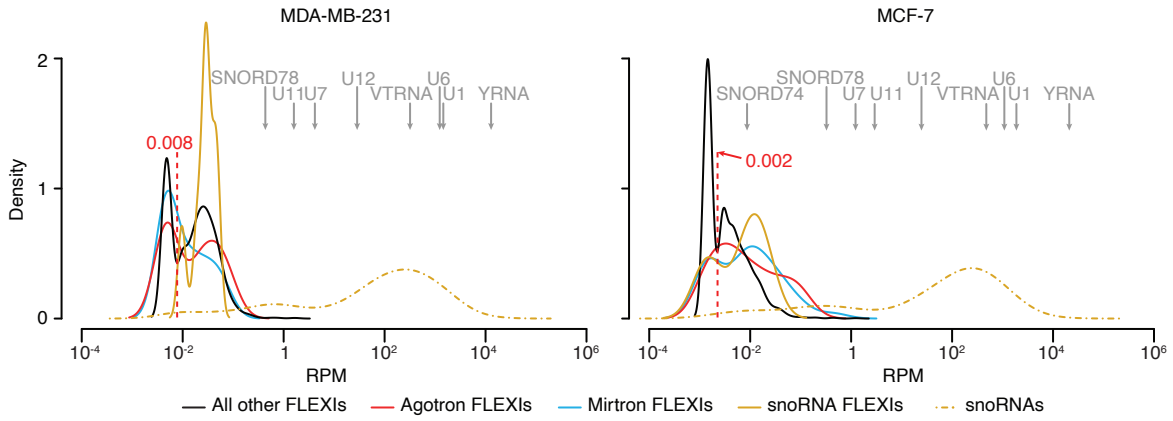

MDA-MB-231
MCF-7
Density
RPM
0.008
0.002
SNORD78
U11
U7
U12
VTRNA
U6
U1
YRNA
SNORD74
All other FLEXIs
Agotron FLEXIs
Mirtron FLEXIs
snoRNA FLEXIs
snoRNAs

Fig. S18

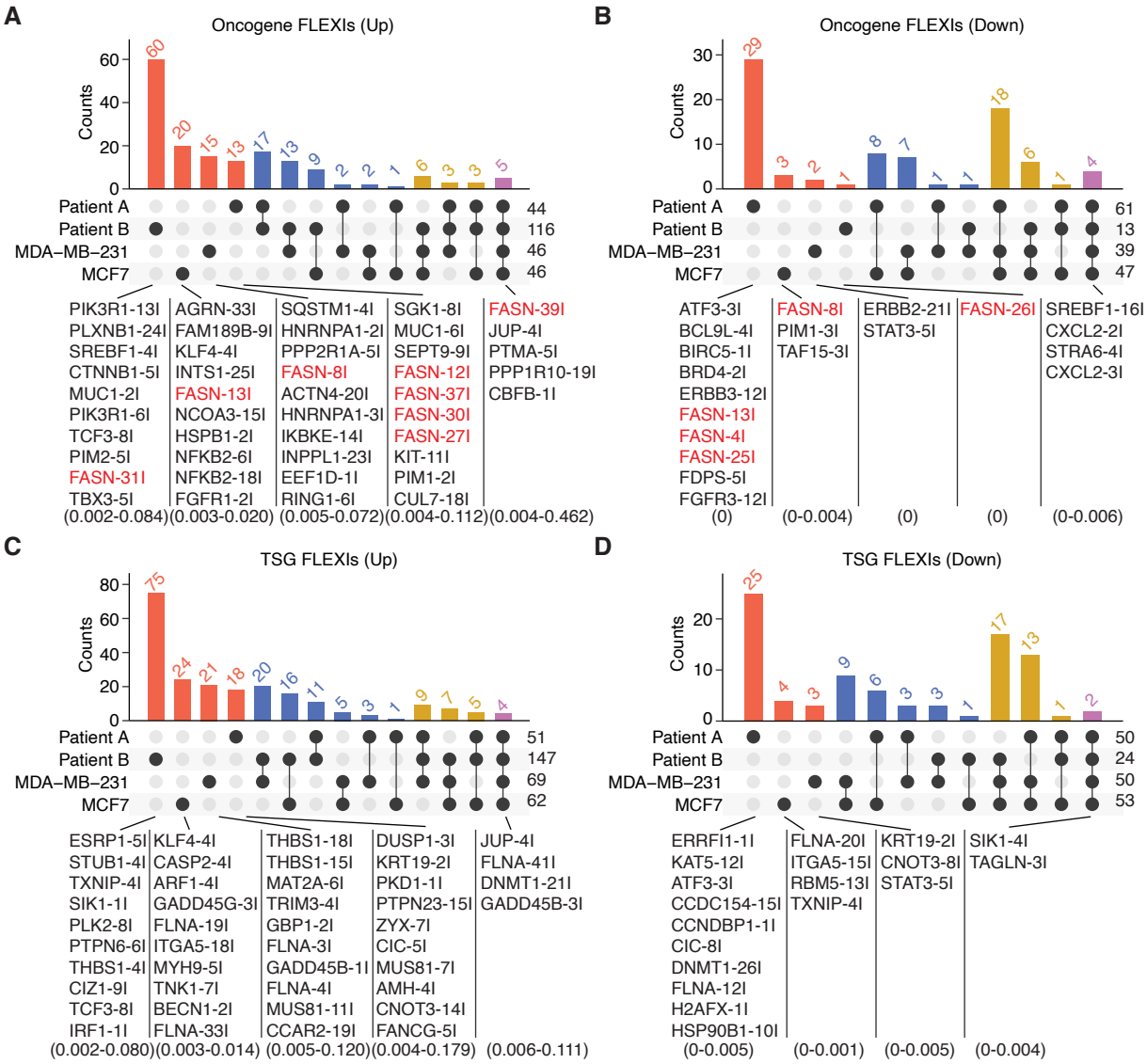

A
Oncogene FLEXIs (Up)
Counts
60 20 15 13 17 13 9 2 2 1 6 3 3 5
Patient A 44
Patient B 116
MDA–MB–231 46
MCF7 46
PIK3R1-13I PLXNB1-24I SREBF1-4I CTNNB1-5I MUC1-2I PIK3R1-6I TCF3-8I PIM2-5I FASN-31I TBX3-5I
(0.002-0.084)
AGRN-33I FAM189B-9I KLF4-4I INTS1-25I FASN-13I NCOA3-15I HSPB1-2I NFKB2-6I NFKB2-18I FGFR1-2I
(0.003-0.020)
SQSTM1-4I HNRNPA1-2I PPP2R1A-5I FASN-8I ACTN4-20I HNRNPA1-3I IKBKE-14I INPPL1-23I EEF1D-1I RING1-6I
(0.005-0.072)
SGK1-8I MUC1-6I SEPT9-9I FASN-12I FASN-37I FASN-30I FASN-27I KIT-11I PIM1-2I CUL7-18I
(0.004-0.112)
FASN-39I JUP-4I PTMA-5I PPP1R10-19I CBFB-1I
(0.004-0.462)
B
Oncogene FLEXIs (Down)
Counts
29 3 2 1 8 7 1 1 18 6 1 4
Patient A 61
Patient B 13
MDA–MB–231 39
MCF7 47
ATF3-3I BCL9L-4I BIRC5-1I BRD4-2I ERBB3-12I FASN-13I FASN-4I FASN-25I FDPS-5I FGFR3-12I
(0)
FASN-8I PIM1-3I TAF15-3I
(0-0.004)
ERBB2-21I STAT3-5I
(0)
FASN-26I
(0)
SREBF1-16I CXCL2-2I STRA6-4I CXCL2-3I
(0-0.006)
C
TSG FLEXIs (Up)
Counts
75 24 21 18 20 16 11 5 3 1 9 7 5 4
Patient A 51
Patient B 147
MDA–MB–231 69
MCF7 62
ESRP1-5I STUB1-4I TXNIP-4I SIK1-1I PLK2-8I PTPN6-6I THBS1-4I CIZ1-9I TCF3-8I IRF1-1I
(0.002-0.080)
KLF4-4I CASP2-4I ARF1-4I GADD45G-3I FLNA-19I ITGA5-18I MYH9-5I TNK1-7I BECN1-2I FLNA-33I
(0.003-0.014)
THBS1-18I THBS1-15I MAT2A-6I TRIM3-4I GBP1-2I FLNA-3I GADD45B-1I FLNA-4I MUS81-11I CCAR2-19I
(0.005-0.120)
DUSP1-3I KRT19-2I PKD1-1I PTPN23-15I ZYX-7I CIC-5I MUS81-7I AMH-4I CNOT3-14I FANCG-5I
(0.004-0.179)
JUP-4I FLNA-41I DNMT1-21I GADD45B-3I
(0.006-0.111)
D
TSG FLEXIs (Down)
Counts
25 4 3 9 6 3 3 1 17 13 1 2
Patient A 50
Patient B 24
MDA–MB–231 50
MCF7 53
ERRFI1-1I KAT5-12I ATF3-3I CCDC154-15I CCNDBP1-1I CIC-8I DNMT1-26I FLNA-12I H2AFX-1I HSP90B1-10I
(0-0.005)
FLNA-20I ITGA5-15I RBM5-13I TXNIP-4I
(0-0.001)
KRT19-2I CNOT3-8I STAT3-5I
(0-0.005)
SIK1-4I TAGLN-3I
(0-0.004)

Table S1. Summary of datasets obtained in this study

Cell / Tissue	Raw reads ($x10^6$)	Trimmed reads ($x10^6$)	Mapped reads ($x10^6$)	Mapped to feature ($x10^6$)
HEK-293T (Bio. 1)*	741.6	726.5 (98.0%)	715.2 (98.5%)	690.9 (96.6%)
HeLa S3†	851.0	803.3 (94.4%)	768.4 (95.7%)	705.6 (92.0%)
K-562 (Bio. 1)§	744.2	725.0 (97.4%)	713.8 (98.4%)	698.5 (97.9%)
UHRR (Bio. 1)‡	712.0	682.2 (95.8%)	666.3 (97.7%)	630.9 (94.7%)
Plasma¶	232.5	122.7 (91.3%)	71.1 (57.9%)	61.7 (87.2%)
MCF7‖	757.8	703.7 (92.9%)	692.1 (98.4%)	673.7 (97.3%)
MDA-MB-231 (Bio. 1)#	226.1	211.3 (93.5%)	207.5 (98.2%)	202.2 (97.5%)
Patient A Healthy**	338.3	312.6 (92.4%)	305.1 (97.6%)	297.6 (97.5%)
Patient A Cancer**	295.8	275.0 (93.0%)	268.2 (97.5%)	256.5 (95.6%)
Patient B Healthy**	281.9	258.2 (91.6%)	251.6 (97.4%)	244.0 (97.0%)
Patient B Cancer**	549.7	492.4 (89.6%)	477.5 (97.0%)	461.6 (96.7%)
Patient A Healthy (Fragmented)††	55.7	52.7 (94.7%)	50.5 (95.8%)	33.2 (60.1%)
Patient A Cancer (Fragmented)††	61.9	59.8 (96.7%)	57.0 (95.3%)	40.5 (68.8%)
Patient B Healthy (Fragmented)††	39.6	35.1 (88.5%)	33.1 (94.3%)	22.0 (57.2%)
Patient B Cancer (Fragmented)††	58.4	56.9 (97.5%)	54.1 (95.1%)	47.1 (85.5%)
HEK-293T (Bio. 2)‡‡	224.1	211.0 (94.2%)	180.8 (85.7%)	162.6 (89.9%)
K-562 (Bio. 2) ¶¶	59.4	54.6 (91.9%)	47.9 (87.7%)	45.5 (95.1%)
MDA-MB-231 (Bio. 2) ##	314.8	274.3 (87.1%)	259.1 (94.5%)	243.7 (94.1%)
UHRR (Bio. 2)§§	416.4	397.3 (95.4%)	359.4 (90.5%)	293.0 (81.5%)

* HEK-293T cell RNA, combined datasets from 8 technical replicates (HEK-rep1 to 8).
† HeLa S3 cell RNA, combined datasets from 10 technical replicates (HeLa-rep1 to 10).
§ K-562 cell RNA, combined datasets from 8 technical replicates (K-562-rep1 to 8).

‡ Universal human reference RNA, combined datasets from 8 technical replicates (UHRR-rep1 to 8, lot# 0006390163).
¶ Commercial human plasma pooled plasma from healthy individuals. Fifteen combined datasets (SRA BioProject accession number PRJNA640428, samples DNaseI_1 to 12, ExoI_1 to 3) (Yao et al. 2020).
‖ MCF7 RNA, combined datasets from 8 technical replicates (MCF-rep1 to 8).
MDA-MB-231 RNA, combined datasets from 2 technical replicates (MDA-rep1 and 2).
** Matched healthy and cancer breast tissue RNA from patients A and B purchased from Origene (Patient A: PR^+, ER^+, $HER2^-$; CR562524/CR543839); Patient B: PR unknown, ER^-, $HER2^-$; CR560540/CR532030). Combined datasets from 3 replicates for each sample type (rep1-3).
†† Chemically fragmented RNAs from matched healthy/cancer tissues from patients A and B.
‡‡ HEK-293T cell RNA, 2 technical replicates (SRA BioProject accession number PRJNA389469, samples WC and YC) (Shurtleff *et al.*, 2017).
¶¶ K-562 cell RNA, Bio. 2, 7 technical replicates (SRA BioProject accession number PRJNA722757, samples UFC1 to 7).
MDA-MB-231 cell RNA, Bio. 2, 5 technical replicates (SRA BioProject accession number PRJNA722757, samples MDA1 to 5).
§§ Universal human reference RNA, 10 technical replicates (SRA BioProject accession number PRJNA722757, samples UHRR0 to 9, lot# 0006227794).

Table S2. Abundance of sncRNAs (RPM) detected by TGIRT-seq in cellular RNA samples compared to literature copy number per cell values for these RNAs

Name	RPM				Copies/cell*	Type
	K-562	HEK-293T	HeLa S3	UHRR		
U1	1,385/6,057	1,266/4,756	1,337/5,732	2,647/7,977	$1x10^6$	Major spliceosomal snRNA
U2	5,764/9,970	9,476/1,1977	28,774/26,539	9,787/12,135	$5x10^5$	Major spliceosomal snRNA
U4	65/371	34/172	540/1,815	76/280	$2x10^5$	Major spliceosomal snRNA
U5	1,146/5,476	2,703/9,161	13,241/26,325	2,715/8,247	$2x10^5$	Major spliceosomal snRNA Minor spliceosomal snRNA
U6	974/3,474	1,766/5,204	923/4,863	1,355/4,947	$4x10^5$	Major spliceosomal snRNA
U7	3/22	4/27	49/197	3/9	$4x10^3$	U7
U11	2/32	3/33	11/108	8/44	$1x10^4$	Minor spliceosomal snRNA
U12	36/157	24/78	192/433	41/110	$5x10^3$	Minor spliceosomal snRNA
U4ATAC	37/183	21/101	1,291/2,883	32/173	$2x10^3$	Minor spliceosomal snRNA
U6ATAC	204/1,104	286/1,481	699/4,236	268/1,350	$2x10^3$	Minor spliceosomal snRNA
RN7SK	1,368/3,585	2,771/4,828	5,043/11,318	7,174/6,428	$2x10^5$	7SK
RN7SL	3,445/10,723	8,080/16,276	4,689/13,770	19,801/17,395	$5x10^5$	7SL
RPPH1	165/474	245/643	257/713	1,293/2,212	$2x10^5$	RNase P RNA component
SNORD3	7,637/13,928	12,700/18,193	23,204/29,379	38,531/30,293	$2x10^5$	C/D box snoRNA, U3
SNORD118	222/758	985/1,617	552/948	745/1,178	$4x10^4$	C/D box snoRNA, U8
SNORD13	436/2,572	367/2,129	817/3,477	1,223/4,786	$1x10^4$	C/D box snoRNA, U13
SNORD14	4,278/6,920	3,887/5,651	9,979/9,149	11,073/9,899	$1x10^4$	C/D box snoRNA, U14
SNORD22	362/1,600	240/963	748/2,526	947/2,547	$1x10^4$	C/D box snoRNA, U22
RMRP	669/1,907	1,776/3,781	2,211/7,294	3,276/6,048	$1x10^5$	MRP

* Copy number per cell values for sncRNAs in vertebrate cells (Tycowski *et al.*, 2006) used in the linear regression analysis of Fig. S6.

Table S3. Proteins that bind 30 or more different FLEXIs

Symbol	Name	Function
AATF	Apoptosis Antagonizing Transcription Factor	Transcriptional cofactor with roles in cell proliferation, apoptosis, DNA damage response and general stress response through regulation of Rb, HDAC1, and p53 functions.
AGO	Argonaute	Central player in RNA silencing processes, essential component of RNA-induced silencing complex (RISC)
BCLAF1	BCL2-2-associated transcription factor 1	Transcriptional repressor; promotes apoptosis through interaction with BCL2; Upregulated in senescence, promotes p53 transcription in response to DNA damage.
DDX24	ATP-dependent RNA helicase DDX24	ATP-dependent RNA helicase and negative regulator of p53
DDX3X	ATP-dependent RNA helicase DDX3X	Multifunctional ATP-dependent RNA helicase with functions in cell cycle control, apoptosis, and innate immunity; critical role in stress granule assembly.
DDX55	ATP-dependent RNA helicase DDX55	ATP-binding RNA helicase
DICER	DICER	Cleaves double-stranded RNA and pre-miRNA to form siRNA or miRNA
DKC1	H/ACA ribonucleoprotein complex subunit DKC1	Catalytic subunit of H/ACA small nucleolar ribonucleoprotein (H/ACA snoRNP) complex; plays an active role in telomerase stabilization
FXR2	Fragile X mental retardation syndrome-related protein 2	RNA-binding protein
G3BP1	Ras GTPase-activating protein-binding protein 1	ATP- and Mg-dependent helicase that plays an essential role in innate immunity; Also functions in stress granule assembly and is associated with cellular senescence. Regulates Ras, TGF-β/Smad, Src/FAK and p53 signaling pathways.
GEMIN5	Gem-associated protein 5	Plays a role in cytoplasmic assembly of snRNPs; component of the SMN complex which plays a role in alternative splicing

GPKOW	G-patch domain and KOW motifs-containing protein	Component of the spliceosome
GRWD1	Glutamate-rich WD repeat-containing protein 1	Histone binding-protein that regulates chromatin dynamics and minichromosome maintenance (MCM) loading at replication origins; negatively regulates p53.
IGF2BP1	Insulin-like growth factor 2 mRNA-binding protein 1	RNA-binding protein that recruits target transcripts to cytoplasmic protein-RNA complexes (mRNPs); Promotes cell cycle progression through regulation of E2F translation.
LARP4	La-related protein 4	RNA binding protein that binds to the polyA tract of mRNA molecules
LIN28B	Lin-28 homolog B	Suppressor of miRNA biogenesis, specifically let7; inhibits apoptosis
LSM11	U7 snRNA-associated Sm-like protein LSm11	Component of the U7 snRNP complex that is involved in the histone 3'-end pre-mRNA processing
METAP2	Methionine aminopeptidase 2	Co-translationally removes the N-terminal methionine from nascent proteins.
NCBP2	Nuclear cap-binding protein subunit 2	Component of cap-binding complex that binds cap of pre-mRNAs; involved in splicing, translation regulation and NMD
NOLC1	Nucleolar and coiled-body phosphoprotein 1	Nucleolar protein that plays a critical role in snoRNP assembly and acts as a regulator of RNA polymerase I by connecting RNA polymerase I with enzymes responsible for ribosomal processing and modification; Stabilizes telomeres by regulating TRF2 retention.
PABPC4	Polyadenylate-binding protein 4	Binds the polyA tail of mRNA
PABPN1	Polyadenylate-binding protein 2	Involved in the 3'-end formation of mRNA precursors (pre-mRNA) by the addition of a poly(A) tail; Regulated by ATM and plays a crucial role in DSB repair.
PCBP2	Poly(rC)-binding protein 2	Binds single stranded nucleic acids; regulates antiviral response mediated by MAVS signaling

RBM15	RNA-binding protein 15	Regulates m6A modification of RNAs; alternative splicing factor
RBM22	RNA-binding protein 22	Required component of the spliceosome
RPS3	40S ribosomal protein S3	Role in regulating transcription; implicated in regulating DNA damage response and apoptosis.
SF3A3	Splicing factor 3a subunit 3	Required splicing factor
SRSF1	Serine/arginine-rich splicing factor 1	Prevents exon skipping, regulates alternative splicing
SUB1	Activated RNA polymerase II transcriptional coactivator p15	General coactivator that functions cooperatively with TAFs and mediates functional interactions between upstream activators and the general transcriptional machinery; critical role in genome integrity and chromatin compaction, regulates transcription in response to stress.
TIA1	TIA1 cytotoxic granule associated RNA binding protein	Alternative splicing factor, regulation of mRNA translation; potential role in apoptosis
TIAL1	TIA1 cytotoxic granule associated RNA binding protein like 1	Binds AU-rich elements; roles in regulating translation, splicing and apoptosis.
U2AF1	U2 small nuclear RNA auxiliary factor 1	Critical role in splicing, recruits U2 snRNP to branch point
U2AF2	U2 small nuclear RNA auxiliary factor 2	Roles in pre-mRNA splicing and 3’ end processing
UCHL5	Ubiquitin carboxyl-terminal hydrolase isozyme L5	Protease
XRN2	5'-3' exoribonuclease 2	May promote transcription termination by RNA polymerase II
YBX3	Y-Box-binding protein 3	Binds also to full-length mRNA and to short RNA sequences containing the consensus site 5'-UCCAUCA-3'.
ZNF622	Zinc finger protein 622	May behave as an activator of the bound transcription factor, MYBL2; positive regulator of apoptosis.
ZNF800	Zinc finger protein 800	May be involved in transcriptional regulation.

Table S4. Oligonucleotides used for construction of TGIRT-seq libraries

Name	Sequence and notes
NTT R2 RNA	5'-AAGAUCGGAAGAGCACACGUCUGAACUCCAGUCAC/3SpC/
NTT R2R DNA	5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTN-3', where N is an equimolar mix of A, C, G, T (obtained by hand mixing of individual oligonucleotides with A, C, G and T at their 3' end).
R1R DNA	R1R DNA: 5'-/5Phos/GATCGTCGGACTGTAGAACTCTGAACGTGT AG/3SpC3/. The R1R oligonucleotide was adenylated. as described in Materials and Methods.
Illumina multiplex PCR primer	5'-AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTA CAGTCCGACGATC-3'
Illumina index PCR primer	5' CAAGCAGAAGACGGCATACGAGAT BARCODE* GTGACTGGA GTTCAGACGTGTGCTCTTCCGATCT-3', where BARCODE* corresponds to the 6 nucleotide Illumina TruSeq barcode sequence.

Table S5. Oligonucleotides used for droplet digital PCR (ddPCR) and Terminator 5' to 3' exonuclease digestion assays.

Name[†]	Sequence
1I_H3F3B (Forward primer-FLEXI)	TAAGACAAAGGCCTGGAGCC
1I_H3F3B (RT primer/Reverse primer-FLEXI)	GGACGAGAGCCGCACTATTA
1I_H3F3B (Forward primer-junction)	CCGACCTGAGGTTTCAGAGC
1I_H3F3B (Reverse primer-junction)	GGCTCCAGGCCTTTGTCTTA
3I_RAN (Forward primer-FLEXI)	TGCTGGAAAACCTTGCTTGTG
3I_RAN (RT primer/Reverse primer-FLEXI)	CTGAAACGGAAACGATGCGA
3I_RAN (Forward primer-junction)	GGTTGGTGATGGTGGTACTG
3I_RAN (Reverse primer-junction)	CACAAGCAAGGTTTTCCAGCA
7I_EIF4A1 (Forward primer-FLEXI)	CCTGGTAGTGAGTTGTTGGGT
7I_EIF4A1 (RT primer/Reverse primer-FLEXI)	TGGAGGTTGGGGGACAAAAAT
7I_EIF4A1 (Forward primer-junction)	TTGACCCTGGAGGGTATCCG
7I_EIF4A1 (Reverse primer-junction)	AACAACTCACTACCAGGCCC
4I_JUP (Forward primer-FLEXI)	GGAGGCTGGGTCCCACT
4I_JUP (RT primer/Reverse primer-FLEXI (Terminator))	AGAGGAAAGGAAAAGCCAGGT
4I_JUP (Reverse primer-FLEXI (ddPCR))	CCAGGTTAAACGTCGGCCAG
4I_JUP (Forward primer-junction)	CACTCATGGAGTTGCTGCAC
4I_JUP (Reverse primer-junction)	AGTGGGACCCAGCCTCC
16I_POLG (Forward primer-FLEXI)	TAGCCAGGCCTTGGGTGG
16I_POLG (RT primer/Reverse primer-FLEXI)	CAGAGCCAGTCCACTAGGGC
16I_POLG (Forward primer-junction)	GATCTGGCCAATGATGCCTG
16I_POLG (Reverse primer-junction)	CAGTCCCCTGCCTAGATCCT
2I_ACTB (Forward primer-FLEXI)	GTGTCTTTCCTGCCTGAGCTG
2I_ACTB (RT primer/Reverse primer-FLEXI)	TGAGGAGGGAAGGGGACAGG
2I_ACTB (Forward primer-junction)	GAAGGAGATCACTGCCCTGG
2I_ACTB (Reverse primer-junction)	CAGCTCAGGCAGGAAAGACA
1I_RPS2 (Forward primer-FLEXI)	GTAGGCTGGTGGCAGGTGAT
1I_RPS2 (RT primer/Reverse primer-FLEXI)	CTGGTGAGGGAAGGAGTCAG
1I_RPS2 (Forward primer-junction)	GAGGCTACTGGGGGAACAAG
1I_RPS2 (Reverse primer-junction)	ATCACCTGCCACCAGCCTAC
U6 (Forward primer)	CGCTTCGGCAGCACATATAC
U6 (RT primer/Reverse primer)	ATTTGCGTGTCATCCTTGCG
U7 (Forward primer)	TACAGCTCTTTTAGAATTTGTCTAGTAGGC
U7 (RT primer/Reverse primer)	GGGGCTTTCCGGTAAAAAGCC
SNORD14B (Forward primer)[*]	CACTGTGATGATGGTTTTCCAAC
SNORD14B (RT primer/Reverse primer)[*]	AGGAAGGTTTACCCAACACTAAG

SNORD44 (Forward primer)	GATAAGCAAATGCTGACTGAACA
SNORD44 (RT primer/Reverse primer)	AGTCAGTTAGAGCTAATTAAGACC
3I_RAN FLEXI synthetic RNA oligo control (ordered with and without a 5'phosphate)	GUAUGUGCUGGAAAACCUUGCUUGUGGAAAUAUGUGAGAAAUGGGUAAGUUCAUCCACUCAAUCGCAUCGUUUCCGUUUCAG

[†]Primers are named according to FLEXI intron number and gene name or sncRNA name.
[*]Dodd et al., 2013