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

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

[1]Departments of Molecular Biosciences and Oncology
University of Texas at Austin
Austin, Texas 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: VantAI, Inc

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

**Abstract**

Previous Thermostable Group II Intron Reverse Transcriptase sequencing (TGIRT-seq) found that human plasma contains short (≤300 nt) structured full-length excised linear intron RNAs (FLEXIs) with potential utility as blood-based biomarkers. Here, TGIRT-seq identified >9,000 different FLEXIs in human cells, including relatively abundant FLEXIs with cell-type-specific expression patterns. Analysis of published CLIP-seq datasets identified 126 RNA-binding proteins (RBPs) that bind different FLEXIs, including 53 RBPs that bind ≥30 FLEXIs. These RBPs included splicing factors, transcription factors, a chromatin remodeling protein, cellular growth regulators, and proteins with cytoplasmic functions. Analysis of published datasets identified subsets of these RBPs that bind at FLEXI splice sites and impact alternative splicing or FLEXI host gene mRNA levels. Hierarchical clustering identified 6 subsets of RBPs whose binding sites were co-enriched in different subsets of FLEXIs: AGO1-4 and DICER, including but not limited to annotated agotrons and mirtron pre-miRNAs; DKC1, NOLC1, SMNDC1, and AATF (Apoptosis Antagonizing Transcription Factor), including but not limited to FLEXIs encoding snoRNAs; two sets of alternative splicing factors; and two sets that included RBPs with cytoplasmic functions (*e.g.*, LARP4, PABPC4, METAP2, and ZNF622) together with nuclear regulatory proteins. The subsets of host genes encoding FLEXIs that bind these RBPs were enriched with non-FLEXI other short and long introns that bind the same RBPs, and TGIRT-seq of nuclear and cytoplasmic fractions from 4 cell lines showed cytoplasmic enrichment of FLEXIs with binding sites for RBPs that function in the cytoplasm. Collectively, our findings suggest that different subsets of RBPs bind FLEXIs and other introns to coordinately regulate the expression of functionally related host genes and then co-localize with stably bound FLEXIs to different intracellular locations. The cell-type specific expression of relatively abundant FLEXIs suggests utility as cellular as well as blood-based RNA biomarkers.

Key words: gene regulation, mitochondrial function, ribosome assembly, ribosomal proteins, RNA splicing

## 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 (1). After splicing, this bond is typically hydrolyzed by the debranching enzyme DBR1, and the resulting linear intron RNA is rapidly degraded by exonucleases (2). Several classes of excised intron RNAs persist after excision. These include trimmed lariat RNAs that lack a 3' tail; circularized intron RNAs whose 5' and 3' nucleotides are putatively linked by a 3'-5' phosphodiester bond; and full-length excised linear intron RNAs (FLEXIs), with members of all 3 classes found to have cellular functions or clinical correlations (3-14). FLEXIs include a group of yeast excised linear intron RNAs that contribute to cell growth regulation by sequestering spliceosomal proteins in stationary phase or under other stress conditions (11, 15). Other FLEXIs include mirtron pre-miRNAs, structured excised intron RNAs that are debranched by DBR1 and processed by DICER into annotated miRNAs, and agotrons, structured excised intron RNAs that bind AGO2 and repress target mRNAs in a miRNA-like manner (16-19).

Previously, we identified 44 different short (≤300 nt) FLEXIs in human plasma by using Thermostable Group II Intron Reverse Transcriptase sequencing (TGIRT-seq), a method that enables continuous end-to-end sequencing reads of structured RNAs (20). Plasma FLEXIs were identified by peak calling as discrete blocks of continuous sequencing reads of excised introns beginning and ending within 3 nucleotides (nt) of annotated splice sites, with relatively large subsets corresponding to annotated agotrons or mirtron pre-miRNAs (20). Almost all of the plasma FLEXIs (>95%) had stable predicted secondary structures ($\Delta G \leq -30$ kcal/mol), with ~50% containing a CLIP-seq-identified binding site for one or more RNA-binding proteins (RBPs) that may help protect FLEXIs from plasma RNases. These findings suggested that FLEXIs might have utility as RNA biomarkers in liquid biopsies (20).

Despite the precedents of agotrons and mirtron pre-miRNAs, human FLEXIs have remained largely unexplored. Here, TGIRT-seq of RNAs present in human cell lines identified >9,000 FLEXIs, subsets of which had cell-type specific expression patterns, suggesting utility as cellular RNA biomarker. Analysis of published CLIP-seq datasets identified subsets of RBPs that bind ≥30 FLEXIs that were enriched in different subsets of functionally related host genes with other

short and long intron that have binding sites for the same RBPs. The latter findings suggest previously unidentified mechanisms for co-regulation of functionally related gene sets and provide a road map for further analysis of these mechanisms as well as FLEXI-RBP interactions and cellular functions.

## Results

### Identification of FLEXIs 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 FLEXIs. To investigate which short introns give rise to FLEXIs, we carried out TGIRT-seq (21, 22) of rRNA-depleted, unfragmented HEK-293T, K-562, and HeLa S3 whole-cell RNAs and Universal Human Reference RNA (UHRR), a commercial mixture of whole-cell RNAs from 10 human cell lines (Fig. 1). TGIRT-seq is well-suited for the identification of FLEXIs as it employs a TGIRT enzyme (GsI-IIC RT aka TGIRT-III) that can efficiently template-switch from an RNA-seq adapter to initiate cDNA synthesis directly at the 3'-nucleotide of target RNAs and then reverse transcribe processively to yield full-length cDNAs that give end-to-end sequencing reads of structured RNAs (23-26). Because this TGIRT enzyme does not efficiently read through the 3'-poly(A) tails of mature mRNAs, most of the reads in the TGIRT-seq datasets of unfragmented whole-cell RNAs corresponded to full-length mature tRNAs and other small non-coding RNAs (sncRNAs; Fig. S1 and S2).

To identify FLEXIs in the whole-cell RNA preparations, we compiled the coordinates of all short introns (≤300 nt) in Ensembl GRCh38 Release 93 annotations into a BED file and searched for intersections in TGIRT-seq datasets for each of the 4 cellular RNA samples. For each cellular RNA sample, the searched TGIRT-seq dataset included merged technical replicates totaling 666 to 768 million mapped paired-end reads, supplemented by datasets for previously published biological replicates to assess reproducibility (Table S1). As done previously for FLEXIs in plasma, we identified FLEXI by continuous intron reads that began and ended within 3 nt of annotated splice sites (20). This definition enabled us to include full-length intron reads with small numbers of non-templated nucleotides added by cellular enzymes to the 3' end of RNAs (27) or by the TGIRT enzyme to the 3' end of completed cDNAs (21, 22). The 300-nt size cutoff for more detailed characterization of FLEXIs was based on the length distribution of FLEXIs found previously in human plasma (20) and encompassed 85-92% of the FLEXIs detected in the 4 cellular

RNA samples, including the most abundant FLEXIs (Fig. S3A). By using the same read mapping pipeline and mapping reads to an updated human genome reference sequence (GRCh38 Release 93), we identified a total 224 FLEXIs in the previously analyzed plasma samples from healthy individuals (20).

By using these approaches, we identified 8,144 different FLEXIs in the initially analyzed cellular and plasma RNA data originating from 3,743 different protein-coding genes, lncRNA genes, or pseudogenes (collectively denoted FLEXI host genes; Supplementary file). The distribution of FLEXIs and their host genes between the different cellular RNA samples and plasma is shown in UpSet plots in Fig. 1A. Fifty-six of the FLEXIs identified in the 4 cellular RNA samples and plasma corresponded to an annotated agotron and 105 corresponded to a pre-miRNAs of an annotated mirtron (Fig. 1B) (16, 19). Notably, the proportion of FLEXIs corresponding to annotated agotron or mirtron pre-miRNAs was higher in plasma (22.8%) than in cells (1.7-2.6%; $p < 1\times10^{-15}$ by Fisher's exact test for all cell types tested; Fig. 1B), suggesting preferential cellular export and/or greater resistance of these classes of FLEXIs to plasma RNases (20). Forty-three of the FLEXIs in the cellular RNA samples encoded a snoRNA, all which were also detected as processed mature snoRNAs in the same samples (Fig. 1B).

In Integrated Genomics Viewer (IGV) alignments of reads mapping to protein-coding genes, cellular FLEXIs stood as blocks of continuous intron reads that began and ended within 3 nt of annotated splice sites (Fig. 2A). The IGV alignments for 4I_JUP, an annotated agotron (19), provided an example of an excised intron RNA for which a high proportion (94%) of the reads satisfied the conservative definition of a FLEXI in whole-cell RNAs from HeLa S3 cells, but had 3' truncations in the other cellular RNA samples that resulted in only small proportions of intron reads in those samples (2-3%) being counted as FLEXIs (Fig. 2A, top left). FLEXI reads showed no indication of impediments at or near a branch-point nucleotide as might be expected for a lariat RNA (Fig. 2A and B).

Most FLEXIs (8,082 of 8,144) identified in the 4 cellular RNA samples and plasma had sequence characteristics of major U2-type spliceosomal introns (8,004 with GU-AG splice sites and 78 with GC-AG splice sites); 36 had sequence characteristics of minor U12-type spliceosomal introns (34 with GU-AG and 2 with AU-AC splice sites); and 26 had non-canonical splice sites (*e.g.*, AU-AG and AU-AU; Fig. 1C and S4) (28, 29). The splice-site and branch-point (BP) consensus sequences of the identified FLEXIs were similar to the literature consensus sequences of canonical human introns (28, 30) as well as those of other short and long introns in

the same TGIRT-seq datasets, with no indication of enrichment for a CC branch-point sequence characteristic of the major class of trimmed lariat introns (Fig. 1C for U2-type FLEXIs and Fig. S4A for U12-type FLEXIs) (10, 12, 30). None of the identified FLEXIs corresponded to a previously identified trimmed lariat intron (10, 12); was present in databases of circular RNAs generated by backsplicing (circRNADb, http://reprod.njmu.edu.cn/circrnadb; circBase, http://www.circbase.org); or was identified by CIRI2 (31) as having back-spliced circular RNA junctions in TGIRT-seq datasets of chemically fragmented UHRR (32). RT-qPCR assays for six relatively abundant FLEXIs confirmed that they were sensitive to 5'- and 3'-exonuclease digestion, whereas a synthetic FLEXI RNA circularized *in vitro* was resistant to digestion by these nucleases (Fig. S5A and B, Supplementary file).

A histogram of the length distribution of all reads that mapped to FLEXIs in the 4 cellular RNA samples showed that the predominant bin corresponded to 90-100% intron length (Fig. 2C, pink bars). By contrast, the reads that mapped to other short introns (≤300 nt) annotated in GRCh38 corresponded to heterogeneously sized RNA fragments distributed across the length of the intron, as expected for rapid degradation of canonical excised intron RNAs (Fig. 2C, blue bars). 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. 2D for all FLEXIs and Fig. S3B for FLEXIs in each cellular RNA samples).

Density distribution plots showed that the FLEXIs identified in cells and plasma differed from those of other short intron in being skewed toward higher GC content and more stable predicted secondary structures (*i.e*., lower MFEs for the most stable secondary structure predicted by RNA fold; $p \leq 0.001$ for both by Wilcoxon signed-rank test; Fig. 1D). These differences were most pronounced for the FLEXIs found in plasma, which were a relatively homogeneous subset with peaks at 90-nt length, 70% GC content, and -40 kcal/mol MFE (Fig. 1D), suggesting preferential export or greater resistance to plasma nucleases of shorter, more stably structured FLEXIs. Similar length distributions were found for different categories of FLEXIs in each of the 4 cellular RNA samples, except those with a binding site for DICER (see below), which showed a secondary peak at ~90% intron length, likely corresponding to DICER cleavage products or intermediates (Fig. S3C).

Collectively, the above findings show that human cells contain large numbers of short and in most cases structured FLEXIs, only a small subset of which correspond to annotated agotrons or mirtron pre-miRNAs (16, 19). Although hydrolytic splicing, which would directly produce an

excised linear intron RNA, cannot be excluded, the largely canonical branch-point sequences in FLEXIs suggest that most if not all were excised as lariat RNAs and debranched after splicing, as found previously for mirtron pre-miRNAs and agotrons (17-19).

**Abundance of FLEXI RNAs**

Density plots of the relative abundance of different FLEXIs in the cellular RNA samples showed separate peaks for lower and higher abundance FLEXIs (<0.002 RPM and ≥0.002 to 0.1 reads per million (RPM) with a tail extending to 6.9 RPM; Fig. 1E). The latter overlapped the lower end of the abundance distribution for snoRNAs (dashed yellow line; Fig. 1E). Lower and higher abundance FLEXIs (dashed and solid black lines, respectively) had similar characteristics (GC contents and MFEs) that differed from those of other short introns (Fig. 1D), suggesting that low abundance FLEXIs were not simply intron RNA turnover intermediates. 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 cellular RNA samples (Fig. 1E). 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 (21) (Fig. S6A; Table S2). This method indicated that moderately abundant FLEXIs (0.01-0.99 RPM) were present at 15 to 239 copies per cell and that the most abundant FLEXIs (≥1 RPM) were present at 363 to 1,863 copies per cell. Based on the linear regression model, FLEXIs corresponding to annotated agotrons were present at up to 1,335 copies per cell; those corresponding to mirtron pre-miRNAs at up to 225 copies per cell; and those encoding snoRNAs at up to 265 copies per cell. Copy number 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 it was used as the standard (Fig. S6B).

**Abundant FLEXI RNAs have cell-type specific expression patterns**

Scatter plots identified subsets of relatively abundant FLEXIs (≥0.01 RPM) that were differentially expressed between two compared human cell lines (Fig. 3A, upper panels) and reproduci-

bly detected at similar abundances in biological replicates of the same cell line (Fig. S7A). The cell-type specific expression patterns of FLEXI RNAs were confirmed by principal component analysis (PCA), t-SNE (33) and ZINB-WaVE (34), which showed clustering of FLEXI expression profiles in TGIRT-seq datasets for unfragmented UHRR and whole-cell RNAs from 5 different human cell lines (HEK-293T, HeLa S3, K-562, MDA-MB-231, and MCF7) for which we had additional biological and technical replicates from this or previous studies (Fig. S8A, Table S1). Notably, the differential expression patterns of FLEXIs were significantly more discriminatory between cell types than those for all RNAs that mapped to the same hosts genes (Fig. 3A, bottom panels), even after down sampling of host genes reads to match the sequencing depth of FLEXI reads (Fig. S7B upper and lower panels; $p<0.0001$ by Fisher's r-to-z transformation in both cases). The clustering of cell-type specific expression patterns appeared to improve slightly when restricted to more abundant FLEXIs detected at ≥0.01 RPM, but not further with higher abundance cutoffs, likely reflecting that clustering was already biased toward higher abundance FLEXIs (Fig. S8B). The more discriminatory expression patterns of FLEXI RNAs, a desirable characteristic for cellular RNA biomarkers, likely reflects that their abundance can be impacted post-transcriptionally by different mechanisms, including alternative splicing.

**Evolutionary conservation of FLEXIs**

PhastCons scores provide a measure of evolutionary conservation, with higher phastCons scores indicating a higher probability that a sequence was conserved across a range of organisms (35). Density plots of phastCons scores for FLEXIs calculated across 27 primates, including humans, plus mice, dogs, and armadillos, showed that most FLEXIs, including those corresponding to annotated agotrons and mirtron pre-miRNAs, 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 than other FLEXIs, likely reflecting conservation of the encoded snoRNA (Fig. 1F, yellow line). The relatively low phastCons scores of most FLEXIs could reflect more recent acquisition and/or more rapid sequence divergence in the human lineage.

Although most FLEXIs were poorly conserved, 158 FLEXIs that did not encode a snoRNA, had phastCons scores ≥0.75, possibly reflecting an evolutionarily conserved, sequence-dependent function (Fig. 3B). 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 an alternatively spliced retained intron (47% compared to 25%), or corresponded to an alternatively spliced retained intron with an in-frame protein-coding sequence (20% compared to 7.8%; Fig. 3B). The proportion of FLEXIs with these characteristics was even higher for the 44 most highly conserved FLEXIs (phastCons scores ≥0.99; Fig. 3B). These findings suggest that in most cases the sequence conservation of FLEXIs with high phastCons scores reflects evolutionary constraints on the sequences of different protein isoforms produced by alternative splicing.

**FLEXI RNAs have binding sites for RBPs with diverse cellular functions**

To explore possible biological functions of FLEXIs, we searched for RBPs that bind FLEXIs in ENCODE eCLIP datasets for K-562 and HepG2 cells (36) and in AGO1-4 and DICER PAR-CLIP datasets for HEK-293 cells (37, 38). 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 (Fig. S9A), with 53 of these RBPs (51 in the eCLIP datasets plus DICER and AGO1-4 in the PAR-CLIP datasets) binding ≥30 different FLEXIs (Fig. 4A).

The 53 RBPs that bind ≥30 different FLEXI RNAs included 6 core spliceosomal proteins (PRPF8, SF3B4, AQR, EFTUD2, BUD13, and PPIG) 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 other cellular processes, including transcriptional regulation, chromatin assembly and disassembly, cellular growth regulation, and stress responses (Fig. 4A; protein functions described in Table S3). Examination of sequence reads in the ENCODE eCLIP datasets showed that all of the RBPs that bind ≥30 different FLEXIs were cross-linked to reads corresponding to ≥95% of the length of the identified FLEXI, with 32 showing a ≥2-fold enrichment of such reads compared to those in control antibody datasets (Fig. S10 and Supplementary file). Analysis of ENCODE and other published datasets found that many of the 53 RBPs that bind ≥30 FLEXIs were bound at pre-mRNA splice sites with subsets impacting alternative splicing of the bound FLEXI or adjacent exons (referred to below as proximate alternative splicing) and/or FLEXI host gene mRNA levels, when knocked down by shRNAs or siRNAs (Fig. 4A, RBP Interactions Grid). Notably all of the FLEXI-binding proteins in Fig. 4A were found in published datasets to be present in the cytoplasm, nucleolus, or cellular condensates (PML bodies, stress granules, or nuclear speckles) in addition to or instead of the nucleus (Fig.

4A, RBP Localization Grid based on ENCODE immunofluorescence data and published findings for AGO1-4 and Dicer (39-43)). The host genes encoding FLEXIs with binding sites for the 53 RBPs that bind ≥30 FLEXIs had a wide range of enriched Gene Ontology (GO) terms for nuclear and cytoplasmic functions (Fig. 4B).

Most of the 53 proteins that bind ≥30 different FLEXIs also had annotated binding sites in hundreds of other short introns and thousands of long introns, respectively (Fig. S9B and C). Scatter plots identified 7 RBPs whose binding sites were relatively abundant in FLEXIs (≥1% of the total) and significantly enriched in FLEXIs compared to both other short and long introns in a merged dataset for the 4 cellular samples (adjusted $p \leq 0.05$; calculated by Fisher's exact test and adjusted by the Benjamini-Hochberg procedure; Fig. 5). These included 4 core spliceosomal proteins (PRPF8, SF3B4, EFTUD2, and BUD13), the miRNA-related proteins AGO1-4, DICER, and RBM22, a spliceosome component that functions in nuclear-cytoplasmic trafficking of ALU-7 and hSlu7 proteins under stress conditions (44) (Fig. 5A and B). Compared to long introns, FLEXIs were also significantly enriched in binding sites for DDX24, a DEAD-box protein that negatively regulates RIG-I and RIG-I-like receptors by sequestering cytosolic activator RNAs (45), and 3 proteins that interact with cytoplasmically localized back-spliced circular exon RNAs: GRWD1, a histone-binding protein that functions in ribosome assembly and regulates chromatin dynamics; UCHL5 (ubiquitin carboxy-terminal hydrolase) (46); and SF3A3 (Splicing Factor 3a subunit 3), whose interaction with a circSCAP activates TP53 (47) (Fig. 5B). The scatter plots also identified RBPs whose binding sites were over-represented in other short or long introns compared to FLEXIs, with larger numbers and greater differences for RBPs with binding sites in long introns (Fig. 5A and B).

**Identification of FLEXIs enriched in binding sites for non-spliceosomal RBPs**

Although FLEXIs as a whole were enriched in CLIP-seq-identified binding sites for 6 core spliceosomal proteins compared to other RBPs (Fig. 5A and B), scatter plots readily identified subsets of FLEXIs that were significantly enriched in binding sites for other subsets of RBPs (≥2% of RBP-binding sites; adjusted $p \leq 0.05$ calculated by Fisher's exact test and adjusted by the Benjamini-Hochberg procedure; Fig. 6). These included: (i and ii) FLEXIs annotated as agotrons (n=61) or pre-miRNAs of annotated mirtrons (n=110), which were significantly enriched in binding sites for AGO1-4 and DICER (Fig. 6A and B); (iii) snoRNA-encoding FLEXIs (n=43),

which were significantly enriched in binding sites for snoRNA-related proteins DKC1 (Dyskerin) and NOLC1 (Nucleolar and Coiled Body Phosphoprotein 1, alias Nopp140) (48, 49); SMNDC1 (Survival of Motor Neuron Domain Containing 1, alias SPF30), better known for binding snRNAs; and AATF (Apoptosis Antagonizing Transcription Factor, as well as several other proteins (Fig. 6C); and (iv) highly conserved, FLEXIs (phastCons scores ≥ 0.75, n=158), which were significantly enriched in binding sites for splicing factors U2AF1 and U2AF2, the splicing regulator TIAL1, ZNF622, which functions in cytoplasmic maturation of 60S ribosomal subunits (50), and the transcription factor BCLAF1 (Fig. 6D). The FLEXIs in these 4 subsets differed from FLEXIs as a whole in length, GC content, and/or MFE for the most stable secondary structure predicted by RNAfold (Fig. 6A-D, right hand panels). AATF, which we found enriched in subsets of FLEXIs encoding snoRNAs, was found previously to bind to both DKC1 mRNA and snoRNAs, potentially linking the ribosome biosynthesis function of snoRNAs to the regulation of cell proliferation and apoptosis by suppressing TP53 activity (51), and BCLAF1, which we found significantly enriched in conserved FLEXIs, was previously suggested to link RNA splicing to transcriptional and other cell regulatory pathways (52, 53).

**Computational identification of FLEXIs co-enriched in binding sites for different subsets of RBPs**

Encouraged by the above findings, we used hierarchical clustering as described in Materials and Methods to computationally identify subsets of RBPs whose binding sites were co-enriched in different subsets of FLEXIs without *a priori* assumptions about the function of the protein or bound FLEXIs. This approach identified 6 clusters of RBPs whose binding sites were significantly co-enriched (relative abundance ≥2%, adjusted p≤0.05) in the same subsets of FLEXI in all or some of the 4 human cell lines (Fig. 7A and Fig. S11). Scatter plots confirmed the enrichment of binding sites for each subset of RBPs relative to those for core spliceosomal proteins in each cluster of FLEXIs (Fig. 6E-J), and scatter plots for each individual RBP in each cluster confirmed co-enrichment of binding sites for the other RBPs in the same cluster (Fig. S12).

Clusters I and II had the largest numbers of FLEXIs and were the most revealing of novel RBP-FLEXI associations. Cluster I was comprised of 386 FLEXIs with two RBPs (LARP4 and PABPC4) that satisfied the relative abundance and significance criteria for association with the cluster in all 4 cell lines and 6 other RBPs (Transcription Regulator SUB1, DDX3X, RPS3, NCBP2, DDX55, and Methionine Aminopeptidase 2 (METAP2)) that satisfied these criteria in

HEK-293T cells but not in some of the other cell lines (Fig. 7A). LARP4 and PABPC4 function in regulating the stability and translation of mRNAs (54, 55), and the other 6 proteins in this cluster either have mRNA synthesis or translation-related functions or could link these functions to other cellular regulatory processes (Table S3). Analysis of ENCODE eCLIP and RBP-knockdown datasets showed that all of the proteins in Cluster I bind at splice sites with knockdown of 5 of the 8 proteins impacting proximate alternative splicing of the bound FLEXI (retained intron (RI) or skipped adjacent exons (SE; Fig. 7A, RBP Interactions Grid; based on data in Figs. S13-15). All but one of these proteins (SUB1) were annotated as being present in the cytoplasm in addition to or instead of the nucleus, with 5 of the 8 proteins also annotated as being present in stress granules (Fig. 7A, RBP Localization Grid).

Cluster II was comprised of 999 FLEXIs with 5 proteins (BCLAF1, UCHL5, ZNF622, TRA2A, ZNF800) satisfying the abundance and significance criteria for association with this cluster in all 4 cell lines, two (GRWD1 and PUM1) satisfying these criteria in 3 of the 4 cell lines, and two others (DDX24 and FXR2) satisfying these criteria in only one of the cell lines (Fig. 7A). All of the proteins associated with this cluster function in regulation of transcription or RNA splicing (Table S3) and bind at splice sites, with knockdown of two of the proteins impacting proximate alternative splicing of the bound FLEXI (Fig. 7A and Fig. S13-15). All but one of these proteins were annotated as being localized in the cytoplasm in addition to or instead of the nucleus, the sole exception being ZNF800, whose cellular localization was not available in ENCODE.

Clusters III and IV were comprised of 100 and 427 FLEXIs, respectively, that were enriched in binding sites for the splicing regulators TIA1 and TIAL1 (Cluster III) or splicing factors U2AF1 and U2AF2 (Cluster IV), with splicing regulator KHSRP satisfying the abundance and significance criteria for association with Cluster IV in HEK-293T cells (Fig. 7A). Three of these proteins (TIAL1, U2AF1, and U2AF2) bind at splice sites and the other two (TIA1 and KHSRP) bind within the introns, with knockdown of all 5 proteins impacting proximate alternative splicing of the bound FLEXI and/or mRNA levels of FLEXI host genes (Fig. 7A). Notably, in addition to or in one case instead of the nucleus, all 5 proteins in these clusters were annotated as being present in nuclear speckles and/or stress granules condensates (56, 57), possibly reflecting post-splicing destinations.

Cluster V was comprised of 160 FLEXIs that were enriched in binding sites for the 4 proteins that were identified in Figure 6C as binding snoRNA-encoding FLEXIs (DKC1, NOLC1,

AATF, and SMNDC1), with the first 3 proteins satisfying the relative abundance and significance criteria for association with this cluster in all 4 of the cell lines and SMNDC1 satisfying these criteria in 2 of the cell lines (Fig. 7A). Only 43 of the 160 FLEXIs in this cluster encoded an annotated snoRNA. All 5 of the RBPs in this cluster bind within the intron with their knockdown affecting proximate alternative splicing of the bound FLEXI (Fig. 7A).

Finally, Cluster VI was comprised of 479 relatively GC-rich FLEXIs that were enriched in binding sites for AGO1-4 and DICER (Fig. 7A). Only 23 of the 250 FLEXIs in this cluster with a binding site for AGO1-4 corresponded to an annotated agotron (19), and only 44 of the 308 FLEXIs in this cluster with a binding site for DICER corresponded to a pre-miRNA of an annotated mirtron (58). The remaining 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 were derived from or could evolve into agotrons or mirtrons, an evolutionary scenario suggested previously for mirtrons (18). Analysis of PAR-CLIP datasets showed that DICER binding sites were localized within FLEXIs, while those for AGO1-4 were at splice sites (Fig. 7), potentially enabling regulation of host gene RNA splicing (59, 60). Although we continue to focus below on the subsets of RBPs that met the criteria for hierarchical clustering in Clusters I-VI, heat maps showed that numerous other RBPs had overlapping binding sites in FLEXIs belonging to these clusters, suggesting more complex regulatory networks to explore in subsequent studies (Fig. S16).

**Clusters of FLEXI RNAs identified by co-enrichment of RBP-binding sites originate from host genes with related biological functions**

To assess if the host genes encoding FLEXIs in each of the 6 identified clusters might be functionally related, we analyzed GO terms for host genes encoding FLEXIs that bind each RBP in each cluster using DAVID bioinformatic tools and displayed the results as a heat map (Fig. 7B). The heat map showed that enriched GO terms for FLEXI host genes within each of the 6 clusters were more congruent with each other than those for the full set of host genes for FLEXIs with binding sites for 53 RBPs that bind ≥30 FLEXIs in Fig. 5B. The host genes for FLEXIs bound by AGO1-4 and DICER (Cluster VI) were enriched in GO terms for cytoplasmic translation and focal adhesion, while those for FLEXIs in Clusters II, III, and IV were enriched in GO terms for RNA splicing, mRNA processing and nuclear speckles, with some host genes for FLEXIs in Clusters I, II, IV, and VI enriched in a GO term for protein binding. The host genes for FLEXIs

bound by the snoRNA-related proteins in Cluster V have functions related to cytosolic ribosomes and cytoplasmic translation, with the host genes of FLEXIs bound by NOLC1 also involved in focal adhesion.

Notably, enriched GO terms for the host genes of FLEXIs in Cluster I included cytosolic ribosome and cytoplasmic translation, which were also enriched in host genes of the snoRNA-related FLEXIs in Cluster V (Fig. 7B). Consistent with these GO terms, we found that the Cluster I and V FLEXIs were enriched in host genes encoding cytoplasmic ribosomal proteins (9.6% and 11.2% of FLEXI host genes, respectively, compared to 1.6% of all FLEXI host genes; $p<1x10^{-11}$ by Fisher's exact test; incorporated into more detailed analysis below). These findings raised the possibility that Clusters I and V RBPs function in parallel pathways for regulating host genes involved in ribosome synthesis and translation. Collectively, the GO term analysis indicated that the host genes for FLEXIs that bind the same subset of RBPs are functionally related and might be coordinately regulated by these RBPs.

**Host genes for FLEXIs are enriched in other short and long introns with binding sites for same subsets of RBPs**

All of the RBPs identified as having annotated binding sites in FLEXIs also had annotated binding sites in other short and long introns (Fig. S9). UpSet plots showed that most FLEXIs containing binding sites for the RBPs in Clusters I to VI were enriched in host genes that also contained other short or long introns with binding sites for the same RBPs, while the majority of other short introns and the vast majority of long introns were present in host genes that lacked FLEXIs with binding sites for Cluster I to VI RBPs (Fig. 8A). Consistent with these findings, hierarchical clustering, done similarly to that for FLEXIs using 10 subsets of 2,000 randomly selected other short or long introns, largely recapitulated Clusters I to VI for FLEXIs with some simulations finding additional RBPs associated with these clusters, as well as additional clusters of other short and long introns co-enriched in binding sites for other RBPs (Clusters VII to X; Fig. 8A). The host genes for Clusters I and V identified by hierarchical clustering of other short or long introns included those with binding sites for the same RBPs identified by hierarchical clustering for FLEXIs and were likewise enriched in GO terms for cytoplasmic translation and cytosolic ribosome (Fig. 9, left and right heat maps, respectively, compare with Fig. 7B). Analysis of CLIP-seq and knockdown datasets showed that in most cases the Cluster I to VI RBPs bound other short and long introns at similar locations at splice sites or within the introns as they did for

FLEXIs, with knockdown of some RBPs impacting proximate alternative splicing of the bound intron or host gene mRNA levels (Fig. S13-15). Collectively, these findings indicate that different subsets of host genes are enriched not only in FLEXIs but also other short introns and long introns that bind similar subsets of RBPs, potentially reinforcing coordinate regulation by these RBPs.

**The intracellular localization of FLEXIs correlates with that of their CLIP-seq-identified RNA-binding proteins**

Most of the RBPs associated with Clusters I and II were annotated as being present in both the cytoplasm and nucleus, with a number of these proteins having known cytoplasmic functions (Fig. 7). To investigate the intracellular location of FLEXIs with binding sites for different RBPs, we used TGIRT-seq to analyze RNAs in nuclear and cytoplasmic fractions prepared from 4 different cell lines (K-562, HeLa S3, MDA-MB-231, and MCF7; Fig. 10). The breast cancer cell lines MDA-MB-231 and MCF7 were included in this analysis to assess the generality of conclusions based on the initially analyzed cellular RNA samples, as well as the potential utility of FLEXIs as breast cancer biomarkers. TGIRT-seq of MDA-MB-231 and MCF7 whole-cell RNAs identified 1,288 additional FLEXIs that were not detected in any of the previously analyzed cellular RNA samples (501 in MCF7, 599 in MDA-MB-231, and 188 in both MCF7 and MDA-MB-231), including cell-type specific FLEXIs from oncogenes and tumor suppressor genes (Fig. S17 and Supplementary file). The clean separation of nuclear and cytoplasmic fractions from the 4 cell lines was confirmed by scatter plots comparing different RNA biotypes in each fraction from each cell type (Fig. S18). PCA analysis and a sample distance heat map showed that the TGIRT-seq datasets clustered mainly by cell type, with nuclear, cytoplasmic, and whole-cell RNAs from each cell type forming separate but closely spaced subclusters (Fig. S19).

Fig. 10A and B show scatter plots for each of the 4 cell lines comparing the relative abundance of CLIP-seq identified RBP-binding sites in FLEXIs enriched in the nuclear or cytoplasmic fractions (fold change (FC)>1.5 via DESeq2 analysis, denoted nuclear or cytoplasmic FLEXIs, respectively). The percentages and numbers of different FLEXIs with a binding site for each RBP associated with Clusters I to VI in the nuclear, cytoplasmic or both fractions are shown in Fig. 11. Both Figures show that cytoplasmic FLEXIs in all 4 cell lines were enriched in binding sites for each of the 8 proteins associated with Cluster I (name labeled red points in scat-

terplots). These included not only proteins with known cytoplasmic functions (LARP4, PABPC4, RPS3, and METAP2), but also DDX3X, which functions in nuclear RNA export (59-61), as well as NCBP2 (nuclear cap binding protein 2) and SUB1 (a transcriptional regulator). Cytoplasmic FLEXIs in all 4 cell lines also appeared enriched in those with binding sites for AGO1-4 and DICER, consistent with the functions of agotrons and mirtrons (Figs. 10A and B, unlabeled brown points, and Fig. 11).

The FLEXIs with binding sites for Cluster II-associated proteins (name labeled orange points in scatter plots) had more variable distributions. Those with binding sites for PUM1, FXR2, and ZNF800 had the highest degree of enrichment in cytoplasmic FLEXIs; those with binding sites for UCHL5 and GRWD1 were more equally distributed between cytoplasmic and nuclear FLEXIs; and those with binding sites for BCLAF1, ZNF622, DDX24, and TRA2A had the highest degree of enrichment in nuclear FLEXIs (Figs. 10 and 11). FLEXIs with binding sites for RBPs associated with Clusters III, IV, and V (light blue, purple and dark blue points in scatter plots) were enriched nuclear FLEXIs (Figs. 10 and 11).

The scatter plots also showed that cytoplasmic FLEXIs were enriched in binding sites for 9 other proteins with functions consistent with their presence in the cytoplasm (36, 62) in one or more of the cell lines (Fig. 10A and B; Table S3). These include 7 proteins that function in nuclear RNA export (AKAP8L, GRSF1, RBM15, SRSF7, UPF1, XPO5, and ZC3H11A) plus SERBP1, which stabilizes inactive ribosomes and is involved in PML-body formation, and YBX3, a non-specific RNA-binding protein and homolog of YBX1, which functions in exosomal packaging of sncRNAs (26).

In contrast to FLEXIs, other short or long introns were comprised of RNA fragments. Scatter plots comparing the distribution of binding sites in RNA fragments of other short and long introns in the nuclear and cytoplasmic showed at most moderate enrichment in one or the other fraction in most cases (Fig. 10C and D). A notable exception was DDX3X, whose binding site were significantly enriched in RNA fragments of other short introns in the cytoplasmic fraction from all 4 cells lines (Fig. 10C). RNA fragments of long introns in the cytoplasmic fraction were significantly enriched in binding sites for the RNA export protein QKI in all 4 cell lines, while RNA fragments with binding sites for the RNA export protein NONO were significantly enriched in the nuclear fraction from 3 of the 4 cell lines (Fig. 10D).

Stacked bar graphs based on CLIP-seq datasets showed that the majority of individual FLEXIs in Clusters I to VI have a binding site for a single RBP associated with that cluster (gray

bars), with smaller percentages having overlapping binding sites for two or more RBPs associated with the same cluster (red and blue bars, respectively) and low percentages having non-overlapping binding sites for 2 or more RBPs associated with the same cluster (orange and purple, respectively; Fig. 12A). A precedent for this situation is the alternative splicing factor RBFOX2, which binds at some splice sites via a primary conserved sequence motif and at others by binding at secondary sites or by interacting with partner proteins bound at those sites (63). Consistent with the latter mechanism, the STRING protein-network database confirmed interactions between nearly all of the proteins associated with each of the hierarchical clusters (Fig. 12B).

**Discussion**

In total for all datasets in this study, we identified 9,429 different FLEXIs expressed from 4,224 host genes, representing ~18% of the 51,645 short introns (≤300 nt) annotated in Ensembl GRCh38 Release 93. FLEXIs were skewed toward higher GC content and more stable predicted secondary structures than those of other short introns (Fig. 1D), likely contributing to the higher stability of FLEXIs in cells and plasma. Subsets of relatively abundant FLEXIs had more discriminatory cell-type specific expression patterns than other RNAs mapping to the same host gene in the same datasets (Fig. 3 and Fig. S7), a desirable feature for RNA biomarkers as discussed further below. The conserved branch-point A residues of FLEXIs (Fig. 1C) suggests that they splice via a lariat intermediate, but does not preclude that they can also splice via a hydrolytic mechanism, as found *in vivo* for certain group II intron RNAs, which also have a conserved branch-point A residue (64).

By searching CLIP-seq datasets, we identified 126 different RBPs that bind FLEXIs, including 53 RBPs that bind ≥30 different FLEXIs (Fig. 4A). These 53 RBPs included not only proteins that function in RNA splicing, but also AGO1-4 and DICER and other proteins that have miRNA-related functions, and a surprising number of proteins that function in other cellular processes, including transcription (AATF, BCLAF1, and SUB1), chromatin remodeling (GRWD1), translation (LARP4, METAP2, PABPC4, RPS3), ribosome maturation (ZNF622), and cellular growth regulation (DDX3X and IGF2BP1; Fig. 4A and Table S3). In addition to being present in the nucleus, all 53 RBPs that bind ≥30 different FLEXIs were also annotated in published datasets as being present in the cytoplasm and/or nuclear or cytoplasmic condensates (Fig. 4). Additionally, a number of RBPs that bind FLEXIs function in nuclear RNA export (BUD13,

NCBP2, PABPC4, PABPN1, RBM15, and U2AF2 among the 53 RBPs that have annotated binding sites in ≥30 FLEXI RNAs in Fig. 4A, and AKAP8L, GRSF1, HNRNPA1, HNRNPC, HNRNPUL1, ILF3, KHDRBS1, NONO, QKI, SRSF7, UPF1, XPO5, ZC3H11A among the remaining RBPs that bind smaller numbers of FLEXIs in Fig. S9). Bound RBPs may contribute to the stability of FLEXIs in cells and plasma as well as shielding against cellular inflammatory responses to structured cytoplasmic RNAs.

By using hierarchical clustering, we identified 6 clusters of FLEXIs (Clusters I to VI) that were significantly co-enriched in binding sites for 6 clusters of RBPs, including regulatory proteins and RBPs whose knockdown affected alternative splicing and/or host gene mRNA levels (Fig. 7). We then found that other short and long introns formed hierarchical clusters similar to Clusters I to VI for FLEXIs and that the host genes for FLEXIs in Clusters I to VI were enriched in other short and long introns with binding sites for the same subsets of RBPs (Fig. 8). Collectively, these findings suggest that different subsets of functionally related host genes are coordinately regulated by different subsets of RBPs whose binding sites are enriched in different classes of introns within that gene set. Particularly suggestive was Cluster I, which included two proteins, LARP4 and PABPC4, that function in regulating the stability and translation of mRNAs (54, 55), along with additional proteins that have translation-related functions (NCBP2, METAP2, and RPS3), or that could link these functions to other cell regulatory pathways (SUB1, RPS3, DDX3X, and DDX55).

Notably, the host genes encoding the FLEXIs in Cluster I were enriched in 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 in parallel to snoRNAs in regulating ribosome synthesis (Fig. 7B). Along the same lines, Cluster II included 3 proteins (UCHL5, ZNF622, and TRA2A) that were distinguished from other proteins associated with the FLEXI-RBP clusters as being localized to the endoplasmic reticulum, cytoskeleton, and cell cortex, possibly reflecting a cytoskeleton or membrane-associated function. The RBPs associated with Clusters III and IV, which function in alternative splicing, were found localized not only in the nucleus but also in stress granules and nuclear speckles, the former possibly a mechanism for muting inflammatory responses to naked cytoplasmic RNAs, and the latter possibly a mechanism for storing alternative splicing factors for subsequent use (65). Although the physiological significance of CLIP-seq identified RNA-protein interactions requires further investigation, the specificity of the CLIP-seq-identified binding sites that formed the basis for our analysis is supported by the findings that host genes

for FLEXIs with binding sites for RBPs in Clusters I to VI were co-enriched with other short and long introns with binding sites for the same RBPs (Fig. 8); that knockdown of subsets of RBPs in each cluster impacted alternative splicing and/or host gene mRNA levels; and that interactions between the RBPs in each cluster, suggested by the hierarchical clustering based on RNA-protein interactions, were confirmed independently by protein-interaction network analysis (Fig. 12B).

Lewis *et al*. (66) recently reported that 5 of the 8 RBPs that we identified as belonging to Cluster I (LARP4, PABPC4, DDX3X, RPS3, and SUB1) have eCLIP-identified binding sites enriched in mRNAs of nuclear-gene encoding mitochondrially localized proteins. They also found that LARP4-binding sites were enriched near the 3' end of mRNAs encoding components of respiratory chain complexes and mitochondrial ribosomal proteins (66). Prompted by these findings, we found that FLEXIs with binding sites for LARP4 and other Cluster I RBPs were also enriched in host genes encoding mitochondrially localized proteins annotated in MitoCarta3.0 (8.3 to 25.5% of FLEXIs host genes compared to 7.5% of all FLEXI host genes and 5.7% of all GRCh38 annotated protein-coding genes; Fig. S20A and B). Extending these findings, we found that FLEXIs with binding sites for LARP4 and other Cluster I RBPs were also enriched in host genes encoding cytoplasmic ribosomal proteins (8.3 to 42.6%), consistent with enrichment for the GO terms cytoplasmic ribosome and cytoplasmic translation for FLEXI host genes in this cluster (Fig. 7A). Some but not all FLEXIs with binding sites for Cluster I RBPs were enriched in genes encoding mitochondrial ribosomal proteins (DDX3X, 6.8%; LARP4, PABPC4, RPS3, 1.9-2.2% compared to 0% for the other 4 RBPs), with similar trends holding for other short and long introns with binding sites for the same RBPs (Fig. S20A). Consistent with the findings of Lewis *et al.* (66), we found that binding sites for LARP4 and Cluster I proteins PABPC4, DDX55 and SUB1 were localized near the 3' end of mRNAs (defined as UTRs + coding sequences (CDS)). Extending these findings, we found that binding sites for these 4 proteins were also localized to 3'-proximal introns, suggesting that they bind initially to unspliced pre-mRNAs and that their localization, possibly as a complex (Fig. 12B), might be dictated by proximity to 3'-poly(A) tails (Fig. S20C). METAP2- and RPS3-binding sites were more uniformly distributed across the length of the mRNAs, but remained concentrated in 3'-proximal introns, while DDX3X- and NCBP2-binding sites were localized near the 5' end of mRNAs and in 5'-proximal introns, with DDX3X still showing a smaller peak for 3'-proximal introns (Fig. S20C). Considered together, these findings suggest that the RBPs that bind FLEXIs in Cluster I function in at least two overlapping regulatory pathways, one for cytoplasmic translation and cy-

tosolic ribosomes parallel to that for snoRNAs, and the other for mitochondrial function, both potentially related to regulation of cell growth and proliferation.

Collectively, our findings suggest a model in which newly synthesized or excess RBPs that are not associated with their primary physiological substrates enter the nucleus where they bind FLEXIs and other short or long introns in pre-mRNAs from subsets of genes with related functions, contributing to the regulation of different gene sets. After splicing, the RBPs remain stably bound to FLEXIs and co-localize with them to the cytoplasm or other intracellular locations, where they may dissociate from FLEXIs to perform different cellular functions. As these RBPs include transcription factors and nuclear export proteins as well as RNA splicing factors, co-regulation of gene sets by these RBPs could occur at multiple levels. After splicing, FLEXIs may serve simply as vehicles that stabilize RNA-binding proteins until they reach an appropriate location to perform their cellular functions or could be actively engaged in those functions. Although only 6 subsets of RBPs met the arbitrary significance criteria for co-enrichment in the same subsets of FLEXIs, a heat map of overlapping RBP-binding sites identified multiple additional RBPs whose binding sites overlapped those of RBPs within each cluster of FLEXIs (Fig. S16), possibly enabling cross talk between different gene sets. Collectively, these findings suggest overarching mechanisms for communicating and coordinating regulation of different cellular processes.

Finally, relatively abundant FLEXIs that have cell-type specific expression patterns may have immediate utility as cellular RNA biomarkers. Besides being linked to the expression of thousands of protein-coding and lncRNA genes, FLEXI RNA levels reflect differences in alternative splicing and intron RNA stability, potentially affording higher resolution of cellular differences than mRNAs transcribed from the same gene. Because most FLEXIs are predicted to fold into stable secondary structures, their initial identification may be best done by TGIRT-seq or other methods using related enzymes that yield full-length, end-to-end sequence reads of structured RNAs. Once identified, the validation of candidate FLEXI biomarkers and their routine monitoring would best be done by methods such as RT-qPCR, microarrays or other hybridization-based assays, or targeted RNA-seq that enable high-throughput quantitative read outs of panels of RNA biomarkers 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, PCR-product sequencing files, and scripts used for data processing and plotting have been deposited in GitHub: https://github.com/reykeryao/FLEXI.

## COMPETING INTEREST STATEMENT

AML is an inventor on patents owned by the University of Texas at Austin for TGIRT enzymes and other stabilized reverse transcriptase fusion proteins and methods for non-retroviral reverse transcriptase template switching. AML, JY, and HX are inventors on a patent application filed by the University of Texas for the use of FLEXIs and other intron RNAs as biomarkers.


## Acknowledgements

We thank Marta Mastroianni for cell culture and RNA preparations, Shelby Winans Bretton for early research contributions, Camilo Molina Romero and Kavyaa Choudhary for help with validating amplicons for exonuclease and droplet digital PCR analysis, Blerta Xhemalce (University of Texas at Austin) for fresh stocks of MDA-MB-231 and MCF7 cells, Carla Van Den Berg (University of Texas at Austin) for comments on the manuscript and use of droplet digital PCR equipment, and Jim Manley (Columbia), Sandra Wolin (NIH/NCI), and Richard Maraia (NIH/NICHD) and for comments on the current and/or an earlier version of the manuscript. The authors acknowledge the University of Texas at Austin Genomic Sequencing and Analysis Facility (GSAF, Center for Biomedical Research Support, RRID#: SCR_021713, https://wikis.utexas.edu/display/GSAF/Home+Page) for providing RNA-seq services and the Texas Advanced Computing Center (TACC; http://www.tacc.utexas.edu) and Biomedical Research Computing Facility (BRCF; https://site.research.utexas.edu/cbrs/cores/cbb/computing-resource) for providing high performance computing resources. This work was supported by NIH grant R35 GM136216 to AML. MA was supported by NIH grant R01 GM040478.


**Figure 1. Characteristics of FLEXIs in human cells and plasma.** (A) UpSet plots of FLEXI RNAs and their host genes identified by TGIRT-seq of ribo-depleted, unfragmented RNA preparations from the indicated human cell lines and plasma. Different FLEXIs from the same host gene were aggregated into one entry for that gene. (B) Numbers and percentages of different FLEXIs in datasets for each sample type corresponding to an annotated agotron, a pre-miRNA of annotated mirtron, or encoding an annotated snoRNA including subsets of small Cajal body-specific snoRNAs (scaRNAs). "All" indicates the total number of different FLEXIs in each cate-

gory in this group of samples, and GRCh38 indicates the number of annotated short introns (≤300 nt) in each category in the GRCh38 human genome reference sequence. (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 4 cellular RNA samples compared to the literature consensus sequences for human U2-type spliceosomal introns (29, 30) and other Ensemble GRCh38-annotated U2-type short (≤300 nt) and long (>300 nt) introns. The number of introns with 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 more abundant (RPM≥0.002, solid black line) and less abundant (RPM <0.002, dashed black line) FLEXI RNAs in a merged dataset for the 4 cellular RNA samples compared to other GRCh38-annotated short introns (≤300 nt, blue line) in same merged dataset and FLEXIs in plasma (red 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 snoRNAs (dashed yellow line) and the abundance of annotated sncRNAs (gray arrows above the plots) in the same samples are shown for comparison. The vertical dashed red line in each plot separate low abundance (<0.002 RPM) from higher abundance (≥0.002 RPM) FLEXIs. (F) Density distribution plots of phastCons scores for different categories of FLEXIs in the cellular and plasma RNA samples compared to those of other annotated short introns (≤300 nt). PhastCons scores were calculated by the average score across all intron bases from multiple sequences alignment of 27 primates, including humans, in addition to mouse, dog, and armadillo.

**Figure 2. FLEXI RNAs are detected by TGIRT-seq as continuous full-length, end-to-end reads.** (A) Integrative Genomics Viewer (IGV) screenshots showing read alignments for FLEXI RNAs in different cellular RNA samples. FLEXIs are named at the top with an arrow indicating 5' to 3' orientation and the tracks below showing gene annotations (exons, thick bars; introns, thin lines) followed by alignments of FLEXI reads in different cellular RNA samples. The length of the intron that is the focus of each panel is shown above the read alignments and the fraction of FLEXI reads (continuous reads that began and ended within 3 nt of the annotated 5'- and 3'-splice sites) of that intron is shown within the read alignments. (B) IGV screenshots showing examples of cell-type specific differences in abundance of FLEXI RNAs from the same host gene due to differences in splicing efficiency, alternative splicing, and/or differential stability of

FLEXI RNAs. Gene maps for different RNA isoforms generated by alternative splicing of FLEXI RNAs are shown at the bottom with introns from the 5' to 3' RNA end named A to F. Reads mapping to exons or non-FLEXI introns were omitted for clarity. In panels A and B, reads for different cellular RNA samples are color coded as shown in the Figure and were down sampled to a maximum of 100. Mismatched nucleotides in boxes at the 5' end of RNA reads due to non-templated nucleotide addition (NTA) to the 3' end of cDNAs during TGIRT-seq library preparation. Two FLEXIs in panel B had indels at short runs of 3' U residues, likely reflecting a previously reported tendency for TGIRT enzyme slippage at such locations (Fig. 2B) (67). (C) Histogram of the length distribution of reads mapping within introns identified as FLEXIs (red) compared to those mapping within other Ensemble GRCh38-annotated short introns (≤300 nt; blue) in a merged dataset for the 4 cellular RNA samples. Percent length was calculated from TGIRT-seq read spans in the indicated intervals normalized to the length of the corresponding intron. FLEXIs encoding embedded snoRNAs/scaRNAs were excluded to avoid contributions from fully or partially processed sno/scaRNAs. (D) Three-dimensional bar graph showing the percentage of FLEXI reads in a merged dataset for the K-562, HEK-293T, HeLa S3, and UHRR cellular RNA samples ending at different positions around intron-exon junctions.

**Figure 3. FLEXI RNA profiles are cell-type specific and alternative splicing characteristics of conserved FLEXIs.** (A) Scatter plots showing pairwise comparisons of $\log_2$-transformed RPM of FLEXI RNA reads (top row) and reads that mapped to FLEXI host genes (bottom row) between different cellular RNA samples. Relatively abundant FLEXIs ( ≥0.01 RPM) and their host genes are colored red, blue, or purple depending upon whether they were reproducibly detected in all replicates of either or both compared cell types (color code in top right panel). Pearson correlation coefficients (*r*) for all detected FLEXIs and their host genes are shown in black, and those for relatively abundant color-coded FLEXIs and their host genes (red and blue) are shown in purple. (B) Characteristics of conserved FLEXIs in a merged dataset for the K-562, HEK-293T, HeLa S3, and UHRR 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 were alternatively spliced to generate different protein isoforms based on Ensemble GRCh38 annotations. 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, and In frame RI indicates alternatively spliced retained FLEXIs that have reading frames in frame with those of flanking exons.

**Figure 4. FLEXI interactions with RNA-binding proteins.** (A) Bar graph showing the number of different FLEXI RNAs that have a CLIP-seq-identified binding site for each of the 53 RBPs that bind ≥30 different FLEXIs in a merged dataset for the K-562, HEK-293T, HeLa S3, and UHRR cellular RNA samples. An extended version of the bar graph for all 126 RBPs that have a CLIP-seq-identified binding site in a FLEXI RNA is shown in Fig. S9A. Bar graph RBPs are color coded by protein function as shown at the top. The grids below the bar graph show reported RBP Functions (red), Intracellular Localization (green), and FLEXI Interactions and characteristics. The Intracellular Localization Grid is based on a supplemental table in ref. (68), the mammalian RNA granule and stress granule protein databases (56, 69), and the UniProt database (https://www.uniprot.org). The Interaction Grid is based on analysis of PAR-CLIP datasets for AGO1-4 and DICER, ENCODE eCLIP and knockdown datasets for other RBPs, and the characteristics of FLEXIs identified in the cellular RNA datasets. RNA-binding sites: I, within the intron; SS-I and SS-E, across splice sites with the peak of density plots for binding sites localized within the intron or a flanking exon, respectively (Fig. S13). Changes in mRNA level based on ENCODE RBP-knockdown datasets: blue and red, significantly biased toward higher ($\log_2$FC>0) or lower ($\log_2$FC<0) mRNA levels, respectively, compared to all genes; gray, significantly changed mRNA levels, but not significantly biased in either direction (Fig. S14). Alternative Splicing (Alt. Spl.) based on ENCODE RBP-knockdown datasets compared to the control dataset (Fig. S15): RI, significant changes in retained introns; SE, significant changes in skipped exons. Lower GC%, RBPs whose binding sites were significantly enriched in FLEXIs having lower GC content than other FLEXIs (Fig. S12), phastCons scores ≥0.75, RBPs whose binding sites were significantly enriched in evolutionarily conserved FLEXIs with phastCons ≥0.75 (Fig. 6D). Changes in mRNA level and alternative splicing for AGO1-4 and DICER were not available in ENCODE. (B) 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. The color scale shown to the right was based on -$\log_{10}$-transformed false discovery rate (FDR).

**Figure 5. Relative abundance of annotated RBP-binding sites in FLEXIs compared to other short (≤300 nt) or long (>300 nt) introns.** RBPs whose binding sites constituted ≥1% of the total RBP-binding sites and were significantly different between the compared groups (adjusted p≤0.05; calculated by Fisher's exact test and adjusted by the Benjamini-Hochberg procedure) are indicated by the name of the protein and color coded by protein function as indicated at the bottom of the Figure. For each panel, the box within the scatter plot on the left delineates a region that is expanded in the scatter plot on the right. The circle at the bottom right of the scatterplot in panel B encompasses 12 additional RBPs (FAM120A, GTF2F1, HNRNPC, HNRNPK, HNRNPL, HNRNPM, ILF3, MATR3, NONO, PCBP2, SUGP2, and TARDBP) whose binding sites met the abundance and significance criteria for enrichment in long introns compared to FLEXIs but were too numerous to name in the Figure.

**Figure 6. Identification and characteristics of subsets of FLEXIs that were enriched in binding sites for different subsets of RBPs.** Different subsets of FLEXIs are named at the top left with the number of FLEXIs in that subset indicated in parentheses. The leftmost panel shows a scatter plot for CLIP-seq-identified binding sites in the Subset of FLEXIs (y-axis) compared to all Other FLEXIs (x-axis) in a merged dataset for the K-562, HEK-293T, HeLa S3, and UHRR cellular RNA samples, and the right three panels show density plots comparing different characteristics of the subset FLEXIs to other FLEXIs in the same merged datasets. (A) FLEXIs annotated as agotrons; (B) FLEXIs that are pre-miRNAs of annotated mirtrons; (C) FLEXIs encoding an annotated snoRNA. Binding sites for AATF were found only in FLEXIs encoding C/D-box snoRNAs, while those for DKC1, a core H/ACA-box snoRNA-binding protein (70), were found in FLEXIs encoding both H/ACA- and C/D-box snoRNAs. (D) Conserved FLEXIs with phastCons score ≥0.75; (E-J) FLEXIs identified by hierarchical clustering as being enriched in binding sites for different subsets of RBPs (see Fig. 7). RBP-binding site annotations in the scatter plots were based on the ENCODE eCLIP datasets for 150 RBPs (68) and AGO1-4 and DICER PAR-CLIP datasets (37, 38). 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

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 ($p<0.01$ by Kolmogorov–Smirnov test and reproducible in ≥95% of 1,000 Monte-Carlo simulations; see Methods). Scatter plots and density plots for FLEXIs with binding sites for each RBP associated with these clusters are shown in Fig. S12.

**Figure 7. Identification of RBPs whose binding sites were co-enriched in clusters of FLEXI RNAs.** (A) Hierarchical clustering of FLEXI RNAs based on patterns of over- and under-represented RBP-binding sites (see Methods and Fig. S11). Clusters I to VI are delineated at the top of panel, and the names of the RBPs associated with each cluster are shown at the bottom of panel B and were color coded by protein function as indicated at the bottom of the Figure. The RBPs that met the abundance and significance criteria for association with the clusters in HEK-293T, HeLa S3, K-562, and UHRR are indicated by filled boxes at the top. The grids below summarize RBP Functions, Localization, and Interactions with FLEXs based on data sources indicated in the legend of Fig. 4. (B) Heatmap of GO terms enriched in host genes of FLEXI RNAs containing binding sites for cluster-associated RBPs. GO enrichment analysis was performed with DAVID bioinformatics tools, using all FLEXI host genes as the background. The color scale to the right was based on $-\log_{10}$-transformed false discovery rate (FDR).

**Figure 8. Host genes for FLEXIs in Clusters I to VI are enriched in other short and long introns that have binding sites for the same subsets of RBPs.** (A) UpSet plots showing the host genes of FLEXIs, other short introns (≤300 nt) and long (>300 nt) introns containing CLIP-seq-identified binding sites for the subsets of RBPs associated with FLEXIs in Clusters I to VI. (B) Hierarchical clustering for co-enrichment of RBP-binding sites in other short and long introns was done as in Fig. 7 for 10 subsets of 2,000 randomly selected other short introns or long introns (Set1-10). Hierarchical clustering recapitulated Clusters I to VI found for FLEXIs with additional RBPs associated with these clusters as well as additional clusters (VII to X) found for some subsets of other short or long introns. The subsets of FLEXIs in HEK-293T, HeLa S3, K-562, and UHRR are shown at the top for comparison. Names of RBPs are color coded by protein function as indicated at the bottom of the Figure.

**Figure 9. GO terms for host genes of other short and long introns that contain binding sites for RBPs in Clusters I to VI are similar to those for FLEXIs in Clusters I to VI.** GO enrichment analysis was done using DAVID bioinformatics tools for host genes containing other short (left) or long (right) introns. The color scales at the bottom of each heat map were based on -$\log_{10}$-transformed FDR. RBP names are color coded by protein function as indicated at the bottom of the Figure.

**Figure 10. The intracellular localization of FLEXI RNAs correlates with that of their CLIP-seq-identified RNA-binding proteins.** Scatter plots comparing the relative abundance of annotated RBP-binding sites in FLEXIs and fragments of other short introns and long intron RNAs in nuclear or cytoplasmic fraction from K-562, HeLa S3, MDA-MB-231, and MCF7 cells. (A and B) Scatter plots for FLEXIs. The box within the scatter plot in panels A delineates a region that is expanded in panels B. (C and D) Scatter plots for other short and long introns, respectively. Each point in the scatter plots represents the relative abundance of RBP-binding sites in intron RNAs in the nuclear and cytoplasmic with points for different categories of RBPs color coded as shown below the plots. In panels A and B, only FLEXIs that were differentially enriched in the nuclear or cytoplasmic fractions (FC>1.5 by DESeq2 analysis; denoted nuclear and cytoplasmic FLEXIs) were used for the analysis. RBPs that belong to Clusters I and II or function in nuclear RNA export are indicated by named points in the scatter plots. Also named are other RBPs whose abundance was ≥2% of all RBP-binding sites and were significantly enriched in intron RNAs in the nuclear or cytoplasmic fractions (adjusted $p \leq 0.05$; calculated by Fisher's exact test and adjusted by the Benjamini-Hochberg procedure).

**Figure 11. Distribution of FLEXI RNAs detected in nuclear, cytoplasmic or both fractions.** Stacked bar graphs show the percentages of FLEXI RNAs containing binding sites for RBPs associated with Clusters I to VI in nuclear, cytoplasmic, or both fractions (color coded as shown bottom right) in combined (left bar graph) or separate (right bar graphs) in TGIRT-seq datasets for 4 different cell lines. The number of FLEXIs identified in either or both fractions is indicated within each section of the stacked bar graph.

**Figure 12. Numbers of CLIP-seq-identified binding sites for different RBPs in individual FLEXI RNAs belonging to Clusters I-VI.** (A) Stacked bar graphs organized by cell type (top

row) or cluster (bottom row) showing the percentage of FLEXI RNAs that have an annotated CLIP-seq-identified binding site for a single RBP, 2 RBPs, or ≥3 RBPs associated with the same cluster, with FLEXI RNAs having binding sites for ≥2 RBPs subdivided by whether or not the CLIP-seq-identified binding sites were overlapping or non-overlapping. (B) Protein-protein interactions of RBPs associated with Clusters I to VI based on protein-interaction network analysis using the STRING database (https://string-db.org/). Interacting RBPs are connected by lines whose thickness indicates the confidence level of the interaction score.

**Figure S1. Classes of RNAs identified by TGIRT-seq of rRNA-depleted unfragmented cellular RNAs.** Stacked bar graphs showing the percentages of sequencing reads or bases that mapped to different categories of Ensembl GRCh38 Release 93 annotated genomic features in combined technical replicates for the indicated cellular RNA samples. (A) Percentages of sequencing reads that mapped to different classes of cellular RNAs. rRNA includes cellular and mitochondrial (Mt) rRNAs, and protein-coding gene RNAs includes all transcripts of protein-coding genes from both the nuclear and mitochondrial genomes. (B) Percentages of sequencing reads that mapped to different sncRNAs. misc RNA: miscellaneous RNAs, such as ribozymes, small NF90-associated RNAs (snaRs), promoter-associated RNAs (pRNAs), and other ncRNAs that do not fall into other categories. (C) Percentage of bases that mapped to different regions of the sense strand of protein-coding genes in the nuclear genome. Because the cellular RNAs had not been chemically fragmented and the variety of TGIRT enzyme used in this study does not efficiently read through poly(A) at the 3' end of mature mRNAs, protein-coding gene RNAs comprised only a low percentage of total reads (0.7-5.3%). Abbreviations: CDS, coding sequences; intergenic, regions upstream or downstream of transcription start and stop sites of protein-coding genes annotated in Ensembl; Intron, intronic regions; UTR, 5'- or 3'-untranslated regions.

**Figure S2. Full-length, end-to-end sequencing reads of sncRNAs in TGIRT-seq cellular datasets.** Integrative Genomics Viewer (IGV) screenshots showing coverage tracks 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 tracks (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 matched the reference base, and other colors indicate bases in the read that did not match the reference base (red, thymidine; green, adenosine; blue, cytidine; and brown, guanosine). Misincorporation at known sites of post-transcriptionally modified bases in tRNAs are highlighted in the alignments: $m^1A58$: 1-methyladenosine at position 58; I: inosine. CCA indicates the post-transcriptionally added 3' CCA sequences of tRNAs mapped against a reference set of mature tRNA sequences.

**Figure S3. Characteristics of FLEXI RNAs.** (A) Density plots showing the length distribution of FLEXI RNAs (≤300 nt, red) and small numbers of longer FLEXI RNAs (>300 nt, blue) in combined datasets for each of the 4 cellular RNA samples. The inset density plots compare the abundance (RPM) of FLEXIs ≤300 nt and >300 nt in the same datasets. (B) Three-dimensional bar chart showing the percentage of FLEXI RNA reads ending at different positions around intron-exon junctions in the combined datasets for each of the four cellular RNA samples. (C) Density plots showing the length distribution of all reads of any length mapping within different categories of FLEXIs 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, 6 core spliceosome proteins (AQR, BUD13, EFTUD2, PPIG, PRPF8 and SF3B4), other annotated RBPs, FLEXIs without (w/o) annotated RBP-binding sites, and other GRCh38-annotated short introns (≤300 nt). Percent length was calculated from the read spans of TGIRT-seq reads normalized to the length of the corresponding intron based on the Ensembl GRCh38 Release 93 human genome reference. FLEXIs encoding embedded snoRNAs/scaRNAs were excluded from the comparisons to avoid contributions from full or partially processed sncRNAs.

**Figure S4. FLEXI RNAs corresponding to U12-type introns or introns 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 in a combined datasets for the 4 cellular RNA samples (HEK-293T, HeLa S3, K-562, and UHRR) compared to literature consensus sequences for human U12-type introns (top) (28, 30) and consensus sequences for other U12-type short (≤300 nt) and long (>300 nt) introns in the same combined datasets. The number of introns for each consensus sequence is indicated to the right of the sequence. (B) IGV screen-

shots showing examples of coverage tracks 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 tracks 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 coverage tracks and are color coded by sample type as indicated in the key at the upper right of the panel. Boxes indicate non-templated addition (NTA) by TGIRT-III. Reads were down sampled to a maximum of 100 for display in IGV.

**Figure S5. FLEXI RNAs are degraded by 5'- and 3'-exonucleases.** DNase-treated, size-selected (≤200 nt) HEK-293T and UHRR RNAs were incubated with (+) or without (mock treatment; -) Terminator and RNase R exonucleases, and the products for 6 different FLEXI RNAs and 3 sncRNAs products were analyzed by RT-qPCR using SYBR Green with gene-specific DNA oligonucleotide primers near the 3' end of the RNA (Table S5; see Methods). (A) Cycle threshold (CT) values for qPCRs of exonuclease- and mock-treated FLEXIs and sncRNAs (top panel) and calculated percentage of RNAs remaining after exonuclease treatment relative to the parallel mock treatment (bottom panel). Three experimental replicates of exonuclease digestion and matched mock-treatments were performed for each RNA type. Each experimental replicate was quantified as the CT mean of 3 qPCR technical replicates and displayed as a box plot. The specificity of PCR products was confirmed by extracting gel bands corresponding to representative qPCR amplicons followed by Sanger sequencing directly or after TOPO-TA cloning (Supplementary file). (B) Denaturing PAGE of synthetic 3I_RAN FLEXI RNA that was circularized *in vitro* by T4 RNA Ligase 1, linear with a 5' phosphate required for circularization, or linear with a 5' OH after mock (-) or exonuclease treatment (+) with Terminator and RNase R exonucleases. Size markers in the left-most lane are the Low range ssRNA ladder (New England Biolabs), and those in the right-most lane are the RNA Century Ladder (Ambion). Circular and linear forms are indicated by arrowheads color coded as shown to the right of the gel.

**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 molecules 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 (71) and $\log_{10}$ transformed RPM of sncRNAs in combined datasets for each of the 4 cellular RNA samples before (top panels) or after (bottom) bias correction of TGIRT-seq reads using a random forest regression model (21). 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 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) (71). Pearson (*r*) correlation coefficients are shown in the upper left of each panel. A test of the linear regression model (*lrm*) for 3 sncRNAs (U7, SNORD14B, and SNORD44), used as digital droplet PCR (ddPCR) standards described in panel B, gave copy number per cell estimates within the same order of magnitude as literature values (71), but with copy number per cell values varying between different cell types (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 for cellular RNA samples. The right-hand columns show copy number per cell values based on ddPCR abundance relative to 3 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 the latter 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). 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. The ddPCR values relative to U7 RNA, which has the closest cellular abundance to FLEXIs by TGIRT-seq reads, 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.

**Figure S7. Cell specificity and reproducibility of detected FLEXI RNAs.** (A) Scatter plots comparing $log_2$-transformed RPM of relatively abundant (≥0.01 RPM) FLEXI RNAs in biological replicates (Bio 1 and Bio 2) for 4 different cellular RNA samples. Bio 1 is a merged dataset of 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 UHRR; see Table S1). FLEXIs that were detected in all Bio 1 technical replicates are shown as red dots. (B) Scatter plots comparing $log_2$-transformed RPM of FLEXI RNA reads in different cellular RNA samples (top row) to those mapping to FLEXI host genes in the same datasets down sampled to match the sequencing depth of FLEXI reads (bottom row). Relatively abundant FLEXIs ( ≥0.01 RPM) detected in all replicates of each cell type and their host genes are color coded as shown in the top right scatter plot. Pearson correlation coefficients (*r*) calculated for all detected FLEXIs and their host genes are shown in black, and those for relatively abundant color-coded FLEXIs and their host genes (red and blue) are shown in purple.

**Figure S8. Clustering analysis of detected FLEXI RNAs.** (A) and (B) PCA, PCA-initialized t-SNE (33) and ZINB-WaVE (34) analysis for FLEXI RNAs detected at ≥1 read (panel A) or ≥0.01 RPM (panel B) in biological and technical replicates of rRNA-depleted unfragmented cellular RNAs (Table S1). Both t-SNE and ZINB-WaVE are widely used for the analysis of single cell RNA-seq datasets with zero-inflated counts. The top row of each panel shows clustering of datasets for technical replicates for different cell types used in this study. The middle and bottom rows show clustering of datasets for rRNA-depleted unfragmented HEK-293T, K-562, MDA-MB-231, and UHRR RNAs from this study (biological replicate 1 (Bio 1, circles) and other studies (biological replicate 2 (Bio 2), triangles) before (middle) and after (bottom) batch effect correction by ZINB-WaVE. All of the compared datasets are listed in Table S1.

**Figure S9. Quantitation of RBP-binding sites in different classes of intron RNAs.** Bar graphs showing the number of (A) FLEXI RNAs, (B) Other Ensembl GRCh38-annotated short introns (≤300 nt), and (C) Ensembl GRCh38-annotated long introns (>300 nt) that have a CLIP-seq-identified binding site for the indicated RBP in a merged dataset for the K-562, HEK-293T, HeLa S3, and UHRR cellular RNA samples. Bars graphs are color coded by RBP function as

shown at the top. Asterisks above the bars in panels B and C indicate the 53 proteins identified as binding 30 or more different FLEXIs in Fig. 4A.

**Figure S10. Sequencing reads corresponding to FLEXIs in ENCODE eCLIP datasets.** eCLIP datasets for HepG2 and K-562 cells were obtained from ENCODE for all RBPs that bind ≥30 FLEXIs, and reads that were ≥95% of the length of FLEXIs were identified by intersecting the eCLIP reads with FLEXI coordinates. Data are shown as fold enrichment of ≥95% full-length intron reads for each RBP, calculated as a ratio of such reads in the eCLIP dataset versus the control antibody dataset (**** $p<0.0001$, * $p<0.05$, Fisher's least significant difference test). The dashed line indicates 2-fold enrichment. Names of proteins are color coded by protein function as indicated in the Figure. For those RBPs in which more than one dataset was available, mean values were calculated for the fold enrichment in eCLIP replicates over controls with error bars indicating the standard deviation (Supplementary file).

**Figure S11. Hierarchical clustering of FLEXIs in different cell lines based on over- and under-represented RBP-binding sites.** FLEXIs containing at least one annotated binding site for any RBP in any of the 4 cell lines (1,261 to 2,404 FLEXIs) were used in this analysis. For hierarchical clustering, subsets of FLEXIs (6 to 1,086) that have binding sites for each of the 47 non-core spliceosomal RBPs that bind ≥30 different FLEXI in merged TGIRT-seq datasets for the 4 cell lines were extracted for each of the cellular RNA sample types. Contingency tables were then generated by comparing the frequency of binding sites for all 126 RBPs that have an identified binding site in a FLEXI RNA in each of these subsets to those in all FLEXI RNAs. Multiple binding sites for the same RBP in the same FLEXI were counted as one binding site. p-values for over- and under-representation of the RBP-binding sites calculated by Fisher's exact test were 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, adjusted $p\leq0.05$ calculated by Fisher's exact test and adjusted by the Benjamini-Hochberg procedure) in the tested subset of FLEXIs compared to all FLEXIs, and the key-coded information was used to construct matrices of the Gower's distance between the 47 subsets in each of the cellular RNA samples that were used as input for hierarchical clustering by the complete linkage clustering method (72). The results were displayed as a two-dimensional 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 at the bottom 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, differently shaded boxes. RBP names are color coded by protein function as indicated in the Figure.

**Figure S12. Enriched RBP-binding sites and characteristics of subsets of FLEXIs in Clusters I-VI.** Scatter plots and density plots are shown for FLEXIs that bind each protein in (A) Cluster I, (B) Cluster II, (C) Cluster III, (D) Cluster IV, (E) Cluster V, and (F) Cluster VI. In the scatter plots (left), RBP-binding sites for other RBPs whose relative abundance was significantly over- or under-represented in the subset of FLEXIs compared to all 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 (right) compare the length, GC content, and MFE for the most stable secondary structure predicted by RNAfold for subsets of FLEXIs bound by each RBP associated with Clusters I to VI (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≤0.05 as determined by 1,000 Monte-Carlo simulation).

**Figure S13. Relative positions of RBP-binding sites in FLEXIs and other classes of introns.** Density plots showing the location of eCLIP-identified RBP-binding sites in K-562 or HepG2 cells within FLEXIs for the 53 RBPs that bind ≥30 FLEXIs in merged K-562, HEK-293T, HeLa S3, and UHRR datasets compared to other short and long introns. RBP-binding sites overlapping introns (≥1 nt) were identified by intersecting the annotated RBP-binding sites 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, and 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. RBP names are color coded by protein function as indicated at the bottom of the Figure. Blank spaces were left when datasets were not available for an RBP in one of the two cell lines used to obtain the eCLIP datasets.

**Figure S14. Changes in mRNA levels of host genes of FLEXIs, other short, and long introns with binding sites for different RBPs in RBP knockdown versus control datasets.** Volcano plots show -$\log_{10}$-transformed adjusted p-values versus $\log_2$-transformed fold changes for ENSEMBL-annotated genes in ENCODE knockdown versus control datasets for the indicated RBPs in K-562 and HepG2 cells. Host genes of (A) FLEXIs, (B) other short introns, and (C) long introns that contain a binding site for the indicated RBP that have significant expression changes (adjusted $p \leq 0.05$, $|LFC| \geq 1$) in the knockdown datasets are shown as red dots. Other genes with or without significant expression changes are shown as black or gray dots, respectively. RBP knockdowns that had a significant bias towards increased or decreased mRNA levels in the group of host genes encoding FLEXIs, other short introns, and long introns with a binding site for the RBP are indicated by up (light blue) or down (red) arrows next to the RBP name (p-value $\leq 0.05$ determined by Fisher's exact test comparing the ratio of significantly up ($\log_2 FC > 0$) or down ($\log_2 FC < 0$) 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, other short introns, and long introns with a binding site for the knocked down RBP, but no significant directional bias, are indicated by a (gray) bi-directional arrow next to the RBP name. Plots are shown only from those RBPs whose knockdown resulted in a significant change ($p \leq 0.05$). Datasets that were not available for an RBP in one of the two cell lines were left as a blank space. RBP names are color coded by protein function as shown at the bottom of the Figure.

**Figure S15. Splicing changes at or adjacent to different classes of introns in RBP-knockdown versus control datasets**. Splicing changes for (A) FLEXIs, (B) other short introns, or (C) long introns in ENCODE knockdown datasets for HepG2 and K-562 cells were calculated using rMATS (https://rnaseq-mats.sourceforge.io). Empirical cumulative distribution function (ECDF) plots showing inclusion of retained introns (RI) or skipped exons (SE) adjacent to introns that have (red) or do not have (gray) a CLIP-seq-identified binding site for the indicated RBP in ENCODE knockdown datasets. Red curves shifting to the right or left of the control

(gray) indicate an increase or decrease in retained introns and skipped exons as indicated at the top of each set of plots. Statistical significance was calculated by Kolmogorov-Smirnov test. Plots are shown only from those RBPs whose knockdown resulted in a significant change ($p \leq 0.05$). Names of RBPs are color coded by protein function as indicated in the Figure. Blank spaces were left for datasets that were not available for an RBP in one of the two cell lines. Axes labels are shown in the key at the bottom right.

**Figure S16. Patterns of overlapping RBP-binding sites in FLEXI RNAs.** Heat map of FLEXI RNAs associated with Clusters I-VI that having overlapping binding sites for each of the 47 non-core spliceosomal RBPs that bind ≥30 FLEXIs (columns and rows, respectively). The number of FLEXIs containing overlapping binding sites for each RBP pair (columns and rows) were $\log_{10}$ transformed and clustered within boxes color coded by the number of different RBPs that bound the same FLEXIs. Names of RBPs primarily associated with Clusters I-VI at the significance levels used for hierarchical clustering and those not associated with Clusters I-VI (Other) are color coded as shown at the bottom of the Figure.

**Figure S17. Cell-type specific FLEXI RNAs and host genes in MCF7 and MDA-MB-231 cancer cell lines.** UpSet plots showing the distribution of (A) FLEXI RNAs, (B) FLEXI host genes, (C) FLEXI host oncogenes, and (D) FLEXI host tumor suppressor genes in MCF7 cells, MDA-MB-231 cells, and a combined dataset for 4 other cellular RNA samples (HEK-293T, HeLa S4, K-562 and UHRR). FLEXI RNAs from known oncogenes (73) or tumor suppressor genes (74) that were detected only in MCF7 and/or MDA-MB-231 are listed below the plots. ATF3 and KLF4 were annotated as both oncogenes and tumor suppressor genes (73, 74).

**Figure S18. Analysis of RNAs in nuclear and cytoplasmic fractions from cultured cells.** Scatter plots comparing enrichment of all and subtypes of cellular RNAs in cytoplasmic and nuclear fractions from HeLa S3, K-562, MDA-MB-231, and MCF7 cells. RNA subtypes are color coded as shown to the right of the scatter plots. Mitochondrial RNAs, tRNAs, Vault RNAs, Y RNA, and 7SL RNA were enriched in the cytoplasmic RNA fraction, while 7SK RNA and snRNAs were enriched in the nuclear RNA fraction. Most snoRNAs (357 to 457 in different cell types) had higher normalized counts in nuclear fractions than in cytoplasmic fractions with only 20-49 snoRNAs having higher normalized counts in cytoplasm fractions.

**Figure S19. Clustering analysis of RNAs in nuclear and cytoplasmic fractions from cultured cells.** (A) and (B) PCA plots for RNAs detected by TGIRT-seq in whole-cell (squares), nuclear fractions (Nucleus; triangles), and cytoplasmic fractions (Cytoplasm; circles) for HeLa S3, K-562, MCF7 and MDA-MB-231 cells color coded as shown in the Figure. (C) Heatmap of Euclidean distance between datasets shaded as being more (darker blue) or less (lighter blue) similar as shown by color scale in the Figure.

**Figure S20. Characterization of RBP-binding sites in intron RNAs encoded by different categories of host genes.** (A) Numbers and percentages of host genes encoding mitochondrial proteins annotated in MitoCarta3.0 or cytoplasmic or mitochondrial ribosomal proteins annotated in GRCh38 that contain FLEXIs (top), other short introns (middle), or long introns (bottom) with binding sites for Cluster I RBPs. (B) Numbers and percentages of all host genes for different subsets of introns compared to those for all GRCh38 annotated protein-coding genes. (C) Density plots showing the average coverage of RBP-binding sites across the gene body (Exons CDS + UTRs; black) or intron body (red) calculated using Picard tools normalized by percent length. In the gene body plots, 0% and 100% indicate the 5' and 3' ends the mRNAs, respectively. In the intron body plots, 0% indicates the 5' end of the first intron and 100% indicates the 3' end of the last intron of the same gene.

## Materials and Methods

### RNA samples and cultured cells

RNA samples used in this work are summarized in Table S1. Universal Human Reference RNA (UHRR) was purchased from Agilent under the product name Quantitative PCR Human Reference Total RNA). HeLa S3 and MCF7 RNAs used for initial FLEXI analysis (Figs. 1 to 9) were purchased from Thermo Fisher. All other cellular RNAs, including those for cell fractionation experiments (Fig. 10), were isolated from cultured cells by using a mirVana miRNA Isolation Kit (Thermo Fisher).

K-562 cells (ATCC CTL-243) were propagated from an ATCC source vial and maintained in Iscove's Modified Dulbecco's Medium (IMDM, Thermo Fisher 12440-053) supplemented with 10% Fetal Bovine Serum (FBS; Gemini Bioscience). HEK-293T/17 cells (ATCC CRL-11268) were propagated from an ATCC source vial and maintained in Dulbecco's Modified Eagle Medium (DMEM, Thermo Fisher 11965-065) supplemented with 10% FBS. HeLa S3 cells (ATCC CCL-2.2) were propagated from an ATCC source vial and were maintained in F-12K (Thermo Fisher 21127-022) supplemented with 10% FBS. MDA-MB-231 cells (ATCC HTB-26) and MCF7 cells (ATCC HTB-22) were obtained from Dr. Blerta Xhelmaçe (University of Texas at

Austin) and were maintained in DMEM (Thermo Fisher 11965-092) supplemented with 10% FBS. All cultures were grown in an Eppendorf CellXpert C170i incubator at 37°C with 5% $CO_2$ atmosphere.

For RNA isolation, 2-4x$10^6$ cells were harvested by centrifugation alone (K-562) or after trypsinization (all other cell lines) at 300 x g for 10 min at 4 °C and washed twice by centrifugation with cold Dulbecco's Phosphate Buffered Saline (D-PBS; Thermo Fisher 14190-144). Cell pellets were 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.

**DNA and RNA oligonucleotides**

DNA and RNA oligonucleotides were purchased from Integrated DNA Technologies (IDT). 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 as reverse transcriptase or PCR primers for the exonuclease and ddPCR experiments are listed in Table S5. The sequence specificity of all primer sets was confirmed by gel purifying a representative subset of qPCR amplicons and either Sanger sequencing amplicons directly or after TOPO-TA cloning (ThermoFisher). 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.

**Exonuclease assay**

Unfragmented 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 small size-selected RNAs were scaled for optimal FLEXI detection in each sample (HEK-293T, 100 ng; UHRR, 70 ng). The size-selected RNAs or synthetic versions of FLEXIs were subjected to exonuclease treatment (1 U Terminator exonuclease (Lucigen), 20 U RNase R (Abcam), 20 U murine RNase Inhibitor (New England Biolabs), 1X RNase R reaction buffer (Abcam), supplemented with additional $MgCl_2$ for a final concentration of 1 mm $MgCl_2$) or mock-treated in the same reaction conditions without the enzymes, for 1 h at 42°C, with 3 replicates performed for each RNA type. The exonuclease reactions were cleaned up with the RNA Clean & Concentrator kit (8X ethanol method; Zymo Research). An aliquot of this material for synthesized FLEXIs was analyzed on a Novex 10% TBE-Urea polyacrylamide gel (Thermo Fisher) to separate linear and circular RNAs and imaged with SYBR Gold (Thermo Fisher). 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 (run in triplicate) were carried out with 1 µL cDNA, 0.3 µM primers for all primer sets except 4I_JUP which used 0.8 µM, and PowerUp SYBR Green master mix (Life Technologies) using a ViiA 7 Real Time PCR system (Thermo Fisher). Matched control samples

omitting the reverse transcriptase (RT-) were performed for every reaction. Mean delta cyclic threshold (CT) values were calculated to obtain a fold change of the RNA levels in the treated compared to the mock-treated samples and converted into a percentage of RNA remaining after exonuclease treatment. In two cases, the exonuclease treatment led to no measurable CT value and in these cases a pseudocount CT value of 41 was added to perform the calculation. Primer 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 inputs 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; UHRR, 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 (Thermo Fisher) 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, 0.2 μM primers for all primer sets except 4I_JUP which used 0.5 μM, and QX200 ddPCR EvaGreen Supermix (Biorad) using 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 length 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 primer dimer artifacts 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 (75), and SNORD44 (76). 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.

**Cell fractionation**

Cell cultures were grown to ~80% confluency (HeLa S3, MDA-MB-231 and MCF7) or to a density of $9x10^5$ cells/mL (K-562) in T-75 flasks. To prepare cell pellets, adherent cells were either (1: HeLa S3) washed 3x with cold D-PBS and then scraped in D-PBS and pelleted by centrifugation at 1,000xg for 5 min at 4°C in an Eppendorf FA-24x2-PTFE rotor; or (2: MDA-MB-231 and MCF7) trypsinized and collected by centrifugation at 1,000 RPM for 5 min at room temperature in a Centra CL2 centrifuge with Aerocarrier rotor (Thermo Fisher) and then resuspended in cold D-PBS and washed 3x by centrifugation at 1,000xg for 5 min at 4°C in an Eppendorf FA-24x2-PTFE rotor. Suspension cells (K-562) were collected by centrifugation at 1,000 RPM for 5 min at room temperature in a Centra CL2 centrifuge with Aerocarrier rotor and then placed on ice and resuspended in cold D-PBS and washed 3x by centrifugation at 1,000xg for 5 min 4°C in an Eppendorf FA-24x2-PTFE rotor.

Cellular fractionation was performed as described in Rio *et al*. (77) . Cell pellets were resuspended in 1.0 mL of cell disruption buffer and then incubated for 10 min on ice and monitored for swelling by light microscopy. Swollen cells were transferred to a baked (400°C, 4 h) 2.0-mL glass Dounce homogenizer and broken with 10 strokes of a Type B pestle. Cell breakage was confirmed by microscopy. Cell homogenates were transferred to 1.7-mL tubes and brought to 0.1% Triton X-100 by the addition of 10% (w/v) Triton X-100 (Sigma). Samples were then centrifuged immediately at 1,500 x g for 5 min at 4°C in an Eppendorf FA-24x2-PTFE rotor.

Nuclear fraction pellets and cytoplasmic fraction supernatants were transferred to fresh tubes and placed on ice. RNA was extracted by using a mirVana RNA extraction kit following the manufacturer's instructions. The nuclear fraction pellets were resuspended in 600 µL of kit lysis buffer and processed on a mirVana column. The cytoplasmic samples were split such that 200 µL of cytoplasm was solubilized in 400 µL of kit lysis buffer and then combined and processed together on the same column. RNA samples were evaluated on an Agilent 2100 Bioanalyzer with a Eukaryotic Total RNA Pico kit using the mRNA assay. RNAs were then processed for TGIRT-seq library construction as described below.

**Thermostable Group II Intron Reverse Transcriptase sequencing (TGIRT-seq)**

TGIRT-seq libraries were prepared from unfragmented RNA as described (21, 22). To remove residual DNA, UHRR and HeLa S3 RNAs (1 µg) 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 K-562, MCF7, MDA-MB-231 and HEK-293T RNAs (5 µg) were incubated with 2 U TURBO DNase for 30 min at 37 °C (Thermo Fisher). After DNase 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.

TGIRT-seq libraries were prepared as described (21, 22) using 20-50 ng of ribodepleted unfragmented cellular RNAs. Template-switching and reverse transcription reactions were done with 1 µM TGIRT-III (InGex, currently available upon request from the Lambowitz lab) 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; denaturation at 98°C for 5 s followed by 12 cycles of 98°C for 5 s, 60°C for 10 s, and 72°C for 15 s 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.

**Sequencing data processing**

Illumina TruSeq adapters and PCR primer sequences were trimmed from the reads with Cutadapt v2.8 (https://github.com/marcelm/cutadapt, sequencing quality score cut-off at 20; p-value <0.01) 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 (http://daehwankimlab.github.io/hisat2/) 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 reference 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 (https://github.com/BenLangmead/bowtie2) 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 (https://github.com/samtools/samtools).

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 (https://github.com/arq5x/bedtools2 ). 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 gene 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, https://igv.org). Genes with >100 mapped reads were down sampled to 100 mapped reads in IGV for visualization. Down sampling of reads mapping to FLEXI host genes was conducted by randomly picking mapped reads to match the read depth of FLEXI RNA reads.

**Identification and characterization of FLEXIs**

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. Mapped reads were then intersected with 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 reads 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 (https://jokergoo.github.io/ComplexHeatmap-reference/book/). 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 (58) and agotrons (19). FLEXI RNAs containing embedded snoRNAs were identified by intersecting the FLEXI RNA coordinates with those of annotated snoRNAs and scaRNAs 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 (29, 30). 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 (29) using FIMO (https://meme-suite.org/meme/tools/fimo) 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 nucleotide 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. 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 (https://meme-suite.org/meme/tools/meme ) 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 Ceqlogo from the MEME suite (https://meme-suite.org/meme/doc/ceqlogo.html).

The length and GC content of each FLEXI was calculated from its GRCh38-annotated intron sequence. Minimum free energy (MFE) was calculated for the most stable secondary structure predicted by RNAfold for each FLEXI RNA. The distribution of length, GC content and MFE of FLEXI subsets compare to those of all other detected FLEXIs were compared by Kolmogorov–Smirnov (KS) test. To test if the significance in distribution was due to the small sample size in FLEXI subsets, those with p-values <0.01 (denoted test subset) were further tested by 1,000 Monte-Carlo simulations. In each simulation, a random selected FLEXI subset with the same number of FLEXIs as the test subset was compared to all other FLEXIs by the same KS test. If more than 50 of the 1,000 simulations (≥95%) yield a calculated p<0.01, the significance in the original test subset was considered to be a false hit.

**PCA, t-SNE, and ZINB-WaVE 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. 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 coordinates with ENCODE annotated binding sites of 150 RBPs from eCLIP-seq with irreproducible discovery rate analysis (36) and with crosslink-centered regions identified from the DICER PAR-CLIP (38) and AGO1-4 PAR-CLIP (37) datasets by using BEDTools. Multiple binding sites for the same RBP in an intron were collapsed and counted as one binding 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 ref. (68). RBPs found in stress granules were as annotated in the RNA Granule and Mammalian Stress Granules Proteome (MSGP) databases (56, 69). The functional annotations, localization patterns, and RNA-binding domains of AGO1-4 and DICER were from the UniProt database (https://www.uniprot.org). 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 (https://github.com/thelovelab/DESeq2) after GC correction with Salmon (https://github.com/COMBINE-lab/salmon) and conditional quantile normalization (CQN) (78). The number of significant gene expression changes (DESeq2-calculated adjusted p-value<0.05) for FLEXI host genes was compared to 3 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 generated 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 (https://rnaseq-mats.sourceforge.io). 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 an empirical cumulative density plot (ECDF) 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 that overlapped 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 of the RBP-binding site co-enrichment, subsets of FLEXIs that have a binding site for each of 47 RBPs that have binding sites in ≥30 different FLEXI RNAs were extracted from datasets for each of the cellular RNA samples. Over- and under-represented RBPs in each of the 47 subsets for each cellular RNA sample were identified 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 (79) were calculated using the key-coded information from the above analyses and used as input for hierarchical clustering by the complete linkage clustering method (72). The results were displayed as a heat map plotted using the R package Cluster and ComplexHeatmap, respectively. For hierarchical clustering analysis of the RBP-binding site co-enrichment in other short introns and long intron, 2,000 randomly selected other short or long introns for each category of intron were analyzed as above, with the process was repeated 10 times for each of the cellular RNA samples.

**Analysis of RBP-binding site in FLEXIs and RNA fragments of other short and long introns in nuclear and cytoplasmic fractions**

FLEXIs detected by TGIRT-seq in nuclear and cytoplasmic fractions were analyzed by DESeq2. FLEXIs that have a fold change (FC) value > 1.5 in one fraction compared to the other were considered as differentially enriched in the corresponding fraction, and were denoted as nuclear or cytoplasmic FLEXIs, respectively. RNA fragments derived from other short and long introns were identified by intersecting mapped reads with other short intron or long intron coordinates extracted from Ensemble GRCh38 Release 93 gene annotations using BEDTools. Read pairs (Read 1 and Read 2) within the annotated introns were identified as intron RNA fragments and merged into a single bed file using BEDTools. RBP-binding in these RNA fragments were identified by intersecting the coordinates of FLEXI RNA, merged short intron, or long intron RNA fragments, with ENCODE annotated binding sites of 150 RBPs from eCLIP-seq with irreproduc-

ible discovery rate analysis (36) and with crosslink-centered regions identified from the DICER PAR-CLIP (38) and AGO1-4 PAR-CLIP (37) datasets by using BEDTools. Multiple binding sites for the same RBP in an intron were collapsed and counted as one binding site for that RBP.

**GO term enrichment analysis**

GO term enrichment analysis of host genes encoding FLEXIs bound by different RBPs was performed by using DAVID bioinformatics tools (https://david.ncifcrf.gov) 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 gplots package in R (https://www.rdocumentation.org/packages/gplots/versions/3.1.3.1).

**Protein function and network analysis**

Proteins identified as oncogene, tumor suppressor gene, or mitochondrial localized protein were based on Liu *et al*. (73), Zhao. *et al*.(74), and human MitoCarta3.0 (https://www.broadinstitute.org/mitocarta/mitocarta30-inventory-mammalian-mitochondrial-proteins-and-pathways). Protein network analysis was done using the STRING database (https://string-db.org) with the default setting.

**Data Availability**

All TGIRT-seq datasets analyzed in this manuscript have been deposited in the National Center for Biotechnology Information Sequence Read Archive. These include datasets for: (i) commercial human plasma pooled from multiple healthy individuals from a previous study (20) (accession number PRJNA640428); (ii) biological replicates of unfragmented cellular RNAs from HEK-293T cells from a previous study (26) (samples WC and YC; accession number (PRJNA389469); (iii) TGIRT-seq datasets for biological replicates of unfragmented cellular RNAs from K-562, HEK-293T, HeLa S3, MCF7, MDA-MB-231 and UHRR (accession number PRJNA648481); (iv) K-562 cells, samples UFC1-7; MDA-MB-231, samples MDA1-5; and UHRR, samples UHRR0-9 (accession number PRJNA722757); (v) TGIRT-seq datasets of whole-cell RNA and nuclear and cytoplasmic fractions from K-562, HeLa S3, MD-MBA-231, andMCF7 (accession number PRJNA648481).

## REFERENCES


1. Wilkinson ME, Charenton C, Nagai K. RNA splicing by the spliceosome. Annu Rev Biochem. 2020;89(1): 359-388. doi: 10.1146/annurev-biochem-091719-064225.
2. Chapman KB, Boeke JD. Isolation and characterization of the gene encoding yeast debranching enzyme. Cell. 1991;65(3): 483-492. doi: 10.1016/0092-8674(91)90466-C.
3. Farrell MJ, Dobson AT, Feldman LT. Herpes simplex virus latency-associated transcript is a stable intron. Proc Natl Acad Sci U S A. 1991;88(3): 790-794. doi: 10.1073/pnas.88.3.790.
4. Kulesza CA, Shenk T. Murine cytomegalovirus encodes a stable intron that facilitates persistent replication in the mouse. Proc Natl Acad Sci U S A. 2006;103(48): 18302-18307. doi: 10.1073/pnas.0608718103.
5. Gardner EJ, Nizami ZF, Talbot CC, Gall JG. Stable intronic sequence RNA (sisRNA), a new class of noncoding RNA from the oocyte nucleus of *Xenopus tropicalis*. Genes Dev. 2012;26(22): 2550-2559. doi: 10.1101/gad.202184.112.
6. Moss WN, Steitz JA. Genome-wide analyses of Epstein-Barr virus reveal conserved RNA structures and a novel stable intronic sequence RNA. BMC Genomics. 2013;14(1): 543. doi: 10.1186/1471-2164-14-543.
7. Zhang Y, Zhang X-O, Chen T, Xiang J-F, Yin Q-F, Xing Y-H, et al. Circular intronic long noncoding RNAs. Mol Cell. 2013;51(6): 792-806. doi: 10.1016/j.molcel.2013.08.017.
8. Pek JW, Osman I, Tay ML-I, Zheng RT. Stable intronic sequence RNAs have possible regulatory roles in *Drosophila melanogaster*. J Cell Biol. 2015;211(2): 243-251. doi: 10.1083/jcb.201507065.
9. Taggart AJ, Lin CL, Shrestha B, Heintzelman C, Kim S, Fairbrother WG. Large-scale analysis of branchpoint usage across species and cell lines. Genome Res. 2017;27(4): 639-649. doi: 10.1101/gr.202820.115.
10. Talhouarne GJS, Gall JG. Lariat intronic RNAs in the cytoplasm of vertebrate cells. Proc Natl Acad Sci U S A. 2018;115(34): E7970-E7977. doi: 10.1073/pnas.1808816115.
11. Morgan JT, Fink GR, Bartel DP. Excised linear introns regulate growth in yeast. Nature. 2019;565(7741): 606-611. doi: 10.1038/s41586-018-0828-1.
12. Saini H, Bicknell AA, Eddy SR, Moore MJ. Free circular introns with an unusual branchpoint in neuronal projections. eLife. 2019;8: e47809. doi: 10.7554/eLife.47809.
13. Rasmussen Asta M, Okholm Trine Line H, Knudsen M, Vang S, Dyrskjøt L, Hansen Thomas B, Pedersen Jakob S. Circular stable intronic RNAs possess distinct biological features and are deregulated in bladder cancer. NAR Cancer. 2023;5(3): zcad041. doi: 10.1093/narcan/zcad041.
14. Manuel Ares J, Haller I, Sol K, John PD. Intron-lariat spliceosomes convert lariats to true circles: implications for intron transposition. bioRxiv. 2024: 586863. doi: 10.1101/2024.03.26.586863.
15. Parenteau J, Maignon L, Berthoumieux M, Catala M, Gagnon V, Abou Elela S. Introns are mediators of cell response to starvation. Nature. 2019;565(7741): 612-617. doi: 10.1038/s41586-018-0859-7.
16. Berezikov E, Chung W-J, Willis J, Cuppen E, Lai EC. Mammalian mirtron genes. Mol Cell. 2007;28(2): 328-336. doi: 10.1016/j.molcel.2007.09.028.
17. Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC. The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. Cell. 2007;130(1): 89-100. doi: 10.1016/j.cell.2007.06.028.
18. Ruby JG, Jan CH, Bartel DP. Intronic microRNA precursors that bypass Drosha processing. Nature. 2007;448(7149): 83-86. doi: 10.1038/nature05983.

19. Hansen TB, Venø MT, Jensen TI, Schaefer A, Damgaard CK, Kjems J. Argonaute-associated short introns are a novel class of gene regulators. Nat Commun. 2016;7: 11538. doi: 10.1038/ncomms11538.
20. Yao J, Wu DC, Nottingham RM, Lambowitz AM. Identification of protein-protected mRNA fragments and structured excised intron RNAs in human plasma by TGIRT-seq peak calling. eLife. 2020;9: e60743. doi: 10.7554/eLife.60743.
21. Xu H, Yao J, Wu DC, Lambowitz AM. Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. Sci Rep. 2019;9(1): 7953. doi: 10.1038/s41598-019-44457-z.
22. Xu H, Nottingham RM, Lambowitz AM. TGIRT-seq protocol for the comprehensive profiling of coding and non-coding RNA biotypes in cellular, extracellular vesicle, and plasma RNAs. Bio Protoc. 2021;11(23): e4239. doi: 10.21769/BioProtoc.4239.
23. Katibah GE, Qin Y, Sidote DJ, Yao J, Lambowitz AM, Collins K. Broad and adaptable RNA structure recognition by the human interferon-induced tetratricopeptide repeat protein IFIT5. Proc Natl Acad Sci U S A. 2014;111(33): 12025–12030. doi: 10.1073/pnas.1412842111.
24. Zheng G, Qin Y, Clark WC, Dai Q, Yi C, He C, et al. Efficient and quantitative high-throughput tRNA sequencing. Nat Methods. 2015;12(9): 835–837. doi: 10.1038/nmeth.3478.
25. Qin Y, Yao J, Wu DC, Nottingham RM, Mohr S, Hunicke-Smith S, Lambowitz AM. High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. RNA. 2016;22(1): 111–128. doi: 10.1261/rna.054809.115.
26. Shurtleff MJ, Yao J, Qin Y, Nottingham RM, Temoche-Diaz MM, Schekman R, Lambowitz AM. Broad role for YBX1 in defining the small noncoding RNA composition of exosomes. Proc Natl Acad Sci U S A. 2017;114(43): E8987-E8995. doi: 10.1073/pnas.1712108114.
27. Munoz-Tello P, Rajappa L, Coquille S, Thore S. Polyuridylation in eukaryotes: a 3′-end modification regulating RNA life. Biomed Res Int. 2015;2015: 968127. doi: 10.1155/2015/968127.
28. Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 2000;28(21): 4364-4375. doi: 10.1093/nar/28.21.4364.
29. Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Res. 2006;34(14): 3955-3967. doi: 10.1093/nar/gkl556.
30. Gao K, Masuda A, Matsuura T, Ohno K. Human branch point consensus sequence is yUnAy. Nucleic Acids Res. 2008;36(7): 2257-2267. doi: 10.1093/nar/gkn073.
31. Gao Y, Zhang J, Zhao F. Circular RNA identification based on multiple seed matching. Brief Bioinform. 2018;19(5): 803-810. doi: 10.1093/bib/bbx014.
32. Nottingham RM, Wu DC, Qin Y, Yao J, Hunicke-Smith S, Lambowitz AM. RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase. RNA. 2016;22(4): 597-613. doi: 10.1261/rna.055558.115.
33. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat Commun. 2019;10(1): 5416. doi: 10.1038/s41467-019-13056-x.
34. Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. A general and flexible method for signal extraction from single-cell RNA-seq data. Nat Commun. 2018;9(1): 284. doi: 10.1038/s41467-017-02554-5.
35. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005;15(8): 1034-1050. doi: 10.1101/gr.3715005.

36. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods. 2016;13(6): 508-514. doi: 10.1038/nmeth.3810.
37. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell. 2010;141(1): 129-141. doi: 10.1016/j.cell.2010.03.009.
38. Rybak-Wolf A, Jens M, Murakawa Y, Herzog M, Landthaler M, Rajewsky N. A variety of Dicer substrates in human and *C. elegans*. Cell. 2014;159(5): 1153-1167. doi: 10.1016/j.cell.2014.10.040.
39. Berezhna SY, Supekova L, Sever MJ, Schultz PG, Deniz AA. Dual regulation of hepatitis C viral RNA by cellular RNAi requires partitioning of Ago2 to lipid droplets and P-bodies. RNA. 2011;17(10): 1831-1845. doi: 10.1261/rna.2523911.
40. Pépin G, Perron MP, Provost P. Regulation of human Dicer by the resident ER membrane protein CLIMP-63. Nucleic Acids Res. 2012;40(22): 11603-11617. doi: 10.1093/nar/gks903.
41. Lou Q, Hu Y, Ma Y, Dong Z. RNA interference may suppress stress granule formation by preventing argonaute 2 recruitment. Am J Physiol Cell Physiol. 2019;316(1): C81-c91. doi: 10.1152/ajpcell.00251.2018.
42. Bronisz A, Rooj AK, Krawczyński K, Peruzzi P, Salińska E, Nakano I, et al. The nuclear DICER–circular RNA complex drives the deregulation of the glioblastoma cell microRNAome. Sci Adv. 2020;6(51): eabc0221. doi: 10.1126/sciadv.abc0221.
43. Johnson Krystal C, Kilikevicius A, Hofman C, Hu J, Liu Y, Aguilar S, et al. Nuclear localization of Argonaute 2 is affected by cell density and may relieve repression by microRNAs. Nucleic Acids Res. 2023;52(4): 1930-1952. doi: 10.1093/nar/gkad1155.
44. Janowicz A, Michalak M, Krebs J. Stress induced subcellular distribution of ALG-2, RBM22 and hSlu7. Biochim Biophys Acta. 2011;1813(5): 1045-1049. doi: 10.1016/j.bbamcr.2010.11.010.
45. Ma Z, Moore R, Xu X, Barber GN. DDX24 negatively regulates cytosolic RNA-mediated innate immune signaling. PLoS Pathog. 2013;9(10): e1003721. doi: 10.1371/journal.ppat.1003721.
46. Okholm TLH, Sathe S, Park SS, Kamstrup AB, Rasmussen AM, Shankar A, et al. Transcriptome-wide profiles of circular RNA and RNA-binding protein interactions reveal effects on circular RNA biogenesis and cancer pathway expression. Genome Med. 2020;12(1): 112. doi: 10.1186/s13073-020-00812-8.
47. Chen D, Zhou H, Cai Z, Cai K, Liu J, Wang W, et al. CircSCAP interacts with SF3A3 to inhibit the malignance of non-small cell lung cancer by activating p53 signaling. J Exp Clin Cancer Res. 2022;41(1): 120. doi: 10.1186/s13046-022-02299-0.
48. Wang C, Query CC, Meier UT. Immunopurified small nucleolar ribonucleoprotein particles pseudouridylate rRNA independently of their association with phosphorylated Nopp140. Mol Cell Biol. 2002;22(24): 8457-8466. doi: 10.1128/mcb.22.24.8457-8466.2002.
49. Alawi F, Lin P. Loss of dyskerin reduces the accumulation of a subset of H/ACA snoRNA-derived miRNA. Cell Cycle. 2010;9(12): 2467-2469. doi: 10.4161/cc.9.12.11922.
50. Lv K, Gong C, Antony C, Han X, Ren JG, Donaghy R, et al. HectD1 controls hematopoietic stem cell regeneration by coordinating ribosome assembly and protein synthesis. Cell Stem Cell. 2021;28(7): 1275-1290.e1279. doi: 10.1016/j.stem.2021.02.008.
51. Kaiser RWJ, Ignarski M, Van Nostrand EL, Frese CK, Jain M, Cukoski S, et al. A protein-RNA interaction atlas of the ribosome biogenesis factor AATF. Sci Rep. 2019;9(1): 11071. doi: 10.1038/s41598-019-47552-3.

52. Savage KI, Gorski JJ, Barros EM, Irwin GW, Manti L, Powell AJ, et al. Identification of a BRCA1-mRNA splicing complex required for efficient DNA repair and maintenance of genomic stability. Mol Cell. 2014;54(3): 445-459. doi: 10.1016/j.molcel.2014.03.021.
53. Vohhodina J, Barros EM, Savage AL, Liberante FG, Manti L, Bankhead P, et al. The RNA processing factors THRAP3 and BCLAF1 promote the DNA damage response through selective mRNA splicing and nuclear export. Nucleic Acids Res. 2017;45(22): 12816-12833. doi: 10.1093/nar/gkx1046.
54. Gilbertson S, Federspiel JD, Hartenian E, Cristea IM, Glaunsinger B. Changes in mRNA abundance drive shuttling of RNA binding proteins, linking cytoplasmic RNA degradation to transcription. eLife. 2018;7: e37663. doi: 10.7554/eLife.37663.
55. Mattijssen S, Arimbasseri AG, Iben JR, Gaidamakov S, Lee J, Hafner M, Maraia RJ. LARP4 mRNA codon-tRNA match contributes to LARP4 activity for ribosomal protein mRNA poly(A) tail length protection. eLife. 2017;6: e28889. doi: 10.7554/eLife.28889.
56. Nunes C, Mestre I, Marcelo A, Koppenol R, Matos CA, Nóbrega C. MSGP: the first database of the protein components of the mammalian stress granules. Database. 2019;2019: baz031. doi: 10.1093/database/baz031.
57. Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, et al. A large-scale binding and functional map of human RNA-binding proteins. Nature. 2020;583(7818): 711-719. doi: 10.1038/s41586-020-2077-3.
58. Wen J, Ladewig E, Shenker S, Mohammed J, Lai EC. Analysis of nearly one thousand mammalian mirtrons reveals novel features of Dicer substrates. PLoS Comput Biol. 2015;11(9): e1004441. doi: 10.1371/journal.pcbi.1004441.
59. Alló M, Buggiano V, Fededa JP, Petrillo E, Schor I, de la Mata M, et al. Control of alternative splicing through siRNA-mediated transcriptional gene silencing. Nat Struct Mol Biol. 2009;16(7): 717-724. doi: 10.1038/nsmb.1620.
60. Ameyar-Zazoua M, Rachez C, Souidi M, Robin P, Fritsch L, Young R, et al. Argonaute proteins couple chromatin silencing to alternative splicing. Nat Struct Mol Biol. 2012;19(10): 998-1004. doi: 10.1038/nsmb.2373.
61. Mo J, Liang H, Su C, Li P, Chen J, Zhang B. DDX3X: structure, physiologic functions and cancer. Mol Cancer. 2021;20(1): 38. doi: 10.1186/s12943-021-01325-7.
62. Melick CH, Meng D, Jewell JL. A-kinase anchoring protein 8L interacts with mTORC1 and promotes cell growth. J Biol Chem. 2020;295(23): 8096-8105. doi: 10.1074/jbc.AC120.012595.
63. Zhou D, Couture S, Scott MS, Abou Elela S. RBFOX2 alters splicing outcome in distinct binding modes with multiple protein partners. Nucleic Acids Res. 2021;49(14): 8370-8383. doi: 10.1093/nar/gkab595.
64. Podar M, Chu VT, Pyle AM, Perlman PS. Group II intron splicing in vivo by first-step hydrolysis. Nature. 1998;391(6670): 915-918. doi: 10.1038/36142.
65. Cui H, Diedrich JK, Wu DC, Lim JJ, Nottingham RM, Moresco JJ, et al. Arg-tRNA synthetase links inflammatory metabolism to RNA splicing and nuclear trafficking via SRRM2. Nat Cell Biol. 2023;25(4): 592-603. doi: 10.1038/s41556-023-01118-8.
66. Lewis BM, Cho CY, Her HL, Mizrahi O, Hunter T, Yeo GW. LARP4 is an RNA-binding protein that binds nuclear-encoded mitochondrial mRNAs to promote mitochondrial function. RNA. 2024;30(3): 223-239. doi: 10.1261/rna.079799.123.
67. Wu DC, Lambowitz AM. Facile single-stranded DNA sequencing of human plasma DNA via thermostable group II intron reverse transcriptase template switching. Sci Rep. 2017;7(1): 8421. doi: 10.1038/s41598-017-09064-w.

68. Van Nostrand EL, Pratt GA, Yee BA, Wheeler EC, Blue SM, Mueller J, et al. Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. Genome Biol. 2020;21(1): 90. doi: 10.1186/s13059-020-01982-9.
69. Youn J-Y, Dyakov BJA, Zhang J, Knight JDR, Vernon RM, Forman-Kay JD, Gingras A-C. Properties of stress granule and P-body proteomes. Mol Cell. 2019;76(2): 286-294. doi: 10.1016/j.molcel.2019.09.014.
70. Massenet S, Bertrand E, Verheggen C. Assembly and trafficking of box C/D and H/ACA snoRNPs. RNA Biol. 2017;14(6): 680-692. doi: 10.1080/15476286.2016.1243646.
71. Tycowski K, Kolev N, Conrad N, Fok V, Steitz J. The ever-growing world of small nuclear ribonucleoproteins. In: Gesteland R, Cech T, Atkins J, editors. The RNA World. Third Edition ed. New York: Cold Spring Harbor Laboratory Press; 2006. p. 327-368.
72. Defays D. An efficient algorithm for a complete link method. The Computer Journal. 1977;20(4): 364-366. doi: 10.1093/comjnl/20.4.364.
73. Liu Y, Sun J, Zhao M. ONGene: a literature-based database for human oncogenes. J Genet Genomics. 2017;44(2): 119-121. doi: 10.1016/j.jgg.2016.12.004.
74. Zhao M, Kim P, Mitra R, Zhao J, Zhao Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Res. 2016;44(D1): D1023-D1031. doi: 10.1093/nar/gkv1268.
75. Dodd DW, Gagnon KT, Corey DR. Digital quantitation of potential therapeutic target RNAs. Nucleic Acid Ther. 2013;23(3): 188-194. doi: 10.1089/nat.2013.0427.
76. Sharbati S, Sharbati J, Hoeke L, Bohmer M, Einspanier R. Quantification and accurate normalisation of small RNAs through new custom RT-qPCR arrays demonstrates *Salmonella*-induced microRNAs in human monocytes. BMC Genomics. 2012;13(1): 23. doi: 10.1186/1471-2164-13-23.
77. Rio DC, Ares M, Jr., Hannon GJ, Nilsen TW. Preparation of cytoplasmic and nuclear RNA from tissue culture cells. Cold Spring Harb Protoc. 2010;2010(6): pdb.prot5441. doi: 10.1101/pdb.prot5441.
78. Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012;13(2): 204-216. doi: 10.1093/biostatistics/kxr054.
79. Gower JC. A general coefficient of similarity and some of its properties. Biometrics. 1971;27(4): 857. doi: 10.2307/2528823.

**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[¶] | 134.5 | 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%) |
| 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%) |
| HeLa S3 (Total RNA) [**] | 46.8 | 27.2 (58.2%) | 20.5 (75.1%) | 13.3 (65.0%) |
| HeLa S3 (Nucleus RNA) [**] | 37.4 | 27.5 (73.5%) | 22.4 (81.4%) | 15.3 (68.5%) |
| HeLa S3 (Cytoplasm RNA) [**] | 49.7 | 29.3 (59.1%) | 21.1 (71.9%) | 13.2 (62.7%) |
| K-562 (Total RNA) [**] | 66.2 | 49.7 (75.1%) | 37.5 (75.4%) | 24.3 (64.8%) |
| K-562 (Nucleus RNA) [**] | 75.5 | 66.7 (88.4%) | 56.5 (84.6%) | 34.4 (60.9%) |
| K-562 (Cytoplasm RNA) [**] | 78.4 | 59.4 (75.8%) | 40.9 (68.8%) | 25.4 (62.0%) |
| MDA-MB-231 (Total RNA) [**] | 82.0 | 74.1 (90.3%) | 57.4 (77.5%) | 38.7 (67.3%) |
| MDA-MB-231 (Nucleus RNA) [**] | 108.6 | 102.3 (94.2%) | 87.7 (85.7%) | 54.3 (61.9%) |

| | | | | |
|---|---|---|---|---|
| MDA-MB-231 (Cytoplasm RNA) ** | 108.1 | 92.6 (85.7%) | 65.2 (70.4%) | 42.0 (64.4%) |
| MCF7 (Total RNA) ** | 70.7 | 63.8 (90.2%) | 49.3 (77.3%) | 33.3 (67.6%) |
| MCF7 (Nucleus RNA) ** | 89.7 | 85.0 (94.8%) | 73.7 (86.7%) | 48.2 (65.5%) |
| MCF7 (Cytoplasm RNA) ** | 109.9 | 94.9 (86.3%) | 68.3 (72.0%) | 43.6 (63.9%) |

*HEK-293T cell RNA, combined datasets from 8 technical replicates (SRA BioProject accession number PRJNA648481, HEK-rep1 to 8).
†HeLa S3 cell RNA, combined datasets from 10 technical replicates (SRA BioProject accession number PRJNA648481, HeLa-rep1 to 10).
§K-562 cell RNA, combined datasets from 8 technical replicates (SRA BioProject accession number PRJNA648481, K-562-rep1 to 8).
‡Universal human reference RNA, combined datasets from 8 technical replicates (SRA BioProject accession number PRJNA648481, 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 (SRA BioProject accession number PRJNA648481, MCF-rep1 to 8).
#MDA-MB-231 RNA, combined datasets from 2 technical replicates (SRA BioProject accession number PRJNA648481, MDA-rep1 and 2).
‡‡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).
**Total cellular, nucleus and cytoplasm RNA from 4 cell lines, 2 technical replicates each (SRA BioProject accession number PRJNA648481, samples SRR26807958-81).

**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<br>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 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; snoRNA processing. |
| AGO | Argonaute | Central player in RNA silencing processes, essential component of RNA-induced silencing complex (RISC). |
| AKAP8L | A-kinase anchoring protein 8 like | mRNA splicing, unspliced mRNA export. |
| AQR | RNA helicase aquarius | Spliceosome component. Pre-mRNA splicing and snoRNP biogenesis. |
| 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. |
| BUD13 | BUD13 homolog | Spliceosome component. Pre-mRNA splicing. |
| 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, may be involved translational initiation, splicing, and ribosome and spliceosome assembly. |
| 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. |
| EFTUD2 | 116 kDa U5 small nuclear ribonucleoprotein component | Spliceosome component. Pre-mRNA splicing. |
| FXR2 | Fragile X mental retardation syndrome-related protein 2 | RNA-binding protein. |
| G3BP1 | Ras GTPase-activating protein-binding protein | ATP- and Mg-dependent helicase that plays an essential role in innate immunity; Also functions in |

| | | |
|---|---|---|
| | 1 | 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. |
| GRSF1 | G-rich RNA sequence binding factor 1 | Alternative splicing, mRNA export. |
| 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. |
| HNRNPA1 | Heterogeneous nuclear ribonucleoprotein A1 | Packaging of pre-mRNA into hnRNP. Transport poly(A) mRNA from the nucleus to the cytoplasm. Translational inhibition by binding to IRES. Binds specific miRNA hairpins. Plays role in HCV RNA replication. |
| HNRNPC | Heterogeneous nuclear ribonucleoproteins C1/C2 | Packaging of pre-mRNA into hnRNP. Interacts with poly-U tracts in the 3'/5' UTR and modulates mRNA stability and translation. Splicing regulation. |
| HNRNPUL1 | Heterogeneous nuclear ribonucleoprotein U-like protein 1 | Transcriptional regulator. mRNA processing and transport. |
| 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. |
| ILF3 | Interleukin enhancer-binding factor 3 | circRNAs processing in the nucleus. Innate anti-viral response. |
| KHDRBS1 | KH domain-containing, RNA-binding, signal transduction-associated protein 1 | mRNA nuclear export, stability, and translation regulation. Involved in alternative splicing. |
| KHSRP | KH-type splicing regulatory protein | Regulates alternative pre-mRNA splicing, mRNA degradation, trafficking and localization. |
| 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. |
| NONO | Non-POU domain-containing octamer-binding protein | Pre-mRNA splicing. Involved in formation of nuclear paraspeckles, DNA non-homologous end joining (NHEJ) for double strand repair and V(D)J recombination. Transcriptional regulation. Anti-viral innate response. |
| 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. |
| PCBP1 | Poly(rC)-binding protein 1 | Binds single stranded nucleic acids; regulates antiviral response mediated by MAVS signaling |
| PCBP2 | Poly(rC)-binding protein 2 | Binds single stranded nucleic acids; regulates antiviral response mediated by MAVS signaling |
| PRPF4 | Pre-mRNA processing factor 4 | Component of the U4/U6-U5 tri-snRNP complex that is involved in pre-mRNA splicing. |
| PRPF8 | Pre-mRNA-processing-splicing factor 8 | Spliceosome component. Pre-mRNA splicing. |
| PUM1 | Pumilio RNA binding family member 1 | Post-transcriptional repressor; binds the 3'-UTR of mRNA targets; regulates MIR199 expression. |
| QKI | KH domain-containing RNA-binding protein QKI | RNA reader protein, which recognizes and binds specific RNAs, thereby regulating RNA metabolic processes, such as pre-mRNA splicing, circular RNA (circRNA) formation, mRNA export, mRNA stability and/or translation. |
| RBFOX2 | RNA binding Fox-1 homolog 2 | Regulates alternative splicing by preventing binding of U2AF2 to the 3'-splice site. |

| | | |
|---|---|---|
| RBM15 | RNA-binding protein 15 | Regulates m6A modification of RNAs; alternative splicing factor |
| RBM22 | RNA-binding protein 22 | Required component of the spliceosome |
| RBM5 | RNA-binding protein 5 | Component of the spliceosome; alternative splicing factor |
| RPS3 | 40S ribosomal protein S3 | Role in regulating transcription; implicated in regulating DNA damage response and apoptosis. |
| SERBP1 | SERPINE1 mRNA binding protein 1 | Stabilize inactive ribosome, involved in PML-body formation |
| SF3A3 | Splicing factor 3a subunit 3 | Required splicing factor |
| SF3B4 | Splicing factor 3B subunit 4 | Spliceosome component. Pre-mRNA splicing. |
| SMNDC1 | Survival motor neuron domain containing 1 | Involved in spliceosome assembly; snoRNA processing. |
| SND1 | Staphylococcal nuclease and Tudor domain containing 1 | Transcriptional co-activator of STAT5 and EB virus nuclear antigen 2; endonuclease that mediates miRNA decay. |
| SRSF1 | Serine/arginine-rich splicing factor 1 | Prevents exon skipping, regulates alternative splicing |
| SRSF7 | Serine/arginine-rich splicing factor 7 | mRNA splicing and export |
| 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. |
| TRA2A | Transformer 2 alpha homolog | Regulates pre-mRNA splicing |
| U2AF1 | U2 small nuclear RNA auxiliary factor 1 | Critical role in splicing, recruits U2 snRNP to branch point |
| U2AF2 | U2 small nuclear RNA | Roles in pre-mRNA splicing and 3' end processing |

| | | |
|---|---|---|
| | auxiliary factor 2 | |
| UCHL5 | Ubiquitin carboxyl-terminal hydrolase isozyme L5 | Protease |
| UPF1 | Regulator of nonsense transcripts 1 | NMD and mRNA export |
| XPO5 | Exportin 5 | dsRNA and pre-miRNA export |
| 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'. |
| ZC3H11A | Zinc finger CCCH-type containing 11A | mRNA and viral RNA export |
| ZNF622 | Zinc finger protein 622 | May behave as an activator of the bound transcription factor, MYBL2; positive regulator of apoptosis, cytoplasmic maturation of 60S ribosomal subunits. |
| 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. |

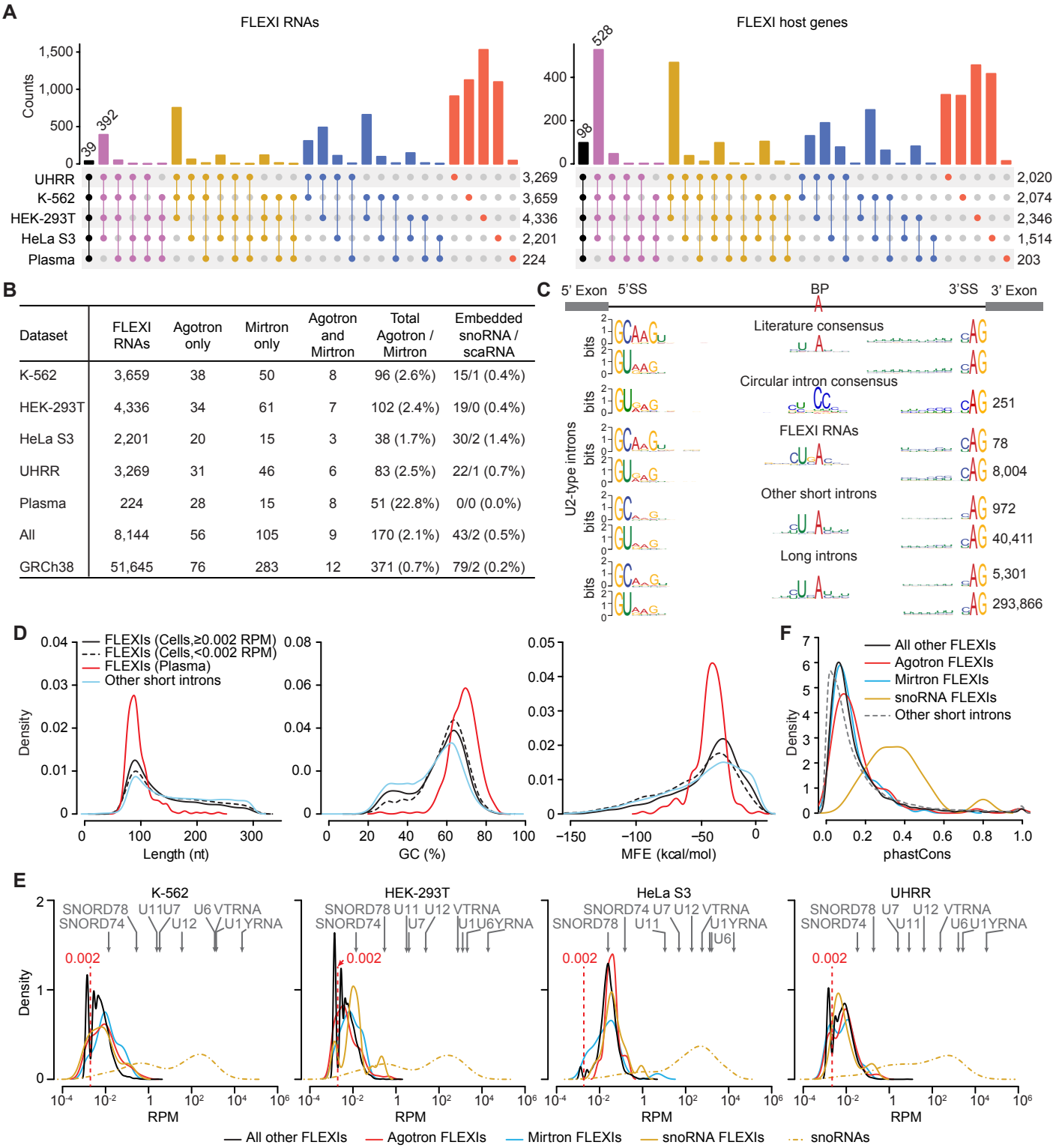
**Table S5. Oligonucleotides used for ddPCR and exonuclease digestion assays.**

| 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) | AGAGGAAAGGAAAAGCCAGGT |
| 4I_JUP (Reverse primer-FLEXI) | 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 |

†Primers are named according to FLEXI intron number and gene name or sncRNA name.

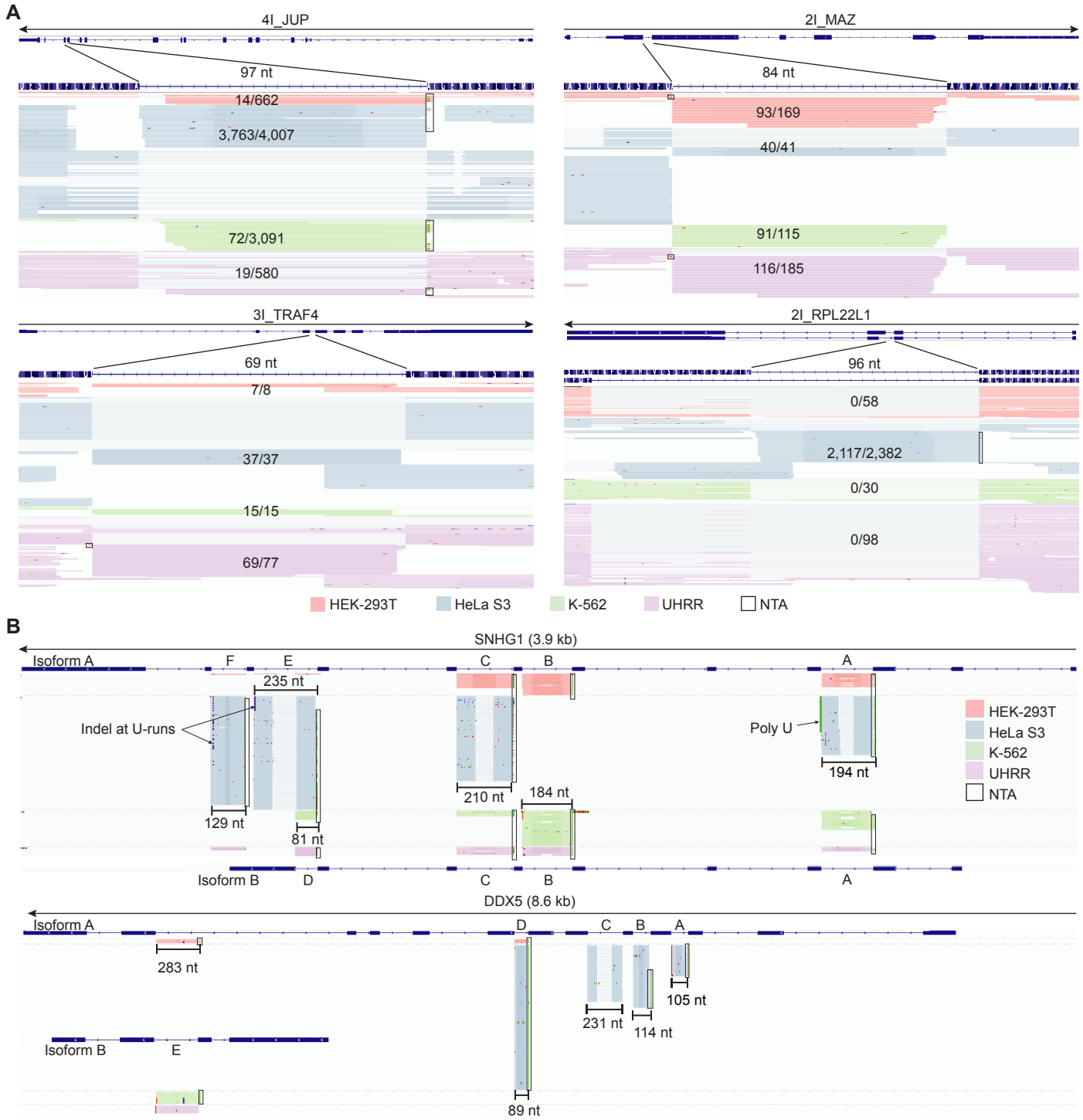
*Dodd *et al*., 2013

# Fig. 1



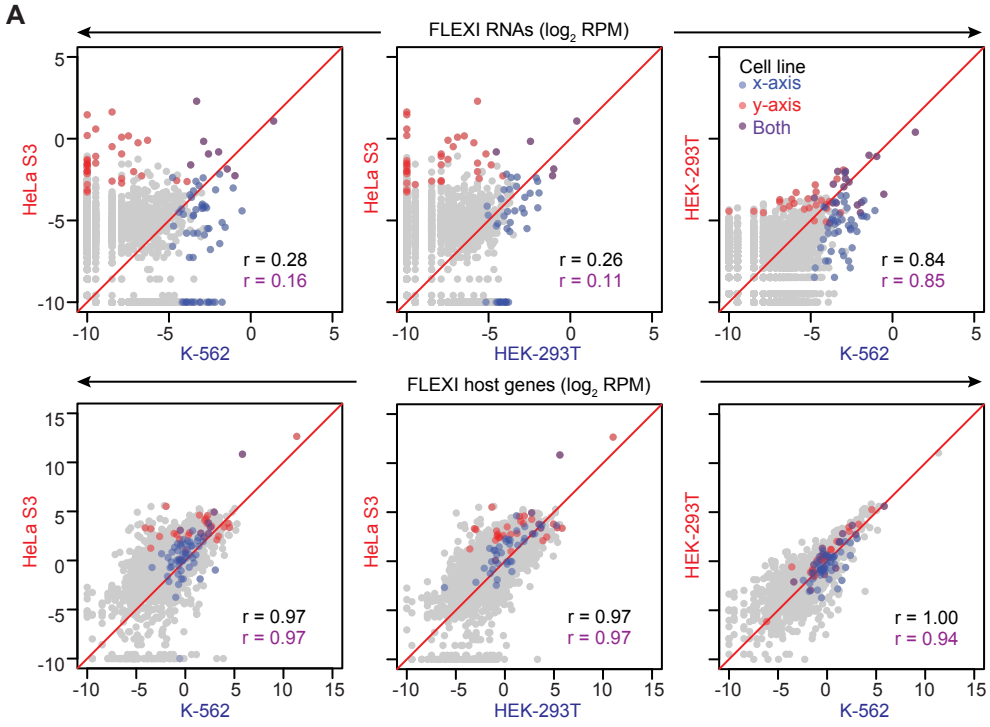

| Dataset | FLEXI RNAs | Agotron only | Mirtron only | Agotron and Mirtron | Total Agotron / Mirtron | Embedded snoRNA / scaRNA |
|---|---|---|---|---|---|---|
| K-562 | 3,659 | 38 | 50 | 8 | 96 (2.6%) | 15/1 (0.4%) |
| HEK-293T | 4,336 | 34 | 61 | 7 | 102 (2.4%) | 19/0 (0.4%) |
| HeLa S3 | 2,201 | 20 | 15 | 3 | 38 (1.7%) | 30/2 (1.4%) |
| UHRR | 3,269 | 31 | 46 | 6 | 83 (2.5%) | 22/1 (0.7%) |
| Plasma | 224 | 28 | 15 | 8 | 51 (22.8%) | 0/0 (0.0%) |
| All | 8,144 | 56 | 105 | 9 | 170 (2.1%) | 43/2 (0.5%) |
| GRCh38 | 51,645 | 76 | 283 | 12 | 371 (0.7%) | 79/2 (0.2%) |

Fig. 2


A
4I_JUP
97 nt
14/662
3,763/4,007
72/3,091
19/580
2I_MAZ
84 nt
93/169
40/41
91/115
116/185
3I_TRAF4
69 nt
7/8
37/37
15/15
69/77
2I_RPL22L1
96 nt
0/58
2,117/2,382
0/30
0/98
HEK-293T
HeLa S3
K-562
UHRR
NTA
B
SNHG1 (3.9 kb)
Isoform A
F
E
C
B
A
235 nt
Indel at U-runs
129 nt
81 nt
210 nt
184 nt
Poly U
194 nt
Isoform B
D
DDX5 (8.6 kb)
283 nt
231 nt
114 nt
105 nt
89 nt
C
FLEXIs
Other short introns
Reads (x10⁵)
Length (%)
D
Reads (%)
42.6
8.5
11.2
2.9
9.6
8.8
5' SS
3' SS

## Fig. 3

**A**




**B**

| 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**

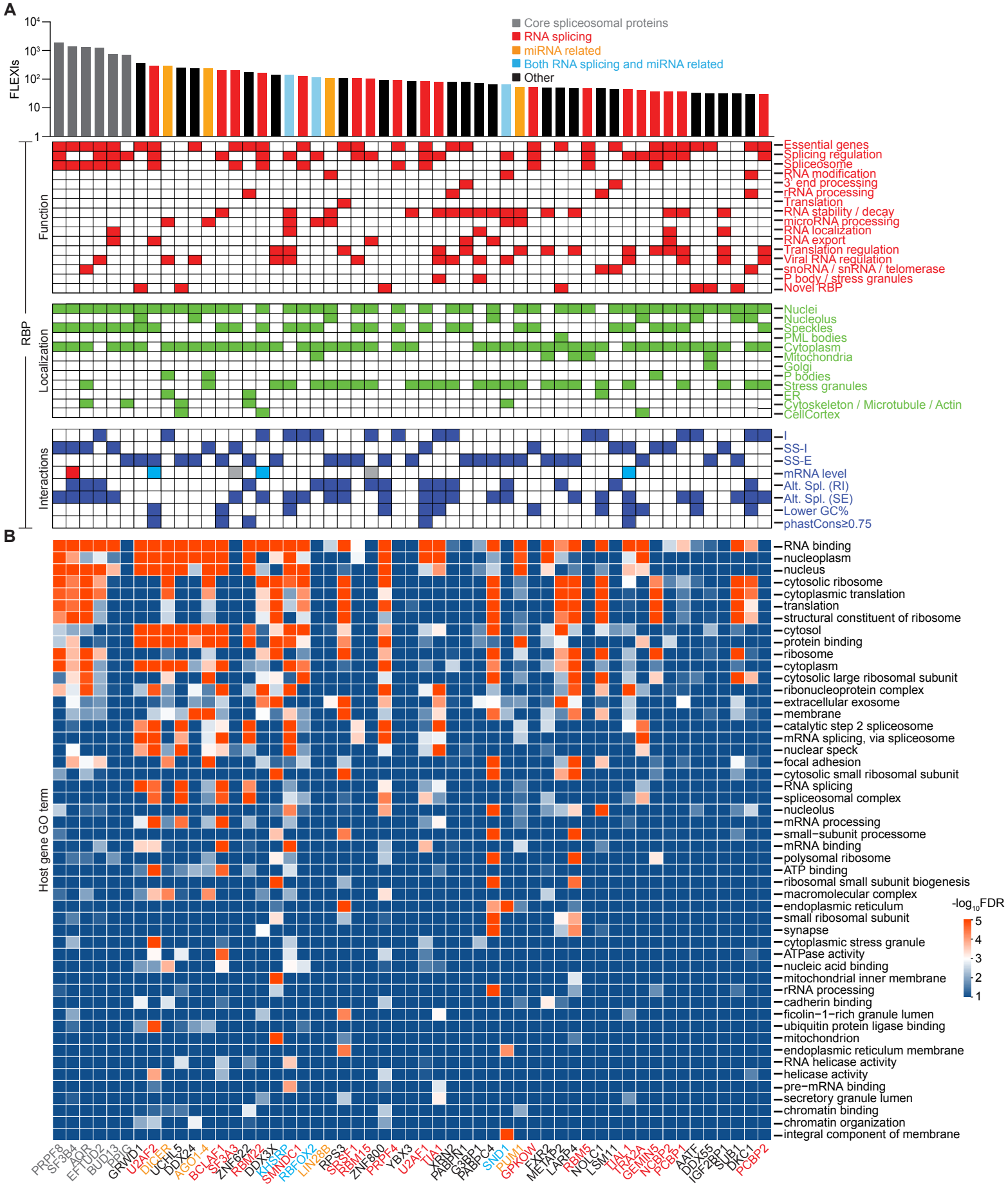

A
FLEXIs
Core spliceosomal proteins
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other
Function
Essential genes
Splicing regulation
Spliceosome
RNA modification
3' end processing
rRNA processing
Translation
RNA stability / decay
microRNA processing
RNA localization
RNA export
Translation regulation
Viral RNA regulation
snoRNA / snRNA / telomerase
P body / stress granules
Novel RBP
RBP
Localization
Nuclei
Nucleolus
Speckles
PML bodies
Cytoplasm
Mitochondria
Golgi
P bodies
Stress granules
ER
Cytoskeleton / Microtubule / Actin
CellCortex
Interactions
I
SS-I
SS-E
mRNA level
Alt. Spl. (RI)
Alt. Spl. (SE)
Lower GC%
phastCons≥0.75
B
Host gene GO term
RNA binding
nucleoplasm
nucleus
cytosolic ribosome
cytoplasmic translation
translation
structural constituent of ribosome
cytosol
protein binding
ribosome
cytoplasm
cytosolic large ribosomal subunit
ribonucleoprotein complex
extracellular exosome
membrane
catalytic step 2 spliceosome
mRNA splicing, via spliceosome
nuclear speck
focal adhesion
cytosolic small ribosomal subunit
RNA splicing
spliceosomal complex
nucleolus
mRNA processing
small−subunit processome
mRNA binding
polysomal ribosome
ATP binding
ribosomal small subunit biogenesis
macromolecular complex
endoplasmic reticulum
small ribosomal subunit
synapse
cytoplasmic stress granule
ATPase activity
nucleic acid binding
mitochondrial inner membrane
rRNA processing
cadherin binding
ficolin−1−rich granule lumen
ubiquitin protein ligase binding
mitochondrion
endoplasmic reticulum membrane
RNA helicase activity
helicase activity
pre−mRNA binding
secretory granule lumen
chromatin binding
chromatin organization
integral component of membrane
-log10FDR
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
TIAL1
TRA2A
GEMIN5
NCBP2
PCBP1
AATF
DDX55
IGF2BP1
SUB1
DKC1
PCBP2

**Fig. 5**

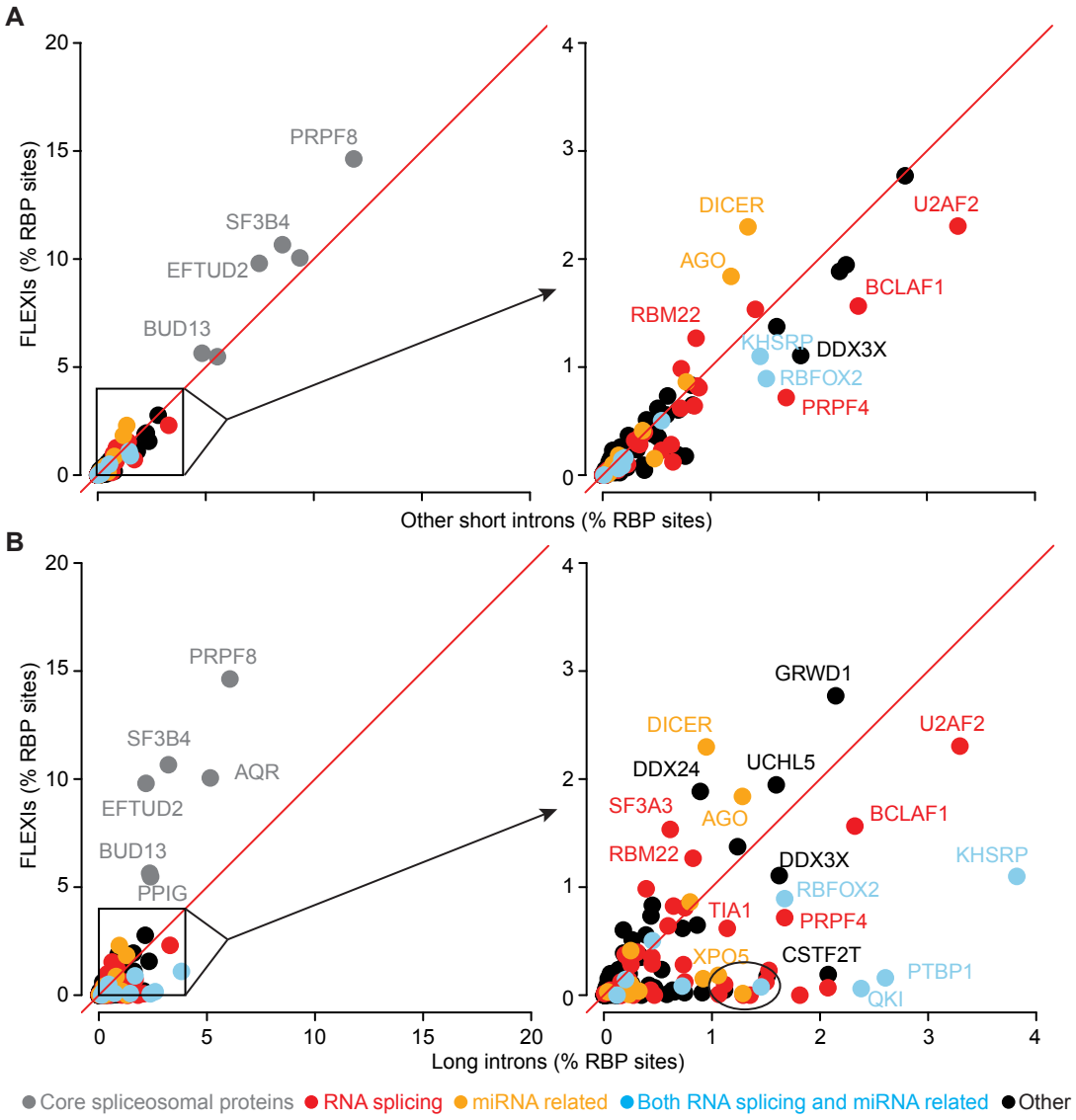

A
FLEXIs (% RBP sites)
PRPF8
SF3B4
EFTUD2
BUD13
DICER
AGO
RBM22
KHSRP
DDX3X
RBFOX2
PRPF4
U2AF2
BCLAF1
Other short introns (% RBP sites)
B
PRPF8
SF3B4
AQR
EFTUD2
BUD13
PPIG
GRWD1
DICER
DDX24
UCHL5
SF3A3
AGO
RBM22
DDX3X
RBFOX2
TIA1
PRPF4
XPO5
CSTF2T
PTBP1
QKI
KHSRP
U2AF2
BCLAF1
Long introns (% RBP sites)
Core spliceosomal proteins
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other

Fig. 6

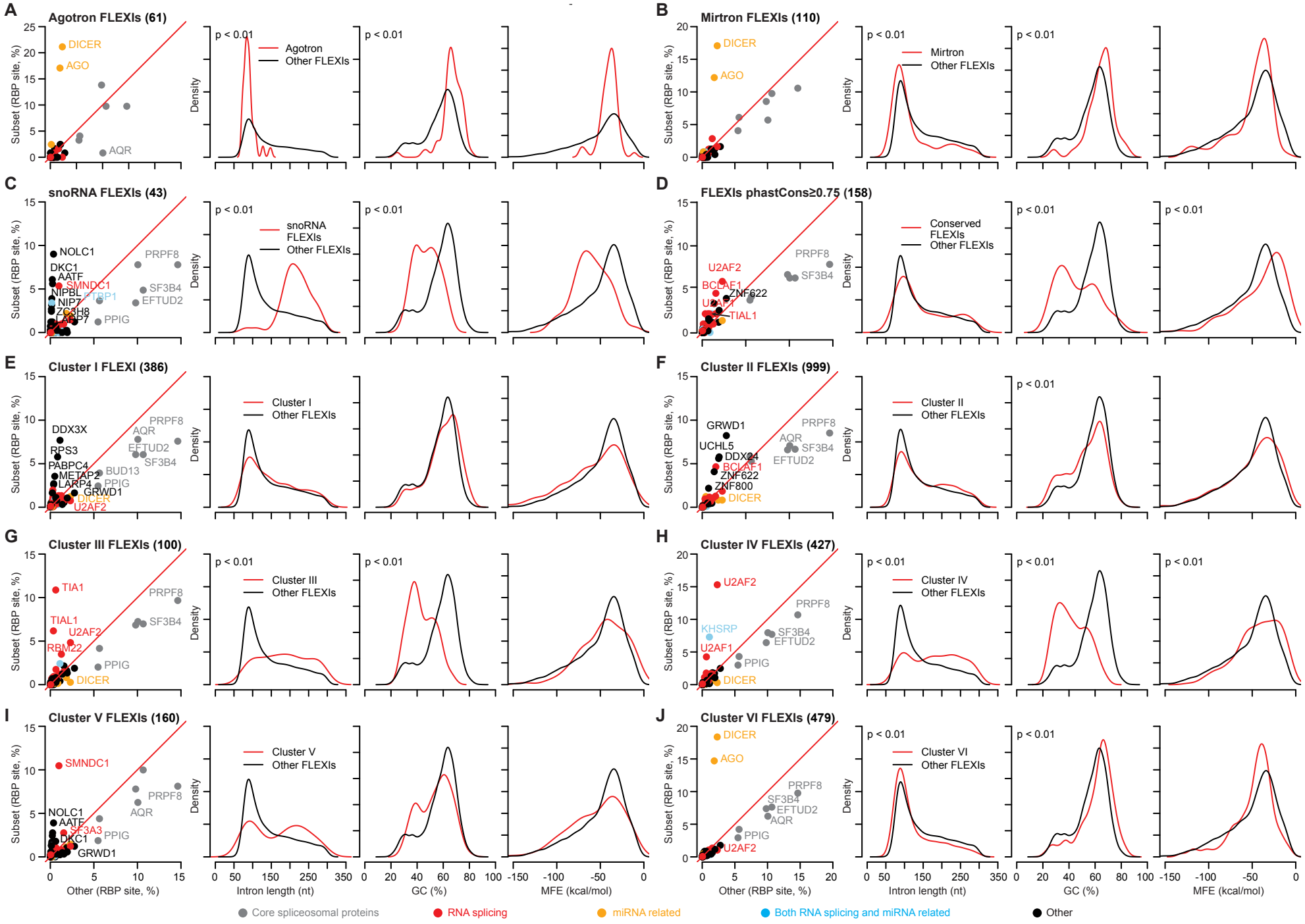

A
Agotron FLEXIs (61)
DICER
AGO
AQR
Agotron
Other FLEXIs
B
Mirtron FLEXIs (110)
Mirtron
C
snoRNA FLEXIs (43)
NOLC1
DKC1
AATF
SMNDC1
NIPBL
PTBP1
NIP7
ZCSH8
RBP7
PRPF8
SF3B4
EFTUD2
PPIG
snoRNA FLEXIs
D
FLEXIs phastCons≥0.75 (158)
U2AF2
BCLAF1
ZNF622
U2AF1
TIAL1
Conserved FLEXIs
E
Cluster I FLEXI (386)
DDX3X
RPS3
PABPC4
METAP2
LARP4
GRWD1
BUD13
Cluster I
F
Cluster II FLEXIs (999)
GRWD1
UCHL5
DDX24
ZNF800
Cluster II
G
Cluster III FLEXIs (100)
TIA1
TIAL1
RBM22
Cluster III
H
Cluster IV FLEXIs (427)
KHSRP
Cluster IV
I
Cluster V FLEXIs (160)
SF3A3
Cluster V
J
Cluster VI FLEXIs (479)
Cluster VI
Subset (RBP site, %)
Other (RBP site, %)
Density
Intron length (nt)
GC (%)
MFE (kcal/mol)
p < 0.01
Other FLEXIs
Core spliceosomal proteins
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other

**Fig. 7**

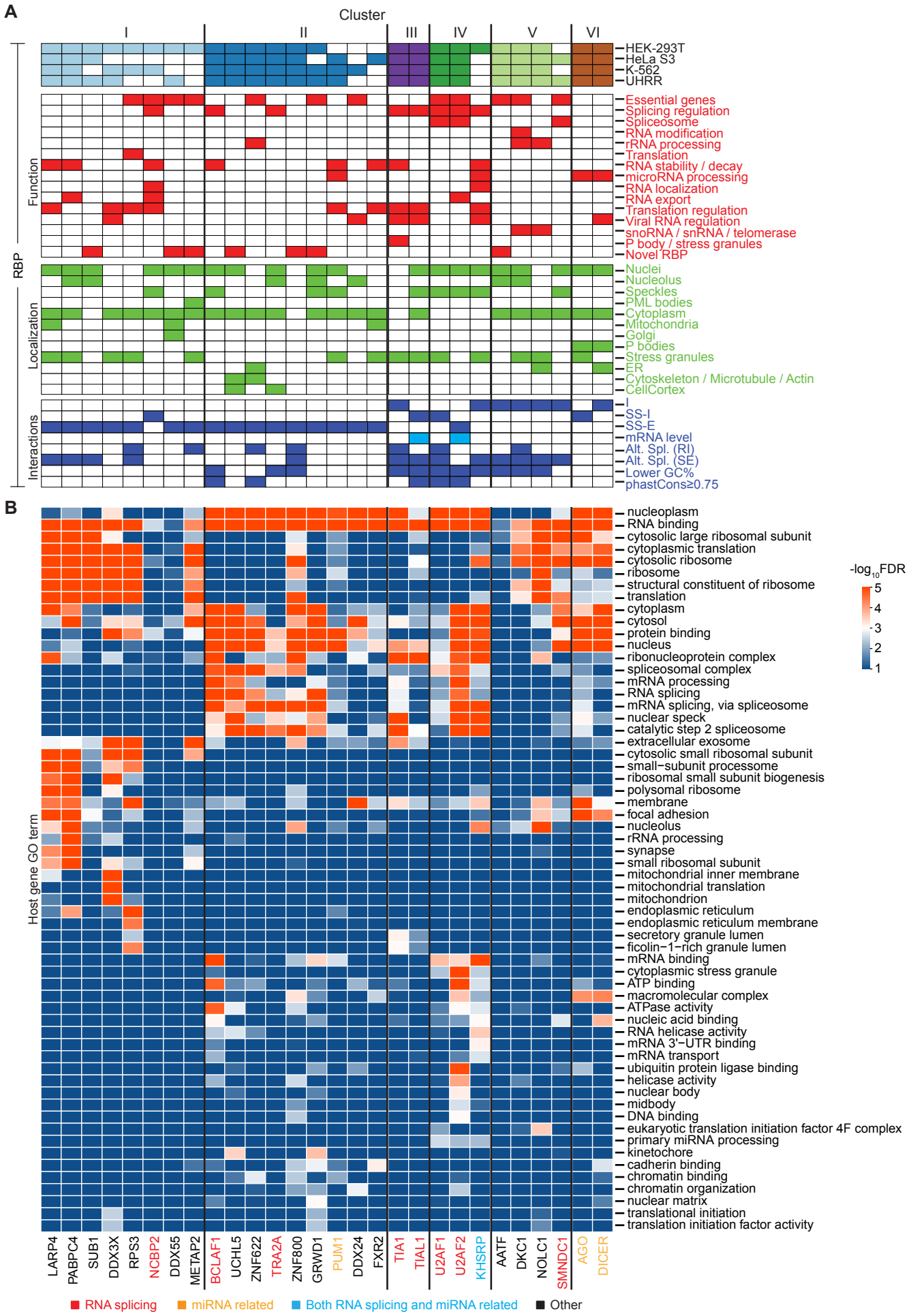

A
Cluster
I
II
III
IV
V
VI
HEK-293T
HeLa S3
K-562
UHRR
Function
Essential genes
Splicing regulation
Spliceosome
RNA modification
rRNA processing
Translation
RNA stability / decay
microRNA processing
RNA localization
RNA export
Translation regulation
Viral RNA regulation
snoRNA / snRNA / telomerase
P body / stress granules
Novel RBP
RBP
Localization
Nuclei
Nucleolus
Speckles
PML bodies
Cytoplasm
Mitochondria
Golgi
P bodies
Stress granules
ER
Cytoskeleton / Microtubule / Actin
CellCortex
Interactions
I
SS-I
SS-E
mRNA level
Alt. Spl. (RI)
Alt. Spl. (SE)
Lower GC%
phastCons≥0.75
B
Host gene GO term
nucleoplasm
RNA binding
cytosolic large ribosomal subunit
cytoplasmic translation
cytosolic ribosome
ribosome
structural constituent of ribosome
translation
cytoplasm
cytosol
protein binding
nucleus
ribonucleoprotein complex
spliceosomal complex
mRNA processing
RNA splicing
mRNA splicing, via spliceosome
nuclear speck
catalytic step 2 spliceosome
extracellular exosome
cytosolic small ribosomal subunit
small−subunit processome
ribosomal small subunit biogenesis
polysomal ribosome
membrane
focal adhesion
nucleolus
rRNA processing
synapse
small ribosomal subunit
mitochondrial inner membrane
mitochondrial translation
mitochondrion
endoplasmic reticulum
endoplasmic reticulum membrane
secretory granule lumen
ficolin−1−rich granule lumen
mRNA binding
cytoplasmic stress granule
ATP binding
macromolecular complex
ATPase activity
nucleic acid binding
RNA helicase activity
mRNA 3'−UTR binding
mRNA transport
ubiquitin protein ligase binding
helicase activity
nuclear body
midbody
DNA binding
eukaryotic translation initiation factor 4F complex
primary miRNA processing
kinetochore
cadherin binding
chromatin binding
chromatin organization
nuclear matrix
translational initiation
translation initiation factor activity
-log10FDR
5
4
3
2
1
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**

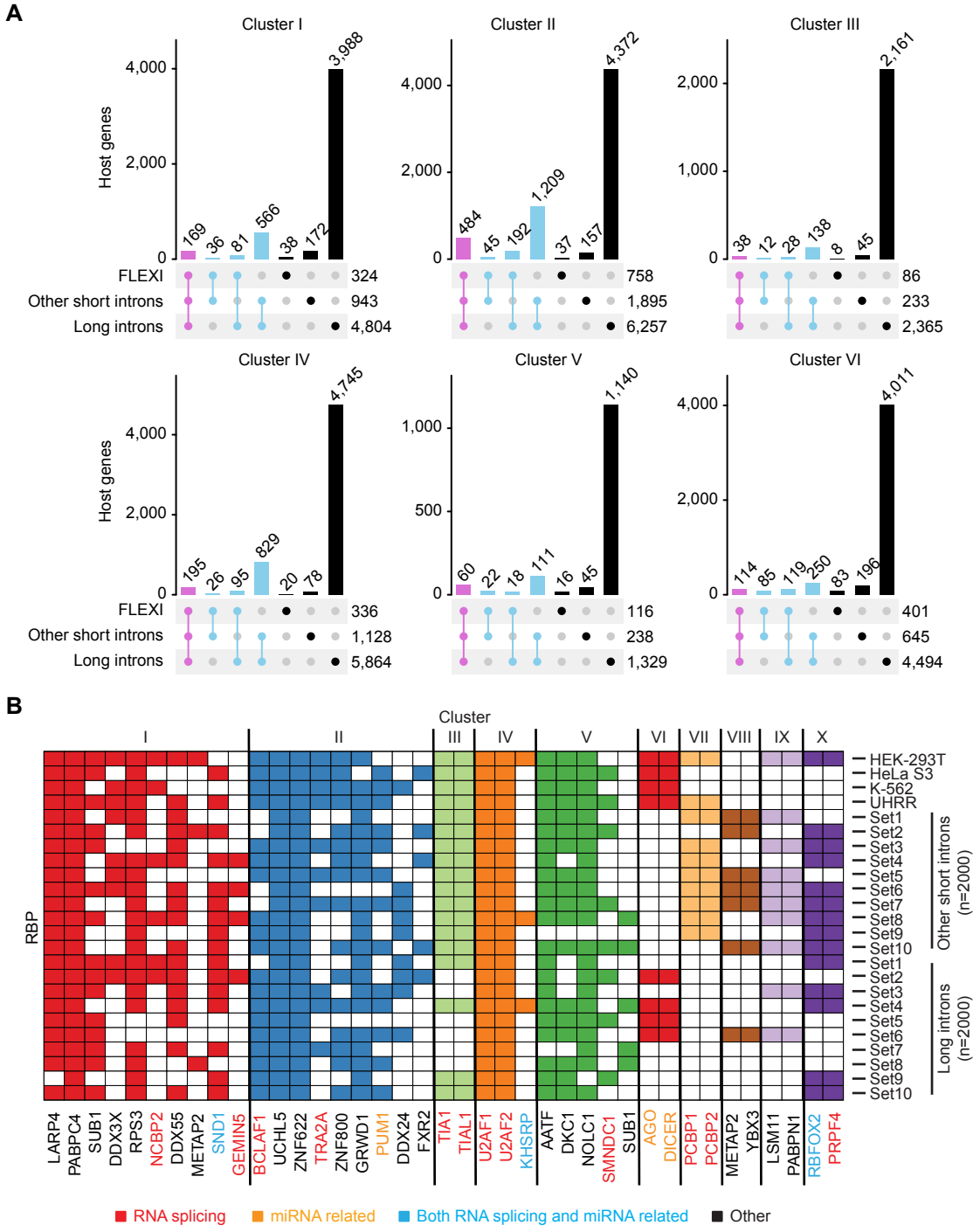

A
Cluster I
Cluster II
Cluster III
Cluster IV
Cluster V
Cluster VI
Host genes
FLEXI
Other short introns
Long introns
169 36 81 566 38 172 3,988
324 943 4,804
484 45 192 1,209 37 157 4,372
758 1,895 6,257
38 12 28 138 8 45 2,161
86 233 2,365
195 26 95 829 20 78 4,745
336 1,128 5,864
60 22 18 111 16 45 1,140
116 238 1,329
114 85 119 250 83 196 4,011
401 645 4,494
B
Cluster
I II III IV V VI VII VIII IX X
RBP
HEK-293T
HeLa S3
K-562
UHRR
Set1 Set2 Set3 Set4 Set5 Set6 Set7 Set8 Set9 Set10
Other short introns (n=2000)
Long introns (n=2000)
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. 9**

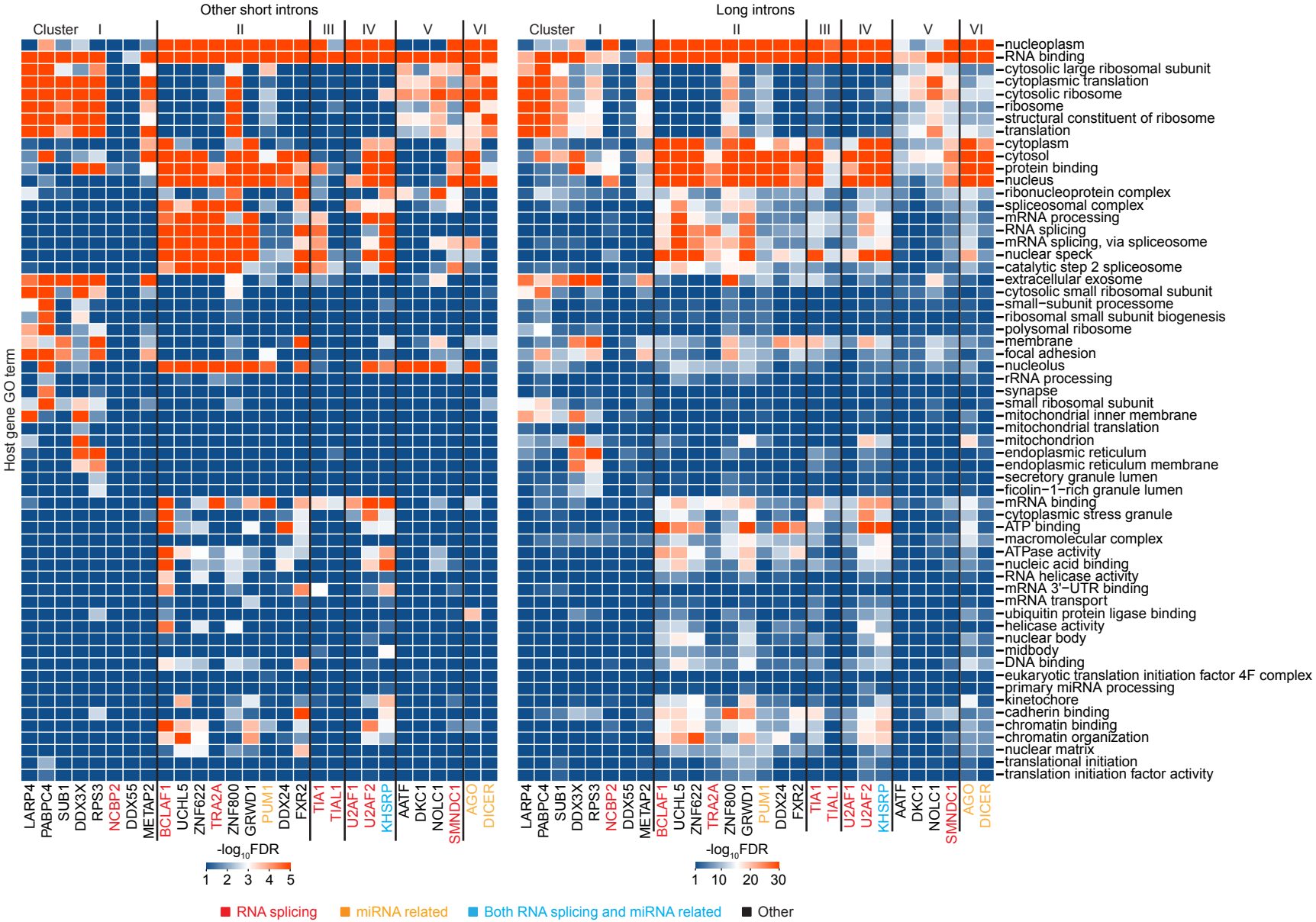

Other short introns
Long introns
Cluster I II III IV V VI
Host gene GO term
nucleoplasm
RNA binding
cytosolic large ribosomal subunit
cytoplasmic translation
cytosolic ribosome
ribosome
structural constituent of ribosome
translation
cytoplasm
cytosol
protein binding
nucleus
ribonucleoprotein complex
spliceosomal complex
mRNA processing
RNA splicing
mRNA splicing, via spliceosome
nuclear speck
catalytic step 2 spliceosome
extracellular exosome
cytosolic small ribosomal subunit
small−subunit processome
ribosomal small subunit biogenesis
polysomal ribosome
membrane
focal adhesion
nucleolus
rRNA processing
synapse
small ribosomal subunit
mitochondrial inner membrane
mitochondrial translation
mitochondrion
endoplasmic reticulum
endoplasmic reticulum membrane
secretory granule lumen
ficolin−1−rich granule lumen
mRNA binding
cytoplasmic stress granule
ATP binding
macromolecular complex
ATPase activity
nucleic acid binding
RNA helicase activity
mRNA 3'−UTR binding
mRNA transport
ubiquitin protein ligase binding
helicase activity
nuclear body
midbody
DNA binding
eukaryotic translation initiation factor 4F complex
primary miRNA processing
kinetochore
cadherin binding
chromatin binding
chromatin organization
nuclear matrix
translational initiation
translation initiation factor activity
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
-log10FDR
1 2 3 4 5
1 10 20 30
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other

Fig. 10

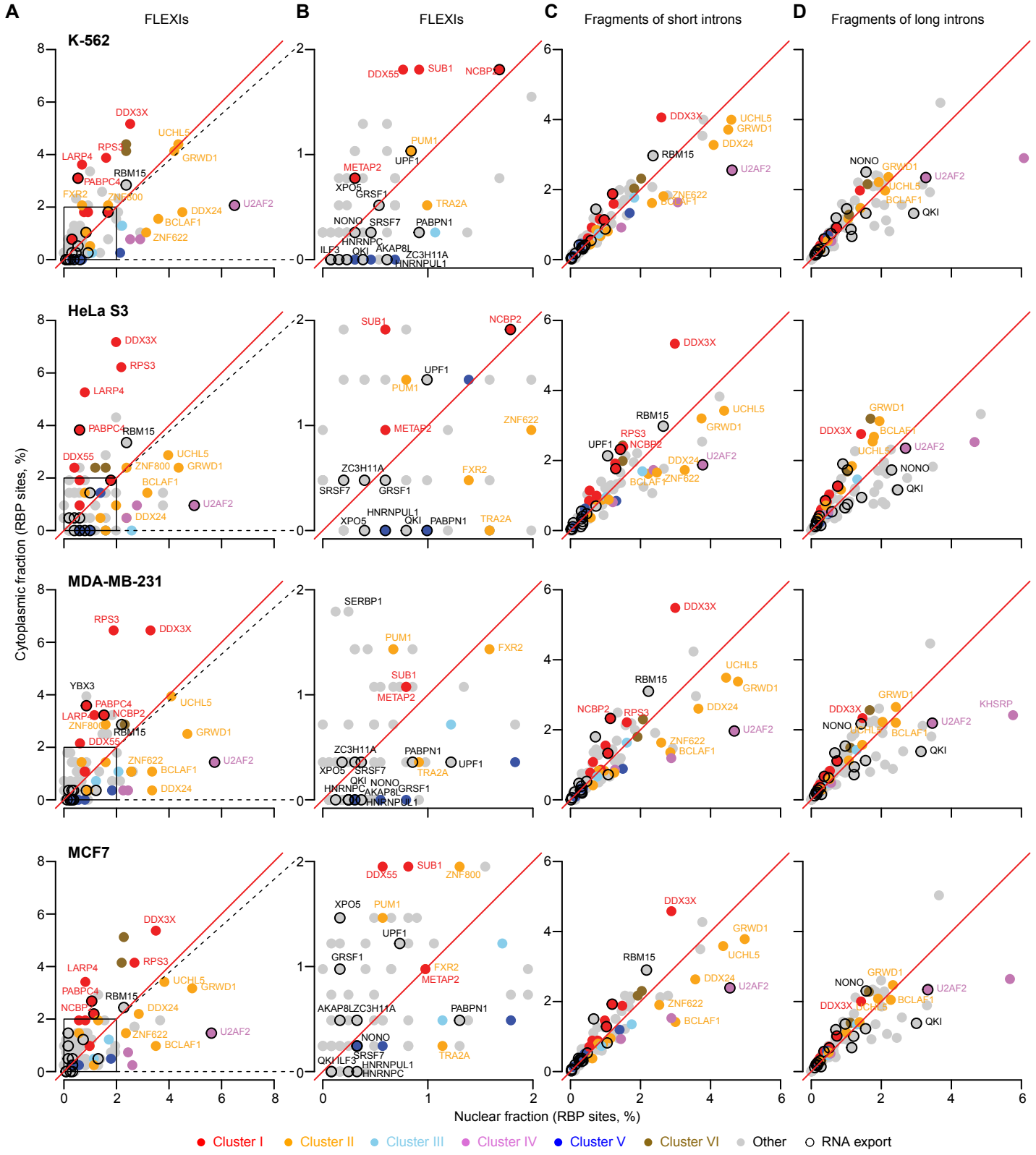

A
FLEXIs
B
FLEXIs
C
Fragments of short introns
D
Fragments of long introns
K-562
HeLa S3
MDA-MB-231
MCF7
Cytoplasmic fraction (RBP sites, %)
Nuclear fraction (RBP sites, %)
Cluster I
Cluster II
Cluster III
Cluster IV
Cluster V
Cluster VI
Other
RNA export

**Fig. 11**

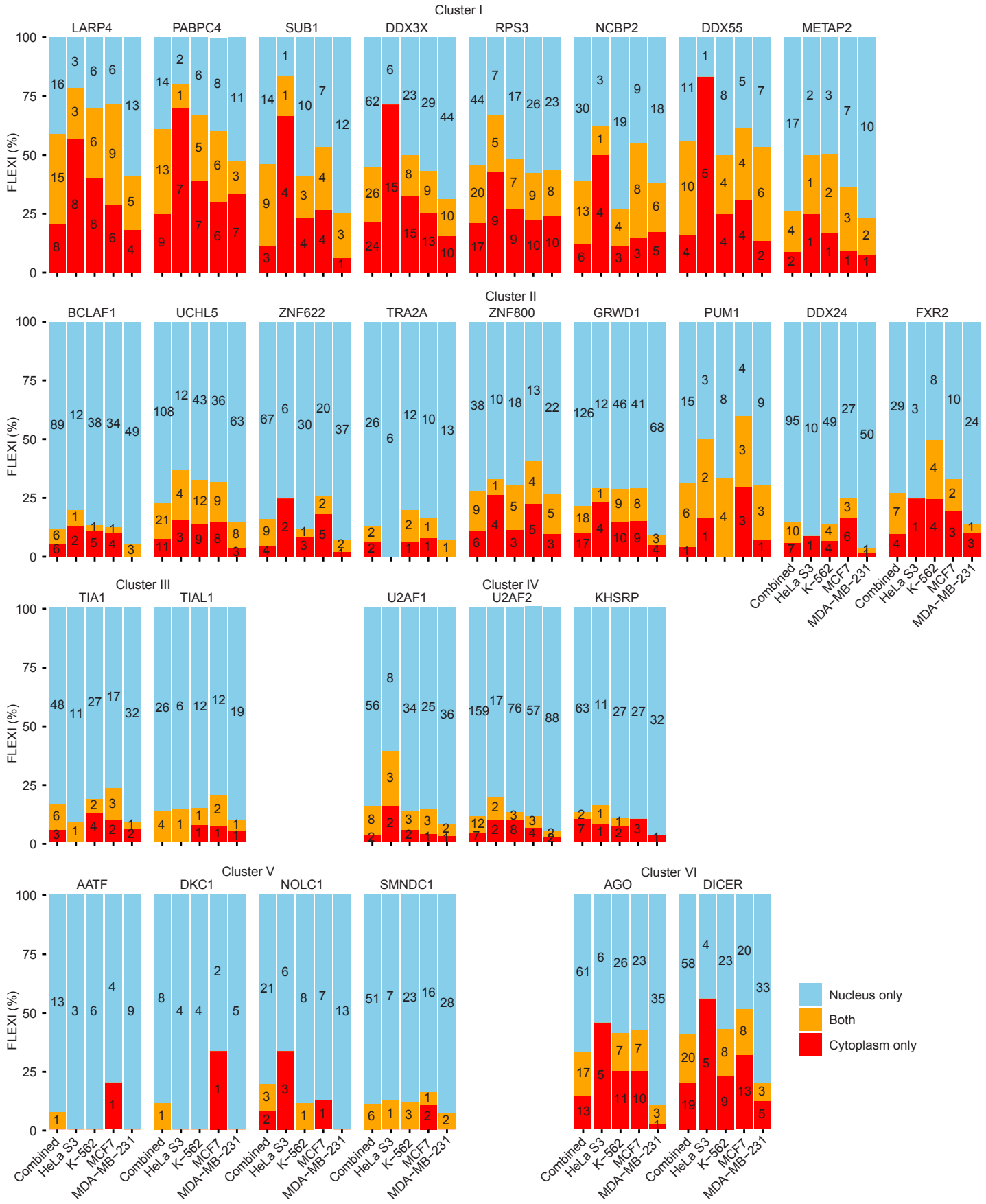
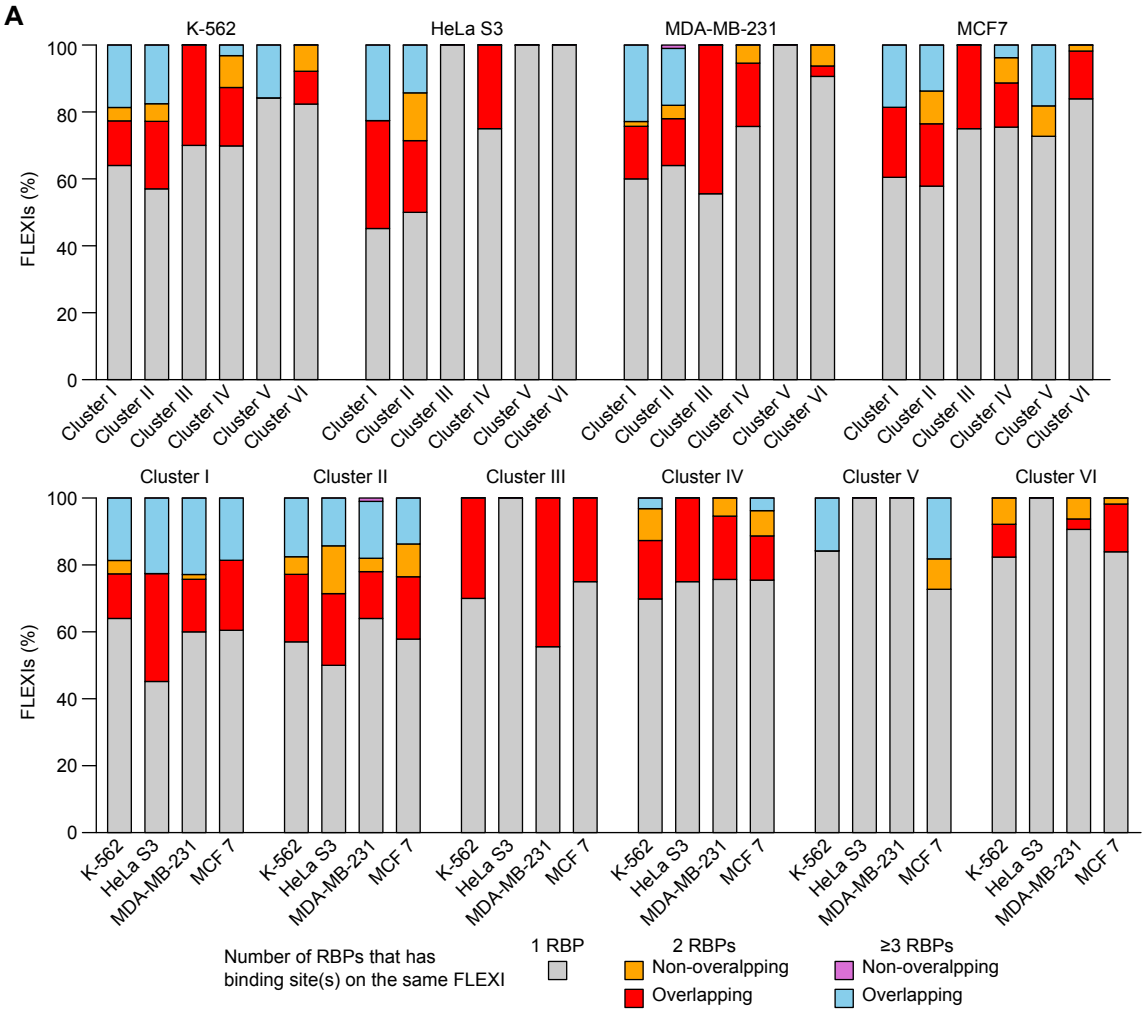

Cluster I
LARP4
PABPC4
SUB1
DDX3X
RPS3
NCBP2
DDX55
METAP2
Cluster II
BCLAF1
UCHL5
ZNF622
TRA2A
ZNF800
GRWD1
PUM1
DDX24
FXR2
Cluster III
TIA1
TIAL1
Cluster IV
U2AF1
U2AF2
KHSRP
Cluster V
AATF
DKC1
NOLC1
SMNDC1
Cluster VI
AGO
DICER
FLEXI (%)
Combined
HeLa S3
K-562
MCF7
MDA-MB-231
Nucleus only
Both
Cytoplasm only

Fig. 12


A
K-562
HeLa S3
MDA-MB-231
MCF7
FLEXIs (%)
100
80
60
40
20
0
Cluster I
Cluster II
Cluster III
Cluster IV
Cluster V
Cluster VI
K-562
HeLa S3
MDA-MB-231
MCF 7
Number of RBPs that has binding site(s) on the same FLEXI
1 RBP
2 RBPs
Non-overalpping
Overlapping
≥3 RBPs
Non-overalpping
Overlapping

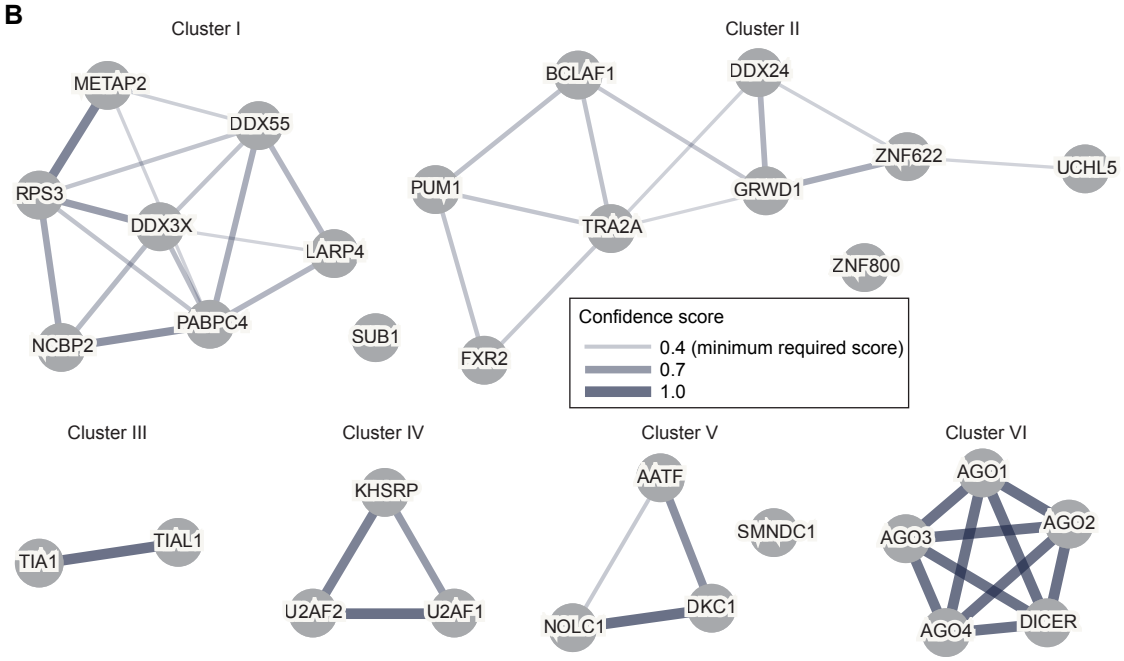

B
Cluster I
METAP2
DDX55
RPS3
DDX3X
LARP4
PABPC4
NCBP2
SUB1
Cluster II
BCLAF1
DDX24
ZNF622
UCHL5
PUM1
GRWD1
TRA2A
ZNF800
FXR2
Confidence score
0.4 (minimum required score)
0.7
1.0
Cluster III
TIA1
TIAL1
Cluster IV
KHSRP
U2AF2
U2AF1
Cluster V
AATF
SMNDC1
NOLC1
DKC1
Cluster VI
AGO1
AGO2
AGO3
AGO4
DICER

**Fig. S1**

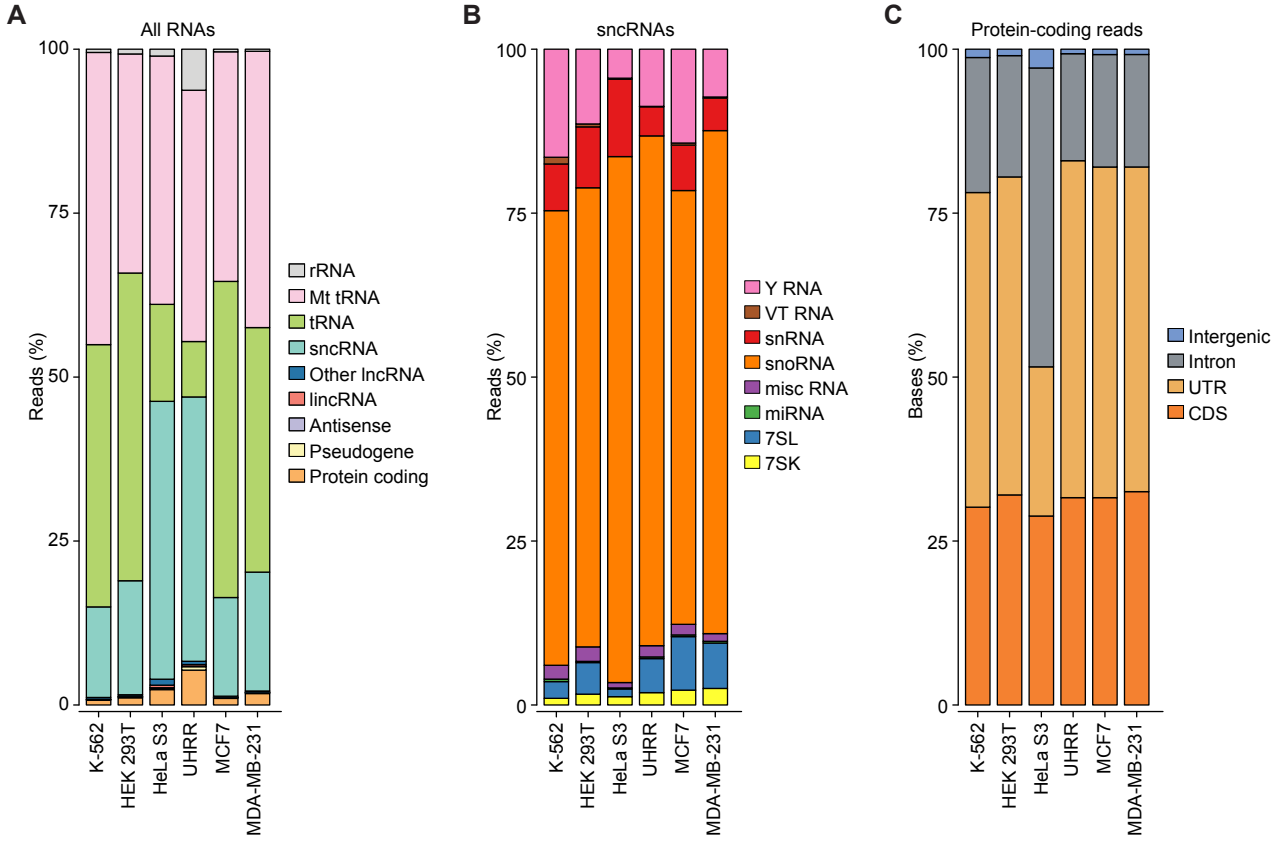

A
All RNAs
Reads (%)
100
75
50
25
0
K-562
HEK 293T
HeLa S3
UHRR
MCF7
MDA-MB-231
rRNA
Mt tRNA
tRNA
sncRNA
Other lncRNA
lincRNA
Antisense
Pseudogene
Protein coding
B
sncRNAs
Reads (%)
Y RNA
VT RNA
snRNA
snoRNA
misc RNA
miRNA
7SL
7SK
C
Protein-coding reads
Bases (%)
Intergenic
Intron
UTR
CDS

**Fig. S2**

snoRA24 (131 nt)
snoRD74 (72 nt)
HEK-293T
HeLa S3
K-562
UHRR
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)
m¹A58
CCA
tRNA-ValAAC (76 nt)
I
m¹A58
CCA

**Fig. S3**

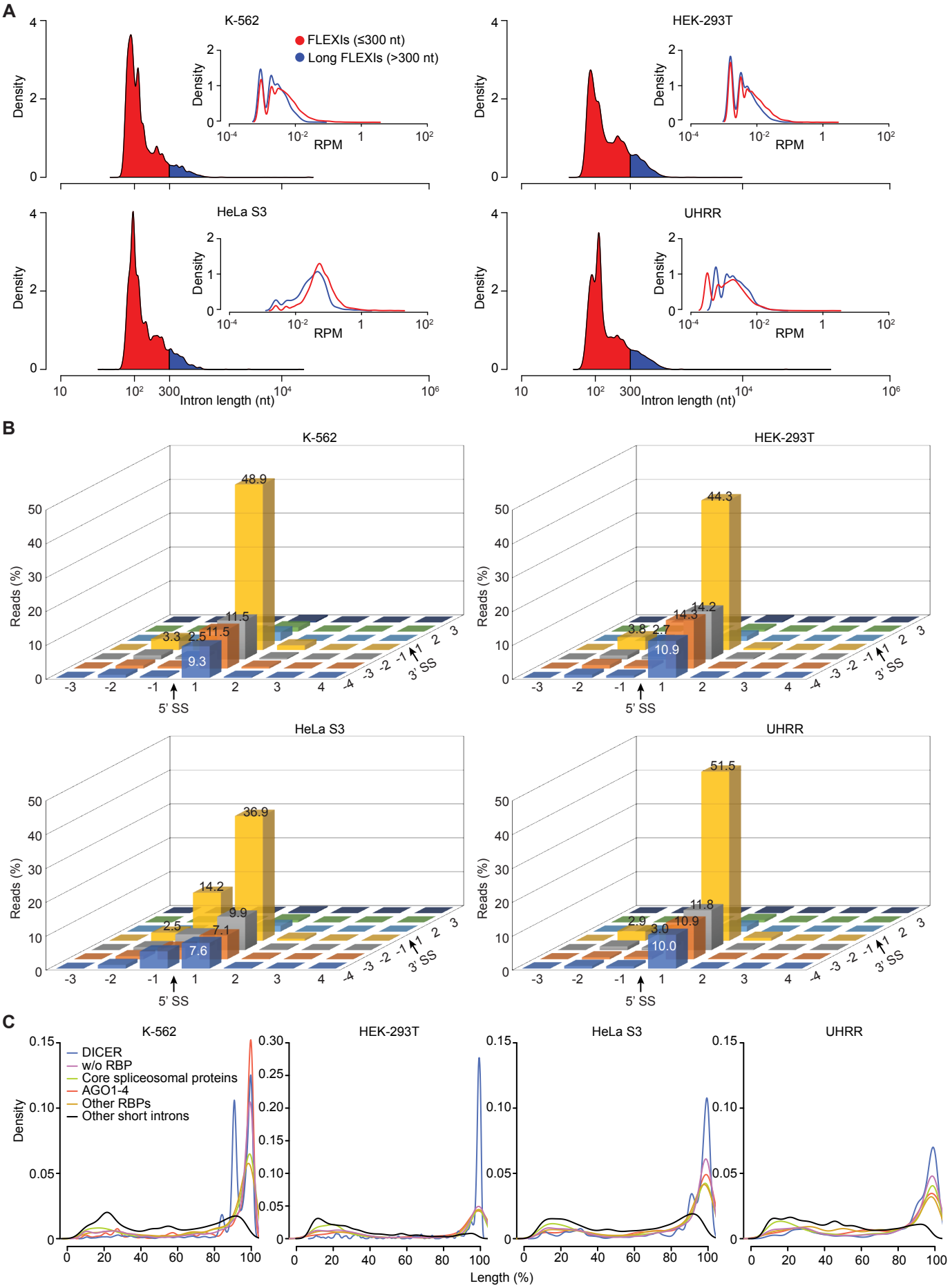

A
K-562
HEK-293T
HeLa S3
UHRR
FLEXIs (≤300 nt)
Long FLEXIs (>300 nt)
Density
RPM
Intron length (nt)
B
Reads (%)
5' SS
3' SS
48.9
11.5
3.3
2.5
9.3
44.3
14.2
14.3
3.8
2.7
10.9
36.9
14.2
9.9
2.5
7.1
7.6
51.5
11.8
10.9
2.9
3.0
10.0
C
DICER
w/o RBP
Core spliceosomal proteins
AGO1-4
Other RBPs
Other short introns
Length (%)

**Fig. S4**

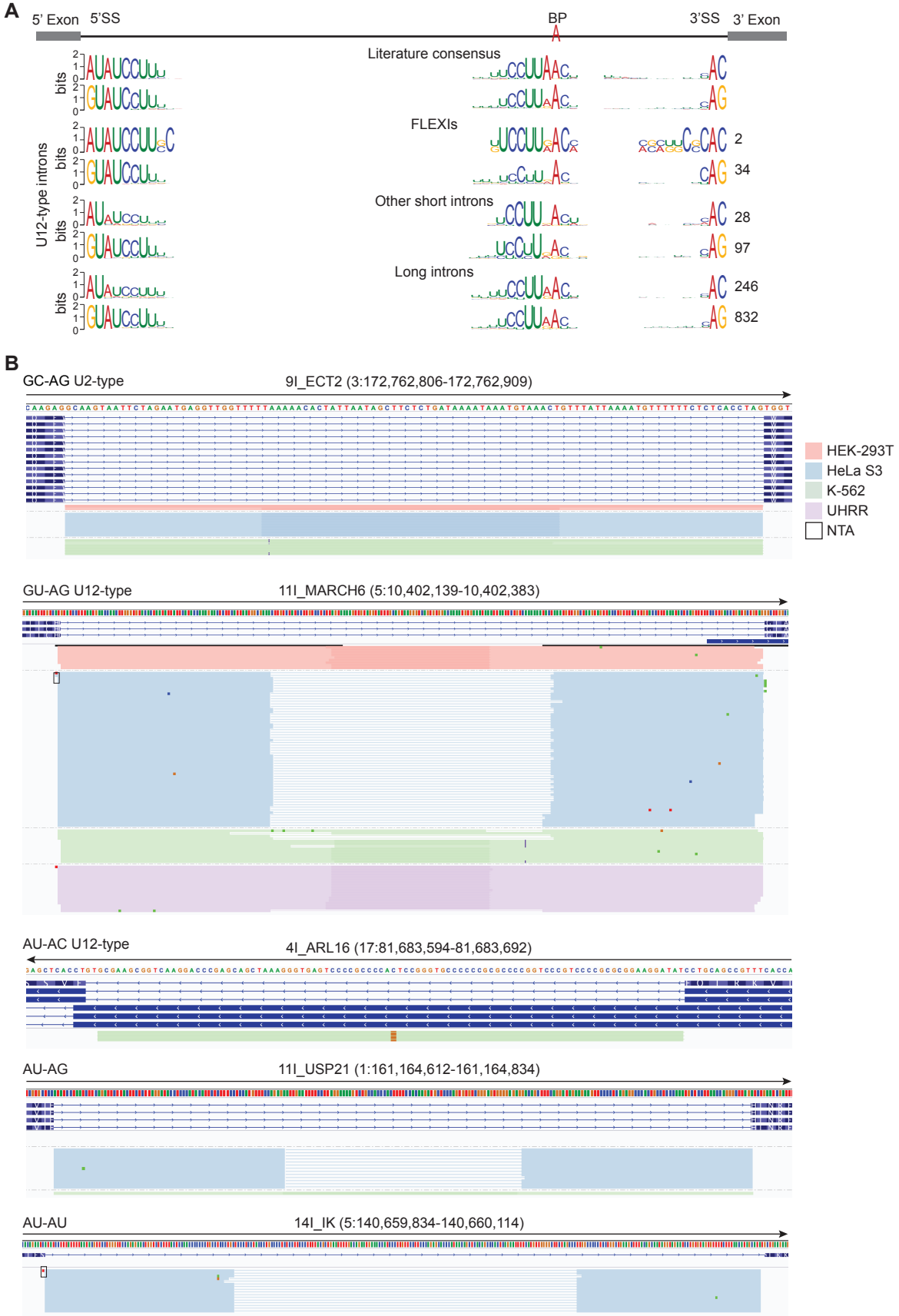
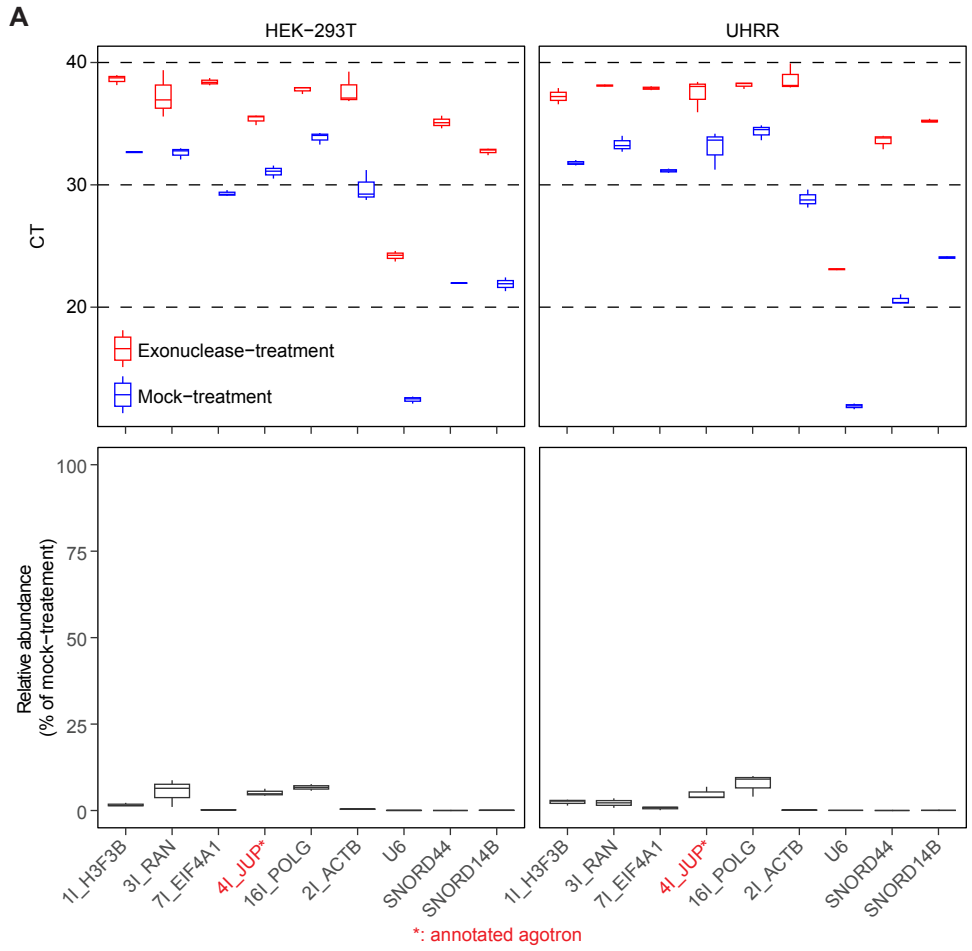
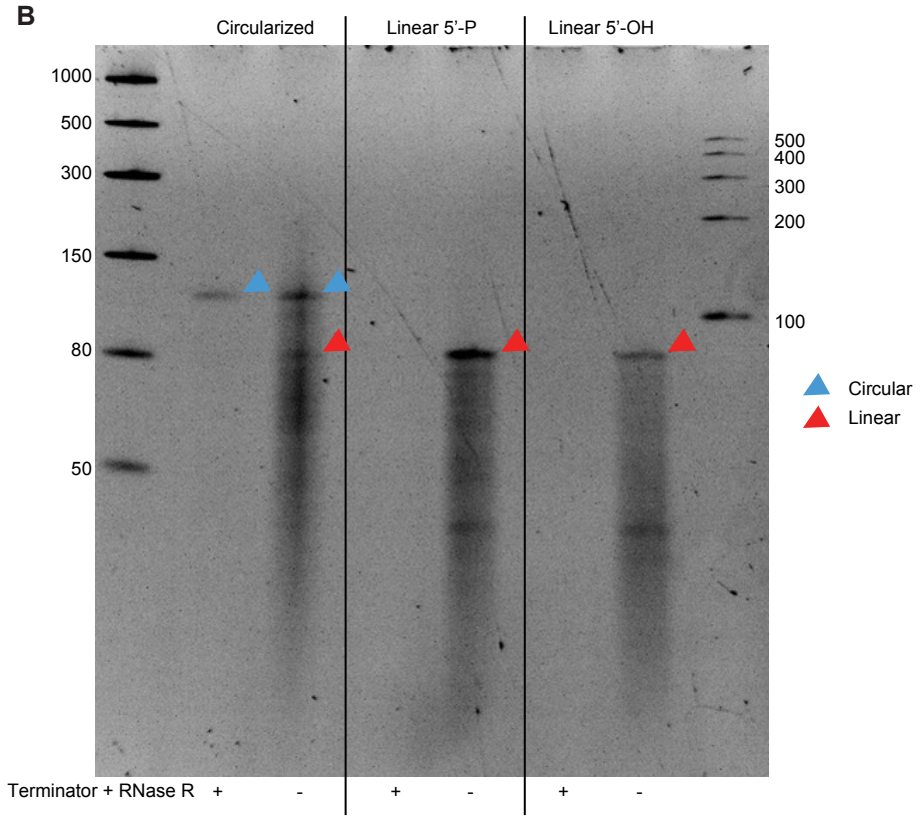

A
5' Exon
5'SS
BP
3'SS
3' Exon
Literature consensus
FLEXIs
Other short introns
Long introns
U12-type introns
bits
2
34
28
97
246
832
B
GC-AG U2-type
9I_ECT2 (3:172,762,806-172,762,909)
HEK-293T
HeLa S3
K-562
UHRR
NTA
GU-AG U12-type
11I_MARCH6 (5:10,402,139-10,402,383)
AU-AC U12-type
4I_ARL16 (17:81,683,594-81,683,692)
AU-AG
11I_USP21 (1:161,164,612-161,164,834)
AU-AU
14I_IK (5:140,659,834-140,660,114)

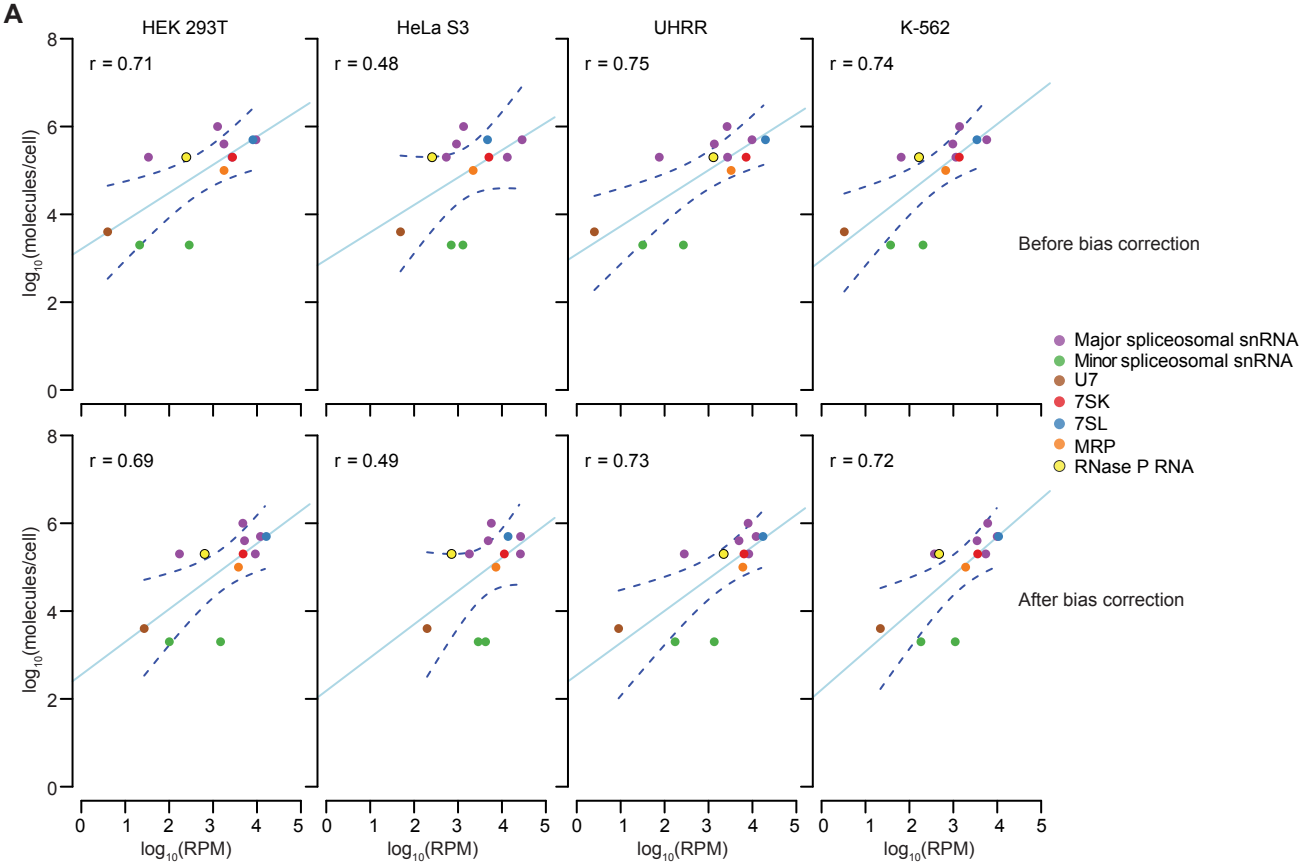
**Fig. S5**


A
HEK−293T
UHRR
CT
40
30
20
Exonuclease−treatment
Mock−treatment
Relative abundance
(% of mock−treatement)
100
75
50
25
0
1l_H3F3B
3l_RAN
7l_EIF4A1
4l_JUP*
16l_POLG
2l_ACTB
U6
SNORD44
SNORD14B
*: annotated agotron


B
Circularized
Linear 5'-P
Linear 5'-OH
1000
500
300
150
80
50
500
400
300
200
100
Circular
Linear
Terminator + RNase R
+
-
+
-
+
-

## Fig. S6

**A**



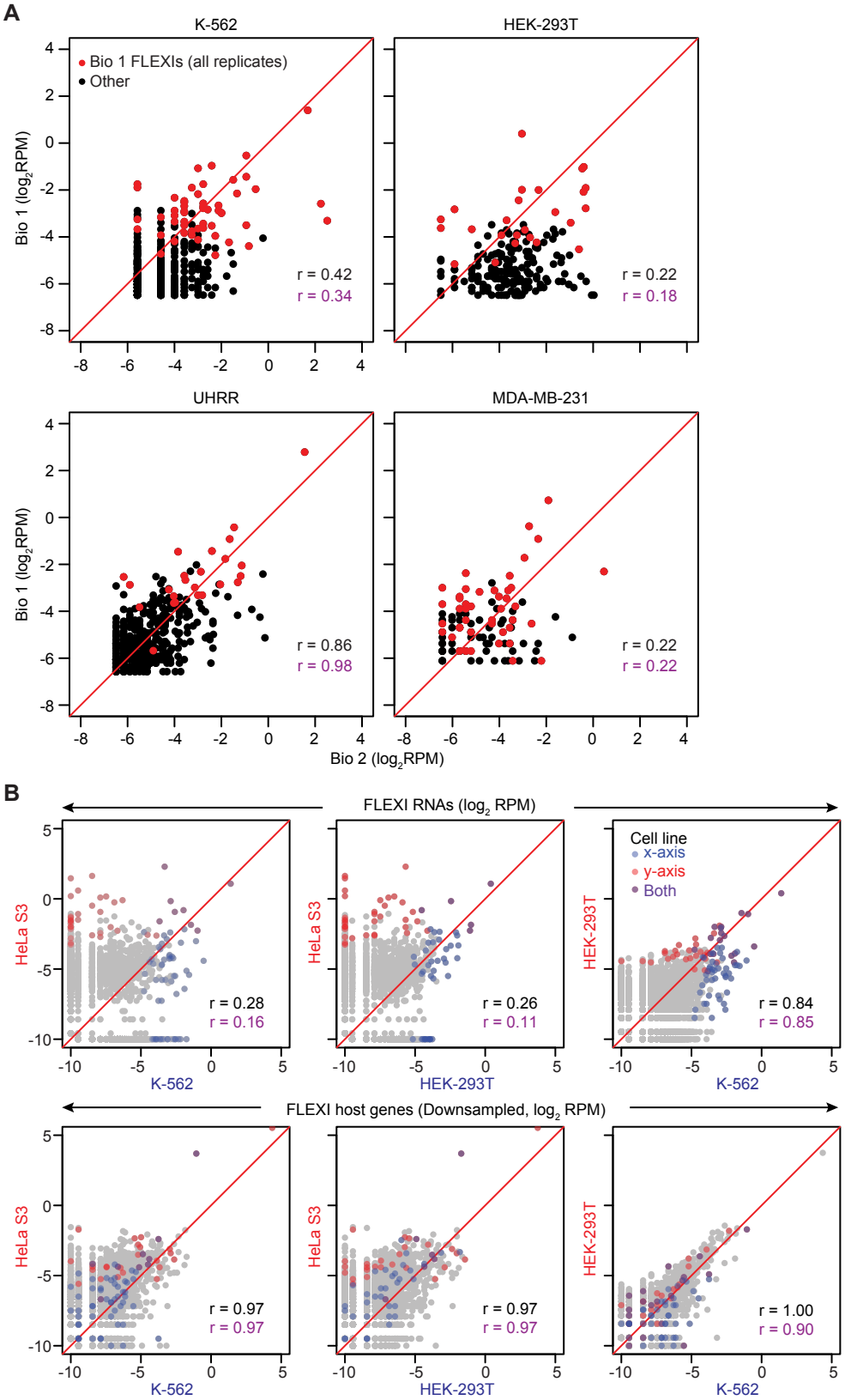

| Copies/cell | Literature value | TGIRT-seq (lrm, before bias correction) | | | TGIRT-seq (lrm, after bias correction) | | |
|---|---|---|---|---|---|---|---|
| | | HEK-293T | HeLa S3 | UHRR | HEK-293T | HeLa S3 | UHRR |
| U7 | 4,000 | 3,918 | 10,459 | 2,225 | 4,171 | 8,355 | 1,736 |
| SNORD14B | 10,000 | 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**

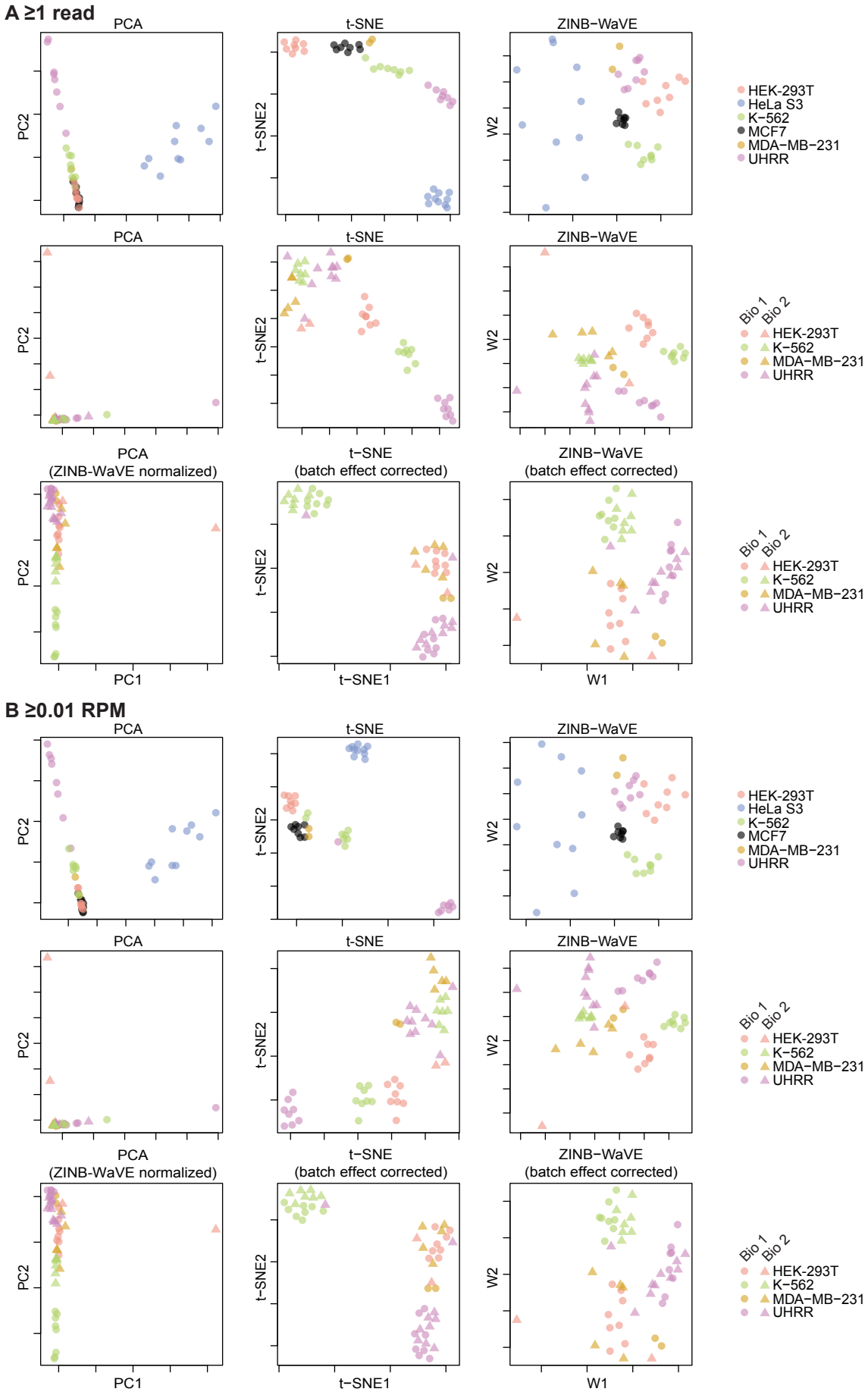
| FLEXIs | TGIRT-seq | | | | ddPCR (U7) | | ddPCR (SNORD14B) | | ddPCR (SNORD44) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | RPM | | Copies/cell | | Copies/cell | | Copies/cell | | Copies/cell | |
| | Original | Bias corrected | Original | Bias corrected | Original | Bias corrected | Original | 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 | 1,335 | 1,422 | 78 | 77 | 68 | 52 |
| 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 | 3,007 | 2,402 | 415 | 322 | 284 | 264 |
| 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 | 460 | 359 | 158 | 149 | 25 | 20 |
| 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**


A
K-562
HEK-293T
UHRR
MDA-MB-231
Bio 1 FLEXIs (all replicates)
Other
Bio 1 (log2RPM)
Bio 2 (log2RPM)
r = 0.42
r = 0.34
r = 0.22
r = 0.18
r = 0.86
r = 0.98
r = 0.22
r = 0.22
B
FLEXI RNAs (log2 RPM)
FLEXI host genes (Downsampled, log2 RPM)
Cell line
x-axis
y-axis
Both
HeLa S3
HEK-293T
K-562
r = 0.28
r = 0.16
r = 0.26
r = 0.11
r = 0.84
r = 0.85
r = 0.97
r = 0.97
r = 0.97
r = 0.97
r = 1.00
r = 0.90

## 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

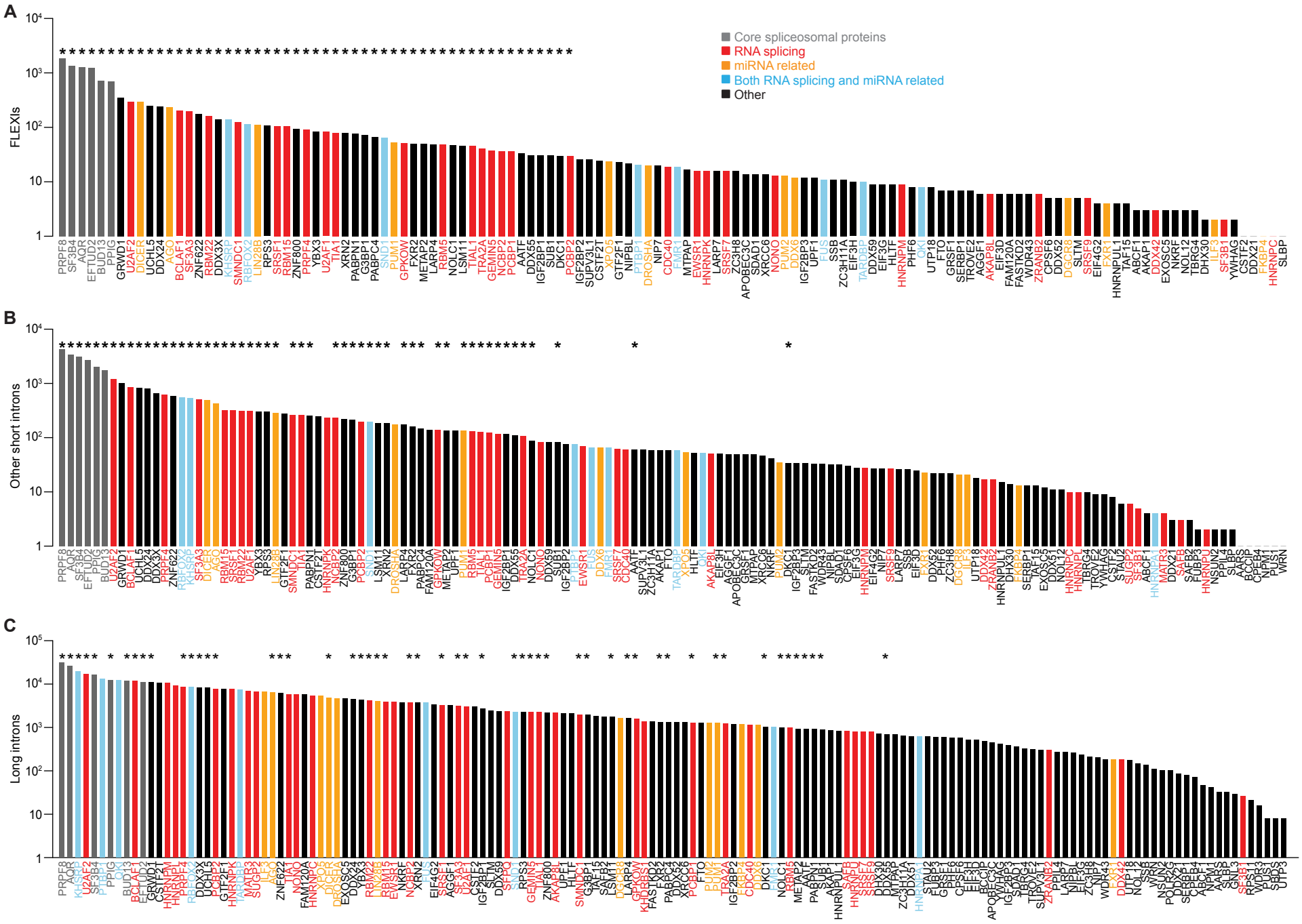

A
FLEXIs
Core spliceosomal proteins
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other
B
Other short introns
C
Long introns

**Fig. S10**

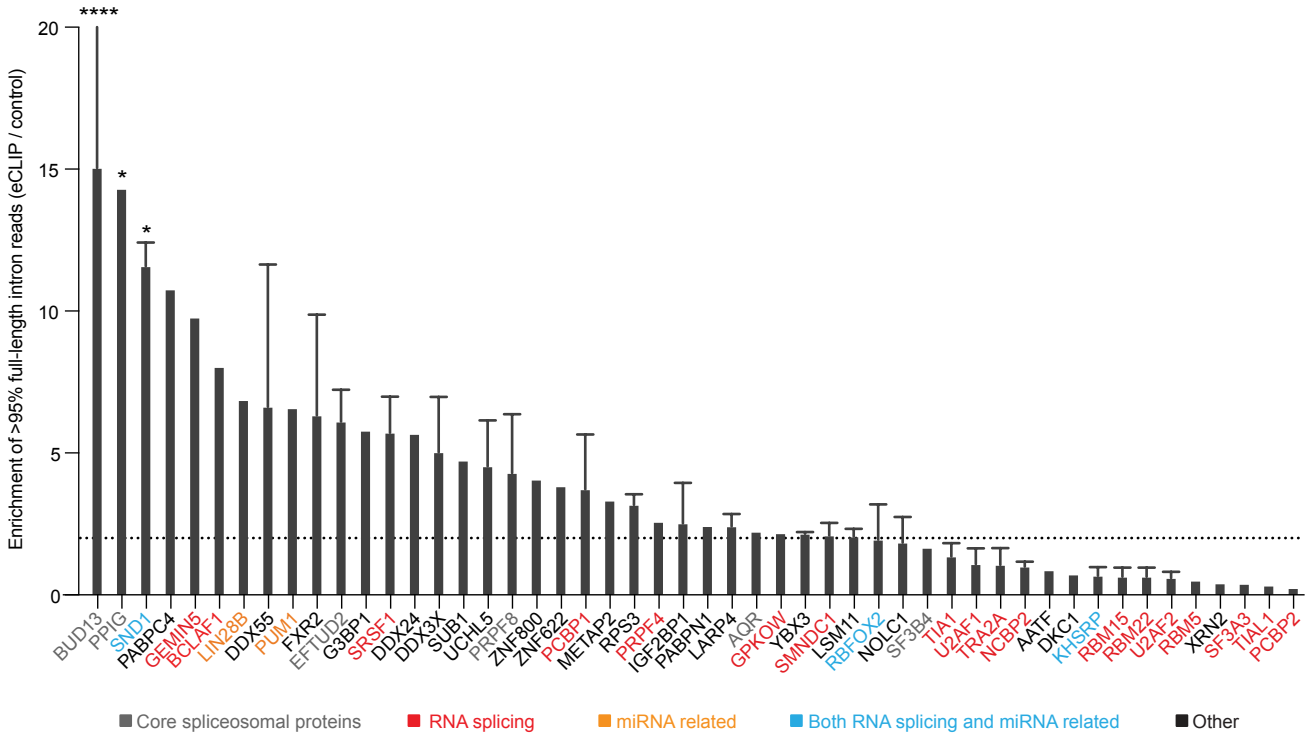
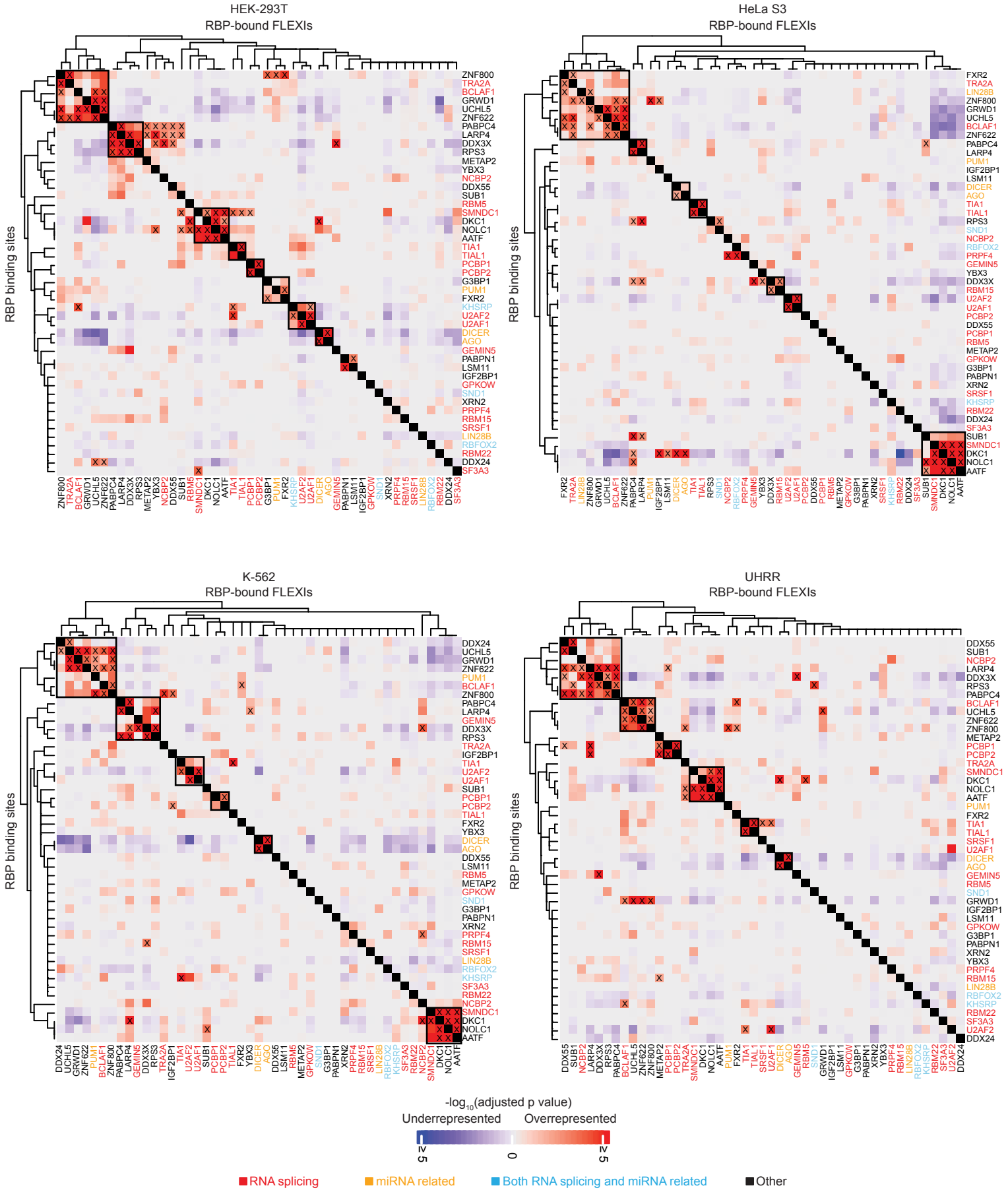
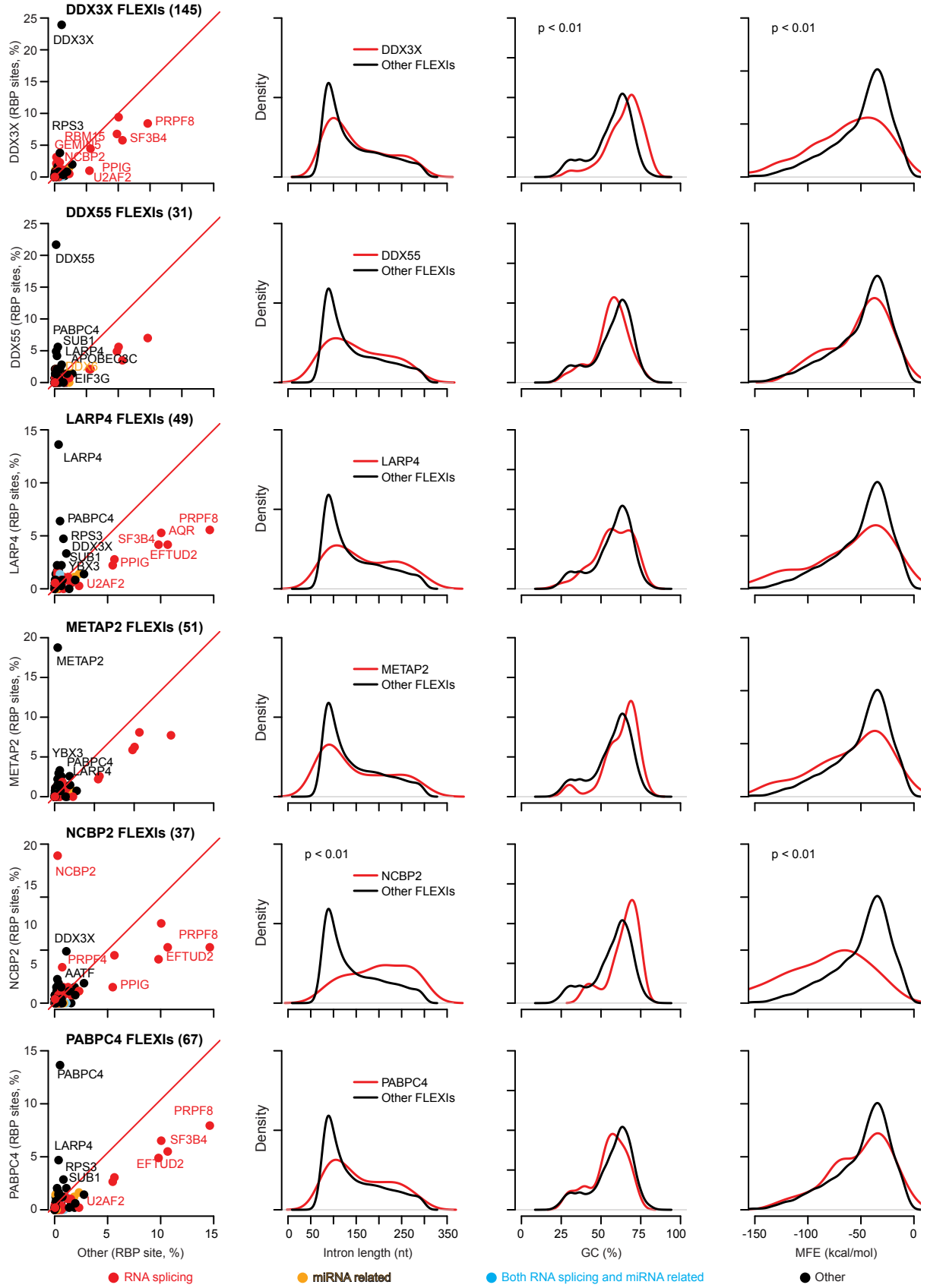

Enrichment of >95% full-length intron reads (eCLIP / control)
20
15
10
5
0
****
*
*
BUD13
PPIG
SND1
PABPC4
GEMIN5
BCLAF1
LIN28B
DDX55
PUM1
FXR2
EFTUD2
G3BP1
SRSF1
DDX24
DDX3X
SUB1
UCHL5
PRPF8
ZNF800
ZNF622
PCBP1
METAP2
RPS3
PRPF4
IGF2BP1
PABPN1
LARP4
AQR
GPKOW
YBX3
SMNDC1
LSM11
RBFOX2
NOLC1
SF3B4
TIA1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

**Fig. S11**


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
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other

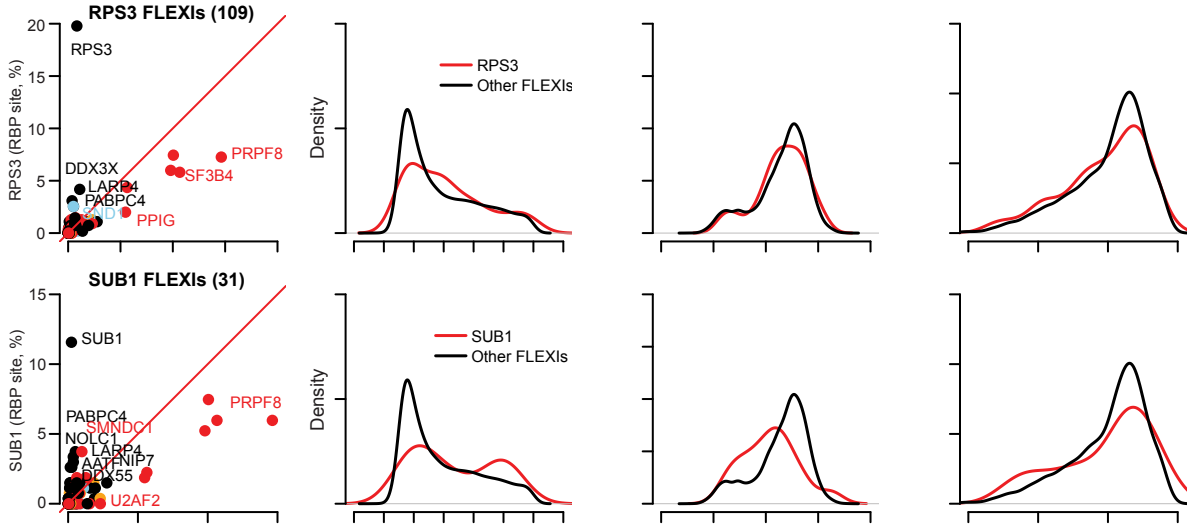
# Fig. S12

## A. Cluster I

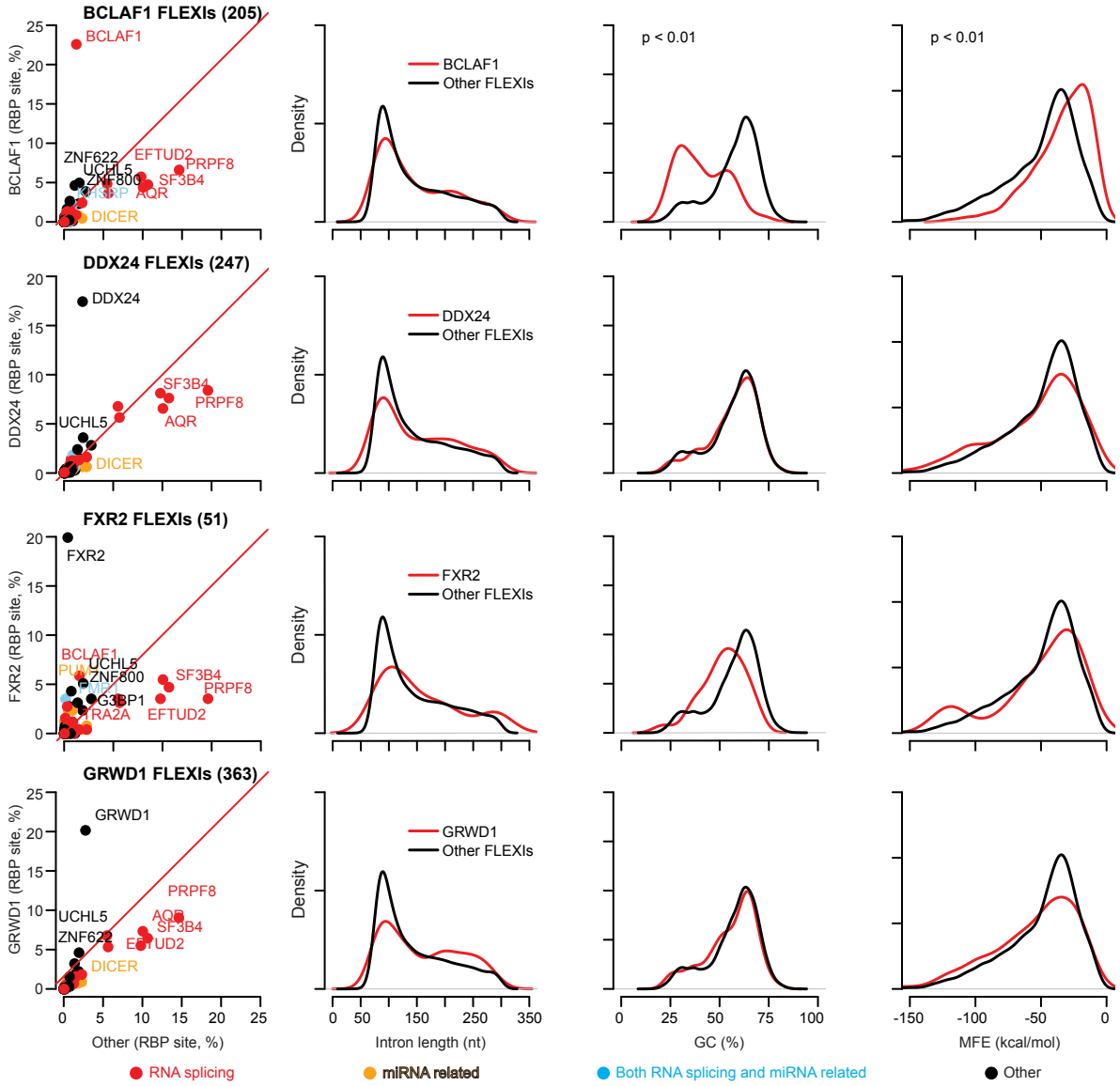


# Fig. S12 (continued)

## A. Cluster I

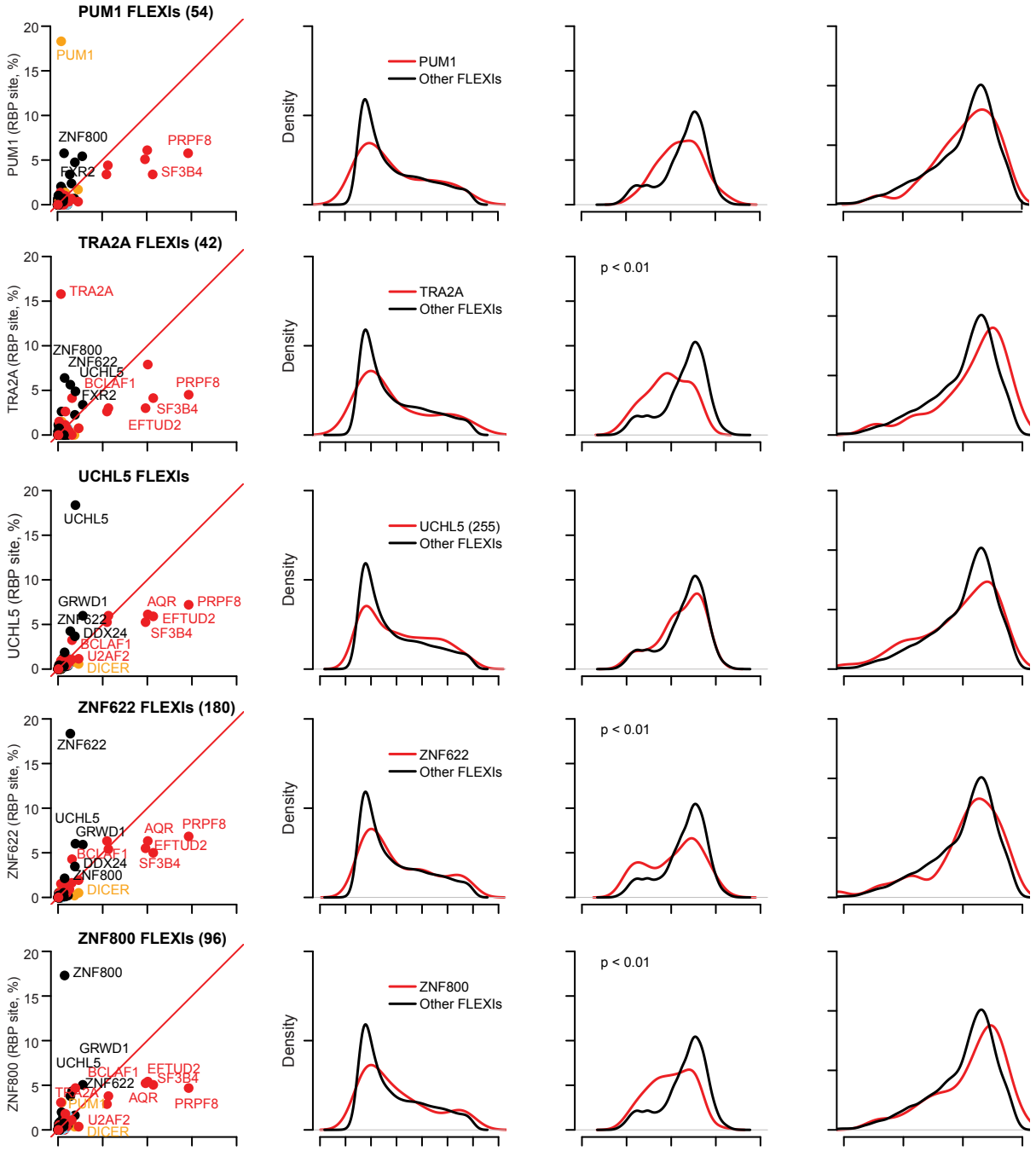



## B. Cluster II

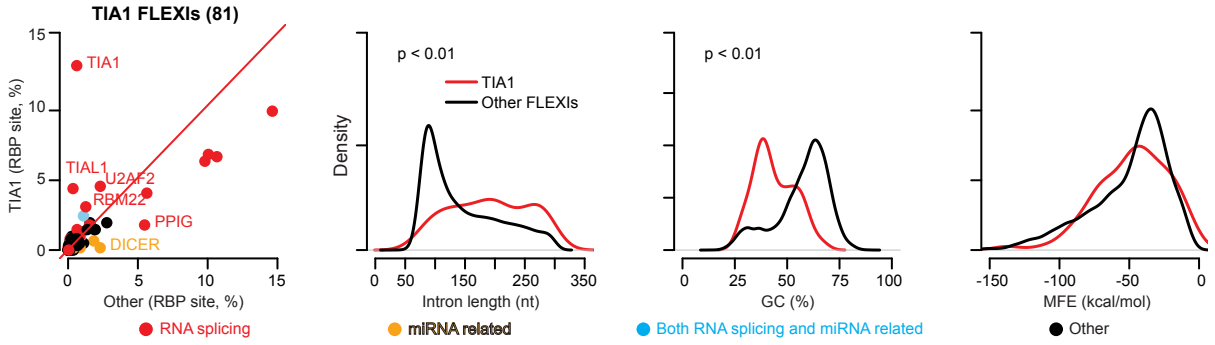


# Fig. S12 (continued)

## B. Cluster II

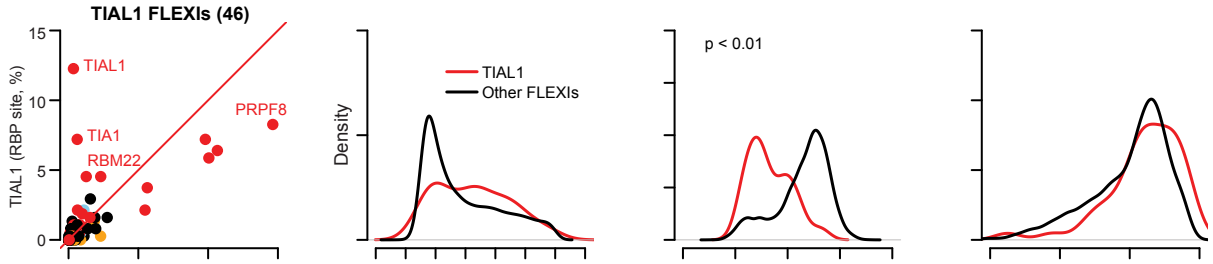



## C. Cluster III

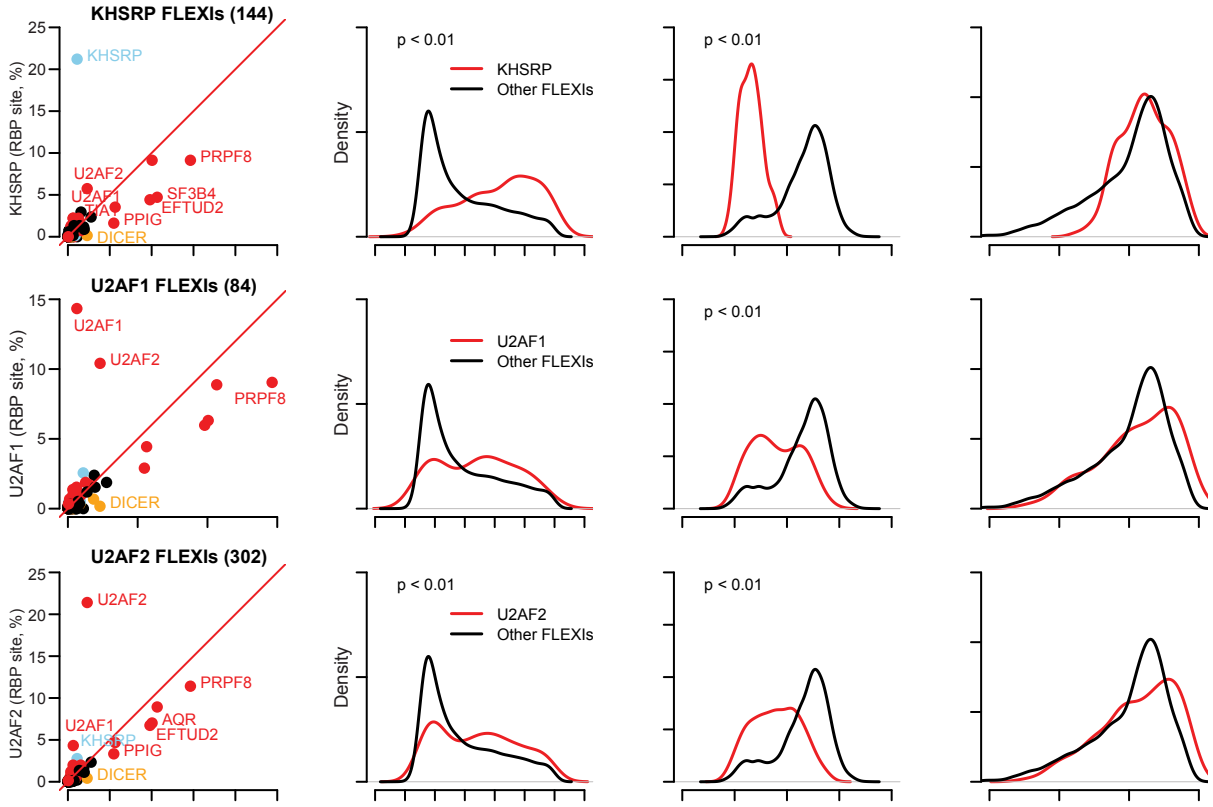


# Fig. S12 (continued)

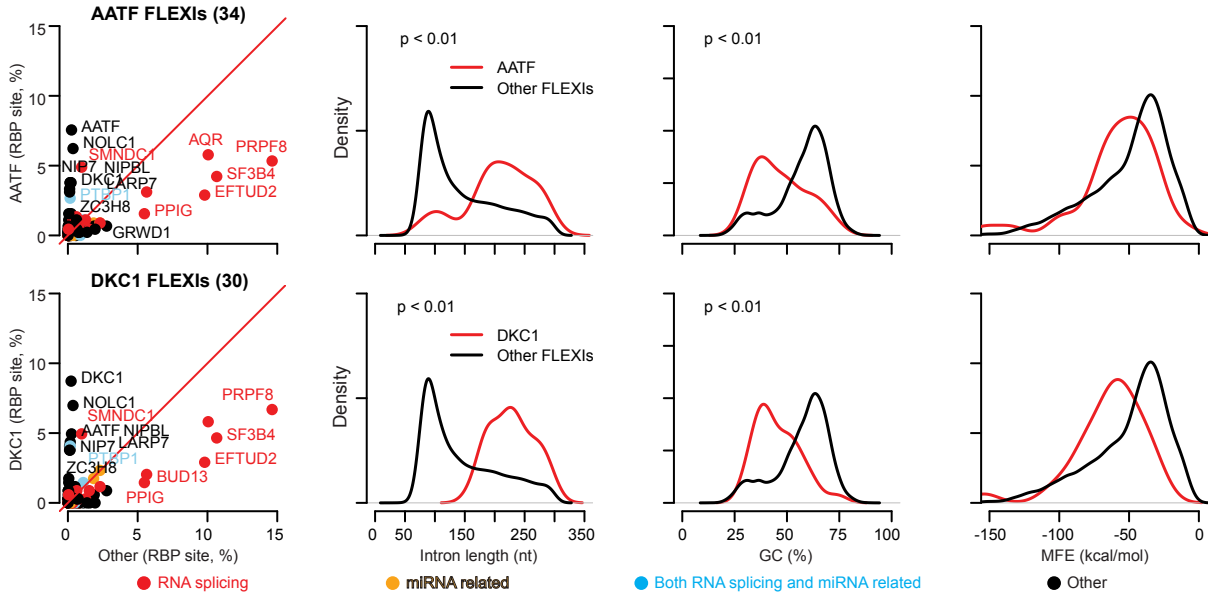
## C. Cluster III




## D. Cluster IV

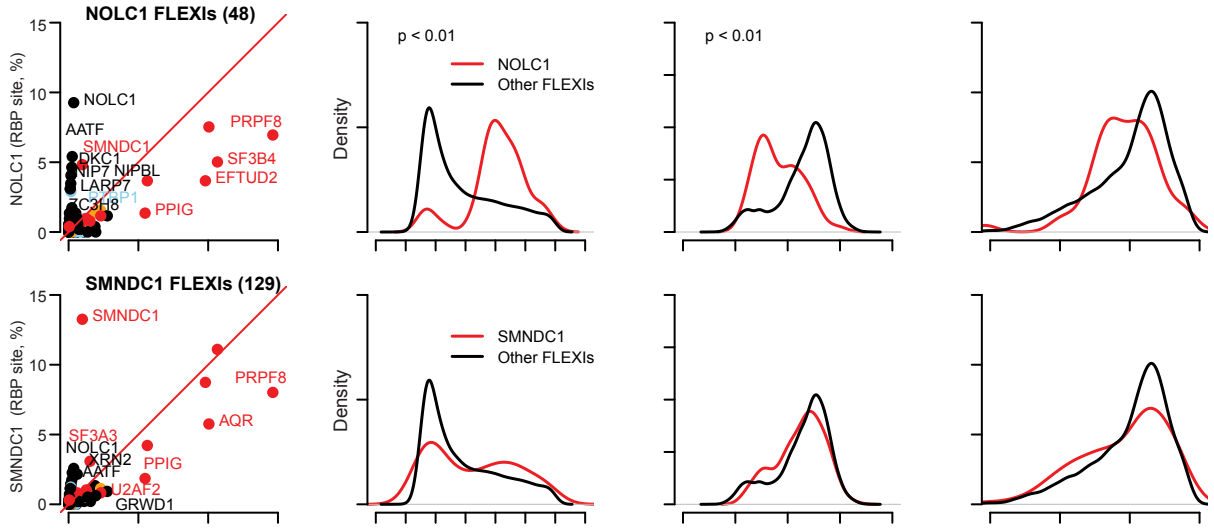



## E. Cluster V



# Fig. S12 (continued)

## E. Cluster V




## F. Cluster VI

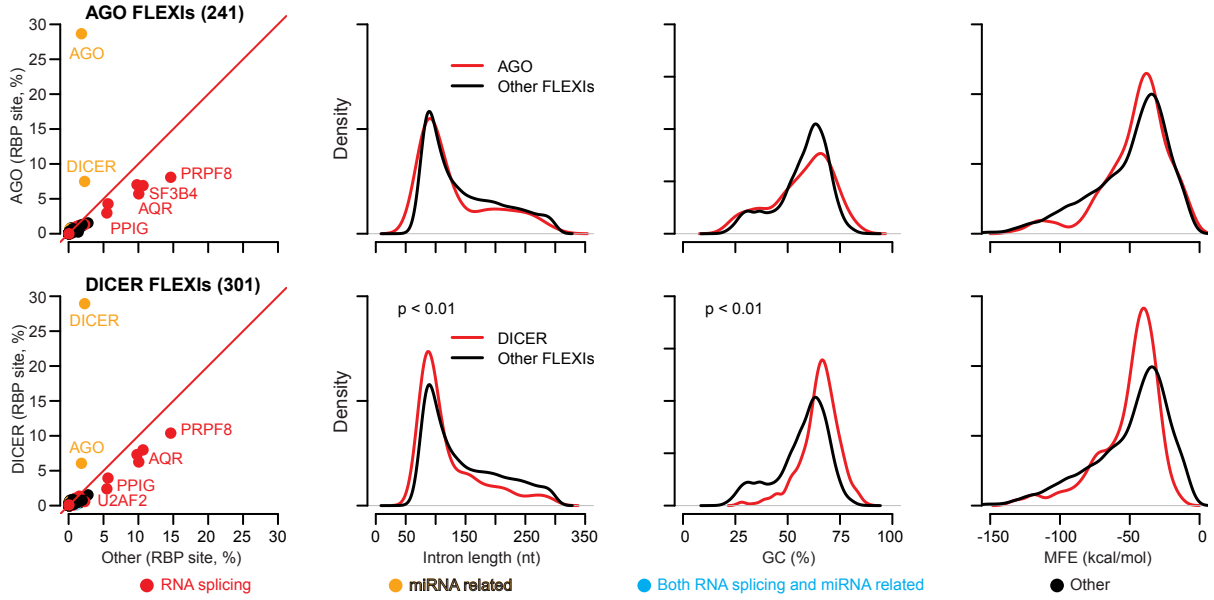
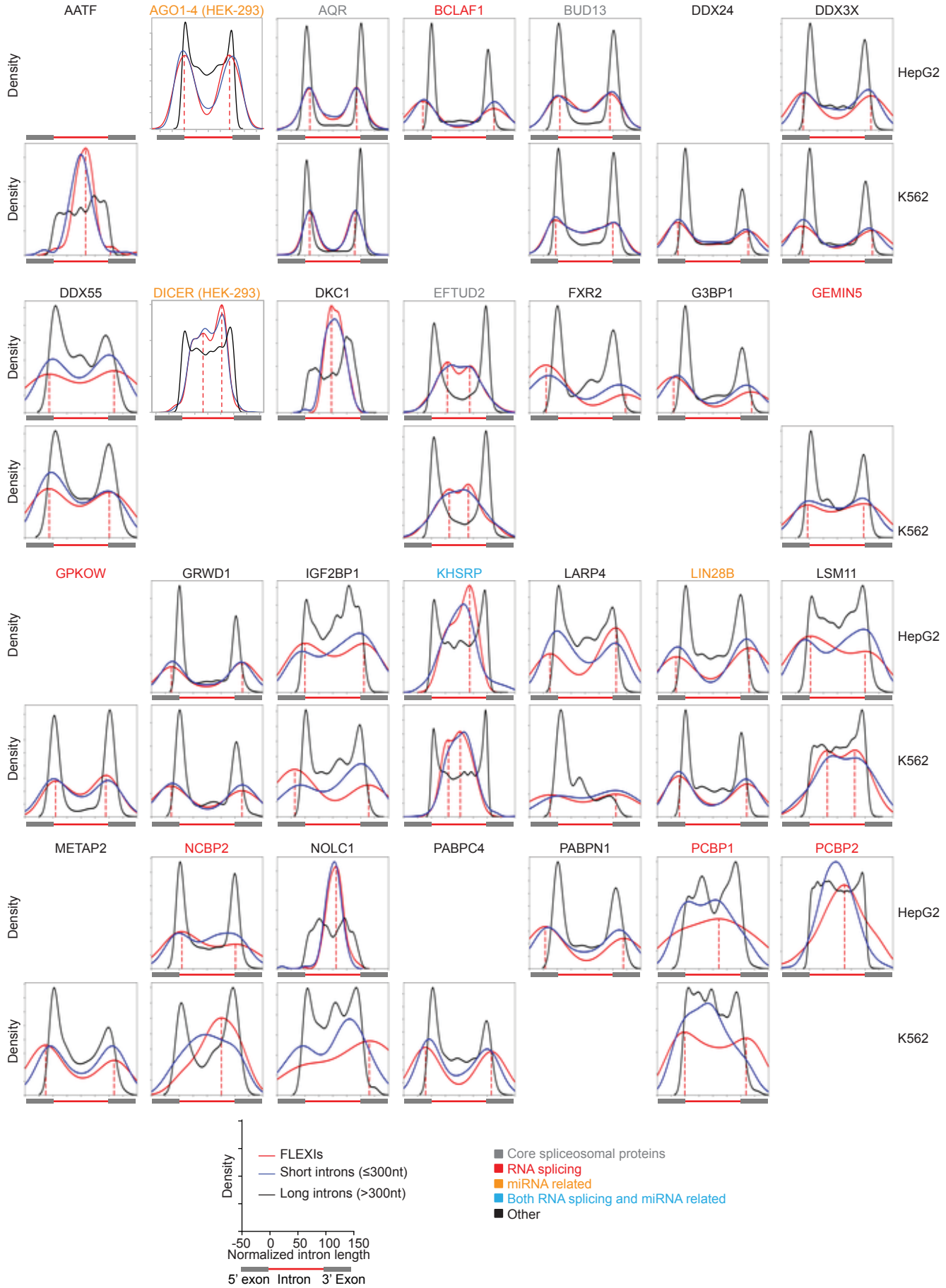
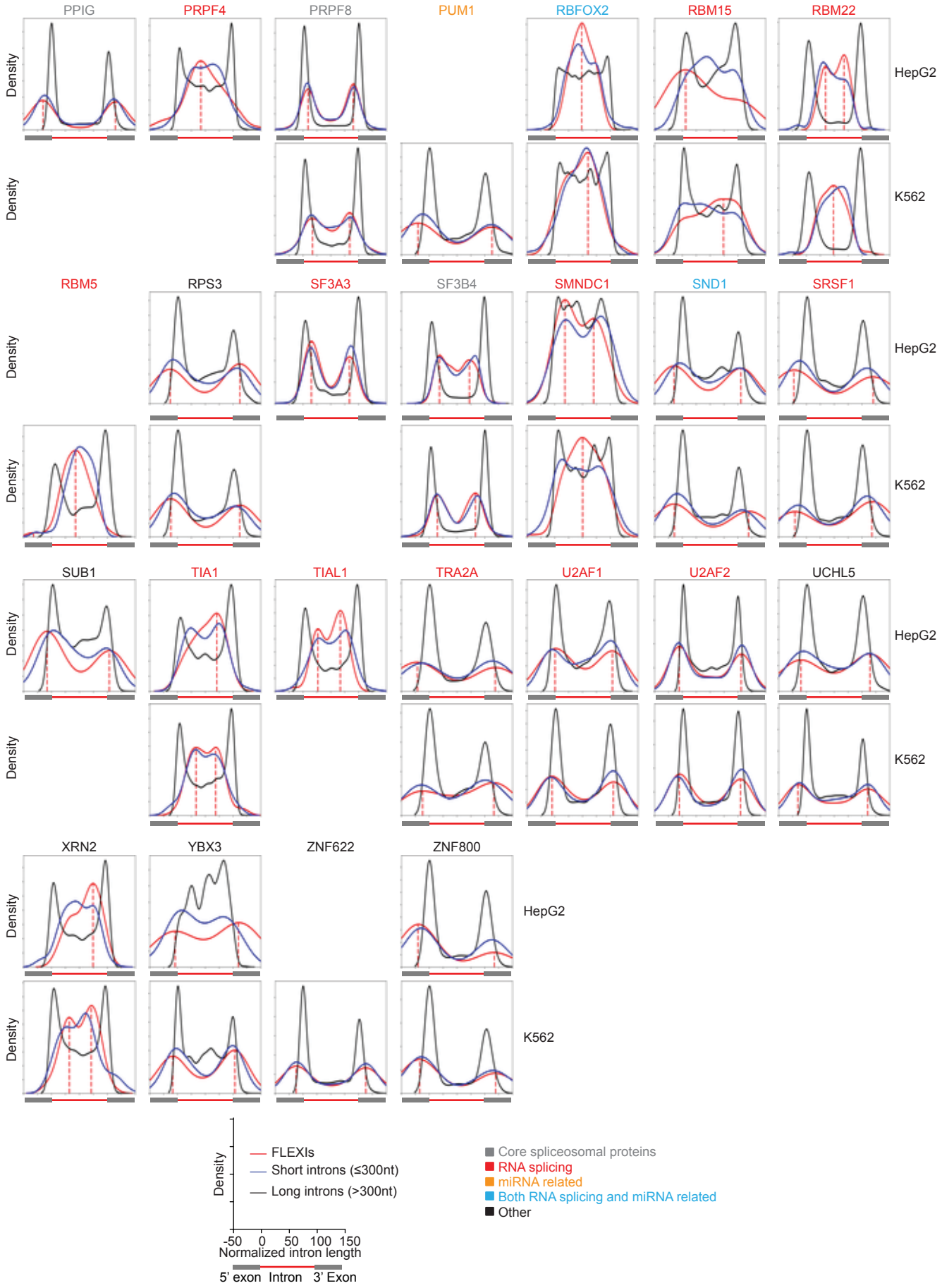


**Fig. S13**

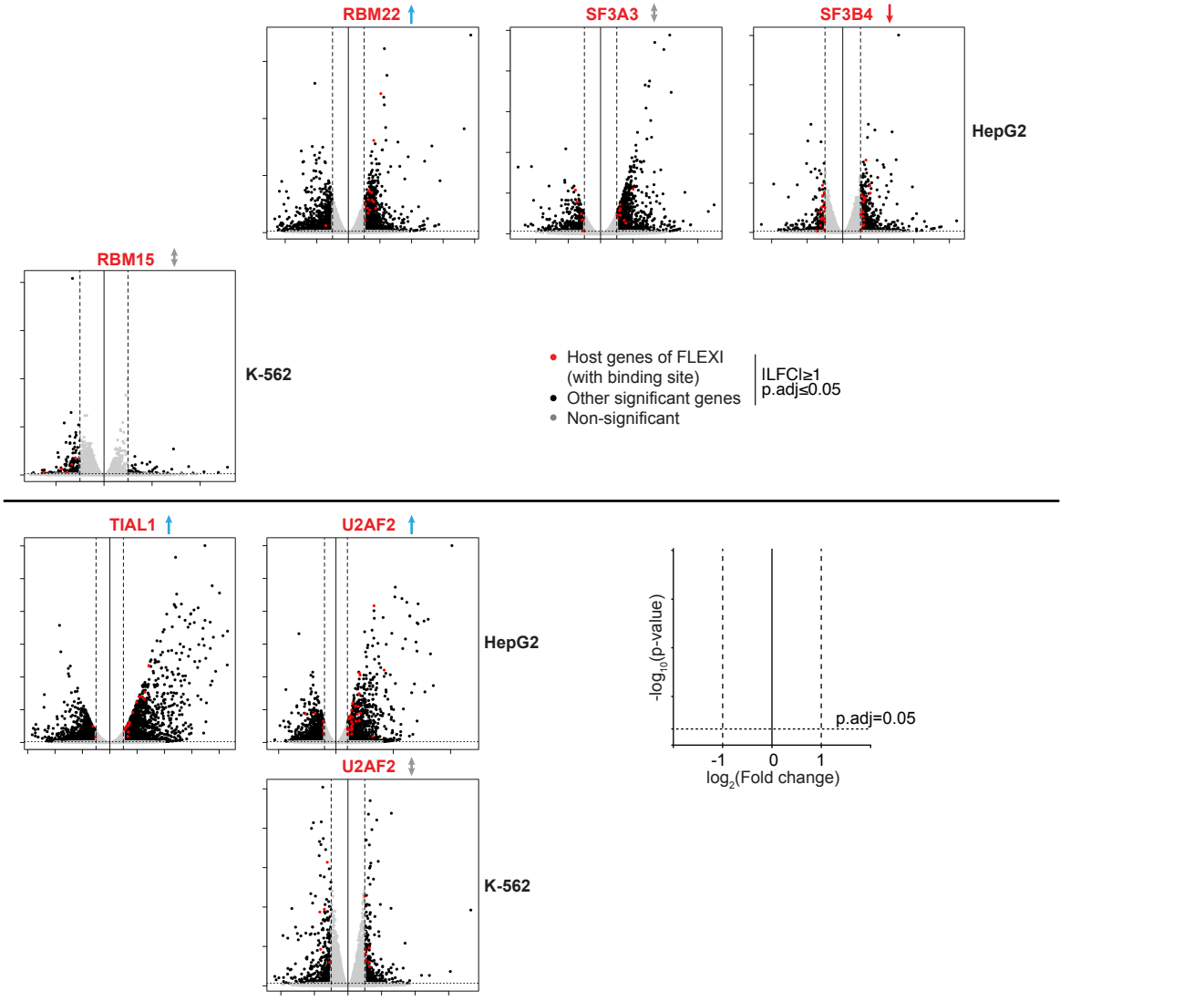

AATF
AGO1-4 (HEK-293)
AQR
BCLAF1
BUD13
DDX24
DDX3X
HepG2
K562
DDX55
DICER (HEK-293)
DKC1
EFTUD2
FXR2
G3BP1
GEMIN5
GPKOW
GRWD1
IGF2BP1
KHSRP
LARP4
LIN28B
LSM11
METAP2
NCBP2
NOLC1
PABPC4
PABPN1
PCBP1
PCBP2
Density
FLEXIs
Short introns (≤300nt)
Long introns (>300nt)
-50 0 50 100 150
Normalized intron length
5' exon Intron 3' Exon
Core spliceosomal proteins
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other

## Fig. S13 (continued)

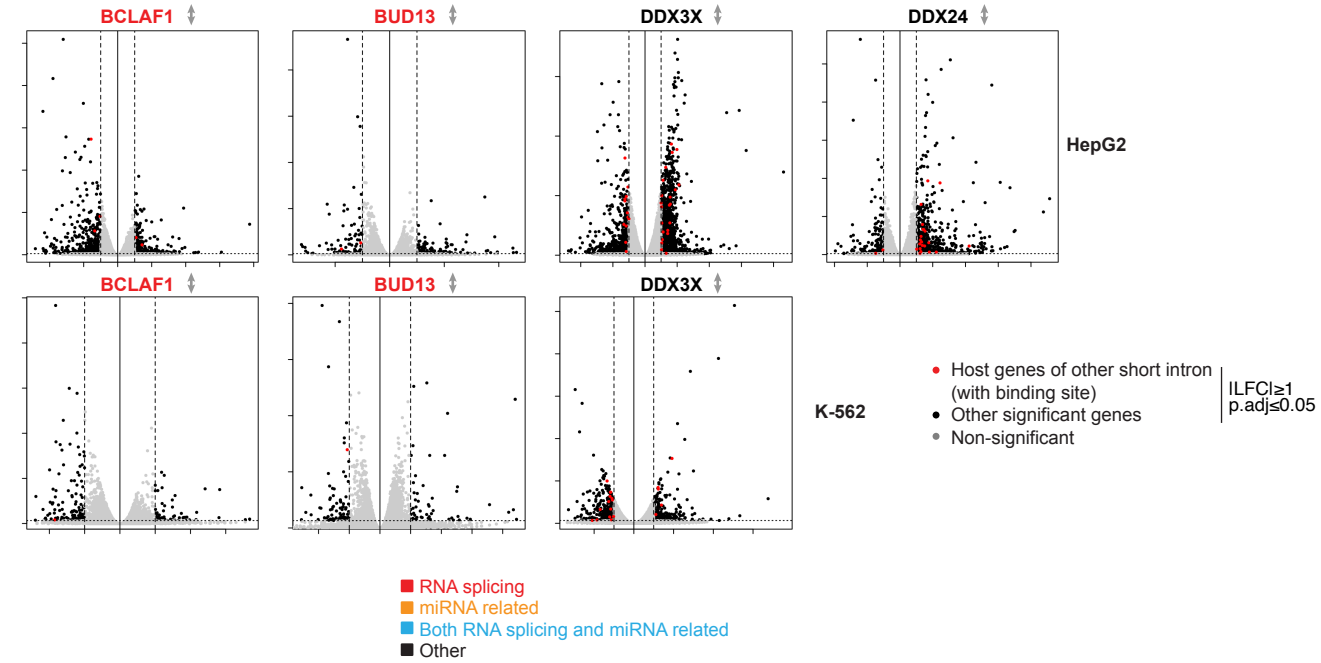


# Fig. S14

## A. FLEXIs

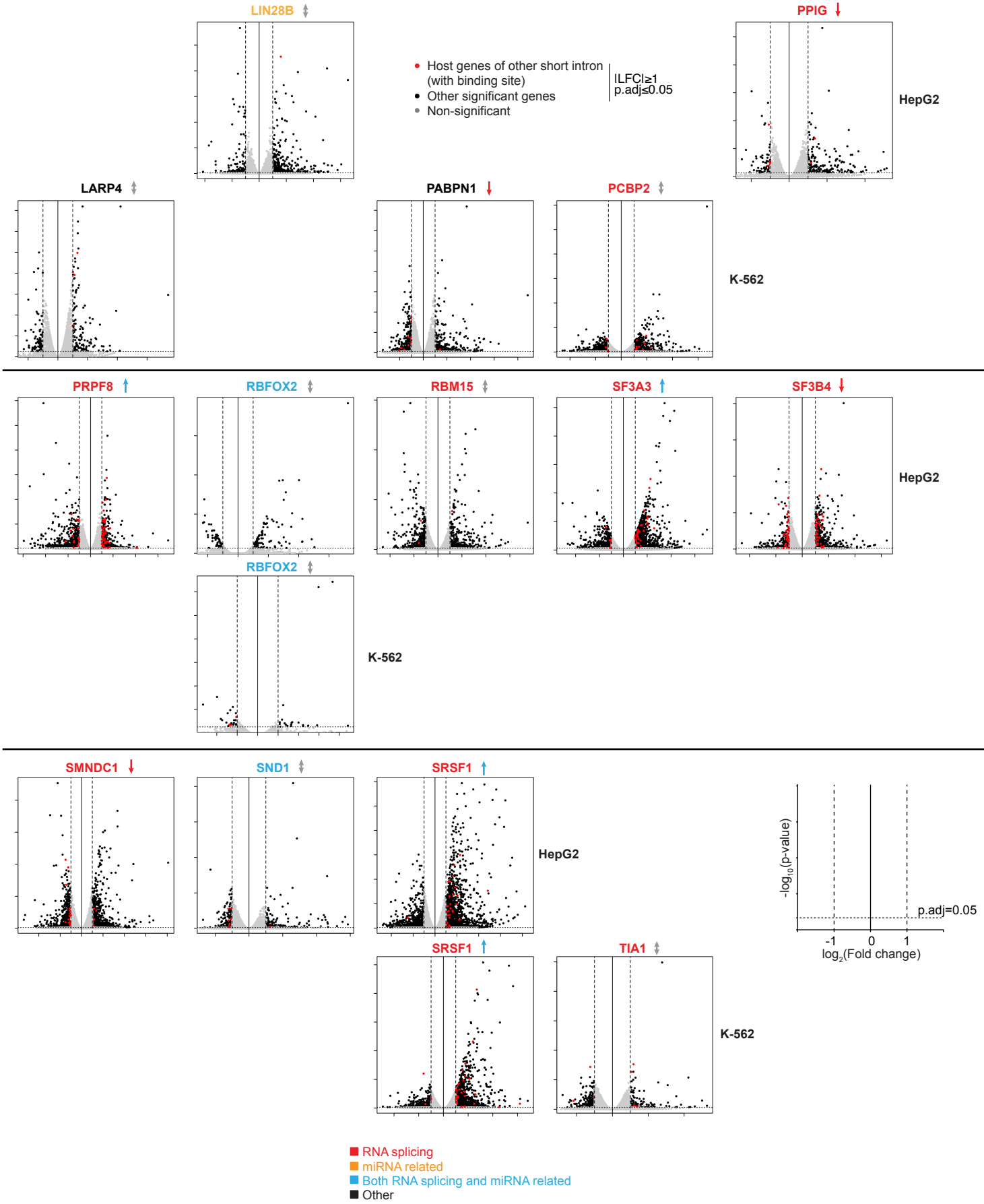



## B. Other short introns



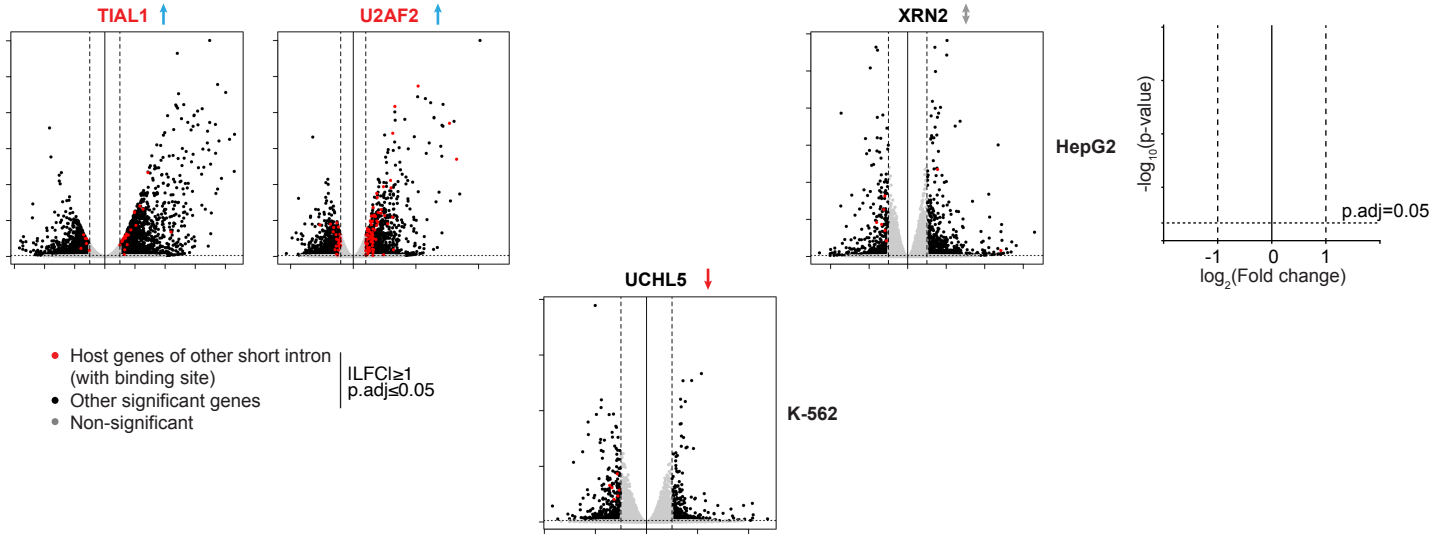
## Fig. S14 (continued)

### B. Other short introns

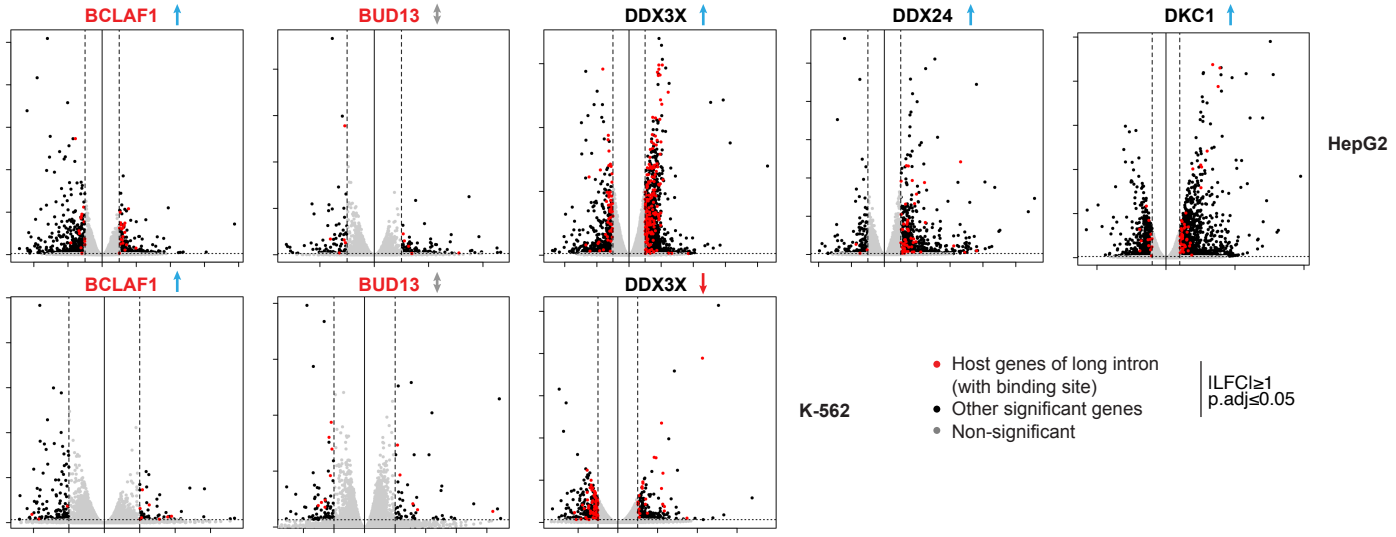


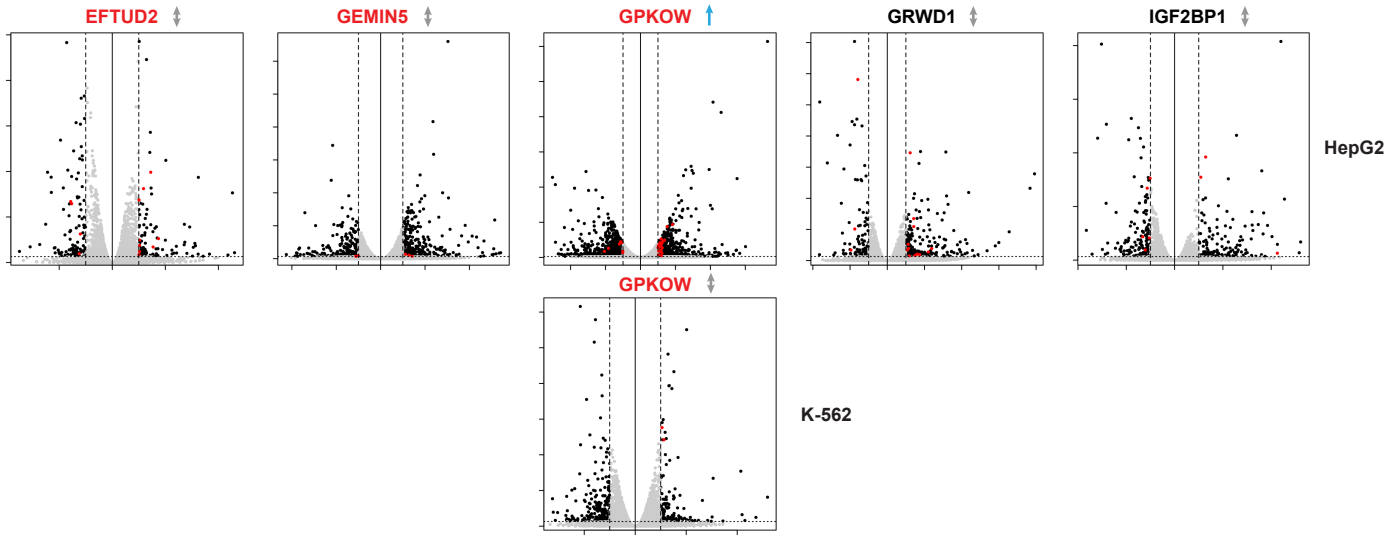
# Fig. S14 (continued)

## B. Other short introns




## C Long introns

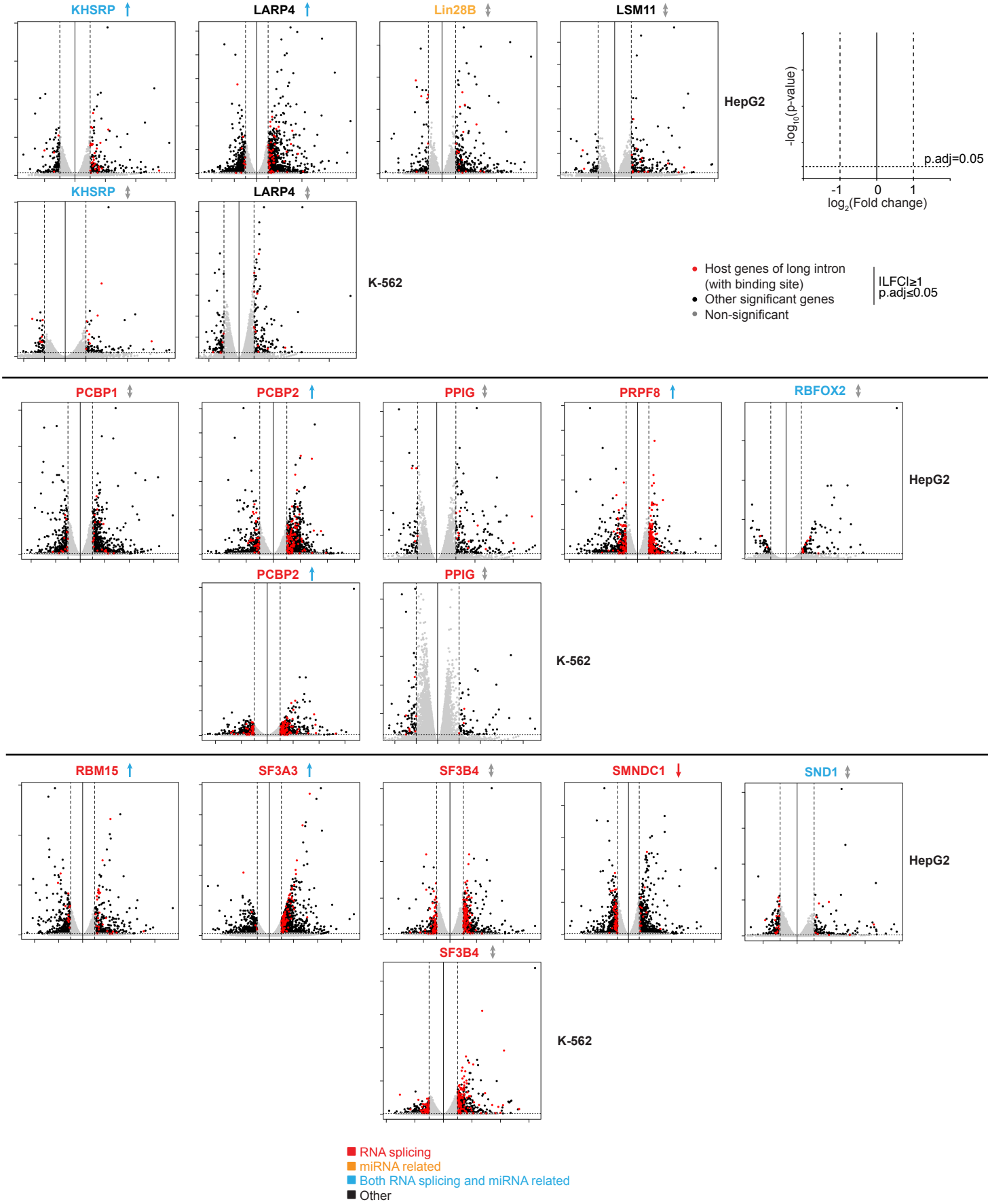



- RNA splicing
- miRNA related
- Both RNA splicing and miRNA related
- Other

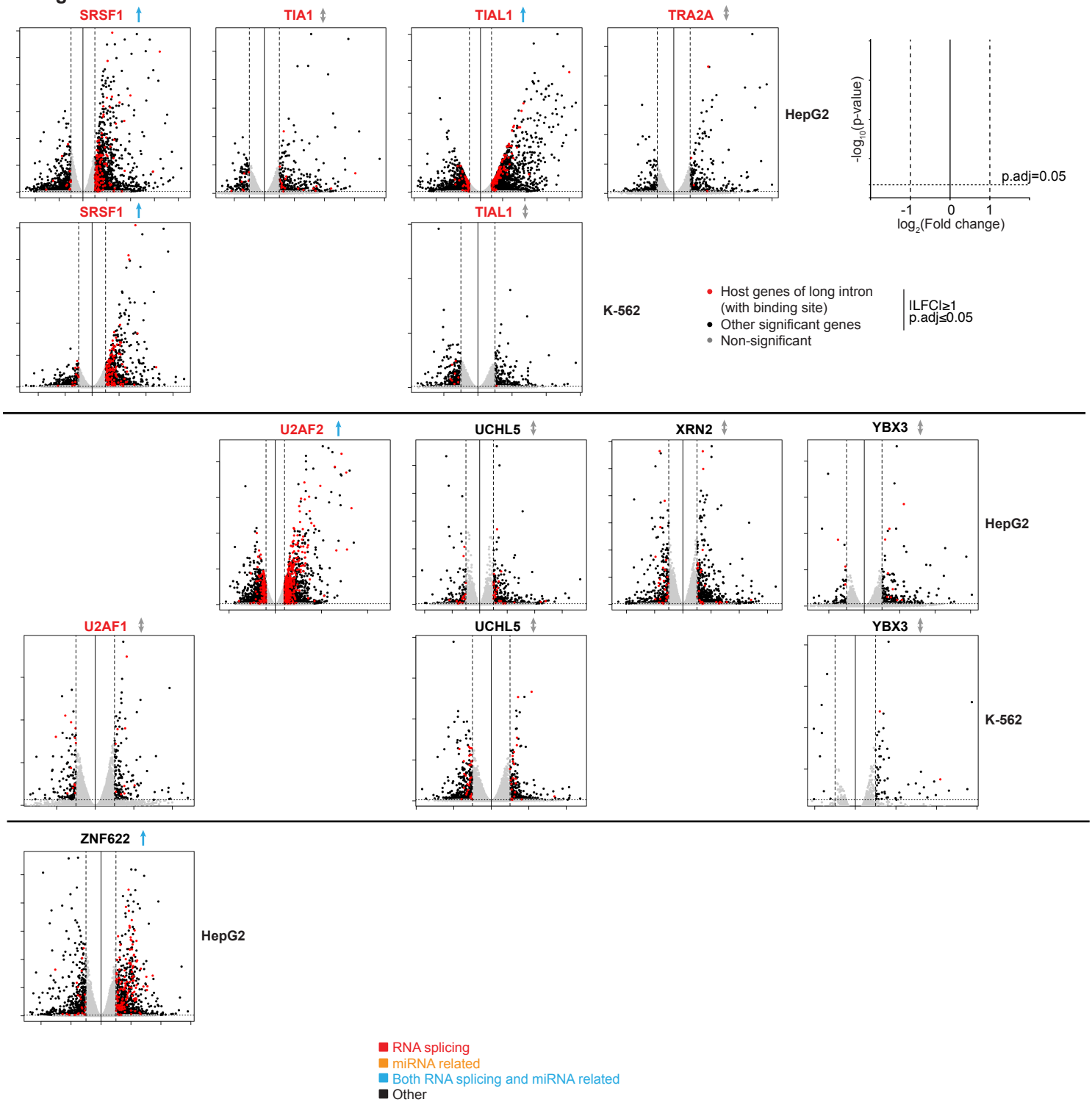
## Fig. S14 (continued)

### C. Long introns



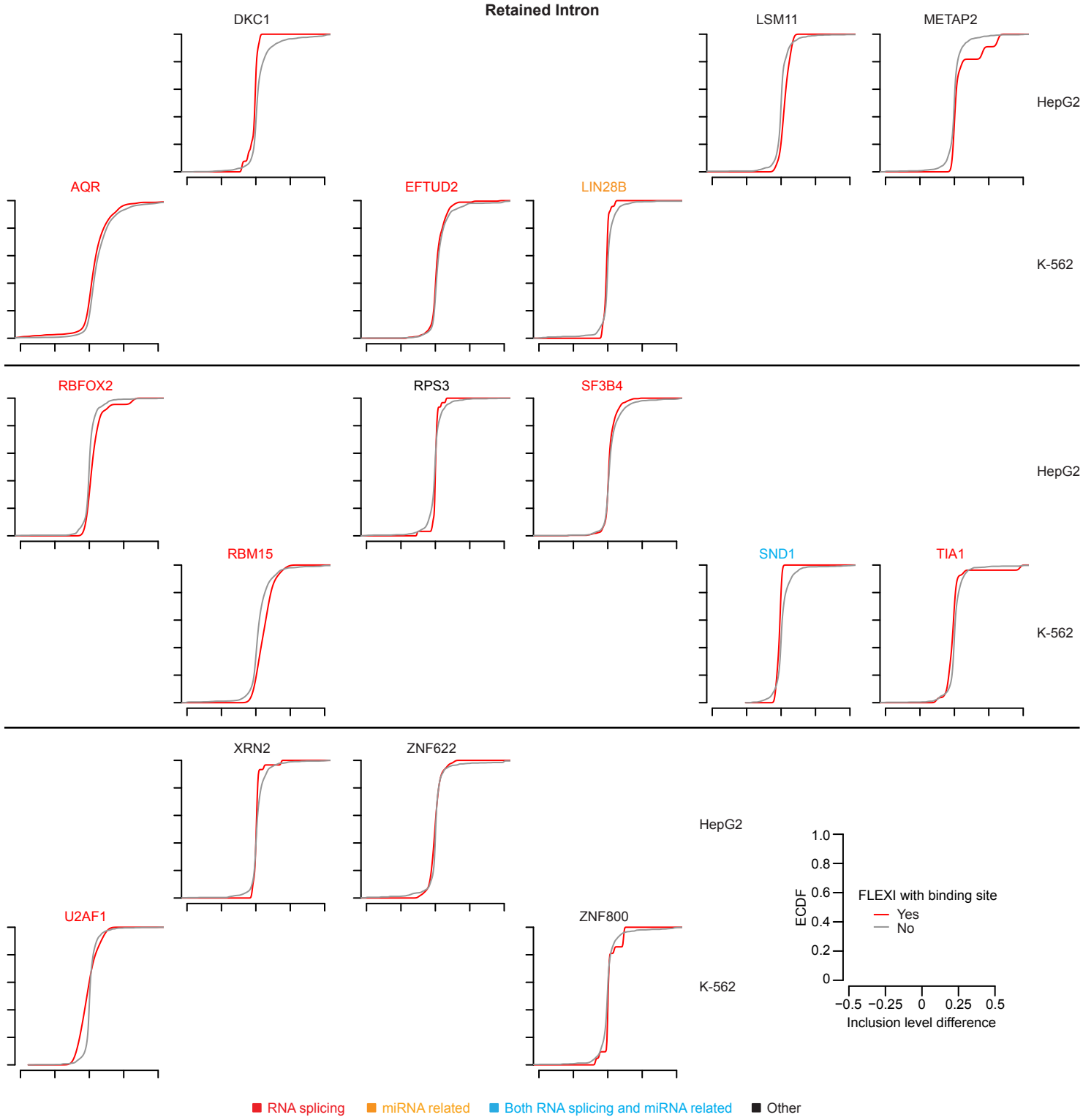
## Fig. S14 (continued)

### C. Long introns



## Fig. S15

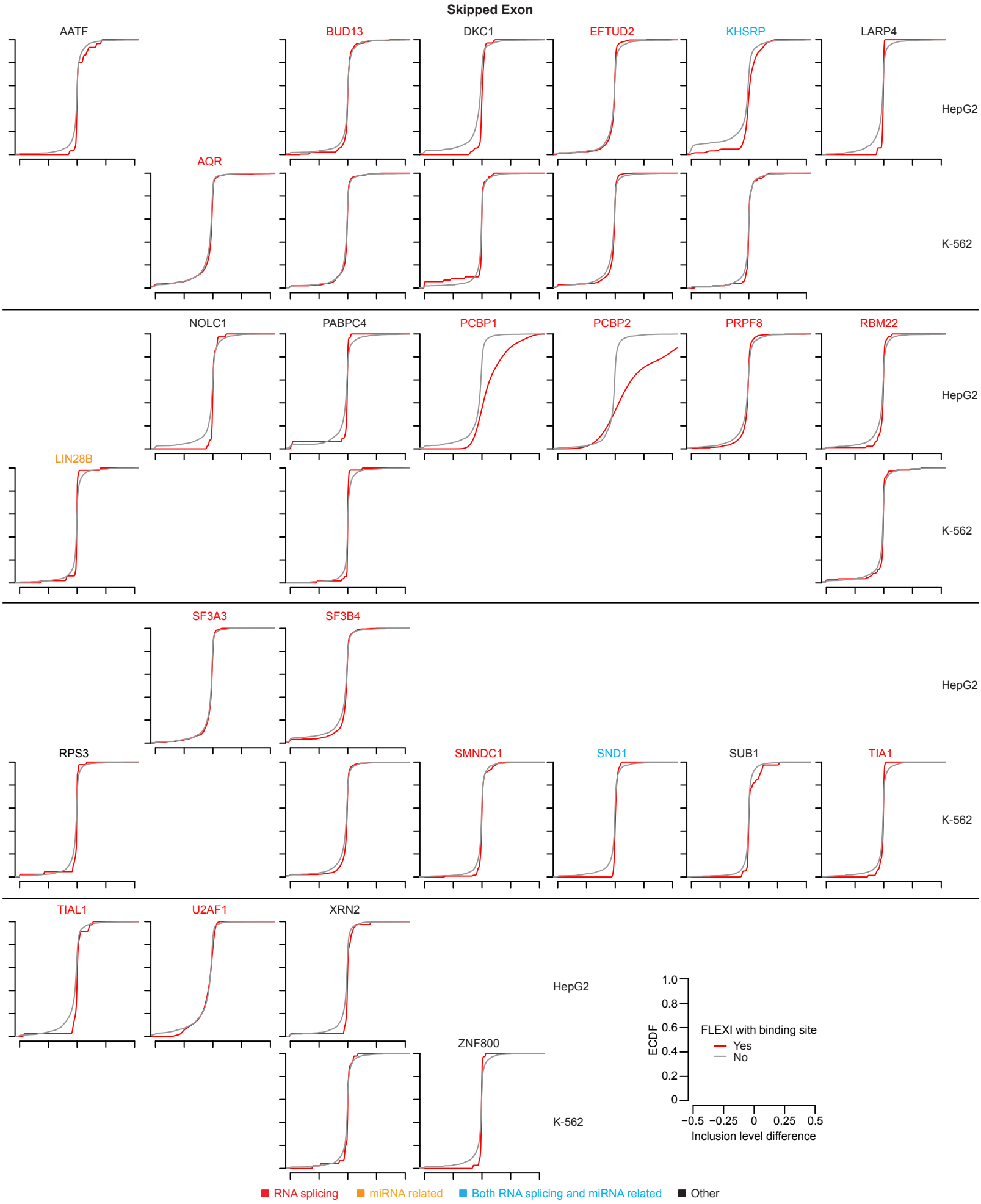
### A. FLEXIs




RNA splicing miRNA related Both RNA splicing and miRNA related Other

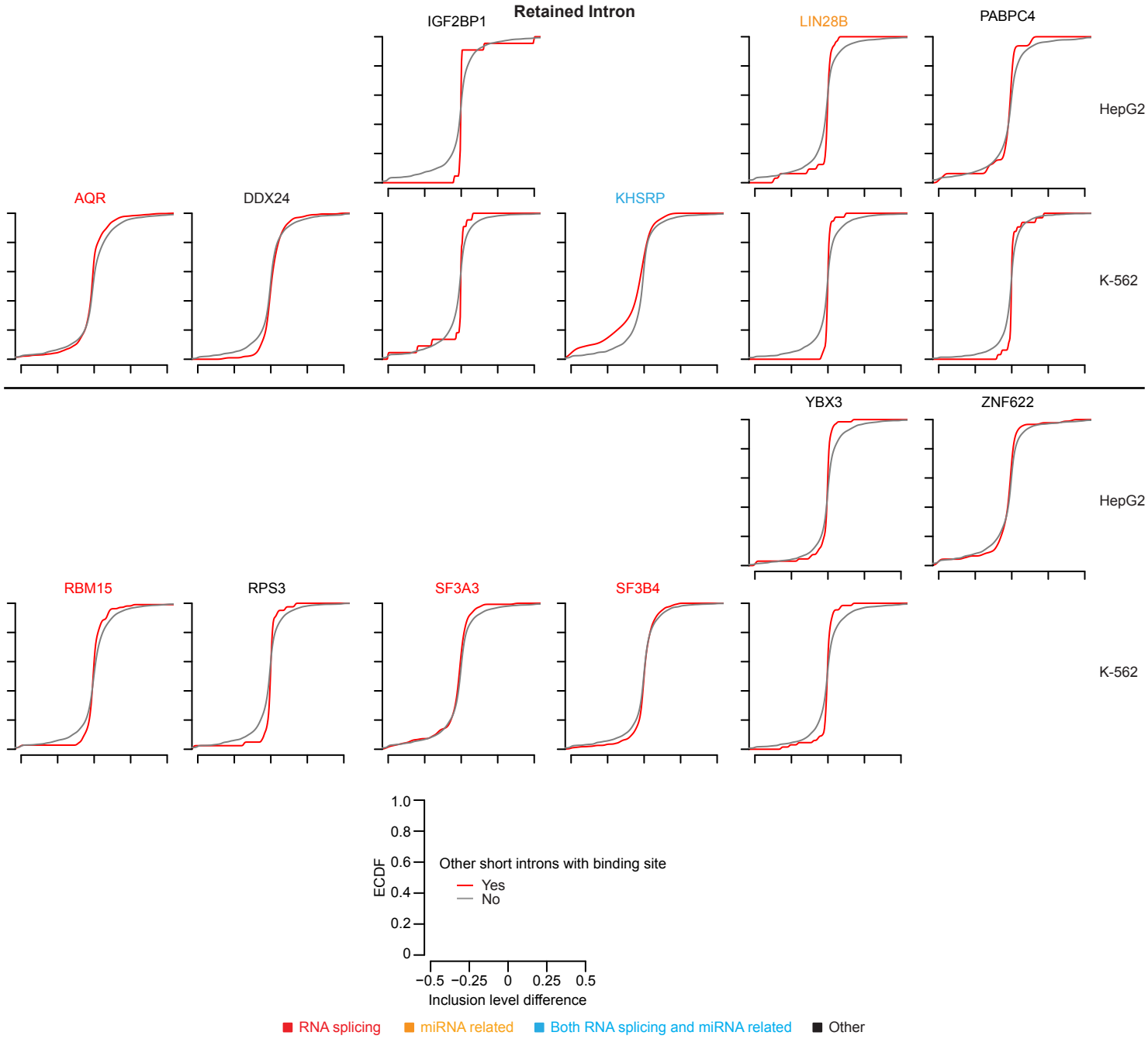
# Fig. S15 (continued)

## A. FLEXIs



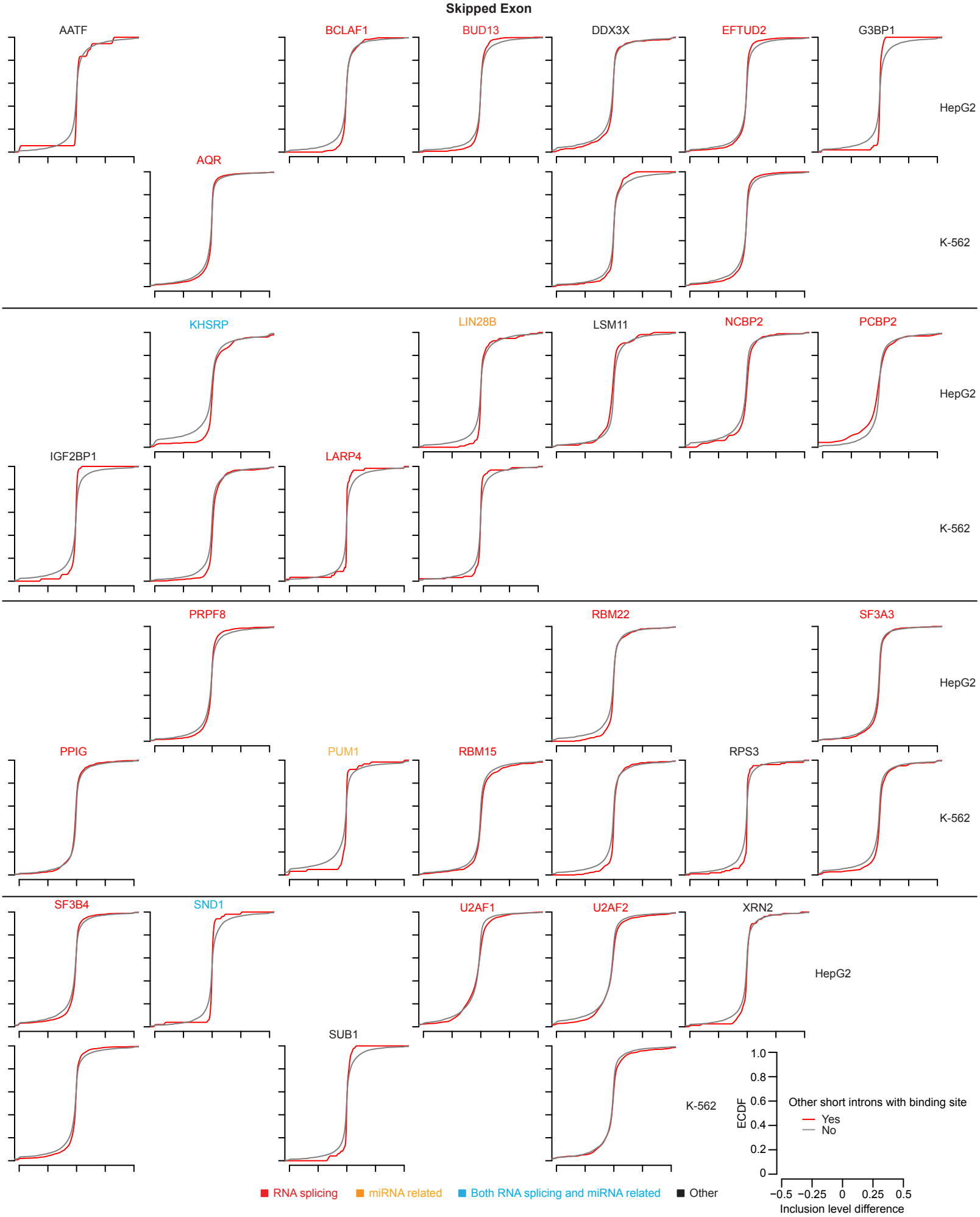
## Fig. S15 (continued)

### B. Other short introns



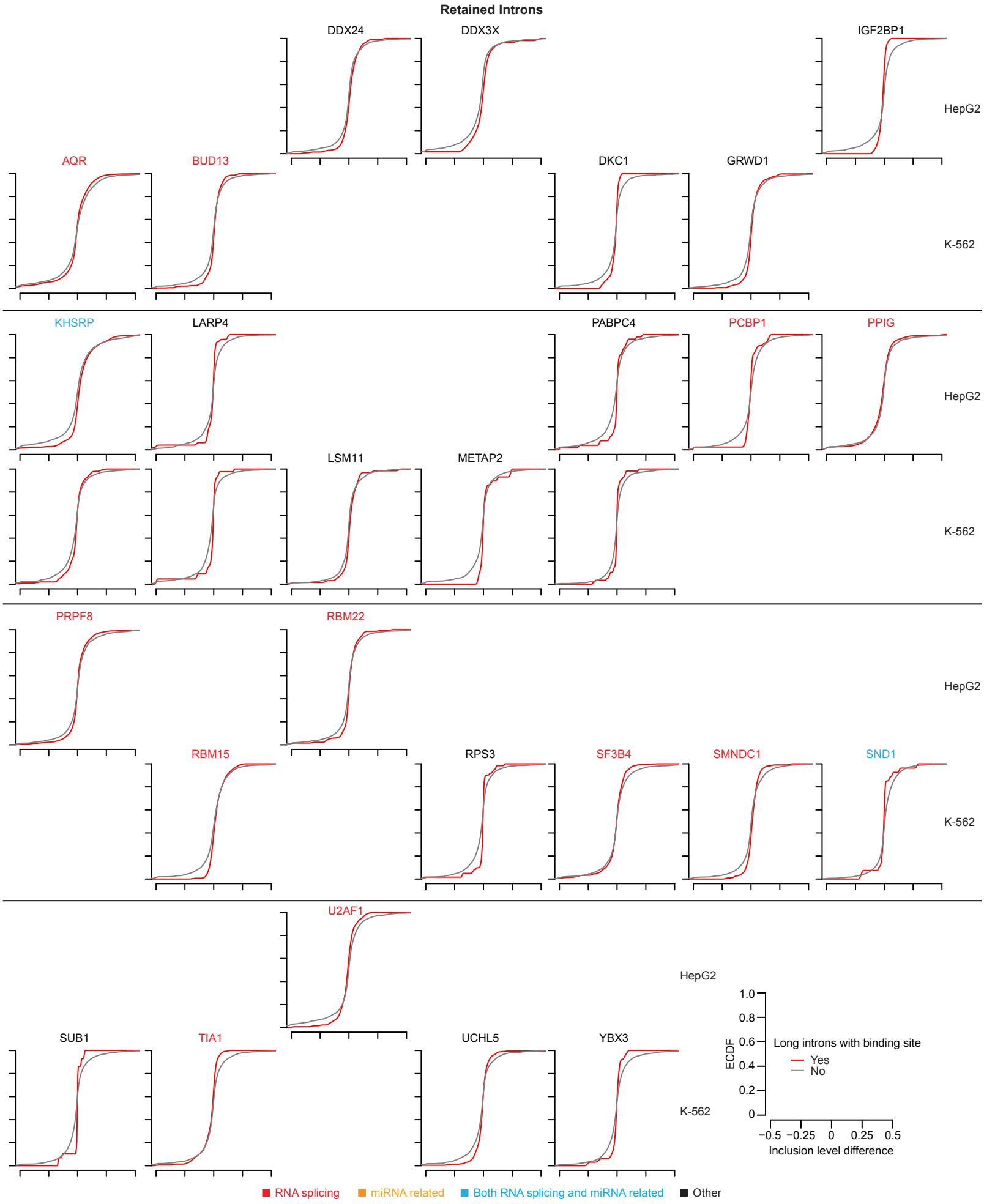
# Fig. S15 (continued)

## B. Other short introns



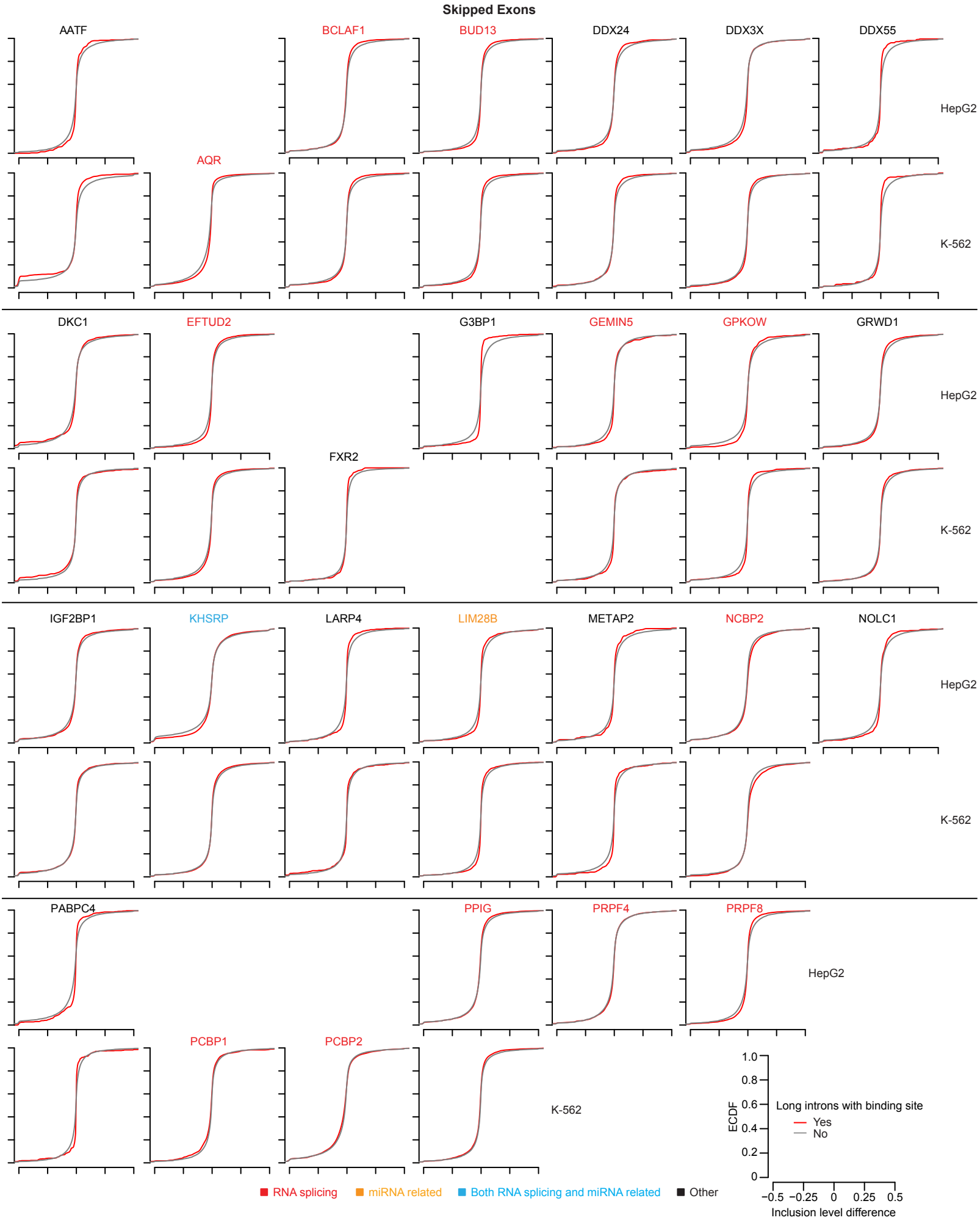
# Fig. S15 (continued)

## C. Long introns



# Fig. S15 (continued)

## C. Long introns



# Fig. S15 (continued)

## C. Long introns

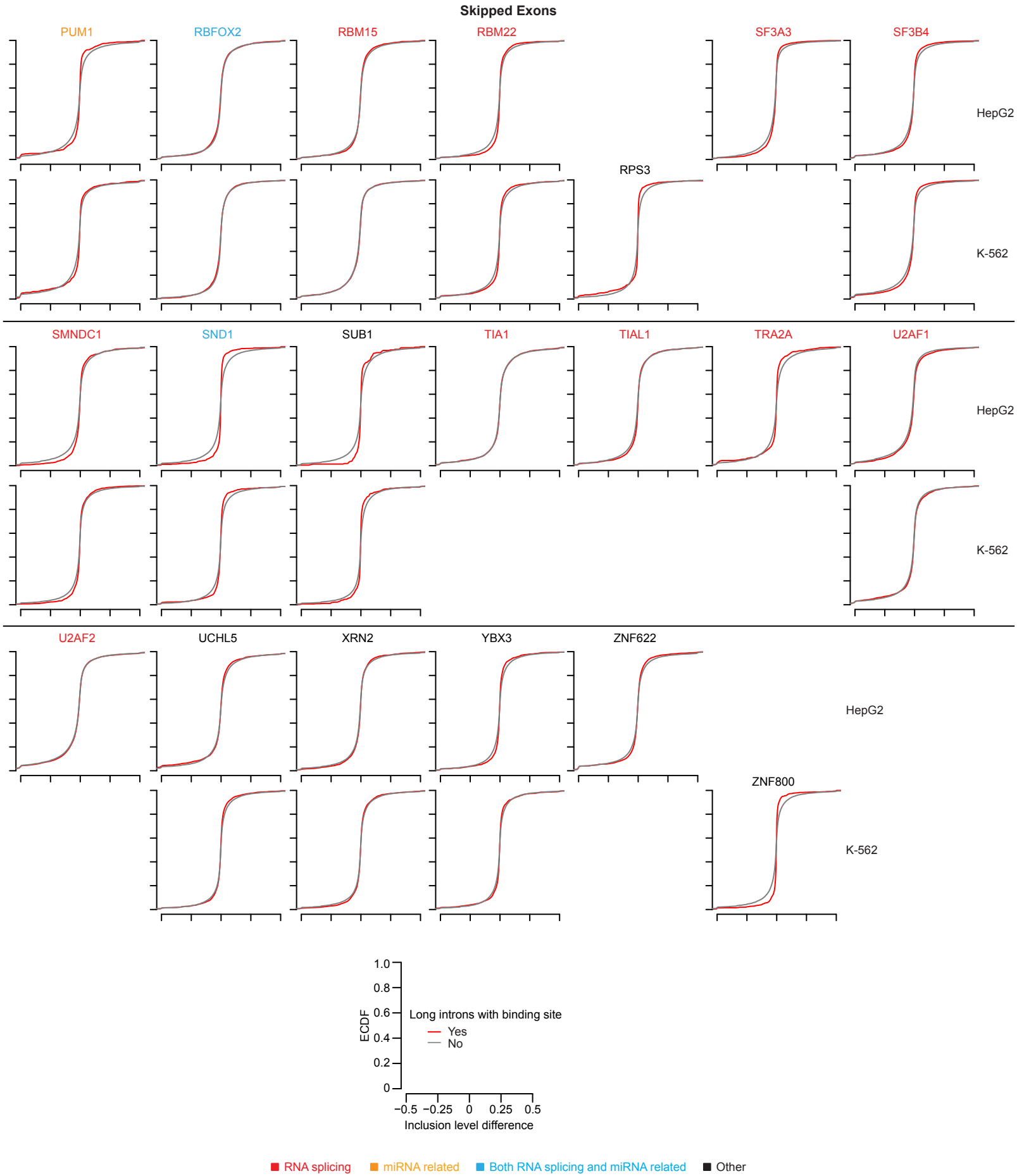


**Fig. S16**

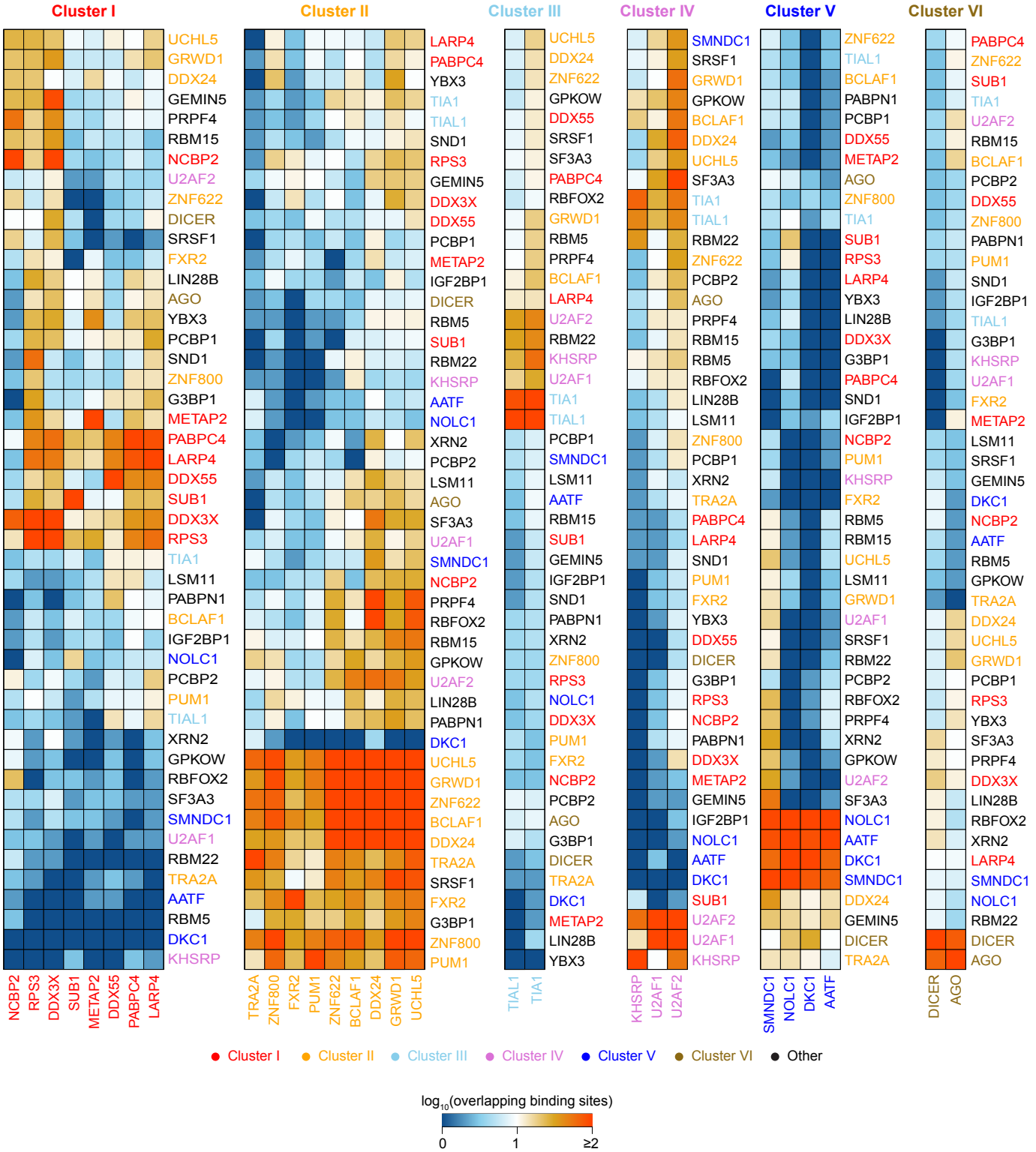

Cluster I
Cluster II
Cluster III
Cluster IV
Cluster V
Cluster VI
UCHL5 GRWD1 DDX24 GEMIN5 PRPF4 RBM15 NCBP2 U2AF2 ZNF622 DICER SRSF1 FXR2 LIN28B AGO YBX3 PCBP1 SND1 ZNF800 G3BP1 METAP2 PABPC4 LARP4 DDX55 SUB1 DDX3X RPS3 TIA1 LSM11 PABPN1 BCLAF1 IGF2BP1 NOLC1 PCBP2 PUM1 TIAL1 XRN2 GPKOW RBFOX2 SF3A3 SMNDC1 U2AF1 RBM22 TRA2A AATF RBM5 DKC1 KHSRP
NCBP2 RPS3 DDX3X SUB1 METAP2 DDX55 PABPC4 LARP4
LARP4 PABPC4 YBX3 TIA1 TIAL1 SND1 RPS3 GEMIN5 DDX3X DDX55 PCBP1 METAP2 IGF2BP1 DICER RBM5 SUB1 RBM22 KHSRP AATF NOLC1 XRN2 PCBP2 LSM11 AGO SF3A3 U2AF1 SMNDC1 NCBP2 PRPF4 RBFOX2 RBM15 GPKOW U2AF2 LIN28B PABPN1 DKC1 UCHL5 GRWD1 ZNF622 BCLAF1 DDX24 TRA2A SRSF1 FXR2 G3BP1 ZNF800 PUM1
TRA2A ZNF800 FXR2 PUM1 ZNF622 BCLAF1 DDX24 GRWD1 UCHL5
UCHL5 DDX24 ZNF622 GPKOW DDX55 SRSF1 SF3A3 PABPC4 RBFOX2 GRWD1 RBM5 PRPF4 BCLAF1 LARP4 U2AF2 RBM22 KHSRP U2AF1 TIA1 TIAL1 PCBP1 SMNDC1 LSM11 AATF RBM15 SUB1 GEMIN5 IGF2BP1 SND1 PABPN1 XRN2 ZNF800 RPS3 NOLC1 DDX3X PUM1 FXR2 NCBP2 PCBP2 AGO G3BP1 DICER TRA2A DKC1 METAP2 LIN28B YBX3
TIAL1 TIA1
SMNDC1 SRSF1 GRWD1 GPKOW BCLAF1 DDX24 UCHL5 SF3A3 TIA1 TIAL1 RBM22 ZNF622 PCBP2 AGO PRPF4 RBM15 RBM5 RBFOX2 LIN28B LSM11 ZNF800 PCBP1 XRN2 TRA2A PABPC4 LARP4 SND1 PUM1 FXR2 YBX3 DDX55 DICER G3BP1 RPS3 NCBP2 PABPN1 DDX3X METAP2 GEMIN5 IGF2BP1 NOLC1 AATF DKC1 SUB1 U2AF2 U2AF1 KHSRP
KHSRP U2AF1 U2AF2
ZNF622 TIAL1 BCLAF1 PABPN1 PCBP1 DDX55 METAP2 AGO ZNF800 TIA1 SUB1 RPS3 LARP4 YBX3 LIN28B DDX3X G3BP1 PABPC4 SND1 IGF2BP1 NCBP2 PUM1 KHSRP FXR2 RBM5 RBM15 UCHL5 LSM11 GRWD1 U2AF1 SRSF1 RBM22 PCBP2 RBFOX2 PRPF4 XRN2 GPKOW U2AF2 SF3A3 NOLC1 AATF DKC1 SMNDC1 DDX24 GEMIN5 DICER TRA2A
SMNDC1 NOLC1 DKC1 AATF
PABPC4 ZNF622 SUB1 TIA1 U2AF2 RBM15 BCLAF1 PCBP2 DDX55 ZNF800 PABPN1 PUM1 SND1 IGF2BP1 TIAL1 G3BP1 KHSRP U2AF1 FXR2 METAP2 LSM11 SRSF1 GEMIN5 DKC1 NCBP2 AATF RBM5 GPKOW TRA2A DDX24 UCHL5 GRWD1 PCBP1 RPS3 YBX3 SF3A3 PRPF4 DDX3X LIN28B RBFOX2 XRN2 LARP4 SMNDC1 NOLC1 RBM22 DICER AGO
DICER AGO
log10(overlapping binding sites)
0
1
≥2
Cluster I
Cluster II
Cluster III
Cluster IV
Cluster V
Cluster VI
Other

**Fig. S17**

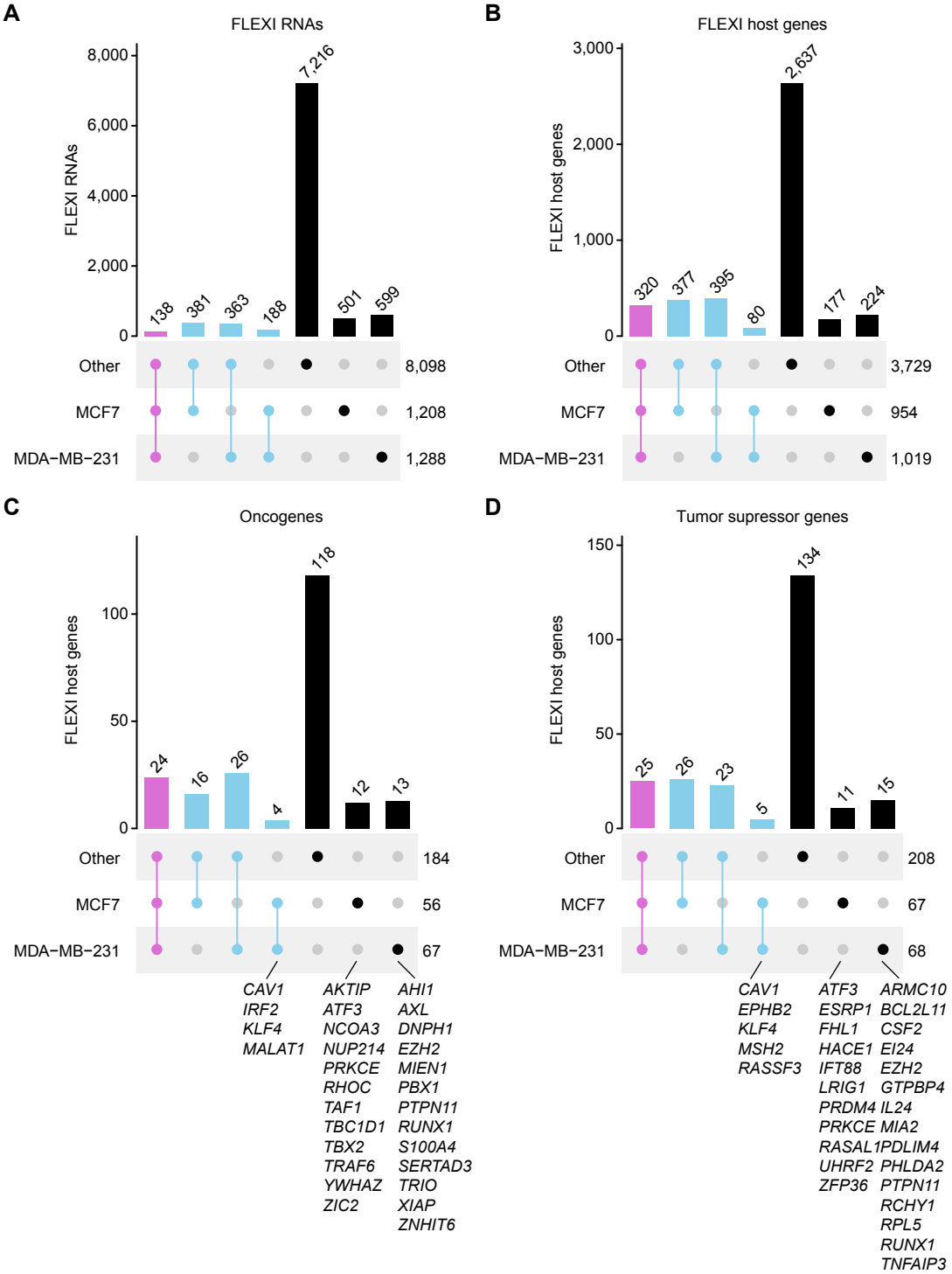

A
FLEXI RNAs
FLEXI RNAs
8,000
6,000
4,000
2,000
0
138
381
363
188
7,216
501
599
Other
MCF7
MDA−MB−231
8,098
1,208
1,288
B
FLEXI host genes
FLEXI host genes
3,000
2,000
1,000
0
320
377
395
80
2,637
177
224
Other
MCF7
MDA−MB−231
3,729
954
1,019
C
Oncogenes
FLEXI host genes
100
50
0
24
16
26
4
118
12
13
Other
MCF7
MDA−MB−231
184
56
67
CAV1
IRF2
KLF4
MALAT1
AKTIP
ATF3
NCOA3
NUP214
PRKCE
RHOC
TAF1
TBC1D1
TBX2
TRAF6
YWHAZ
ZIC2
AHI1
AXL
DNPH1
EZH2
MIEN1
PBX1
PTPN11
RUNX1
S100A4
SERTAD3
TRIO
XIAP
ZNHIT6
D
Tumor supressor genes
FLEXI host genes
150
100
50
0
25
26
23
5
134
11
15
Other
MCF7
MDA−MB−231
208
67
68
CAV1
EPHB2
KLF4
MSH2
RASSF3
ATF3
ESRP1
FHL1
HACE1
IFT88
LRIG1
PRDM4
PRKCE
RASAL1
UHRF2
ZFP36
ARMC10
BCL2L11
CSF2
EI24
EZH2
GTPBP4
IL24
MIA2
PDLIM4
PHLDA2
PTPN11
RCHY1
RPL5
RUNX1
TNFAIP3

**Fig. S18**

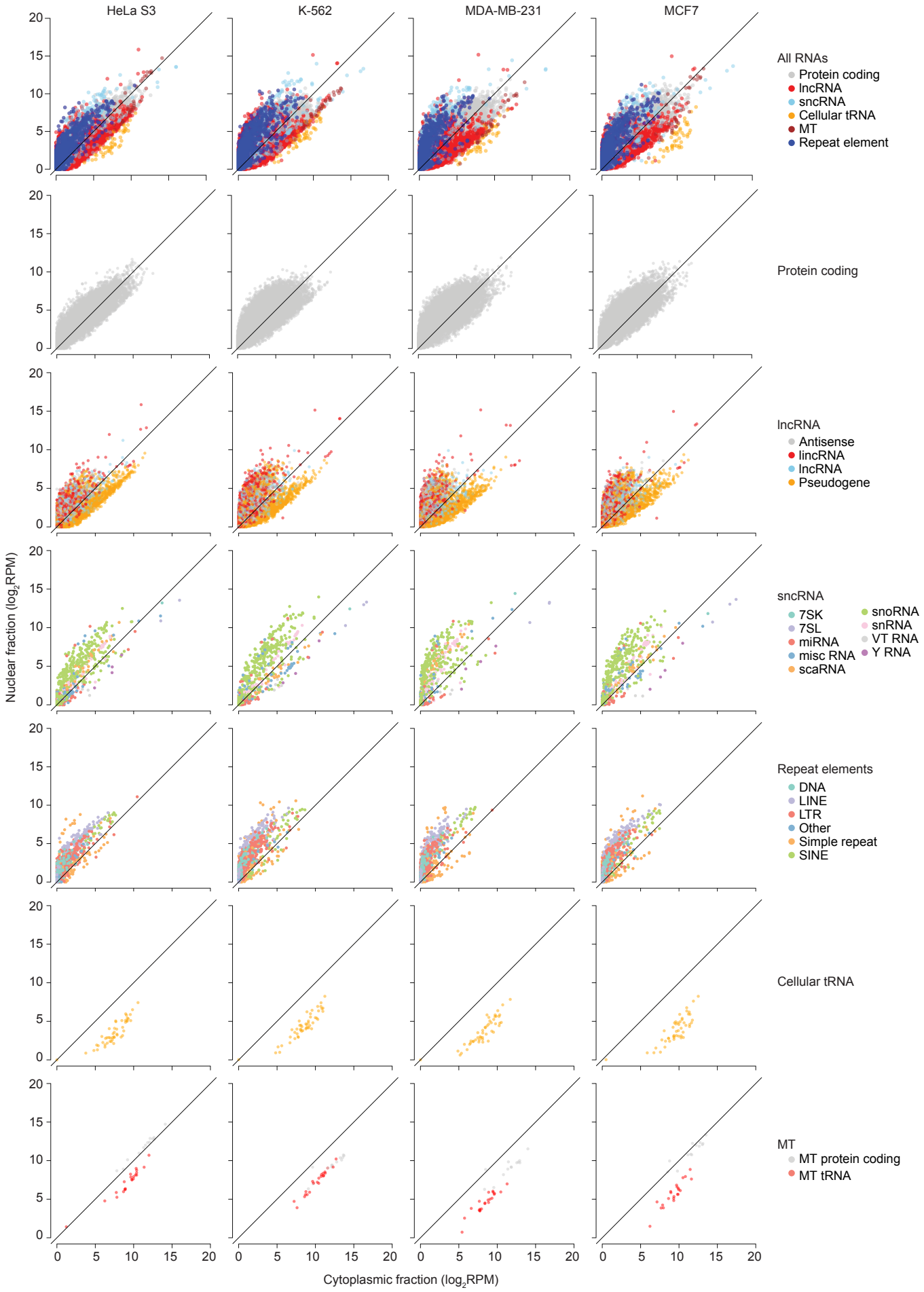
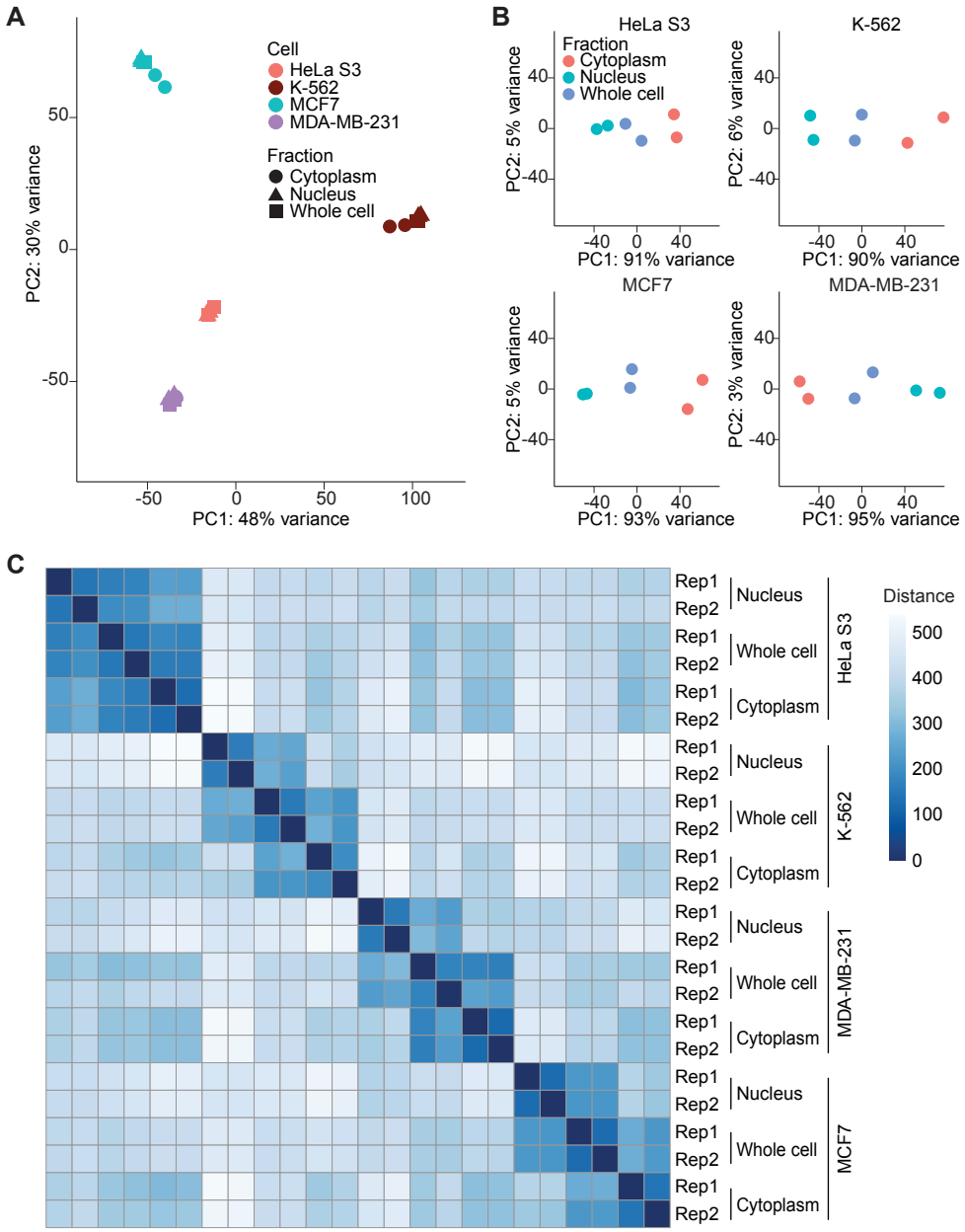

HeLa S3
K-562
MDA-MB-231
MCF7
All RNAs
Protein coding
lncRNA
sncRNA
Cellular tRNA
MT
Repeat element
Protein coding
lncRNA
Antisense
lincRNA
lncRNA
Pseudogene
sncRNA
7SK
7SL
miRNA
misc RNA
scaRNA
snoRNA
snRNA
VT RNA
Y RNA
Repeat elements
DNA
LINE
LTR
Other
Simple repeat
SINE
Cellular tRNA
MT
MT protein coding
MT tRNA
Nuclear fraction (log2RPM)
Cytoplasmic fraction (log2RPM)
0
5
10
15
20

**Fig. S19**


A
Cell
HeLa S3
K-562
MCF7
MDA-MB-231
Fraction
Cytoplasm
Nucleus
Whole cell
PC2: 30% variance
PC1: 48% variance
B
HeLa S3
K-562
MCF7
MDA-MB-231
PC2: 5% variance
PC2: 6% variance
PC2: 3% variance
PC1: 91% variance
PC1: 90% variance
PC1: 93% variance
PC1: 95% variance
C
Rep1
Rep2
Nucleus
Whole cell
Cytoplasm
HeLa S3
K-562
MDA-MB-231
MCF7
Distance
500
400
300
200
100
0

# Fig. S20

**A**

**FLEXIs**

| Host genes / RBP | Total | MitoCarta3.0 | Cyto. ribosomal proteins | Mito. ribosomal proteins |
|---|---|---|---|---|
| LARP4 | 45 | 10 (22.2%) | 15 (33.3%) | 1 (2.2%) |
| PABPC4 | 54 | 8 (14.8%) | 23 (42.6%) | 1 (1.9%) |
| SUB1 | 26 | 3 (11.5%) | 8 (30.8%) | 0 (0.0%) |
| DDX3X | 137 | 35 (25.5%) | 14 (10.2%) | 9 (6.6%) |
| RPS3 | 99 | 14 (14.1%) | 10 (10.1%) | 2 (2.0%) |
| NCBP2 | 36 | 3 (8.3%) | 3 (8.3%) | 0 (0.0%) |
| DDX55 | 28 | 4 (14.3%) | 4 (14.3%) | 0 (0.0%) |
| METAP2 | 48 | 4 (8.3%) | 7 (14.6%) | 0 (0.0%) |

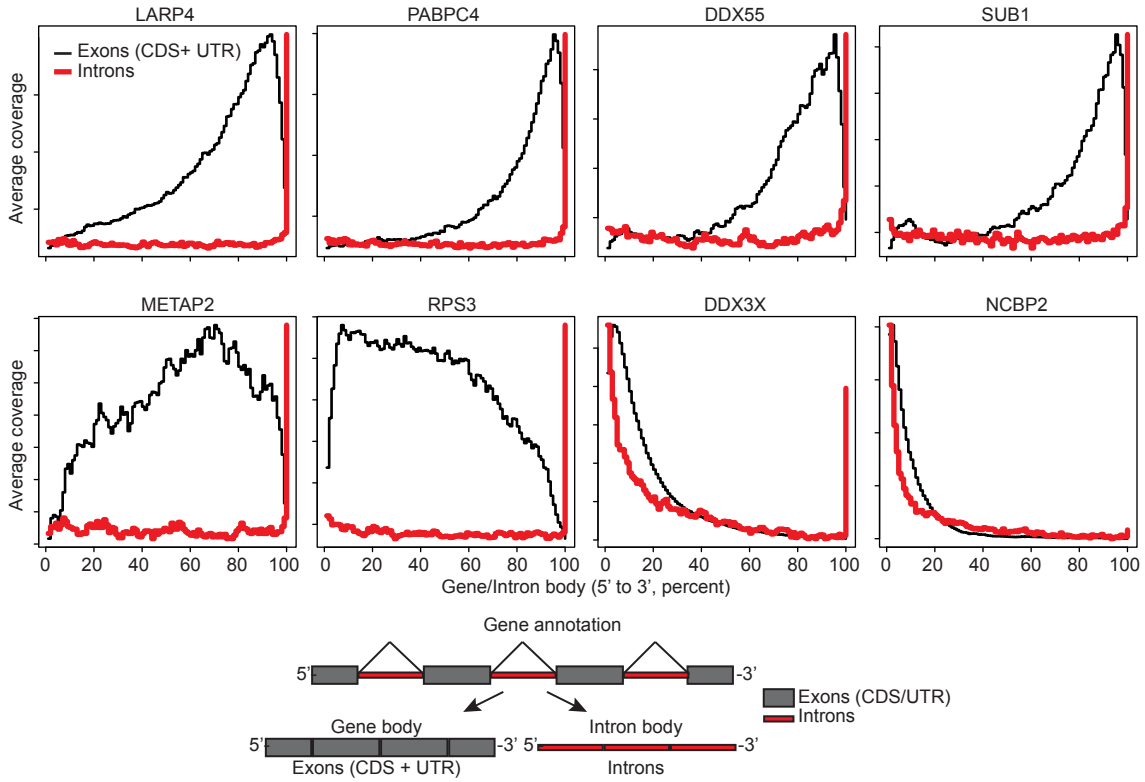
**Other short introns**

| Host genes / RBP | Total | MitoCarta3.0 | Cyto. ribosomal proteins | Mito. ribosomal proteins |
|---|---|---|---|---|
| LARP4 | 132 | 25 (18.9%) | 18 (13.6%) | 4 (3.0%) |
| PABPC4 | 112 | 15 (13.4%) | 25 (22.3%) | 0 (0.0%) |
| SUB1 | 57 | 6 (10.5%) | 9 (15.8%) | 0 (0.0%) |
| DDX3X | 510 | 83 (16.3%) | 23 (4.5%) | 11 (2.2%) |
| RPS3 | 223 | 28 (12.6%) | 16 (7.2%) | 3 (1.3%) |
| NCBP2 | 170 | 19 (11.2%) | 6 (3.5%) | 0 (0.0%) |
| DDX55 | 87 | 15 (17.2%) | 3 (3.4%) | 0 (0.0%) |
| METAP2 | 96 | 7 (7.3%) | 8 (8.3%) | 0 (0.0%) |

**Long introns**

| Host genes / RBP | Total | MitoCarta3.0 | Cyto. ribosomal proteins | Mito. ribosomal proteins |
|---|---|---|---|---|
| LARP4 | 2,011 | 118 (14.7%) | 49 (6.1%) | 14 (1.7%) |
| PABPC4 | 2,286 | 86 (13.0%) | 68 (10.3%) | 9 (1.4%) |
| SUB1 | 1,534 | 62 (13.4%) | 31 (6.7%) | 5 (1.1%) |
| DDX3X | 5,510 | 354 (11.4%) | 47 (1.5%) | 34 (1.1%) |
| RPS3 | 1,587 | 112 (12.8%) | 35 (4.0%) | 4 (0.5%) |
| NCBP2 | 2,584 | 133 (8.6%) | 20 (1.3%) | 3 (0.2%) |
| DDX55 | 1,467 | 44 (10.7%) | 11 (2.7%) | 3 (0.7%) |
| METAP2 | 874 | 39 (8.8%) | 28 (6.3%) | 2 (0.4%) |

**B**

| Host genes | Total | MitoCarta3.0 | Cyto. ribosomal proteins | Mito. ribosomal proteins |
|---|---|---|---|---|
| Cluster I FLEXIs | 324 | 59 (18.2%) | 31 (9.6%) | 10 (3.1%) |
| Cluster V FLEXIs | 116 | 8 (6.9%) | 13 (11.2%) | 0 (0.0%) |
| All FLEXIs | 3,729 | 280 (7.5%) | 60 (1.6%) | 23 (0.6%) |
| Other short introns | 5,051 | 399 (7.9%) | 74 (1.5%) | 29 (0.6%) |
| Long introns | 13,166 | 970 (7.4%) | 95 (0.7%) | 73 (0.6%) |
| GRCh38 genes | 19,780 | 1,122 (5.7%) | 101 (0.5%) | 78 (0.4%) |

**C**