Human cells contain myriad excised linear introns with potential functions in gene regulation and as RNA biomarkers

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

[1]Institute for Cellular and Molecular Biology and
Departments of Molecular Biosciences and Oncology
University of Texas at Austin
Austin TX 78712

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

Running Title: Human excised linear intron RNAs

[3]Present address: Invitae Corporation

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

Abstract

We used thermostable group II intron reverse transcriptase sequencing (TGIRT-seq), which gives full-length end-to-end sequence reads of structured RNAs, to identify > 8,500 short full-length excised linear intron (FLEXI) RNAs (≤ 300 nt) originating from > 3,500 different genes in human cells and tissues. FLEXIs are distinguished from other introns by their accumulation as stable full-length linear RNAs. Subsets of the detected FLEXI correspond to pre-miRNAs of annotated mirtrons (introns that fold into a stem-loop structure and are processed by DICER into functional miRNAs) or agotrons (structured introns that bind AGO2 and function in a miRNA-like manner) and a few encode snoRNAs, but the vast majority had not been identified previously. FLEXI RNA profiles are cell-type specific, reflecting differences in transcription, alternative splicing, and intron RNA turnover, and comparisons of matched tumor and healthy tissues from breast cancer patients and cell lines revealed hundreds of differences in FLEXI RNA expression. About half the detected FLEXI RNAs contained a CLIP-seq identified binding site for one or more RNA-binding proteins. In addition to proteins that have RNA splicing- or miRNA-related functions, proteins that bind groups of 30 or more different FLEXI RNAs include transcription factors, chromatin remodeling proteins, and proteins involved in cellular stress responses and growth regulation, raising the possibility of previously unsuspected connections between intron RNAs and cellular regulatory pathways. Our findings identify a large new class of human introns with potential to serve as RNA biomarkers.

Introduction

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

Recently, while analyzing human plasma RNAs by Thermostable Group II Intron Reverse Transcriptase sequencing (TGIRT-seq), which gives full-length, end-to-end sequence reads of structured RNAs, we identified 44 short (≤ 300 nt) full-length excised linear intron (FLEXI) RNAs, subsets of which corresponded to annotated agotrons or pre-miRNAs of annotated mirtrons (denoted mirtron pre-miRNAs) (Yao et al. 2020). Here, we followed up this discovery by using TGIRT-seq to systematically search for FLEXI RNAs in human cell lines and tissues. We thus identified > 8,500 different FLEXI RNAs, many with stable predicted RNA secondary

structures that would make them difficult to detect by other methods. By combining the newly obtained FLEXI RNA datasets with published CLIP-seq datasets, we identified numerous intron RNA-protein interactions and potential connections to cellular regulatory pathways that had not been seen previously. Finally, we found that FLEXI RNA expression patterns were more discriminatory between cell types than were mRNAs from the corresponding host genes, suggesting utility as biomarkers for human diseases.

Results

Identification of FLEXI RNAs in human cells

A search of the human genome (Ensembl GRCh38 Release 93 annotations; http://www.ensembl.org) identified 51,645 short introns (≤ 300 nt) in 12,020 different genes that could potentially give rise to FLEXI RNAs. To determine which of these short introns might give rise to FLEXI RNAs in biological samples, we carried out TGIRT-seq of ribodepleted, intact (*i.e.*, non-chemically fragmented) human cellular RNAs, including Universal Human Reference RNA (UHRR; a mixture of total RNAs from ten human cell lines) and total cellular RNA from HEK-293T, K-562, and HeLa S3 cells (Fig. 1).

TGIRT-seq is particularly well-suited for the detection of excised linear intron RNAs. In the version of the method used here, the TGIRT enzyme initiates reverse transcription precisely at the 3' nucleotide of a target RNA by template-switching from an RNA-seq adapter and then reverse transcribes to the 5' end of the RNA, yielding a full-length intron cDNA to which a second RNA-seq adapter is ligated for minimal PCR amplification (see Methods). The high processivity and strand displacement activity of TGIRTs together with reverse transcription at elevated temperatures enable full-length end-to-end reads of highly structured RNAs (Katibah et al. 2014; Qin et al. 2016). In TGIRT-seq datasets of the ribodepleted non-chemically fragmented cellular RNAs, such as those obtained in this study, most of the reads correspond to full-length mature tRNAs and other structured sncRNAs (Supplemental Fig. S1 and S2). After ribodepletion, only a small percentage of the total reads (0.5-6.3%) corresponded to cellular or

mitochondrial (Mt) rRNAs (Supplemental Fig. S1). Additionally, because TGIRT-enzymes do not read through long stretches of poly(A), mRNA reads from protein-coding genes comprised a relatively low percentage of total reads (0.7-5.3%) and corresponded largely to nascent transcripts and non- or minimally polyadenylated mRNA sequences (Supplemental Fig. S1).

To search for human FLEXIs in the TGIRT-seq datasets on intact cellular RNAs, we compiled the coordinates of all short introns (≤ 300 nt) in Ensembl GRCh38 Release 93 annotations into a BED file and searched for intersections with full-length excised linear intron RNAs, which were defined for these searches and all subsequent analyses as continuous intron reads whose 5' and 3' ends were within three nucleotides of annotated splice sites. For each sample type, the searches were done by using combined datasets obtained from multiple replicate libraries totaling 666 to 768 million mapped reads for the cellular RNA samples (Supplemental Table S1). In addition to human cellular RNAs, we used this approach to search remapped datasets of human plasma RNAs from healthy individuals (Yao et al. 2020). We thus identified 8,144 different FLEXI RNAs represented by at least one read in any of the cellular or plasma RNA datasets. These FLEXI RNAs originated from 3,743 different protein-coding genes, lncRNA genes, or pseudogenes (collectively denoted FLEXI host genes; Fig. 1).

UpSet plots and pairwise scatter plots comparing different cell lines showed that both FLEXI RNA and FLEXI host gene expression patterns were cell type-specific (Fig. 1A and B). Notably, the scatter plots for FLEXI RNAs showed greater discrimination between cell types than did those for all transcripts from the corresponding host genes, with numerous FLEXIs of abundances up to 1 to 7 RPM detected in only one or the other of two compared cell types (Fig. 1B). Principal component analysis (PCA), as well as PCA-initialized t-SNE (Kobak and Berens 2019) and ZINB-WaVE (Risso et al. 2018), both of which are widely used for the analysis of single cell RNA-seq datasets with zero inflated counts, showed clustering of cell-type-specific FLEXI RNA profiles in 44 different replicates of the cellular RNA datasets obtained in this study (Supplemental Fig. S3).

Density plots of the length distribution of all reads that mapped to introns in a combined dataset for the UHRR, K-562, HEK-293T and HeLa S3 RNA samples showed a peak at near 100% of the full intron length for the detected FLEXIs, whereas reads mapping to other short introns (≤ 300 nt) annotated in GRCh38 corresponded largely to heterogeneously sized fragments, as expected for the more typical situation of intron RNAs that turnover rapidly after RNA splicing (Fig. 1C, left panel). Similar patterns were seen in UHRR and the individual cell types for different subgroups of FLEXIs described further below, except for a subgroup of FLEXIs containing annotated binding sites for DICER, which showed additional peaks corresponding to discrete shorter intron RNA fragments, as expected for DICER cleavage (Supplemental Fig. S4). These findings identify FLEXIs as a distinct class of short introns that are stable in cells as full-length linear RNA molecules.

Sequence and structural characteristics of detected FLEXI RNAs

Integrative Genomic Viewer (IGV) alignments showed that the FLEXIs detected by TGIRT-seq are full-length linear intron RNAs, with reads extending continuously from the 5'- to 3'-splice site even for highly structured FLEXIs and no stops or base substitutions that might suggest the presence of a branched nucleotide residue (Fig. 2A-C). Analysis of FLEXI RNA expression in different cell-types indicated that differences in FLEXI RNA abundance reflect differences in host gene transcription, alternative splicing, or stability of the excised intron RNAs, the latter suggested by differences in the relative abundance of non-alternatively spliced FLEXIs transcribed from the same gene (examples shown in Fig. 2D).

Most of the detected FLEXI RNAs had sequence characteristics of major U2-type spliceosomal introns (8,082, 98.7% with canonical GU-AG splice sites and 1.3% with GC-AG splice sites), with only 36 FLEXI RNAs having sequence characteristics of minor U12-type spliceosomal introns (34 with GU-AG and 2 with AU-AC splice sites), and 23 having non-canonical splice sites (*e.g.*, AU-AG and AU-AU; Fig. 3A and Supplemental Fig. S5) (Burset et al. 2000; Sheth et al. 2006). The identified FLEXI RNAs had a canonical branch-point (BP)

consensus sequence (Fig. 3A) (Gao et al. 2008), suggesting that most if not all were excised as lariat RNAs and debranched after splicing, as found for mirtron pre-miRNAs (Okamura et al. 2007; Ruby et al. 2007).

In our previous analysis of human plasma RNAs, we identified 44 different FLEXI RNAs of which 13 corresponded to annotated agotrons and 10 corresponded to pre-miRNAs of annotated mirtrons, with 7 annotated as both an agotron or mirtron (Yao et al. 2020). Of the > 8,000 different FLEXI RNAs detected here in the human cellular RNA and remapped plasma RNA datasets, 65 corresponded to an annotated agotron (Hansen et al. 2016) and 114 corresponded to a pre-miRNA for an annotated mirtron (Fig. 3B and C) (Berezikov et al. 2007). Notably, the proportion of FLEXI RNAs corresponding to annotated agotrons or mirtron pre-mRNAs in plasma (22.8%) was considerably higher than that in the cellular RNA preparations (1.7-2.6%; Fig. 3C), possibly reflecting preferential cellular export or greater stability of these RNAs in plasma (Yao et al. 2020). A small number of FLEXI RNAs found in cells but not plasma (43, 0.5% of the total) encode snoRNAs, all of which were also detected as mature snoRNAs in the same RNA samples (Fig. 3C).

Analysis of other characteristics showed that the 224 FLEXI RNAs detected in human plasma were a relatively homogeneous subset with peaks at 90-nt length, 70% GC content, and -40 kcal/mole minimum free energy (MFE; ΔG) for the most stable RNA secondary structure predicted by RNAfold (Fig. 1C). By comparison, the FLEXI RNAs detected in cells were more heterogeneous, with similar peaks but larger shoulders extending to longer lengths, lower GC contents, and less stable predicted secondary structures (≥ -25 kcal/mole; Fig. 1C). The more homogeneous subset of FLEXI RNAs found in plasma could reflect preferential export or greater resistance to plasma RNases of shorter, more stably structured FLEXI RNAs.

Abundance of FLEXIs in cellular RNA samples

Fig. 1D shows density plots of the abundance (RPMs) of different categories of FLEXIs in the different cellular RNA samples compared to those of sncRNAs spanning a range of different

cellular abundances in the same samples (Supplemental Table S2). The large numbers of newly identified FLEXIs (denoted All other FLEXIs) showed two major peaks: one at ~0.001 RPM and the other between 0.002 and 0.1 RPM with a tail extending to 1.3-6.9 RPM in the different cellular RNA samples. In HeLa S3 cells, the peak between 0.01 and 0.1 RPM was predominant with only small peaks at lower abundances. In each of the cellular RNA samples, the FLEXIs previously annotated as agotrons, mirtrons, or containing embedded snoRNAs overlapped the peak of more abundant FLEXI RNAs. The abundances of most FLEXIs overlapped the lower end of the abundance distribution for snoRNAs (dashed yellow line) in the same samples. By using sncRNAs of known cellular abundance to produce a linear regression model for the relationship between $\log_{10}$ transformed copy number per cell and RPM values in the TGIRT-seq datasets, we estimated that the most abundant FLEXIs (≥ 1 RPM) may be present at ≥ 1-2 $\times 10^3$ molecules per cell and that substantial numbers of FLEXIs with RPMs ≥ 0.01 RPM (20-87% in different cellular RNA samples) may be present at 150-187 copies per cell (Supplemental Fig. S6).

FLEXIs exhibit different degrees of evolutionary conservation and highly conserved FLEXIs are associated with a distinct set of RNA-binding proteins

Fig. 1C right panel shows density plots of phastCons scores for all FLEXIs detected in a combined dataset for the human cellular RNA samples, with those corresponding to mirtrons, agotrons, and snoRNAs again split out as separate categories from all other FLEXIs. Most FLEXIs, including those corresponding to mirtron pre-miRNAs or agotrons, had low phastCons scores with peaks at 0.06-0.09 compared to 0.02 for other annotated short introns in GRCh38 and with tails extending to higher phastCons scores. As might be expected, FLEXIs encoding snoRNAs had higher phastCons scores (four at ≥ 0.5) than did other FLEXIs (Fig. 2C, right panel, yellow line). The low phastCons scores for most FLEXIs, including those with biological functions as agotrons or mirtron pre-miRNAs, indicates that they were acquired recently in the human lineage or have undergone rapid sequence divergence.

Five percent (399) of the detected FLEXIs that were not annotated as mirtrons, agotrons or encoding snoRNAs, had phastCons scores > 0.47 and 2% (159) had phastCons scores > 0.74), possibly reflecting an evolutionarily conserved sequence-dependent function. At the high end of the spectrum, 44 FLEXI RNAs had phastCons scores (≥ 0.99). Forty-one of these highly conserved FLEXIs were within protein-coding sequences and 37 were known to be alternatively spliced to generate different protein isoforms, with 26 sharing 5'- or 3'-splice sites with a longer intron and 16 containing in-frame protein-coding sequences that would be expressed if the intron was retained in a mRNA (examples in *HNRNPL*, *HNRNPM*, and *FXR1*; UCSC genome browser; Supplemental File). A FLEXI in the human *EIF1* gene with phastCons score = 1.00 resulted from acquisition of a 3'-splice site in its highly conserved 3' UTR and is spliced to encode a novel human-specific EIF1 isoform (chr17:41,690,818-41,690,902; Supplemental File) (Kim et al. 2020).

A search of CLIP-seq datasets (Hafner et al. 2010; Rybak-Wolf et al. 2014; Van Nostrand et al. 2016) identified a group of RNA-binding proteins (RBPs), whose binding sites were significantly enriched ($p \leq 0.05$ calculated by Fisher's exact test) in highly conserved FLEXIs (phastCons scores ≥ 0.99), including alternative splicing regulators (KHSRP, TIAL1, TIA1, PCBP2), extrinsic splicing factors (SFRS1, U2AF1, U2AF2), and a number of protein with no known RNA splicing- or miRNA-related function described further below (Fig. 3D). By contrast, annotated binding sites for core spliceosomal proteins (AQR, BUD13, EFTUD2, PRPF8, SF3B4) were under-represented in these highly conserved FLEXI RNAs (Fig. 3D).

FLEXI RNAs contain experimentally identified binding sites for a variety of RNA-binding proteins

The finding above that highly conserved FLEXIs are enriched in CLIP-seq identified binding sites for a distinct set of RBPs prompted us to do a comprehensive search for RBPs associated with different FLEXI RNAs in high confidence eCLIP (Van Nostrand et al. 2016), DICER PAR-CLIP (Rybak-Wolf et al. 2014), and AGO1-4 PAR-CLIP (Hafner et al. 2010) datasets. We found

that more than half of the detected FLEXI RNAs (4,505; 55%) contained an experimentally identified binding site for one or more of 126 different RBPs (Fig. 4, Supplemental Fig. S7, and Supplemental File). These 126 RBPs included spliceosome components and proteins that function in RNA splicing regulation; DICER, AGO1-4 and other proteins that function in the processing or function of miRNAs; and a surprising number of proteins whose primary functions are unrelated to RNA splicing or miRNAs. Notably, 121 of the identified RBPs had CLIP-seq-identified binding sites in multiple different FLEXI RNAs (Supplemental Fig. S7), with 53 RBPs having CLIP-seq-identified binding sites in 30 or more different FLEXI RNAs (Fig. 4A).

Overall, compared to longer introns > 300 nt or all RBP-binding sites in the CLIP-seq datasets, the detected FLEXI RNAs were significantly enriched ($p \leq 0.05$ calculated by Fisher's exact test) in CLIP-seq-identified binding sites for six spliceosomal proteins (AQR, BUD13, EFTUD2, PPIG, PRPF8, SF3B4), with these proteins found associated with 740 to 1,922 different FLEXI RNAs (Fig. 4). The enrichment of CLIP-seq identified binding sites for this set of spliceosomal proteins in this large group of FLEXI RNAs suggests that they may dissociate more slowly from spliceosomal complexes than do longer introns.

Many of the detected FLEXIs also contained CLIP-seq-identified binding sites for proteins with miRNA-related functions, with 250 containing a binding site for AGO1-4, 308 containing a binding site for DICER, and 66 containing binding sites for both AGO1-4 and DICER (Fig. 4A). However, only 23 of the 250 FLEXI RNAs identified as a binding site for AGO1-4 in the AGO1-4 PAR-CLIP dataset corresponded to an annotated agotron (Hansen et al. 2016) and only 44 of the 308 FLEXIs identified as binding site for DICER in the DICER PAR-CLIP dataset corresponded to a pre-miRNA for an annotated mirtron (Fig. 3C and Supplemental File) (Wen et al. 2015). The large numbers of additional FLEXIs containing AGO1-4 or DICER binding sites could be unannotated agotrons or mirtrons. Alternatively, they could be processed by DICER into other types of short regulatory RNAs, function as sponges for AGO1-4 and DICER, or affect the subcellular localization of these proteins, as found recently for a circular RNA linked to aberrant nuclear localization of DICER in glioblastoma (Bronisz et al. 2020). As noted

previously, the FLEXI RNAs with annotated DICER-binding sites differed from other FLEXIs in showing discrete size classes of relatively abundant shorter RNA fragments, as expected for DICER cleavage (Supplemental Fig. S4).

Surprisingly, 23 RBPs that bind 30 to 365 different FLEXI RNAs have no known RNA splicing- or miRNA-related function (Fig. 4A, protein names in black). They instead function in a variety of other cellular processes, including regulation of transcription, apoptosis, stress responses, cellular growth regulation, and histone assembly and disassembly (summarized in Table 1), potentially linking FLEXI RNA binding to the regulation of these processes. In general, the binding of these protein to FLEXI RNAs could contribute to the regulation of cellular processes by regulating the splicing and expression of the FLEX host genes, by forming an RNP complex that functions directly in the process or its regulation; or by changing the intracellular localization or level or free protein, particularly for those proteins that bind large numbers of different FLEXI RNAs.

Subsets of FLEXIs bind RBPs that perform specialized biological functions

Although the majority of FLEXI RNAs have annotated binding sites for spliceosomal proteins, the findings above that highly conserved FLEXIs were enriched in CLIP-seq-identified binding sites for other types of proteins and under-represented in binding sites for spliceosomal proteins (Fig. 3D) suggested that there could be different classes of FLEXIs that bind different RBPs, possibly in order to perform specialized biological functions. Precedents for the latter are agotrons and mirtrons, which presumably dissociate from the spliceosome and preferentially bind AGO1-4 to downregulate mRNAs or DICER to function as miRNA precursors (Berezikov et al. 2007; Hansen et al. 2016).

To search for such FLEXIs, we focused on RBPs with no known RNA splicing- or miRNA-related function that bind 30 or more different FLEXIs and examined UpSet plots to assess the extent to which these RBPs are associated with FLEXIs that lack CLIP-seq-identified binding sites for the five most ubiquitous core spliceosomal proteins (PRPF8, SF3B4, AQR, EFTUD2,

and BUD13), which collectively bind thousands of other FLEXI RNAs (Fig. 4A). Using agotrons and mirtrons as standards for FLEXI RNAs with specialized biological functions, we found that 51% of the FLEXIs with a CLIP-seq-identified binding site for AGO1-4 and 44% of those with a CLIP-seq-identified binding site for DICER lacked annotated binding sites for any of the five ubiquitous spliceosomal proteins (Fig. 5A and B). Similar UpSet plots for the 23 RBPs that bind 30 or more different FLEXIs but have no known RNA splicing- or miRNA-related function identified 16 for which substantial proportions (29-55%) of the bound FLEXIs lacked annotated binding sites for any of the five spliceosomal proteins (examples shown in Fig. 5C-H; others listed in the legend of Fig. 5 and indicated by a † in Fig. 6 below). These findings suggest that after splicing, some groups of FLEXIs may preferentially bind other non-splicing related RBPs with a variety of cellular functions.

FLEXIs that bind the same RBP originate from host genes with related biological functions

To further explore the biological significance of these FLEXI RNA-RBP interactions, we performed hierarchical clustering based on GO terms for the host genes encoding FLEXI RNAs bound by the same RBP. Focusing again on those RBPs that bind 30 or more different FLEXI RNAs, we identified five major clusters of FLEXIs whose host genes showed significant enrichment for different sets of biological processes (Fig. 6). By contrast, a control group consisting of the 3,639 detected FLEXI RNAs that did not contain an annotated RBP-binding site in the CLIP-seq datasets originated from host genes that showed no similar enrichment for GO terms associated with biological processes (Fig. 6, right lane), nor did randomly sampled subsets of host genes for all FLEXIs, FLEXIs that did not contain an annotated RBP-binding site, all human genes, or all human genes than contain short introns (≤ 300 nt) over a range of different randomly sampled pool sizes (Supplemental Fig. S8). These controls indicate that the GO term enrichment for FLEXIs bound by different RBPs is not merely a byproduct of random sampling of the genes encoding FLEXIs. Collectively, these findings suggest that the host genes

encoding FLEXI RNAs bound by the same RBP have related biological functions and thus might be coordinately regulated to produce different subsets of FLEXI RNPs.

FLEXI RNAs may function in diverse cellular regulatory pathways

The GO term clustering and biological functions of the host genes encoding FLEXI RNAs bound by different RNPs identify previously unsuspected interactions and connections to cellular regulatory pathways. Cluster I comprised of FLEXI RNAs that bind the five ubiquitous core spliceosomal proteins (SF3B4, BUD13, EFTUD2, AQR, PRPF8) plus AGO1-4 originated from host genes associated with the widest variety of biological processes, whereas clusters II to V were comprised of FLEXI RNAs whose host genes were associated with different subsets of these processes. The host genes for FLEXI RNAs bound by the RBPs in cluster II were enriched for GO terms involved with rRNA processing, translation, and mRNA splicing, while those in cluster III were enriched for a smaller set of GO terms involved with rRNA processing and translation.

Notably, cluster III includes three RBPs (DKC1, NOLC1, and AATF; denoted with § in Fig. 6) that have annotated binding sites in overlapping sets of FLEXIs that also contained annotated binding sites for the five core spliceosomal proteins (Supplemental Fig. S9). The FLEXI RNAs bound by these RBPs were distinguished by relatively low GC content (peak at 30-40% GC) and above average phastCons scores (peaks at 0.3 to 0.4; Supplemental Fig. S10A), and upon further examination were found to include 41 of the 43 FLEXIs that encode snoRNAs (Supplemental File). DKC1 (dyskerin) and NOLC1 (nucleolar and coiled-body phosphoprotein 1) are components of snoRNPs that bind intronic snoRNA sequences co-transcriptionally to delineate these regions for snoRNA processing (Kufel and Grzechnik 2019), possibly accounting for the occurrence of CLIP-seq identified spliceosomal protein binding sites in the same FLEXIs (Supplemental Fig. S9). DKC1 also stabilizes telomerase RNA (MacNeil et al. 2019), and NOLC1 interacts with TRF2 (Telomeric Repeat-Binding Factor 2) to mediate its trafficking between the nucleolus and nucleus (Yuan et al. 2017). The third protein, AATF (Apoptosis

Antagonizing Transcription Factor), is a Pol II-interacting protein that regulates the function of the p53 and Rb oncogenes and is over produced in many cancers to inhibit apoptosis (Iezzi and Fanciulli 2015; Kaiser et al. 2019). A recent study found that AATF binds 45S precursor rRNA, as well as mRNAs encoding ribosome biogenesis factors and both H/ACA- and C/D-box snoRNAs, leading to the hypothesis that AATF involvement in ribosome biogenesis might be linked to its role in apoptosis (Kaiser et al. 2019). However, 11 of the 34 FLEXIs bound by AATF are short introns that neither encode snoRNAs nor are in genes related to ribosome biogenesis, and AATF also binds to 925 long introns (> 300 nt) of which only 295 encode snoRNAs, suggesting a broader role for AATF linked to RNA splicing or intron binding. Cluster III also includes RPS3, which has been implicated in regulating transcription, DNA damage response, and apoptosis (Gao and Hardwidge 2011); DDX3X, a DEAD-box RNA helicase with functions in regulating stress granule formation and apoptosis (Schröder 2010; Hilliker et al. 2011); and YBX3, a homolog of YBX1, a low specificity RBP that plays a role in regulating stress granule assembly, sorting small non-coding RNAs into extracellular vesicles, and a variety of other cellular processes (Somasekharan et al. 2015; Shurtleff et al. 2017).

The host genes for the FLEXI RNAs in cluster IV are enriched in many of the same GO terms related to RNA splicing as cluster II plus additional GO terms related to transcription and chromatin (Fig. 6). Surprisingly, 8 of the 12 RBPs that comprise this cluster corresponded to those identified above (Fig. 3D) as binding FLEXI RNAs with very high phastCons scores (≥ 0.99; BCLAF1, GRWD1, SRSF1, TIA1, UCHL5, U2AF1, U2AF2, and ZNF622; denoted with asterisks in Fig. 6), although these proteins also bind many additional FLEXIs with lower phastCons scores (Supplemental Fig. S10B and Supplemental File). Five of these proteins as well as TRA2 in cluster IV bind FLEXIs with significantly lower GC content than other FLEXIs (Supplemental Fig. 6B). Four of the proteins that bind highly conserved FLEXIs (TIA1, SRSF1, U2AF1, and U2AF2) function in the regulation of alternative splicing, as does TRA2A and possibly PPIG, which is also found in this cluster, potentially providing examples of subsets of FLEXI RNPs that result from alternative splicing regulation. TIA1 also plays a key role in stress

granule formation (Kedersha et al. 1999); GRWD1is a histone-binding protein that regulates chromatin dynamics (Sugimoto et al. 2015); and BCLAF1 (BCL2-2-associated transcriptional factor) and ZNF622 are positive regulators of apoptosis (Vohhodina et al. 2017), increasing to five the number of FLEXI RNA-binding proteins connected to this process (see above).

The host genes for the FLEXI RNAs in cluster V are enriched in only a few GO terms for each RBP. Four of these proteins function in splicing regulation (LSM11, PCBP2, RBFOX1, and GPKOW) and three others are notable regulatory proteins: IGF2BP1 (insulin-like growth factor 2 mRNA-binding protein1), which functions in cell cycle regulation (Müller et al. 2020); G3BP1 (Ras GTPase-activating protein binding protein 1), a helicase that plays an essential role in innate immunity, functions in stress granule assembly, is associated with cellular senescence, and regulates important signaling pathways (Zhang et al. 2019; Eiermann et al. 2020; Omer et al. 2020); and PABPN1 (polyadenylate-binding protein 2), a crucial player in double-strand-break repair (Gavish-Izakson et al. 2018). Three of these proteins (IGF2BP1, LSM11, and G3BP1) were among those binding to substantial subsets of FLEXIs that lacked annotated binding sites for the five core spliceosomal proteins (see above). Collectively, these findings suggest that the binding of non-spliceosomal RBPs to different subsets of FLEXIs could be linked to multiple cellular regulatory pathways.

FLEXI RNAs as potential cancer biomarkers

The cell- and tissue-specific expression patterns of FLEXI RNAs suggested that they might be useful as biomarkers to distinguish normal and abnormal cellular states. To test this idea, we compared FLEXI RNAs and FLEXI host genes in matched tumor and neighboring healthy tissue from two breast cancer patients (patients A and B; PR^+, ER^+, $HER2^-$ and PR unknown, ER^-, $HER2^-$, respectively) and two breast cancer cell lines (MDA-MB-231 and MCF7). UpSet plots showed hundreds of differences in FLEXI RNAs and FLEXI host genes between the cancer and healthy samples (Fig. 7A for FLEXIs detected at ≥ 0.01 RPM and Supplemental Fig. S11 for FLEXI RNAs detected at ≥ 1 read). The discriminatory ability of FLEXIs was also evident in

scatter plots comparing the FLEXI RNAs detected in the matched healthy and tumor samples from patients A and B, which showed a wider spectrum of differences than did those for mRNAs from the same host genes quantitated in chemically fragmented RNA preparations (Fig. 7B). The scatter plots identified multiple candidate FLEXI RNA biomarkers, including 18 and 16 in patients A and B, respectively, that were detected at relatively high abundance (0.05-0.16 RPM) and in at least two replicate libraries from the cancer patient, but not detected in the matched healthy tissue (red dots in Fig. 7B, genes listed to the right).

GO enrichment analysis of FLEXI RNA host genes in the four cancer samples but not healthy tissues showed significant enrichment ($p \leq 0.05$) of hallmark gene sets (Liberzon et al. 2015) that may be disregulated in many cancers (*e.g.*, glycolysis, G2M checkpoint, UV response up, and PI3K/AKT/MTOR signaling; Fig. 7C, pathway names in red). Gene sets that were significantly enriched in one or more of the cancer samples but not in the healthy controls included mitotic spindle, MYC targets V1 and V2, estrogen response early and late, androgen response, oxidative phosphorylation, mTORC1 signaling, apical junction, and cholesterol homeostasis (Fig. 7C, pathway names in orange).

Only a small number of the potential FLEXI RNA biomarkers identified in the UpSet and scatter plots in Fig. 7 corresponded to previously identified oncogenes (names with asterisks in Fig. 7A and B). This reflects a combination of factors, including that some prominent oncogenes (*e.g.*, *CD24*, *ERAS*, and *MYC*) as well as hormone receptors genes (*ERBB2*, *ESR1*, and *PR*) do not encode FLEXIs; that FLEXI RNA abundance is dictated by alternative splicing and intron RNA turnover in addition to transcription; and that those FLEXI RNAs that best discriminate between cancer and healthy samples arise from genes that are strongly up or downregulated in response to oncogenesis but are not oncogenes that drive this process.

To directly examine the relationship between FLEXI RNAs and oncogene expression, we identified FLEXI RNAs from known oncogenes ($n = 803$) (Liu et al. 2017) that were ≥ 2-fold up- or downregulated in any of the cancer samples. UpSet plots identified 169 FLEXI RNAs from known oncogenes that were up upregulated in any of the cancer samples compared to the

healthy controls, with 13 to 60 upregulated in only one of the cancer samples and 5 upregulated in all four cancer samples (Fig. 8A). Another 81 FLEXI RNAs from known oncogenes were downregulated by ≥ 2-fold in any of the cancer samples compared to healthy controls, with 1 to 29 downregulated in only one of the cancer samples and 4 downregulated in all four cancer samples (Fig. 8B). Similar patterns were seen in UpSet plots for up- and downregulated tumor suppressor genes (n = 1,217) (Zhao et al. 2016) (Fig. 8C and D).

The up and down patterns for FLEXI RNAs from both oncogenes and tumor suppressor genes again reflect that FLEXI RNA abundance is dictated by factors other than transcription (Fig. 2). The *FASN* (fatty acid synthase) gene, for example, contains 8 FLEXIs that were upregulated and 5 that were down regulated in the cancer samples (*e.g.*, FASN-13I and FASN81 in MCF7 cells and FASN-31I and FASN-26I in patient B*; FASN* FLEXIs highlighted in red in Fig. 8A and B). This situation does not preclude these introns from serving as a biomarker for a specific cancer, so long as they are found to be reproducibly up or downregulated in that cancer.

Notably, the RBP-binding sites enriched in oncogene FLEXIs were also potentially informative, as illustrated by scatter plots for the RBP-bindings sites that were enriched in oncogene FLEXIs that were ≥ 2-fold upregulated in MCF-7 or MDA-MB-231 cells or in all four cancer samples. These included proteins that function in or regulate transcription (GTF2F1), RNA splicing (KHSRP), RNA processing (CSTF2T), nuclear cap binding (NCBP2), cell cycle progression (TBRG4), and cell division (CDC40; Fig. 8E),

Discussion

Here, we characterized a new large class of human RNAs, short full-length excised linear intron RNAs at the transcriptome level. In total, including FLEXIs detected in both the initial cellular and plasma samples and the subsequent cancer samples, we identified 8,687 different FLEXI RNAs expressed from 3,923 host genes representing ~17% of the 51,645 short introns (≤ 300 nt) annotated in Ensembl GRCh38 Release 93 annotations (http://www.ensembl.org). Most FLEXI RNAs have relatively high GC content (60-70%) and are predicted to fold into stable RNA

secondary structures (-20 to -50 kcal/mole). The detected FLEXI RNAs had cell- and tissue-specific expression patterns, reflecting differences in host gene transcription, alternative splicing, and intron RNA turnover, and they contained experimentally identified binding sites for diverse proteins, including transcription factors, chromatin remodeling proteins, and proteins that function in cellular stress responses, apoptosis, and cell proliferation, potentially linking FLEXI RNA binding to regulation of these processes. Their cell-specific expression patterns and origin from thousands of different protein-coding genes suggest that FLEXI RNAs may have utility as RNA biomarkers for human diseases.

Some of the FLEXI RNAs that we detected in human cells and plasma were known to have specialized biological functions as agotrons or as mirtron- or snoRNA-precursors and to bind specific non-splicing related RBPs needed to carry out these functions. Such binding could occur either before or after dissociation from the spliceosome, which may occur at different rates for different excised intron RNAs. To explore whether the much larger number of newly detected FLEXIs might include others that have other specialized biological functions, we searched for FLEXI RNA-binding proteins in published CLIP-seq datasets. We thus identified 126 different proteins that have annotated binding sites in FLEXI RNAs, with 121 of these proteins binding multiple different FLEXI RNAs and 53 binding 30 or more different FLEXIs (Fig. 4 and Supplemental Fig. S7). Based on the CLIP-seq datasets, spliceosomal proteins have annotated binding sites in the largest numbers of FLEXI RNAs, followed by AGO1-4 and DICER, with the latter including but not limited to annotated agotrons and mirtron pre-miRNAs (Fig. 4).

Surprisingly, 23 proteins that have CLIP-seq identified binding sites in 30 to 365 different FLEXIs have no known RNA splicing- or miRNA-related functions (Table 1), and 16 of these proteins were associated with distinct subsets of FLEXI RNAs that were under-represented in binding sites for spliceosomal proteins (Fig. 3D and Fig. 5). These RBPs included 16 that function in other processes unrelated to RNA splicing, including seven transcription regulators, four chromatin-binding or remodeling proteins, and two proteins that function in protein modification. Five of these proteins function in the regulation of apoptosis (AATF, BCLAF1,

DDX3X, RPS3, ZNF622); four are regulators of p53 transcription or function (AATF, BCLAF1, DDX24, GRWD1); four function in DNA damage responses (AATF, BCLAF1, PABPN1, RPS3); four function in cellular stress responses (BCLAF1, G3BP1, SUB1, AATF); three function in cell growth regulation (AATF, DDX3X, IGF2BP1); and three play key roles in stress granule formation (DDX3X, TIA1, and G3BP1).

In general, FLEXI RNAs could contribute to cellular regulation by serving as substrates for DICER- or RNase III-cleavage to generate as yet unannotated small regulatory RNAs; by forming an RNP complex that functions in or regulates a process; by regulating of FLEXI host gene splicing, as found for alternative splicing factors that bind highly conserved FLEXI RNAs within protein-coding sequences (Fig. 3D and Fig. 6); by altering the subcellular localization of the bound protein, as found for a circular RNA linked to aberrant nuclear localization of DICER in glioblastoma (Bronisz et al. 2020); or by sequestering proteins, as suggested for the yeast linear intron RNAs that accumulate in stationary phase (Morgan et al. 2019; Parenteau et al. 2019).

Pertinent to their ability to function in cellular regulation, more than half of the newly identified FLEXIs were as abundant as mirtrons, agotrons, or biologically relevant snoRNAs in the same datasets (Martens-Uzunova et al. 2015; Hansen et al. 2016; Oliveira et al. 2021), with quantitative estimates based on sncRNAs with known copy number per cell values in the same datasets suggesting that the most abundant FLEXIs may be present at 1-2 x 10^3 copies per cell and substantial numbers (20-87% in the different cellular RNA samples) may be present at ≥ 150 copies per cell (Fig. 1D and Supplemental Fig. S6). Although most individual FLEXIs were not present in sufficiently high abundance to bind a substantial fraction of a typical intracellular target protein, collectively thousands of different FLEXIs have annotated binding sites for a small set of spliceosomal proteins and hundreds of different FLEXIs have annotated binding sites for DICER and AGO1-4. Thus in aggregate, FLEXI abundance could be sufficient to affect intracellular protein levels in response to stimuli that globally affect FLEXI RNA turnover, as found for the collective of yeast linear introns that accumulate under stress conditions (Morgan et

al. 2019). The findings that key cellular regulatory proteins bind groups of 30 to 365 different FLEXIs (Fig. 4A) and that FLEXI RNAs bound by the same RBP have related and possibly coordinately regulated biological functions (Fig. 6) further suggest how the effects of FLEXI binding on the intracellular protein concentrations might be amplified to compensate for the relatively low abundance of some FLEXI RNAs.

With respect to evolution, although more conserved than other short introns, most of the detected FLEXI RNAs, including those corresponding to mirtron pre-miRNAs or agotrons, had relatively low PhastCons scores (< 0.2; Fig. 1C, right panel), indicating either recent acquisition or rapid sequence divergence. FLEXI RNAs with relatively high PhastCons scores included those encoding snoRNAs, which contain binding sites for proteins involved in snoRNA biogenesis (Supplemental Fig. S9), and those that are alternatively spliced to generate different protein isoforms, which contain binding sites for a distinct set of non-spliceosomal RBPs, including proteins known to function in alternative splicing (Fig. 3D).

In general, FLEXI RNAs could arise either by splice-site acquisition, as found for an *EIF1* intron whose acquisition resulted in a novel human EIF1 isoform (Kim et al. 2020), or by an active intron transposition process, as found for spliceosomal introns in fungi, algae, and yeast (van der Burgt et al. 2012; Simmons et al. 2015; Lee and Stevens 2016). Most of the short introns in the human genome (97%) have unique sequences, with the remainder (1,719 introns with 693 unique sequences) arising by external or internal gene duplications, as described in detail for an abundant FLEXI RNA found in human plasma (Yao et al. 2020). Thus, if intron transposition is involved in the origin of FLEXIs, it must either be relatively rare or followed by rapid sequence divergence. The large number of FLEXIs encoded in the human genome may reflect that short introns within protein-coding sequences are more easily acquired or less deleterious for gene function than longer introns. Additional functions of FLEXI RNAs may arise secondarily and would be favored by stable predicted secondary structures that facilitate splicing by bringing splice sites closer together, contribute to the formation of protein-binding

sites, and/or stabilize the intron RNA from turnover by cellular RNases, enabling them to persist long enough to perform their function after debranching.

Regardless of their function or origin, FLEXI RNAs constitute a large previously unidentified class of potential RNA biomarkers, with genome coverage comparable to mRNAs or miRNA. In addition to being linked to the transcription of thousands of protein-coding and lncRNA genes, FLEXI RNA levels can also reflect differences in alternative splicing and intron RNA stability (Fig. 2), potentially providing higher resolution of cellular differences than mRNAs transcribed from the same gene. FLEXI RNAs may have particular utility as biomarkers in bodily fluids such as plasma, where they are enriched compared to other RNA species and their stable secondary structures and/or bound proteins may protect them from extracellular RNases (Yao et al. 2020). As many FLEXIs are predicted to fold into stable RNA secondary structures, their initial identification as candidate biomarkers seems best done by TGIRT-seq, which can yield full-length, end-to-end sequence reads of structured RNAs. Once identified, the validation of candidate FLEXI biomarkers and their routine monitoring in clinical samples would best be done by methods that give quantitative read outs for specific RNAs, such as RT-qPCR, microarrays, other hybridization-based assays, or targeted RNA-seq. Targeted RNA panels of FLEXI RNAs by themselves or together with other RNA biomarkers or analytes could provide a rapid cost-effective method for the diagnosis and routine monitoring of progression and response to treatment of a wide variety of human diseases.

Methods

DNA and RNA oligonucleotides

The DNA and RNA oligonucleotides used for TGIRT-seq on the Illumina sequencing platform are listed in Supplemental Table S3. Oligonucleotides were purchased from Integrated DNA Technologies (IDT) in RNase-free, HPLC-purified form. 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.

RNA preparations

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

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

For RNA isolation, cells were harvested by centrifugation (after trypsinization for HEK-293T/17 and MDA-MB-231 cells) at 300 x g for 10 min at 4 °C and washed twice by centrifugation with cold Dulbecco’s Phosphate Buffered Saline (Thermo Fisher). The indicated number of cells (see above) was then resuspended in 600 µL of mirVana Lysis Buffer and RNA was isolated according to the kit manufacturer’s protocol with elution in a final volume of 100 µL. To remove residual DNA, UHRR and HeLa S3 RNAs (1 µg) and patients A and B healthy and cancer tissue RNAs (500 ng) were treated with 20 U exonuclease I (Lucigen) and 2 U Baseline-ZERO DNase (Lucigen) in Baseline-ZERO DNase Buffer for 30 min at 37 °C. K562, MDA-MB-231 and HEK-293T cell RNAs (5 µg) were incubated with 2 U TURBO DNase

(Thermo Fisher). After DNA digestion, RNA was cleaned up with an RNA Clean & Concentrator kit (Zymo Research) with 8 volumes of ethanol (8X ethanol) added to maximize the recovery of small RNAs. The eluted RNAs were ribodepleted by using the rRNA removal section of a TruSeq Stranded Total RNA Library Prep Human/Mouse/Rat kit (Illumina), with the supernatant from the magnetic-bead separation cleaned-up by using a Zymo RNA Clean & Concentrator kit with 8X ethanol. After checking RNA concentration and length by using an Agilent 2100 Bioanalyzer with a 6000 RNA Pico chip, RNAs were aliquoted into ~20 ng portions and stored at -80 °C until use.

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

TGIRT-seq

TGIRT-seq libraries were prepared as described (Xu et al. 2019) using 20-50 ng of ribodepleted unfragmented RNA or 4-10 ng of ribodepleted chemically fragmented RNA. The template-switching and reverse transcription reactions were done with 1 μM TGIRT-III (InGex) and 100 nM pre-annealed R2 RNA/R2R DNA starter duplex in 20 μl of reaction medium containing 450

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

TGIRT-seq of RNA from commercial human plasma pooled from multiple healthy individuals was described previously (Yao et al. 2020), and the resulting datasets were previously deposited in the National Center for Biotechnology Information Sequence Read Archive under accession number PRJNA640428.

Bioinformatics

All data analysis used combined TGIRT-seq datasets obtained from multiple replicates of different sample types (Supplemental Table S1). Illumina TruSeq adapters and PCR primer sequences were trimmed from the reads with Cutadapt v2.8 (sequencing quality score cut-off at 20; p-value <0.01) (Martin 2011) and reads <15-nt after trimming were discarded. To minimize mismapping, a sequential mapping strategy was used. First, reads were mapped to the human

mitochondrial genome (Ensembl GRCh38 Release 93) and the *Escherichia coli* genome (GeneBank: NC_000913) using HISAT2 v2.1.0 (Kim et al. 2019) with customized settings (-k 10 --rfg 1,3 --rdg 1,3 --mp 4,2 --no-mixed --no-discordant --no-spliced-alignment) to filter out reads derived from mitochondrial and *E. coli* RNAs (denoted Pass 1). Unmapped read from Pass1 were then mapped to a customized set of references sequences for genes encoding human sncRNAs (miRNA, tRNA, Y RNA, Vault RNA, 7SL RNA, 7SK RNA genes) and rRNAs (the 2.2-kb 5S rRNA repeats from the 5S rRNA cluster on chromosome 1 (1q42, GeneBank: X12811) and the 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) and splice aware mapping (-k 10 --rdg 1,3 --rfg 1,3 --mp 4,2 --no-mixed --no-discordant --dta) (denoted Pass 4). Finally, the remaining unmapped reads were mapped to Ensembl GRCh38 Release 93 by Bowtie 2 v2.2.5 (Langmead and Salzberg 2012) using local alignment (with settings as: -k 10 --rdg 1,3 --rfg 1,3 --mp 4 --ma 1 --no-mixed --no-discordant --very-sensitive-local) to improve the mapping rate for reads containing post-transcriptionally added 5’ or 3’ nucleotides (poly(A) or poly(U)), short untrimmed adapter sequences, or non-templated nucleotides added to the 3’ end of the cDNAs by TGIRT-III during TGIRT-seq library preparation (denoted Pass 5). For reads that map to multiple genomic loci with the same mapping score in passes 3 to 5, the alignment with the shortest distance between the two paired ends (*i.e.*, the shortest read span) was selected. In the case of ties (*i.e.*, reads with the same mapping score and read span), reads mapping to a chromosome were selected over reads mapping to scaffold sequences, and in other cases, the read was assigned randomly to one of the tied choices. The filtered multiply mapped reads were then combined with the uniquely mapped reads from Passes 3-5 by using SAMtools v1.10 (Li et al. 2009) and intersected with

gene annotations (Ensembl GRCh38 Release 93) with the *RNY5* gene and its 10 pseudogenes, which are not annotated in this release, added manually to generate the counts for individual features. Coverage of each feature was calculated by BEDTools v2.29.2 (Quinlan 2014). To avoid miscounting reads with embedded sncRNAs that were not filtered out in Pass 2 (*e.g.*, snoRNAs), reads were first intersected with sncRNA annotations and the remaining reads were then intersected with the annotations for protein-coding genes RNAs, lincRNAs, antisense RNAs, and other lncRNAs to get the read count for each annotated feature.

Coverage plots and read alignments were created by using Integrative Genomics Viewer v2.6.2 (IGV) (Robinson et al. 2011). Genes with >100 mapped reads were down sampled to 100 mapped reads in IGV for visualization.

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

UpSet plots of FLEXI RNAs from different sample types were plotted by using the ComplexHeatmap package v2.2.0 in R (Gu et al. 2016), and Venn diagrams were plotted by using the VennDiagram package v1.6.20 in R (Chen and Boutros 2011). For 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 of $\log_2$ transformed RPM of the detected FLEXI RNAs and FLEXI host genes were plotted by using R. PCA analysis of cell-type specific FLEXI RNA profiles in replicate cellular RNA datasets was plotted using R, and PCA initialized t-SNE and ZINB-WaVE analyses of these datasets were plotted using the Rtsne (https://github.com/jkrijthe/Rtsne) and zinbwave (https://github.com/drisso/zinbwave) packages in R (Risso et al. 2018).

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 (Sheth et al. 2006; Gao et al. 2008). Splice-site consensus sequences of FLEXI RNAs were calculated from nucleotides frequencies of the first and last 10 nt from the intron ends. FLEXI RNAs corresponding to U12-type introns were identified by searching for (i) FLEXI RNAs with AU-AC ends and (ii) the 5'-splice site consensus sequence of U12-type introns with GU-AG ends (Sheth et al. 2006) using FIMO (Grant et al. 2011) with the following settings: FIMO --text --norc <GU_AG_U12_5SS motif file> <sequence file>. The branch-point consensus sequence of U2-type FLEXI RNAs was determined by searching for motifs enriched within 40 nt of the 3' end of the introns using MEME (Bailey et al. 2009) with settings: meme <sequence file> -rna -oc <output folder> -mod anr -nmotifs 100 -minw 6 -minsites 100 -markov order 1 -evt 0.05. The branch-point consensus sequence of U12-type FLEXI RNAs (2 with AU-AC ends and 34 with GU-AG matching the 5' sequence of GU-AG U12-type introns) was identified by manual sequence alignment and calculation of nucleotide frequencies. Motif logos were plotted from the nucleotide frequency tables of each motif using scripts from MEME suite (Bailey et al. 2009).

FLEXI RNAs corresponding to annotated mirtrons, agotrons, and RNA-binding-protein (RBP) binding sites were identified by intersecting the FLEXI RNA coordinates with the coordinates of annotated mirtrons (Wen et al. 2015), agotrons (Hansen et al. 2016), 150 RBPs (eCLIP, GENCODE, annotations with irreproducible discovery rate analysis) (Van Nostrand et al. 2016), DICER PAR-CLIP (Rybak-Wolf et al. 2014), and Ago1-4 PAR-CLIP (Hafner et al. 2010) datasets by using BEDTools. The functional annotations, localization patterns, and predicted RNA-binding domains of the 150 RBPs in the ENCODE eCLIP dataset were based on Table S2 of (Van Nostrand et al. 2020). RBPs found in stress granules were as annotated in the RNA Granule and Mammalian Stress Granules Proteome (MSGP) databases (Nunes et al. 2019; Youn et al. 2019). The functional annotations, localization patterns, and RNA-binding domains of AGO1-4 and DICER were retrieved from the UniProt database (The UniProt Consortium 2018). FLEXI RNAs containing embedded snoRNAs were identified by intersecting the FLEXI

RNA coordinates with the coordinates of annotated snoRNA and scaRNA from Ensembl GRCh38 annotations.

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

Data access

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

Competing Interests Statement

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

Acknowledgements. We thank Marta Mastroianni and Ryan Nottingham for cell culture and RNA preparations, Blerta Xhemalce (University of Texas at Austin) for a fresh stock of MDA-

MB-231 cells, and Carla Van Den Berg (University of Texas at Austin) for comments on the manuscript. The authors acknowledge the Texas Advanced Computing Center (TACC; http://www.tacc.utexas.edu) at the University of Texas at Austin for providing high performance computing resources that have contributed to the research results reported within this paper (URL: http://www.tacc.utexas.edu). This work was supported by NIH grants R01 GM37949 and R35 GM136216 and Welch Foundation grant F-1607 to A.M.L.

Figure 1. Characteristics of FLEXI RNAs in human cells and plasma. (*A*) UpSet plots of FLEXI RNAs and their host genes detected at ≥ 1 read in unfragmented RNA preparations from the indicated samples. (*B*) Scatter plots comparing $\log_2$-transformed RPM of FLEXI RNAs and all transcripts of FLEXI host genes in different cellular RNA samples. r and $r_{s,}$ are Pearson and Spearman correlation coefficients, respectively. (*C*) Density plots of different characteristics of FLEXI RNAs in combined datasets for the UHRR, K-562, HEK-293T and HeLa S3 cellular RNA samples. Left panel, density plot showing the distribution of RNA fragment lengths mapping to FLEXIs in the cellular RNA samples (red line) compared to those mapping to other Ensemble GRCh38-annotated short introns ≤ 300 nt in the same samples (dashed black line). Percent intron length was calculated from the read spans of TGIRT-seq reads normalized to the length of the corresponding intron. Introns encoding embedded snoRNAs or scaRNAs were removed for these comparisons to avoid interference from mature snoRNA reads. Middle panels, density distribution plots of length, GC content, and minimum free energy calculated for the most stable RNA secondary structure predicted by RNAfold for all FLEXIs detected in the cellular RNA (red) and plasma cell-free RNA (purple) samples, compared to other Ensemble GRCh38-annotated short introns ≤ 300 nt detected in the same samples (dashed black line). Right panel, density distribution plots of phastCons scores of different categories of FLEXIs detected in the cellular RNA samples compared to all other Ensemble GRCh38 annotated short introns ≤ 300 nt (color codes as indicated in the Figure). PhastCons scores were the average PhastCons score across all intron bases calculated from multiple sequences alignment of 27

primates, including humans plus mouse, dog, and armadillo. (*D*) Density distribution plots of the abundance (RPM) of different categories of FLEXI RNAs color coded as indicated in the Figure in different cellular RNA samples. Only full-length FLEXI RNA reads with 5' and 3' ends within 3 nts of annotated splice sites were used in calculating abundances. The abundance distribution of annotated mature snoRNAs in the same samples is shown for comparison (dashed yellow line), as are the positions (arrows) of different sncRNAs detected by TGIRT-seq in the same samples, including two low abundance, biologically relevant C/D box snoRNAs (SNORD74: 0.01-0.1 RPM; SNORD78: 0.02-0.3 RPM) (Martens-Uzunova et al. 2015; Oliveira et al. 2021).

Figure 2. IGV screenshots showing read alignments for FLEXI RNAs. Gene names are at the top with the arrow below indicating the 5’ to 3’ orientation of the encoded RNA followed by tracts showing gene annotations (exons, thick bars; introns, thin lines), sequence, and read alignments for FLEXI RNAs color coded by sample type as indicated in the Figure (bottom right). (*A*) Long and short FLEXI RNAs; (*B*) FLEXI RNAs having high and low GC content; (*C*) FLEXI RNAs having low and high minimum free energies (MFEs) for the most stable RNA secondary structure predicted by RNAfold; (*D*) FLEXI RNAs showing cell-type specific differences due to alternative splicing and differential stability of FLEXI RNAs encoded by the same gene. The most stable secondary structure predicted by RNAfold is shown below the read alignments (panels A-C only) along with intron length, GC content, calculated MFE, and PhastCons score for 27 primates and three other species. In panel D, gene maps for the different RNA isoform generated by alternative splicing of FLEXI RNAs are shown at the bottom. Mismatched nucleotides in boxes at the 5' end of the RNA sequence are due to non-templated nucleotide addition (NTA) to the 3' end cDNAs by TGIRT-III during TGIRT-seq library preparation. Some *MAZ* FLEXIs (panel B) have a non-coded 3' A or U tail.

Figure 3. FLEXI RNA splice-site and branch-point consensus sequences, FLEXI RNAs annotated as mirtrons or agotrons or encoding an embedded snoRNAs in different sample types,

and RBP-binding sites enriched in highly conserved FLEXI RNAs. (*A*) 5'- and 3'-splice sites (5'SS and 3'SS, respectively) and branch-point (BP) consensus sequences of FLEXI RNAs compared to those of human major (U2-type) and minor (U12-type) spliceosomal introns. The number of FLEXIs matching each consensus sequence is indicated to the right. The remaining FLEXIs have non-canonical 5'- and 3'-splice site sequences. (*B*) Venn diagrams showing the relationships between FLEXI RNAs corresponding to annotated agotrons (left) or mirtrons (right) detected in different sample types. FLEXI RNAs annotated as both a mirtron and an agotron are included in both Venn diagrams. (*C*) Numbers and percentages of detected FLEXI and short introns ≤ 300 nt in the human genome (GRCh38) corresponding to annotated agotrons or mirtrons or encoding embedded snoRNAs in different sample types. "Agotron and Mirtron" indicates introns annotated as both an agotron or mirtron, and "Agotron or Mirtron" indicates the total number and percentage of introns annotated as either or both an agotron or mirtron. The number of embedded snoRNAs that are small Cajal body-specific snoRNAs (scaRNAs) is also indicated. (*D*) Scatter plots showing the relative abundance (percentage) of annotated binding sites for different RBPs in highly conserved FLEXI RNAs (phastCons score ≥ 0.99; n = 44) compared to that in all detected FLEXIs RNAs in the cellular and plasma samples. RBP-binding site annotations are from the ENCODE 150 RBP eCLIP dataset with irreproducible discovery rate (IDR) and AGO1-4 and DICER PAR-CLIP datasets. The scatter plot on the right is an enlargement of the 0 to 4% abundance region of the scatter plot on the left. RBPs whose relative abundance was significantly different between the highly conserved FLEXIs and all FLEXIs ($p \leq 0.05$ calculated by Fisher's exact test) are labeled with the name of the RBP color coded by protein function: red, RNA splicing related; orange, miRNA related; blue, both RNA splicing and miRNA related; black, Other, RBPs whose primary function is not RNA splicing or miRNA related.

Figure 4. Protein-binding sites in FLEXI RNAs. (*A*) Bar graph showing the number of detected FLEXIs in the cellular and plasma RNA datasets that have an experimentally identified RBP

binding site for the indicated RBP. Only RBPs that bind 30 or more different FLEXIs are shown; a bar graph for the complete set of detected FLEXIs is shown in Supplemental Fig. S7. (*B* and *C*) Scatter plots comparing the relative abundance of RBP-binding sites in the detected FLEXI RNAs with those in all annotated longer introns > 300 nt in GRCh38 (panel B) or all RBP-binding sites in the ENCODE 150 RBP eCLIP dataset with IDR plus the AGO1-4 and DICER PAR-CLIP dataset using GRCh38 as the reference sequence (panel C). RBPs whose relative abundance was ≥ 4% and significantly different between the compared groups ($p \leq 0.05$ calculated by Fisher's exact test) are indicated by the name of the RBP color coded by protein function as indicated in the keys in panels A and B.

Figure 5. UpSet plots identifying RBPs that bind FLEXI RNAs lacking annotated binding sites for core spliceosomal proteins. (*A*) and (*B*) AGO1-4 and DICER, respectively. (*C-I*) RBPs that have no known RNA splicing- or miRNA-related function. Each plot compares the FLEXI RNAs in the cellular and plasma RNA datasets that contained an annotated binding site for the RBP of interest in the CLIP-seq datasets (red) to those that contained annotated binding sites for any of five ubiquitous core spliceosomal proteins (AQR, BUD13, EFTUD2, PRPF8, and SF3B4) in those datasets (black). In each case, a substantial proportion (29-55%) of the FLEXI RNAs bound by the RBP of interest lacked an annotated binding site for any of the five core spliceosomal proteins. The inset in the top left UpSet plot shows the total number of different FLEXIs that contained an annotated binding site for each of the RBPs. Similar distinct classes of FLEXIs that bind the indicated RBP but lack annotated binding sites for the spliceosomal proteins were found for DDX55 (55%), IGF2BP1 (52%), FRX2 (47%), ZNF800 (33%), LARP4 (33%), RPS3 (33%), UCHL5 (32%), METAP2 (31%), LSM11 (30%), and GRWD1 (29%).

Figure 6. Heatmap of GO terms enriched in host genes of FLEXI RNAs containing binding sites for different RBPs. GO enrichment analysis was performed using DAVID bioinformatics tools. and clustering was performed based on the adjusted p-value for each enriched category using

Seaborn ClusterMap. The function, cellular localization, and protein motif information for the RBPs are summarized below using information from (Van Nostrand et al. 2020) supplemented by information from mammalian RNA granule and stress granule protein databases (Nunes et al. 2019; Youn et al. 2019), and AGO1-4 and DICER information from the UniProt database (The UniProt Consortium 2018). RBP are color coded by protein function: red, RNA splicing-related function; orange, miRNA-related functions blue, both an RNA splicing- and a miRNA-related function; black, RBPs whose primary function is not RNA splicing- or miRNA-related. *: RBPs that bind FLEXI RNAs with phastCons ≥ 0.99; §: three RBPs that bind FLEXIs with relatively low GC content including 41 of 43 FLEXIs that encode embedded snoRNAs; †: RBPs that bind a substantial proportion of the FLEXI RNAs (29-55%) that lacked annotated binding sites for any of the five most ubiquitous core spliceosomal proteins (AQR, BUD13, EFTUD2, PRPF8, and SF384).

Figure 7. FLEXI RNAs in breast cancer tumors and cell lines. (*A*) UpSet plots of FLEXI RNAs and FLEXI host genes detected at ≥ 0.01 RPM in unfragmented RNA preparations from matched cancer/healthy breast tissues from patients A (PR^+, ER^+, $HER2^-$) and B (PR unknown, ER^-, $HER2^-$) and breast cancer cell lines MDA-MB-231 and MCF7. Different FLEXI RNAs from the same host gene were aggregated into one entry for that gene. FLEXI RNAs and FLEXI host genes are listed below some sample groups in descending order of RPM, with the RPM indicated in parentheses at the bottom. (*B*) Scatter plots comparing $\log_2$ transformed RPM of FLEXI RNAs in unfragmented RNA preparations and all transcripts from FLEXI host genes in chemically fragmented RNA preparations from cancer and healthy breast tissues from patients A and B. FLEXI RNAs present at ≥ 0.05 RPM and detected in at least two replicate libraries from the cancer tissue but not in the matched healthy tissue indicated in red and listed to the right of the scatter plots. (*C*) GO enrichment analysis of genes encoding detected FLEXI RNAs in 50 hallmark gene sets (MSigDB) in cancer samples and combined patient A + B healthy tissue samples. Names of pathways significantly enriched ($p \leq 0.05$) in all or at least one cancer sample

are in red and orange, respectively. In panels A and B, oncogenes and FLEXI RNAs originating from oncogenes are denoted with an asterisk.

Figure 8. Oncogene FLEXI RNAs in breast cancer tumors and cell lines. (*A* and *B*) UpSet plots of upregulated (A) and downregulated (B) oncogene FLEXI RNAs in unfragmented RNA preparations from matched cancer/healthy breast tissues from patients A and B and breast cancer cell lines MDA-MB-231 and MCF7. FLEXIs originating from the *FASN* gene are highlighted in red. (*C* and *D*) UpSet plots of upregulated (C) and downregulated (D) tumor suppressor gene (TSG) FLEXI RNAs in the same unfragmented RNA preparations. Up and down regulated FLEXI RNAs were defined as those with an RPM-fold change ≥ 2. The most abundant oncogene or TSG FLEXI RNAs (up to a limit of 10) are listed below some sample groups, with the range of RPM values indicated in parentheses at the bottom. (*E*-*G*) Scatter plot comparing the relative abundance (percentage) of different RBP-binding sites in oncogene FLEXIs that are upregulated only in MCF7 cells, only in MDA-MB-231 cells, or in all four cancer samples compared to the abundance of all detected FLEXIs in the same sample or samples. For each pair of plots, the RBPs whose relative abundance is significantly different ($p \leq 0.05$ calculated by Fisher's exact test) are shown in red with names labeled.

Figure S1. TGIRT-seq of ribodepleted unfragmented cellular RNA. (*A*) Stacked bar graphs showing the percentage of reads in the combined datasets for the indicated samples in this study that mapped to different categories of annotated genomic features in the GRCh38 human genome reference sequence. Genomic features follow Ensembl GRCh38 Release 93 annotations. rRNA includes cellular and mitochondrial (Mt) rRNAs; protein coding includes protein-coding transcripts from both the nuclear and Mt genomes. (*B*) Stacked bar graphs showing the percentage of bases that mapped to different regions of the sense strand of protein-coding genes. CDS, coding sequences; intergenic, regions upstream or downstream of transcription start and stop sites annotated in RefSeq; intron, intronic regions; and UTR, 5'- or 3'-untranslated regions;

C, tumor tissue from breast cancer patients A or B; H, neighboring healthy tissue from the same patient.

Figure S2. Integrative Genomics Viewer (IGV) screenshots showing examples of sncRNAs detected in ribodepleted intact (non-chemically fragmented) cellular RNAs by TGIRT-seq. After mapping individual datasets to a customized set of sncRNA references sequences that included mature tRNAs with post-transcriptionally added 3' CCAs and to the Ensembl GRCh38 Release 93 human genome reference sequence, as described in Methods, individual datasets were combined and alignments for the indicated sncRNAs were displayed in IGV. Coverage at each position along the gene is shown in the top tract, and read alignments are shown below with reads down sampled to a maximum of 100 reads for display when necessary. Gray represents bases that match the reference base. Other colors indicate bases that do not match the reference base (red, green, blue, and brown indicate thymidine, adenosine, cytidine, and guanosine, respectively). Misincorporation at known sites of tRNA post-transcriptionally modified bases are highlighted in the alignments: m^1A58: 1-methyladenosine at position 58; I: inosine.

Figure S3. PCA, PCA-initialized t-SNE, and ZINB-WaVE analysis of FLEXI RNAs detected in different replicates of all ribodepleted intact cellular RNA datasets in this study (Supplemental Table S1). The plots show sample clustering based on all FLEXI RNAs detected at ≥ 1 read in these datasets. Different cell types are color-coded as indicated in the Figure, with each dot of the same color representing a replicate for that cell type.

Figure S4. Density plots showing the distribution of RNA fragment lengths for different subcategories of FLEXIs in each of the cellular RNA samples. % intron length was calculated from the read span of TGIRT-seq reads normalized for the length of each intron. FLEXIs and short introns with embedded snoRNA or scaRNA were removed prior to calculating the distributions to avoid interference from mature snoRNA reads. Separate plots are shown for

FLEXI RNAs containing annotated binding sites for AGO1-4, DICER, five core spliceosome proteins (AQR, BUD13, EFTUD2, PRPF8 and SF3B4), other annotated RBPs, FLEXIs without annotated RBP-binding sites in the searched datasets, and all other GRCh38-annotated short introns (≤ 300 nt). The different intron types are color coded as shown in the key in the top left plot.

Figure S5. IGV screenshots showing read alignments for FLEXI RNAs having non-GU-AG 5'- and 3'-splice sites. Dinucleotides at the 5'- and 3'-ends of the intron are indicated at the upper left with the number of introns in that category indicated in parentheses. Gene name and genomic coordinates of the FLEXI are shown at the top with the arrow below indicating the 5' to 3' orientation of the encoded RNA followed by tracts show the genomic sequence and gene annotations for different transcript isoforms (exons, thick bars; introns, thin lines). Read alignments for FLEXI RNAs are shown below the tracts and are color coded by sample type as indicated in the key at the upper right.

Figure S6. Scatter plots showing the relationship between the abundance of sncRNAs detected by TGIRT-seq and copy number per cell values for human sncRNAs reported in the literature. The abundance of sncRNAs detected by TGIRT-seq (RPM) and literature values for their copy number per cell (Tycowski et al. 2006) were $\log_{10}$ transformed and plotted. Pearson (r) and Spearman (r_s) correlation coefficients are shown in the upper left. Linear regression was modeled for each cell type and plotted as light blue line, with the 95% confidence interval of the linear regression plotted as blue dashed lines. Major spliceosomal snRNAs used in the linear regression were U1, U2, U4, U5, and U6; minor spliceosomal snRNAs were U11, U12, U4ATAC, and U6ATAC; and C/D box snoRNAs were SNORD3, SNORD13, SNORD14, SNORD22, and SNORD118 (Table S2) (Tycowski et al. 2006).

Figure S7. Bar graph showing the number of detected FLEXIs that have an experimentally identified RBP-binding site for the indicated RBP. FLEXIs are color coded by type as indicated in the key at the upper right.

Figure S8. GO term enrichment of randomly sampled FLEXI host and other genes. Random samples were taken from lists of all FLEXIs, FLEXIs without RBP-binding sites, all annotated genes, or all genes containing short introns (≤ 300 nt) in GRCh38. For each category, the number of included introns in each randomly selected list corresponded to the minimum, maximum, and quartile numbers of FLEXIs bound by different RBPs in Fig. 6 plus lists of 500 and 1,000 introns to better sample the full range of list sizes. Enrichment for the same GO terms as in Fig. 6 is shown as a heatmap of average p-values for three replicates of each randomly sampled list, with red corresponding to significant p-values ≤ 0.05 and blue indicating non-significance.

Figure S9. UpSet plots of FLEXI RNAs bound by AATF, DKC1, NOLC1. Each plot compares the FLEXI RNAs in the cellular RNA datasets that contained an annotated binding site for one of the above RBPs to those that contained an annotated binding site for any of five most ubiquitous core spliceosomal proteins (AQR, BUD13, EFTUD2, PRPF8, and SF384) grouped as one entry.

Figure S10. Density distribution plots comparing different characteristics of FLEXI RNAs containing a binding site for the indicated RBP (red) to those for all other detected FLEXIs (black). (*A*) Three RBPs in cluster III of Fig. 6 that bind FLEXI RNAs with relatively low GC content and above average phastCons scores. (*B*) RBPs in cluster IV whose binding sites are enriched in highly conserved FLEXIs and/or bind FLEXI RNAs with relatively low GC content. For each plot, the same number of FLEXI RNAs as those bound by the specific RBP of interest were randomly selected from other detected FLEXIs and used the calculate a density distribution for the same characteristic, and a p-value comparing the two density distributions was calculated

by Wilcoxon test. This process was repeated 100 times, and a false discovery rate (FDR) was calculated as the probability of p-value ≤ 0.05. FDRs ≤ 0.01 are highlighted in red.

Figure S11. UpSet plots of FLEXI RNAs detected at ≥ 1 read and their host genes in unfragmented RNA preparations from tumor and healthy breast tissues from patients A and B and breast cancer cell lines. Different FLEXI RNAs detected at ≥ 1 read from the same host gene were aggregated into one entry for that gene. FLEXI RNAs and FLEXI host genes are listed below some sample groups in descending order of RPM, with the RPM range of the detected FLEXIs indicated in parentheses at the bottom. Oncogenes and FLEXIs originating from oncogenes are indicated with an asterisk.

Table 1. Proteins with no known RNA splicing- or miRNA-related function that bind 30 or more different FLEXIs.

Symbol	Name	Function
AATF	Apoptosis Antagonizing Transcription Factor	Transcriptional cofactor with roles in cell proliferation, apoptosis, DNA damage response and general stress response through regulation of Rb, HDAC1, and p53 functions.
BCLAF1	BCL2-2-associated transcription factor 1	Transcriptional repressor; promotes apoptosis through interaction with BCL2; Upregulated in senescence, promotes p53 transcription in response to DNA damage.
DDX24	ATP-dependent RNA helicase DDX24	ATP-dependent RNA helicase and negative regulator of p53
DDX3X	ATP-dependent RNA helicase DDX3X	Multifunctional ATP-dependent RNA helicase with functions in cell cycle control, apoptosis, and innate immunity; critical role in stress granule assembly.
DDX55	ATP-dependent RNA helicase DDX55	Probable ATP-binding RNA helicase
DKC1	H/ACA ribonucleoprotein complex subunit DKC1	Catalytic subunit of H/ACA small nucleolar ribonucleoprotein (H/ACA snoRNP) complex; plays an active role in telomerase stabilization
FXR2	Fragile X mental retardation syndrome-related protein 2	RNA-binding protein
G3BP1	Ras GTPase-activating protein-binding protein 1	ATP- and Mg-dependent helicase that plays an essential role in innate immunity; Also functions in stress granule assembly and is associated with cellular senescence. Regulates Ras, TGF-β/Smad, Src/FAK and p53 signaling pathways.
GRWD1	Glutamate-rich WD repeat-containing protein 1	Histone binding-protein that regulates chromatin dynamics and minichromosome maintenance (MCM) loading at replication origins; negatively regulates p53.
IGF2BP1	Insulin-like growth factor 2 mRNA-binding protein 1	RNA-binding protein that recruits target transcripts to cytoplasmic

		protein-RNA complexes (mRNPs); Promotes cell cycle progression through regulation of E2F translation.
LARP4	La-related protein 4	RNA binding protein that binds to the polyA tract of mRNA molecules
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.
NOLC1	Nucleolar and coiled-body phosphoprotein 1	Nucleolar protein that plays a critical role in snoRNP assembly and acts as a regulator of RNA polymerase I by connecting RNA polymerase I with enzymes responsible for ribosomal processing and modification; Stabilizes telomeres by regulating TRF2 retention.
PABPC4	Polyadenylate-binding protein 4	Binds the polyA tail of mRNA
PABPN1	Polyadenylate-binding protein 2	Involved in the 3'-end formation of mRNA precursors (pre-mRNA) by the addition of a poly(A) tail; Regulated by ATM and plays a crucial role in DSB repair.
RPS3	40S ribosomal protein S3	Role in regulating transcription; implicated in regulating DNA damage response and apoptosis.
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.
UCHL5	Ubiquitin carboxyl-terminal hydrolase isozyme L5	Protease
XRN2	5'-3' exoribonuclease 2	May promote the termination of transcription by RNA polymerase II
YBX3	Y-Box-binding protein 3	Binds also to full-length mRNA and to short RNA sequences containing the

		consensus site 5'-UCCAUCA-3'.
ZNF622	Zinc finger protein 622	May behave as an activator of the bound transcription factor, MYBL2; positive regulator of apoptosis.
ZNF800	Zinc finger protein 800	May be involved in transcriptional regulation.

Supplemental Table S1. Summary of datasets

RNA origin	Raw reads ($x10^6$)	Trimmed reads ($x10^6$)	Mapped reads ($x10^6$)	Mapped to feature ($x10^6$)
HEK-293T*	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%)
UHRR‡	712.0	682.2 (95.8%)	666.3 (97.7%)	630.9 (94.7%)
K-562§	744.2	725.0 (97.4%)	713.8 (98.4%)	698.5 (97.9%)
Plasma¶	232.5	122.7 (91.3%)	71.1 (57.9%)	61.7 (87.2%)
MDA-MB-231#	226.1	211.3 (93.5%)	207.5 (98.2%)	202.2 (97.5%)
MCF7‖	757.8	703.7 (92.9%)	692.1 (98.4%)	673.7 (97.3%)
Patient A Healthy**	338.3	312.6 (92.4%)	305.1 (97.6%)	297.6 (97.5%)
Patient A Cancer**	295.8	275.0 (93.0%)	268.2 (97.5%)	256.5 (95.6%)
Patient B Healthy**	281.9	258.2 (91.6%)	251.6 (97.4%)	244.0 (97.0%)
Patient B Cancer**	549.7	492.4 (89.6%)	477.5 (97.0%)	461.6 (96.7%)
Patient A Healthy (Fragmented)††	55.7	52.7 (94.7%)	50.5 (95.8%)	33.2 (60.1%)
Patient A Cancer (Fragmented) ††	61.9	59.8 (96.7%)	57.0 (95.3%)	40.5 (68.8%)
Patient B Healthy (Fragmented) ††	39.6	35.1 (88.5%)	33.1 (94.3%)	22.0 (57.2%)
Patient B Cancer (Fragmented) ††	58.4	56.9 (97.5%)	54.1 (95.1%)	47.1 (85.5%)

* HEK-293T cell RNA, combined datasets from 8 replicates (HEK-rep1 to 8).
† HeLa S3 cell RNA, combined datasets from 10 replicates (HeLa-rep1 to 10).
‡ Universal human reference RNA, combined datasets from 8 replicates (UHRR-rep1 to 8).
§ K-562 cell RNA, combined datasets from 8 replicates (K-562-rep1 to 8).
¶ Commercial human plasma pooled plasma from healthy individuals. Fifteen combined datasets (SRA BioProject accession number PRJNA640428, samples DNaseI_1-12, ExoI_1-3) (Yao et al. 2020).
# MDA-MB-231 RNA, combined datasets from 2 replicates (MDA-rep1 and 2).
‖ MCF7 RNA, combined datasets from 8 replicates (MCF-rep1 to 8).
** Matched healthy and cancer breast tissue RNA from patients A and B purchased from Origene (Patient A: PR^+, ER^+, $HER2^-$; CR562524/CR543839); Patient B: PR unknown, ER^-, $HER2^-$; CR560540/CR532030). Combined datasets from 3 replicates for each sample type (rep1-3).

[††] Chemically fragmented RNAs from matched healthy/cancer tissues from patients A and B.

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

Name	RPM				Copies/cell*	Type
	K-562	HEK-293T	HeLa S3	UHRR		
U1	1,386	1,267	1,337	2,649	$1x10^6$	Major spliceosomal snRNA
U2	5,768	9,482	28,781	9,794	$5x10^5$	Major spliceosomal snRNA
U4	65	34	540	76	$2x10^5$	Major spliceosomal snRNA
U5	1,147	2,704	13,244	2,717	$2x10^5$	Major spliceosomal snRNA Minor spliceosomal snRNA
U6	974	1,767	923	1,356	$4x10^5$	Major spliceosomal snRNA
U7	3	4	49	3	$4x10^3$	U7
U11	2	3	11	8	$1x10^4$	Minor spliceosomal snRNA
U12	36	24	192	41	$5x10^3$	Minor spliceosomal snRNA
U4ATAC	37	21	1,292	32	$2x10^3$	Minor spliceosomal snRNA
U6ATAC	204	286	700	268	$2x10^3$	Minor spliceosomal snRNA
RN7SK	1,369	2,773	5,044	7,180	$2x10^5$	7SK
RN7SL	3,447	8,085	4,690	19,815	$5x10^5$	7SL
RPPH1	165	245	257	1,294	$2x10^5$	RNase P RNA component
SNORD3	7,637	12,699	23,195	38,533	$2x10^5$	C/D box snoRNA, U3
SNORD118	212	976	546	741	$4x10^4$	C/D box snoRNA, U8
SNORD13	435	366	817	1,221	$4x10^4$	C/D box snoRNA, U13
SNORD14	4,281	3,890	9,982	11,081	$1x10^4$	C/D box snoRNA, U14
SNORD22	362	241	748	948	$1x10^4$	C/D box snoRNA, U22
RMRP	670	1,777	2,211	3,279	$1x10^5$	MRP

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

Supplemental Table S3. Oligonucleotides used in this study 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/GATCGTCGGACTGTAGAACTCTGAACGTGTAG/3SpC3/. The R1R oligonucleotide was adenylated. as described in Materials and Methods.
Illumina multiplex PCR primer	5’-AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGACGATC-3’
Illumina index PCR primer	5’ CAAGCAGAAGACGGCATACGAGAT BARCODE* GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3’, where BARCODE* corresponds to the 6 nucleotide Illumina TruSeq barcode sequence.

References

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res* **37**: W202-W208.

Berezikov E, Chung W-J, Willis J, Cuppen E, Lai EC. 2007. Mammalian mirtron genes. *Mol Cell* **28**: 328-336.

Bronisz A, Rooj AK, Krawczyński K, Peruzzi P, Salińska E, Nakano I, Purow B, Chiocca EA, Godlewski J. 2020. The nuclear DICER–circular RNA complex drives the deregulation of the glioblastoma cell microRNAome. *Sci Adv* **6**: eabc0221.

Burset M, Seledtsov IA, Solovyev VV. 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res* **28**: 4364-4375.

Chapman KB, Boeke JD. 1991. Isolation and characterization of the gene encoding yeast debranching enzyme. *Cell* **65**: 483-492.

Chen H, Boutros PC. 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12**: 35.

Eiermann N, Haneke K, Sun Z, Stoecklin G, Ruggieri A. 2020. Dance with the devil: stress granules and signaling in antiviral responses. *Viruses* **12**.

Farrell MJ, Dobson AT, Feldman LT. 1991. Herpes simplex virus latency-associated transcript is a stable intron. *Proc Natl Acad Sci USA* **88**: 790-794.

Gao K, Masuda A, Matsuura T, Ohno K. 2008. Human branch point consensus sequence is yUnAy. *Nucleic Acids Res* **36**: 2257-2267.

Gao X, Hardwidge PR. 2011. Ribosomal protein s3: a multifunctional target of attaching/effacing bacterial pathogens. *Front Microbiol* **2**: 137-137.

Gardner EJ, Nizami ZF, Talbot CC, Gall JG. 2012. Stable intronic sequence RNA (sisRNA), a new class of noncoding RNA from the oocyte nucleus of *Xenopus tropicalis*. *Genes Dev* **26**: 2550-2559.

Gavish-Izakson M, Velpula BB, Elkon R, Prados-Carvajal R, Barnabas GD, Ugalde AP, Agami R, Geiger T, Huertas P, Ziv Y et al. 2018. Nuclear poly(A)-binding protein 1 is an ATM target and essential for DNA double-strand break repair. *Nucleic Acids Res* **46**: 730-747.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017-1018.

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**: 2847-2849.

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

Hansen TB, Venø MT, Jensen TI, Schaefer A, Damgaard CK, Kjems J. 2016. Argonaute-associated short introns are a novel class of gene regulators. *Nat Commun* **7**: 11538.

Hilliker A, Gao Z, Jankowsky E, Parker R. 2011. The DEAD-box protein Ded1 modulates translation by the formation and resolution of an eIF4F-mRNA complex. *Mol Cell* **43**: 962-972.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44-57.

Iezzi S, Fanciulli M. 2015. Discovering Che-1/AATF: a new attractive target for cancer therapy. *Front Genet* **6**.

Kaiser RWJ, Ignarski M, Van Nostrand EL, Frese CK, Jain M, Cukoski S, Heinen H, Schaechter M, Seufert L, Bunte K et al. 2019. A protein-RNA interaction atlas of the ribosome biogenesis factor AATF. *Sci Rep* **9**: 11071.

Katibah GE, Qin Y, Sidote DJ, Yao J, Lambowitz AM, Collins K. 2014. Broad and adaptable RNA structure recognition by the human interferon-induced tetratricopeptide repeat protein IFIT5. *Proc Natl Acad Sci USA* **111**: 12025–12030.

Kedersha NL, Gupta M, Li W, Miller I, Anderson P. 1999. RNA-binding proteins TIA-1 and TIAR link the phosphorylation of eIF-2 alpha to the assembly of mammalian stress granules. *J Cell Biol* **147**: 1431-1442.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907-915.

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

Kobak D, Berens P. 2019. The art of using t-SNE for single-cell transcriptomics. *Nat Commun* **10**: 5416.

Kufel J, Grzechnik P. 2019. Small nucleolar RNAs tell a different tale. *Trends Genet* **35**: 104-117.

Kulesza CA, Shenk T. 2006. Murine cytomegalovirus encodes a stable intron that facilitates persistent replication in the mouse. *Proc Natl Acad Sci USA* **103**: 18302-18307.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Lee S, Stevens SW. 2016. Spliceosomal intronogenesis. *Proc Natl Acad Sci USA* **113**: 6514-6519.

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

Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. 2015. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**: 417-425.

Liu Y, Sun J, Zhao M. 2017. ONGene: A literature-based database for human oncogenes. *J Genet Genomics* **44**: 119-121.

MacNeil DE, Lambert-Lanteigne P, Autexier C. 2019. N-terminal residues of human dyskerin are required for interactions with telomerase RNA that prevent RNA degradation. *Nucleic Acids Res* **47**: 5368-5380.

Martens-Uzunova ES, Hoogstrate Y, Kalsbeek A, Pigmans B, Vredenbregt-van den Berg M, Dits N, Nielsen SJ, Baker A, Visakorpi T, Bangma C et al. 2015. C/D-box snoRNA-derived RNA production is associated with malignant transformation and metastatic progression in prostate cancer. *Oncotarget* **6**: 17430-17444.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**: pp. 10-12.

Morgan JT, Fink GR, Bartel DP. 2019. Excised linear introns regulate growth in yeast. *Nature* **565**: 606-611.

Moss WN, Steitz JA. 2013. Genome-wide analyses of Epstein-Barr virus reveal conserved RNA structures and a novel stable intronic sequence RNA. *BMC Genomics* **14**: 543.

Müller S, Bley N, Busch B, Glaß M, Lederer M, Misiak C, Fuchs T, Wedler A, Haase J, Bertoldo JB et al. 2020. The oncofetal RNA-binding protein IGF2BP1 is a druggable, post-transcriptional super-enhancer of E2F-driven gene expression in cancer. *Nucleic Acids Res* **48**: 8576-8590.

Nunes C, Mestre I, Marcelo A, Koppenol R, Matos CA, Nóbrega C. 2019. MSGP: the first database of the protein components of the mammalian stress granules. *Database* **2019**.

Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC. 2007. The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. *Cell* **130**: 89-100.

Oliveira D, Prahm KP, Christensen IJ, Hansen A, Høgdall CK, Høgdall EV. 2021. Noncoding RNA (ncRNA) profile association with patient outcome in epithelial ovarian cancer cases. *Reprod Sci* **28**: 757-765.

Omer A, Barrera MC, Moran JL, Lian XJ, Di Marco S, Beausejour C, Gallouzi IE. 2020. G3BP1 controls the senescence-associated secretome and its impact on cancer progression. *Nat Commun* **11**: 4979.

Parenteau J, Maignon L, Berthoumieux M, Catala M, Gagnon V, Abou Elela S. 2019. Introns are mediators of cell response to starvation. *Nature* **565**: 612-617.

Pek JW, Osman I, Tay ML-I, Zheng RT. 2015. Stable intronic sequence RNAs have possible regulatory roles in *Drosophila melanogaster*. *J Cell Biol* **211**: 243-251.

Qin Y, Yao J, Wu DC, Nottingham RM, Mohr S, Hunicke-Smith S, Lambowitz AM. 2016. High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. *RNA* **22**: 111–128.

Quinlan AR. 2014. BEDTools: the swiss-army tool for genome feature analysis. *Curr Protoc Bioinformatics* **47**: 11.12.11-34.

Risso D, Perraudeau F, Gribkova S, Dudoit S, Vert J-P. 2018. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun* **9**: 284.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24-26.

Ruby JG, Jan CH, Bartel DP. 2007. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**: 83-86.

Rybak-Wolf A, Jens M, Murakawa Y, Herzog M, Landthaler M, Rajewsky N. 2014. A variety of Dicer substrates in human and *C. elegans*. *Cell* **159**: 1153-1167.

Saini H, Bicknell AA, Eddy SR, Moore MJ. 2019. Free circular introns with an unusual branchpoint in neuronal projections. *eLife* **8**: e47809.

Schröder M. 2010. Human DEAD-box protein 3 has multiple functions in gene regulation and cell cycle control and is a prime target for viral manipulation. *Biochem Pharmacol* **79**: 297-306.

Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. 2006. Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res* **34**: 3955-3967.

Shurtleff MJ, Yao J, Qin Y, Nottingham RM, Temoche-Diaz MM, Schekman R, Lambowitz AM. 2017. Broad role for YBX1 in defining the small noncoding RNA composition of exosomes. *Proc Natl Acad Sci USA* **114**: E8987-E8995.

Simmons MP, Bachy C, Sudek S, van Baren MJ, Sudek L, Ares M, Jr, Worden AZ. 2015. Intron invasions trace algal speciation and reveal nearly identical arctic and antarctic micromonas populations. *Mol Biol Evol* **32**: 2219-2235.

Somasekharan SP, El-Naggar A, Leprivier G, Cheng H, Hajee S, Grunewald TGP, Zhang F, Ng T, Delattre O, Evdokimova V et al. 2015. YB-1 regulates stress granule formation and tumor progression by translationally activating G3BP1. *J Cell Biol* **208**: 913-929.

Sugimoto N, Maehara K, Yoshida K, Yasukouchi S, Osano S, Watanabe S, Aizawa M, Yugawa T, Kiyono T, Kurumizaka H et al. 2015. Cdt1-binding protein GRWD1 is a novel histone-binding protein that facilitates MCM loading through its influence on chromatin architecture. *Nucleic Acids Res* **43**: 5898-5911.

Talhouarne GJS, Gall JG. 2018. Lariat intronic RNAs in the cytoplasm of vertebrate cells. *Proc Natl Acad Sci USA* **115**: E7970-E7977.

The UniProt Consortium. 2018. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**: D506-D515.

Tycowski KT, Kolev NG, Conrad NK, Fok V, Steitz JA. 2006. The ever-growing world of small nuclear ribonucleoproteins. In *The RNA World, Third Edition*, (ed. RF Gesteland, et al.), pp. 327-368. Cold Spring Harbor Laboratory Press, NY.

van der Burgt A, Severing E, de Wit Pierre JGM, Collemare J. 2012. Birth of new spliceosomal introns in fungi by multiplication of introner-like elements. *Curr Biol* **22**: 1260-1265.

Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, Surka C, Elkins K et al. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**: 508-514.

Van Nostrand EL, Pratt GA, Yee BA, Wheeler EC, Blue SM, Mueller J, Park SS, Garcia KE, Gelboin-Burkhart C, Nguyen TB et al. 2020. Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biol* **21**: 90.

Vohhodina J, Barros EM, Savage AL, Liberante FG, Manti L, Bankhead P, Cosgrove N, Madden AF, Harkin DP, Savage KI. 2017. The RNA processing factors THRAP3 and BCLAF1 promote the DNA damage response through selective mRNA splicing and nuclear export. *Nucleic Acids Res* **45**: 12816-12833.

Wen J, Ladewig E, Shenker S, Mohammed J, Lai EC. 2015. Analysis of nearly one thousand mammalian mirtrons reveals novel features of Dicer substrates. *PLoS Computat Biol* **11**: e1004441.

Wilkinson ME, Charenton C, Nagai K. 2020. RNA splicing by the spliceosome. *Annu Rev Biochem* **89**: 359-388.

Xu H, Yao J, Wu DC, Lambowitz AM. 2019. Improved TGIRT-seq methods for comprehensive transcriptome profiling with decreased adapter dimer formation and bias correction. *Sci Rep* **9**: 7953.

Yao J, Wu DC, Nottingham RM, Lambowitz AM. 2020. Identification of protein-protected mRNA fragments and structured excised intron RNAs in human plasma by TGIRT-seq peak calling. *eLife* **9**: e60743.

Youn J-Y, Dyakov BJA, Zhang J, Knight JDR, Vernon RM, Forman-Kay JD, Gingras A-C. 2019. Properties of stress granule and P-body proteomes. *Mol Cell* **76**: 286-294.

Yuan F, Li G, Tong T. 2017. Nucleolar and coiled-body phosphoprotein 1 (NOLC1) regulates the nucleolar retention of TRF2. *Cell Death Discov* **3**: 17043.

Zhang C-H, Wang J-X, Cai M-L, Shao R, Liu H, Zhao W-L. 2019. The roles and mechanisms of G3BP1 in tumour promotion. *J Drug Target* **27**: 300-305.

Zhang Y, Zhang X-O, Chen T, Xiang J-F, Yin Q-F, Xing Y-H, Zhu S, Yang L, Chen L-L. 2013. Circular intronic long noncoding RNAs. *Mol Cell* **51**: 792-806.

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

Fig. 1

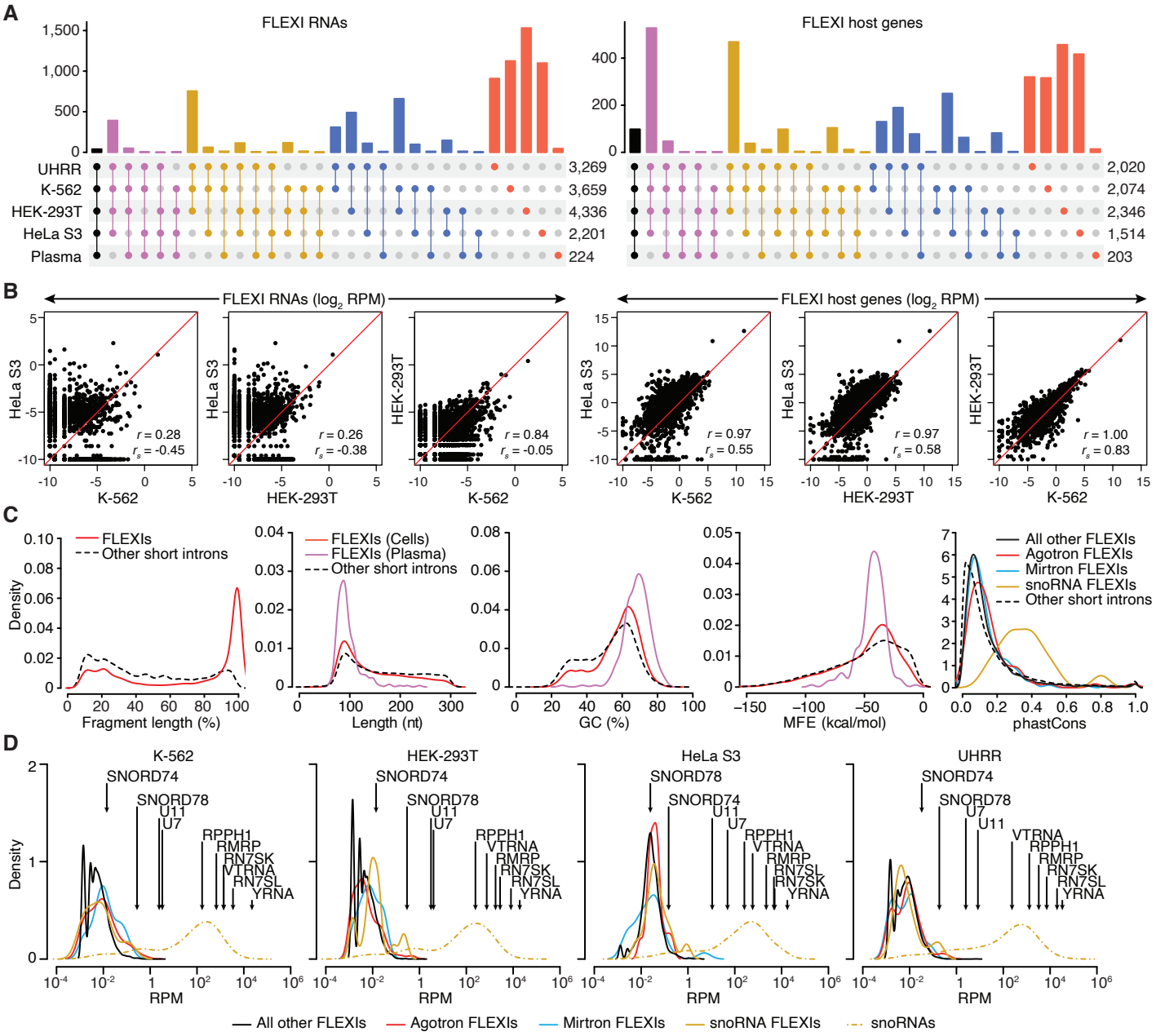

A
FLEXI RNAs
FLEXI host genes
UHRR
K-562
HEK-293T
HeLa S3
Plasma
3,269
3,659
4,336
2,201
224
2,020
2,074
2,346
1,514
203
B
FLEXI RNAs (log2 RPM)
FLEXI host genes (log2 RPM)
r = 0.28
rs = -0.45
r = 0.26
rs = -0.38
r = 0.84
rs = -0.05
r = 0.97
rs = 0.55
r = 0.97
rs = 0.58
r = 1.00
rs = 0.83
C
FLEXIs
Other short introns
FLEXIs (Cells)
FLEXIs (Plasma)
All other FLEXIs
Agotron FLEXIs
Mirtron FLEXIs
snoRNA FLEXIs
Density
Fragment length (%)
Length (nt)
GC (%)
MFE (kcal/mol)
phastCons
D
K-562
HEK-293T
HeLa S3
UHRR
SNORD74
SNORD78
U11
U7
RPPH1
RMRP
RN7SK
VTRNA
RN7SL
YRNA
RPM
All other FLEXIs
Agotron FLEXIs
Mirtron FLEXIs
snoRNA FLEXIs
snoRNAs

Fig. 2

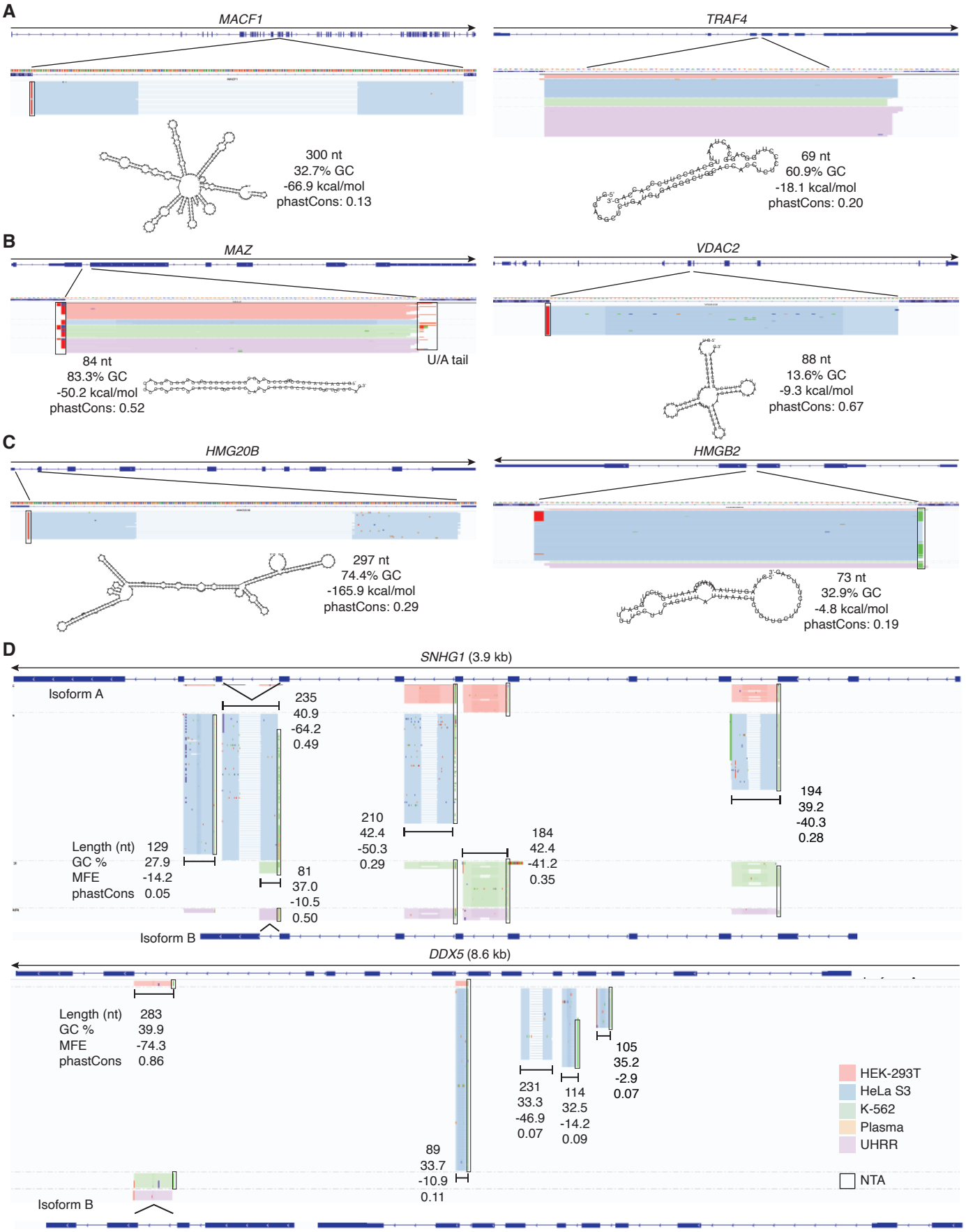
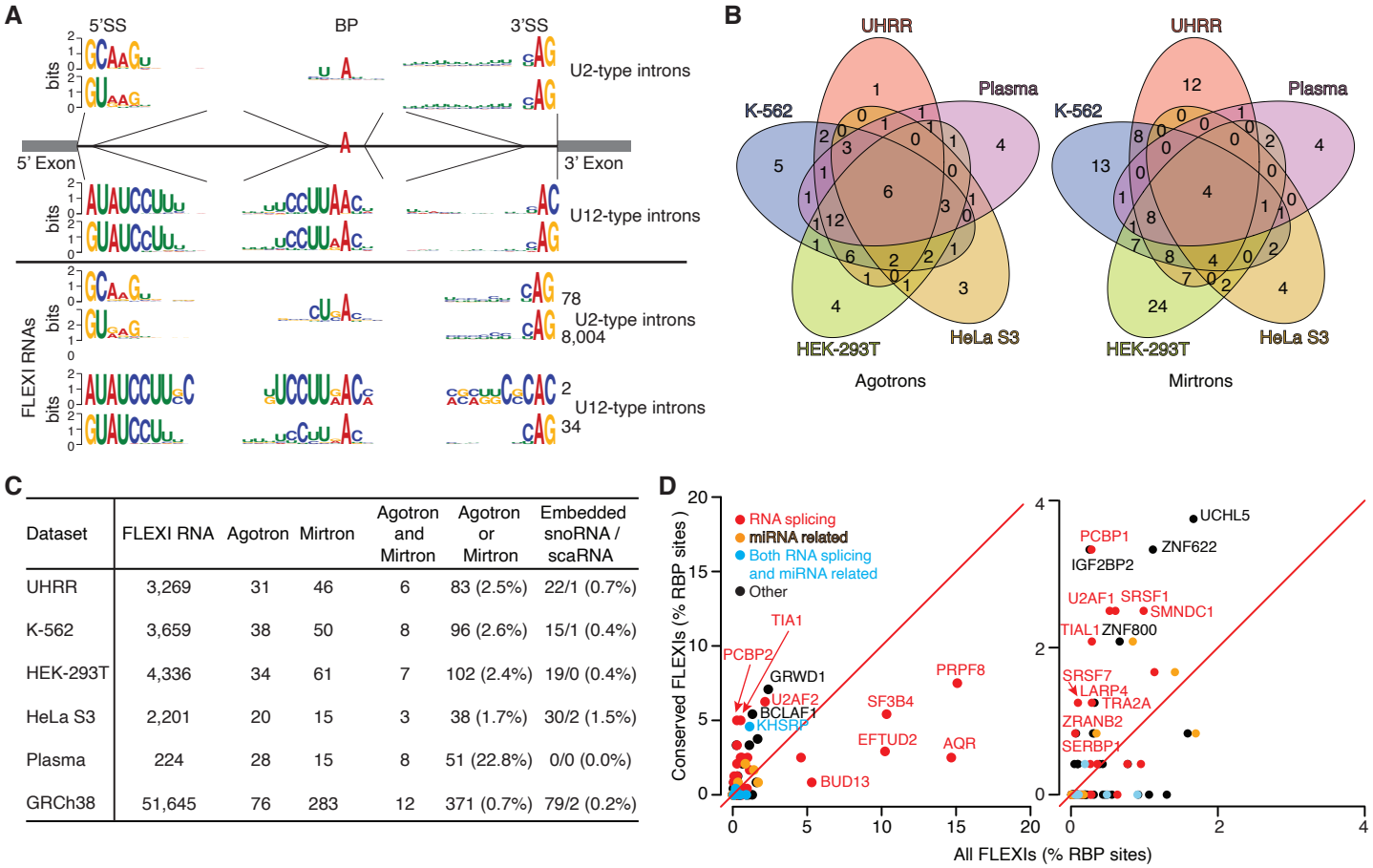

A
MACF1
300 nt
32.7% GC
-66.9 kcal/mol
phastCons: 0.13
TRAF4
69 nt
60.9% GC
-18.1 kcal/mol
phastCons: 0.20
B
MAZ
84 nt
83.3% GC
-50.2 kcal/mol
phastCons: 0.52
U/A tail
VDAC2
88 nt
13.6% GC
-9.3 kcal/mol
phastCons: 0.67
C
HMG20B
297 nt
74.4% GC
-165.9 kcal/mol
phastCons: 0.29
HMGB2
73 nt
32.9% GC
-4.8 kcal/mol
phastCons: 0.19
D
SNHG1 (3.9 kb)
Isoform A
235
40.9
-64.2
0.49
Length (nt) 129
GC % 27.9
MFE -14.2
phastCons 0.05
81
37.0
-10.5
0.50
210
42.4
-50.3
0.29
184
42.4
-41.2
0.35
194
39.2
-40.3
0.28
Isoform B
DDX5 (8.6 kb)
Length (nt) 283
GC % 39.9
MFE -74.3
phastCons 0.86
89
33.7
-10.9
0.11
231
33.3
-46.9
0.07
114
32.5
-14.2
0.09
105
35.2
-2.9
0.07
HEK-293T
HeLa S3
K-562
Plasma
UHRR
NTA
Isoform B

Fig. 3



Dataset	FLEXI RNA	Agotron	Mirtron	Agotron and Mirtron	Agotron or Mirtron	Embedded snoRNA / scaRNA
UHRR	3,269	31	46	6	83 (2.5%)	22/1 (0.7%)
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 (1.5%)
Plasma	224	28	15	8	51 (22.8%)	0/0 (0.0%)
GRCh38	51,645	76	283	12	371 (0.7%)	79/2 (0.2%)

Fig. 4

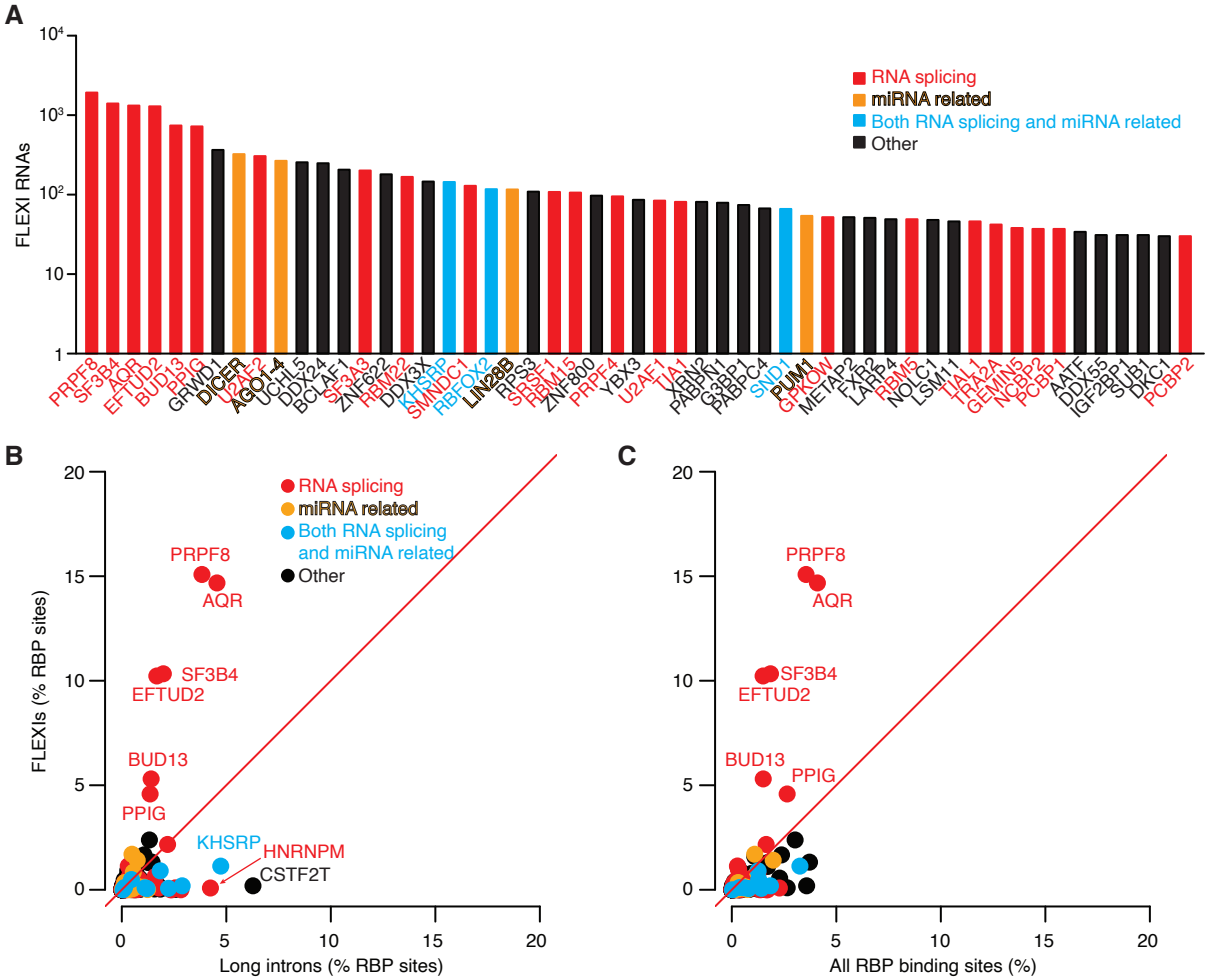

A
FLEXI RNAs
RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other
PRPF8 SF3B4 AQR EFTUD2 BUD13 PPIG GRWD1 DICER U2AF2 AGO1-4 UCHL5 DDX24 BCLAF1 SF3A3 ZNF622 RBM22 DDX3X KHSRP SMNDC1 RBFOX2 LIN28B RPS3 SRSF1 RBM15 ZNF800 PRPF4 YBX3 U2AF1 TIA1 XRN2 PABPN1 G3BP1 PABPC4 SND1 PUM1 GPKOW METAP2 FXR2 LARP4 RBM5 NOLC1 LSM11 UAL1 TRA2A GEMIN5 NCBP2 PCBP1 AATF DDX55 IGF2BP1 SUB1 DKC1 PCBP2
B
FLEXIs (% RBP sites)
Long introns (% RBP sites)
PRPF8
AQR
SF3B4
EFTUD2
BUD13
PPIG
KHSRP
HNRNPM
CSTF2T
C
All RBP binding sites (%)
PRPF8
AQR
SF3B4
EFTUD2
BUD13
PPIG

Fig. 5

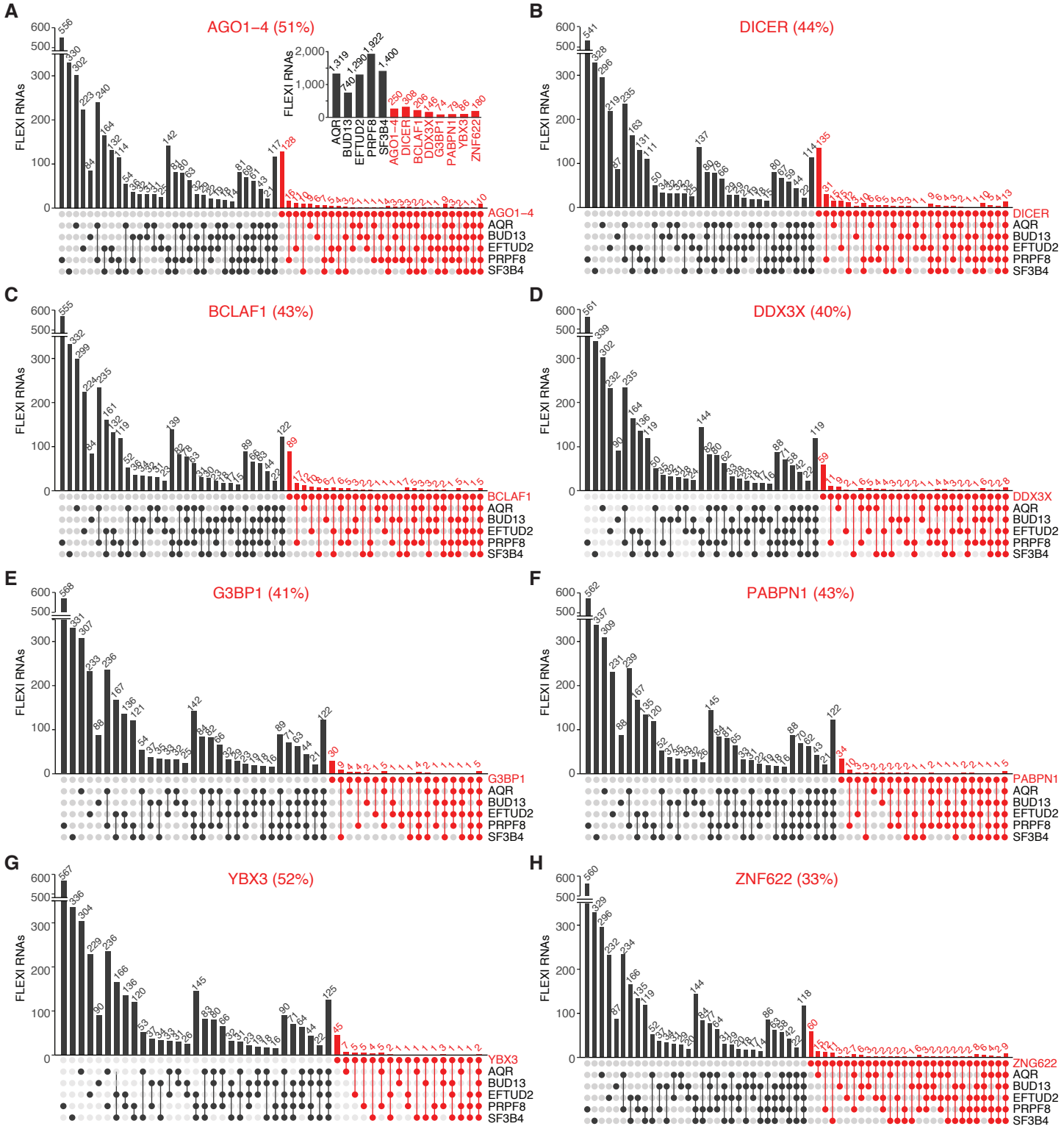

A
AGO1–4 (51%)
B
DICER (44%)
C
BCLAF1 (43%)
D
DDX3X (40%)
E
G3BP1 (41%)
F
PABPN1 (43%)
G
YBX3 (52%)
H
ZNF622 (33%)
FLEXI RNAs
AQR
BUD13
EFTUD2
PRPF8
SF3B4

Fig. 6

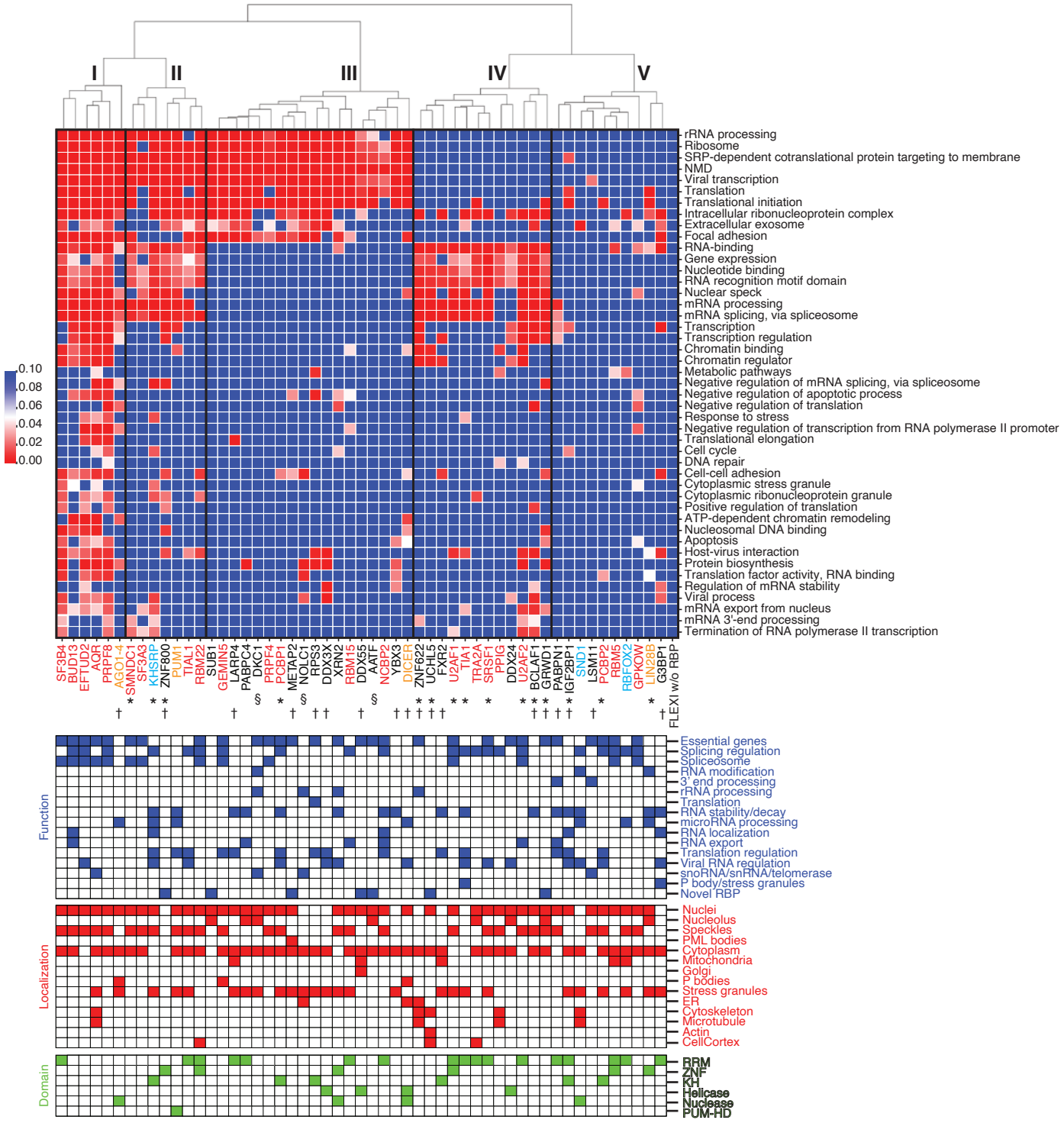

I
II
III
IV
V
0.10
0.08
0.06
0.04
0.02
0.00
rRNA processing
Ribosome
SRP-dependent cotranslational protein targeting to membrane
NMD
Viral transcription
Translation
Translational initiation
Intracellular ribonucleoprotein complex
Extracellular exosome
Focal adhesion
RNA-binding
Gene expression
Nucleotide binding
RNA recognition motif domain
Nuclear speck
mRNA processing
mRNA splicing, via spliceosome
Transcription
Transcription regulation
Chromatin binding
Chromatin regulator
Metabolic pathways
Negative regulation of mRNA splicing, via spliceosome
Negative regulation of apoptotic process
Negative regulation of translation
Response to stress
Negative regulation of transcription from RNA polymerase II promoter
Translational elongation
Cell cycle
DNA repair
Cell-cell adhesion
Cytoplasmic stress granule
Cytoplasmic ribonucleoprotein granule
Positive regulation of translation
ATP-dependent chromatin remodeling
Nucleosomal DNA binding
Apoptosis
Host-virus interaction
Protein biosynthesis
Translation factor activity, RNA binding
Regulation of mRNA stability
Viral process
mRNA export from nucleus
mRNA 3'-end processing
Termination of RNA polymerase II transcription
SF3B4
BUD13
EFTUD2
AQR
PRPF8
AGO1-4
SMNDC1
SF3A3
KHSRP
ZNF800
PUM1
TIAL1
RBM22
SUB1
GEMIN5
LARP4
PABPC4
DKC1
PRPF4
PCBP1
METAP2
NOLC1
RPS3
DDX3X
XRN2
RBM15
DDX55
AATF
NCBP2
YBX3
DICER
ZNF622
UCHL5
FXR2
U2AF1
TIA1
TRA2A
SRSF1
PPIG
DDX24
U2AF2
BCLAF1
GRWD1
PABPN1
IGF2BP1
SND1
LSM11
PCBP2
RBM5
RBFOX2
GPKOW
LIN28B
G3BP1
FLEXI w/o RBP
Function
Essential genes
Splicing regulation
Spliceosome
RNA modification
3' end processing
rRNA processing
Translation
RNA stability/decay
microRNA processing
RNA localization
RNA export
Translation regulation
Viral RNA regulation
snoRNA/snRNA/telomerase
P body/stress granules
Novel RBP
Localization
Nuclei
Nucleolus
Speckles
PML bodies
Cytoplasm
Mitochondria
Golgi
P bodies
Stress granules
ER
Cytoskeleton
Microtubule
Actin
CellCortex
Domain
RRM
ZNF
KH
Helicase
Nuclease
PUM-HD

Fig. 7

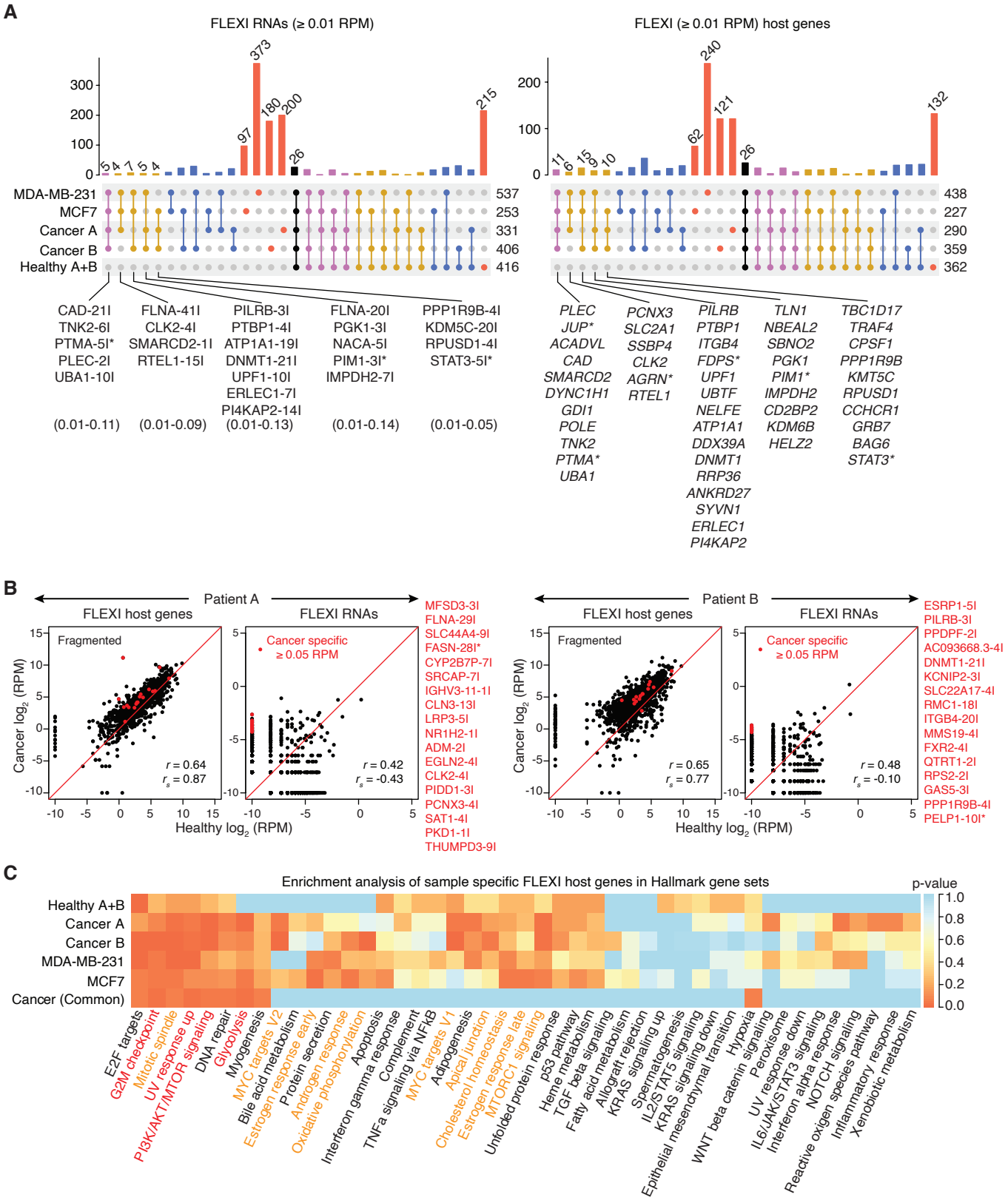

A
FLEXI RNAs (≥ 0.01 RPM)
FLEXI (≥ 0.01 RPM) host genes
MDA-MB-231
MCF7
Cancer A
Cancer B
Healthy A+B
537
253
331
406
416
438
227
290
359
362
CAD-21I
TNK2-6I
PTMA-5I*
PLEC-2I
UBA1-10I
(0.01-0.11)
FLNA-41I
CLK2-4I
SMARCD2-1I
RTEL1-15I
(0.01-0.09)
PILRB-3I
PTBP1-4I
ATP1A1-19I
DNMT1-21I
UPF1-10I
ERLEC1-7I
PI4KAP2-14I
(0.01-0.13)
FLNA-20I
PGK1-3I
NACA-5I
PIM1-3I*
IMPDH2-7I
(0.01-0.14)
PPP1R9B-4I
KDM5C-20I
RPUSD1-4I
STAT3-5I*
(0.01-0.05)
PLEC
JUP*
ACADVL
CAD
SMARCD2
DYNC1H1
GDI1
POLE
TNK2
PTMA*
UBA1
PCNX3
SLC2A1
SSBP4
CLK2
AGRN*
RTEL1
PILRB
PTBP1
ITGB4
FDPS*
UPF1
UBTF
NELFE
ATP1A1
DDX39A
DNMT1
RRP36
ANKRD27
SYVN1
ERLEC1
PI4KAP2
TLN1
NBEAL2
SBNO2
PGK1
PIM1*
IMPDH2
CD2BP2
KDM6B
HELZ2
TBC1D17
TRAF4
CPSF1
PPP1R9B
KMT5C
RPUSD1
CCHCR1
GRB7
BAG6
STAT3*
B
Patient A
FLEXI host genes
FLEXI RNAs
Fragmented
Cancer specific
≥ 0.05 RPM
r = 0.64
r_s = 0.87
r = 0.42
r_s = -0.43
Cancer log2 (RPM)
Healthy log2 (RPM)
MFSD3-3I
FLNA-29I
SLC44A4-9I
FASN-28I*
CYP2B7P-7I
SRCAP-7I
IGHV3-11-1I
CLN3-13I
LRP3-5I
NR1H2-1I
ADM-2I
EGLN2-4I
CLK2-4I
PIDD1-3I
PCNX3-4I
SAT1-4I
PKD1-1I
THUMPD3-9I
Patient B
FLEXI host genes
FLEXI RNAs
Fragmented
Cancer specific
≥ 0.05 RPM
r = 0.65
r_s = 0.77
r = 0.48
r_s = -0.10
Cancer log2 (RPM)
Healthy log2 (RPM)
ESRP1-5I
PILRB-3I
PPDPF-2I
AC093668.3-4I
DNMT1-21I
KCNIP2-3I
SLC22A17-4I
RMC1-18I
ITGB4-20I
MMS19-4I
FXR2-4I
QTRT1-2I
RPS2-2I
GAS5-3I
PPP1R9B-4I
PELP1-10I*
C
Enrichment analysis of sample specific FLEXI host genes in Hallmark gene sets
p-value
Healthy A+B
Cancer A
Cancer B
MDA-MB-231
MCF7
Cancer (Common)
E2F targets
G2M checkpoint
Mitotic spindle
UV response up
PI3K/AKT/MTOR signaling
DNA repair
Glycolysis
Myogenesis
MYC targets V2
Bile acid metabolism
Estrogen response early
Protein secretion
Androgen response
Oxidative phosphorylation
Apoptosis
Interferon gamma response
Complement
TNFa signaling via NFkB
MYC targets V1
Adipogenesis
Apical junction
Cholesterol homeostasis
Estrogen response late
MTORC1 signaling
Unfolded protein response
p53 pathway
Heme metabolism
TGF beta signaling
Fatty acid metabolism
Allograft rejection
KRAS signaling up
Spermatogenesis
IL2/STAT5 signaling
KRAS signaling down
Epithelial mesenchymal transition
Hypoxia
WNT beta catenin signaling
Peroxisome
UV response down
IL6/JAK/STAT3 signaling
Interferon alpha response
NOTCH signaling
Reactive oxigen species pathway
Inflammatory response
Xenobiotic metabolism

Fig. 8

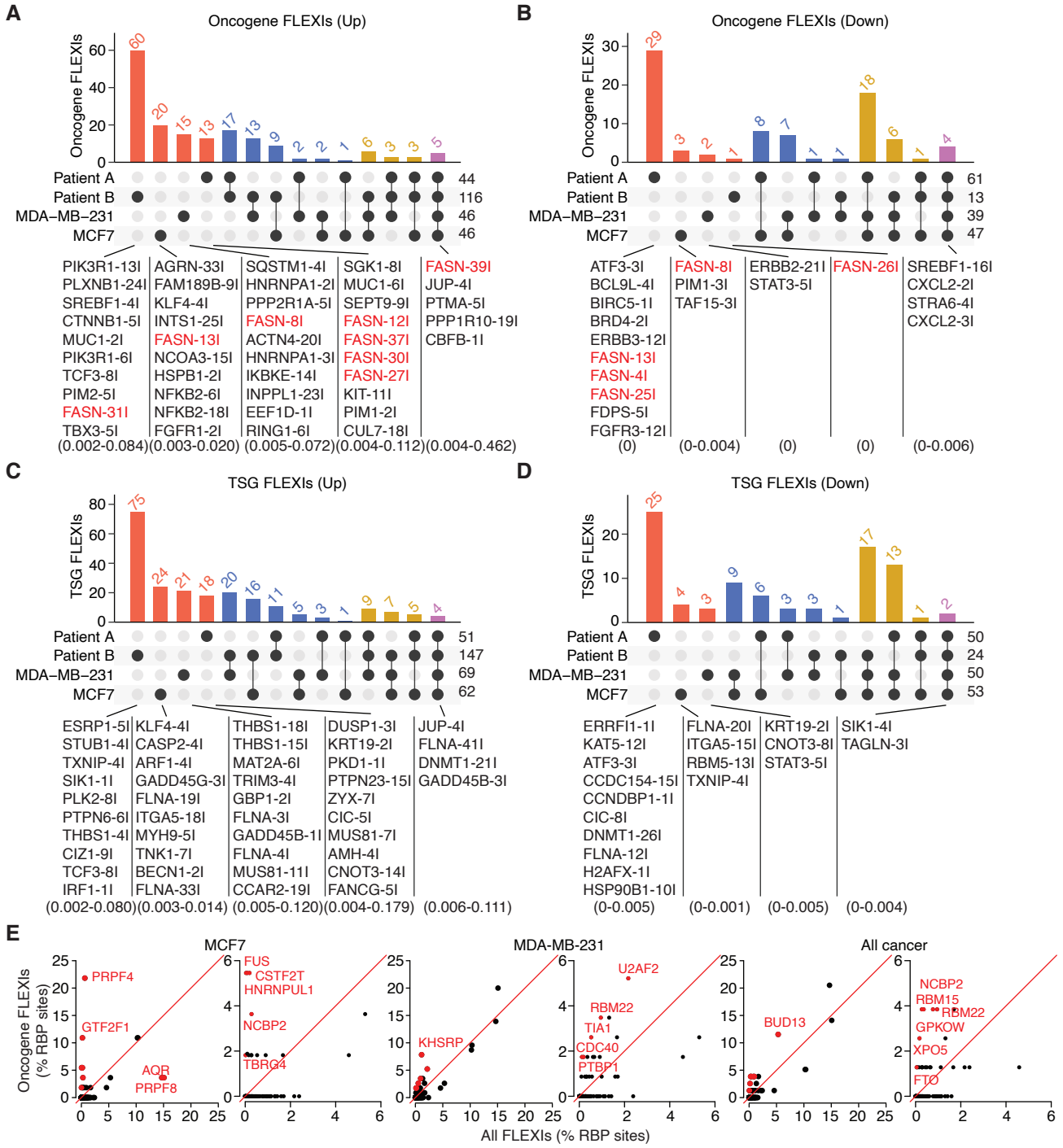

A
Oncogene FLEXIs (Up)
Oncogene FLEXIs
60 20 15 13 17 13 9 2 2 1 6 3 3 5
Patient A 44
Patient B 116
MDA–MB–231 46
MCF7 46
PIK3R1-13I PLXNB1-24I SREBF1-4I CTNNB1-5I MUC1-2I PIK3R1-6I TCF3-8I PIM2-5I FASN-31I TBX3-5I
AGRN-33I FAM189B-9I KLF4-4I INTS1-25I FASN-13I NCOA3-15I HSPB1-2I NFKB2-6I NFKB2-18I FGFR1-2I
SQSTM1-4I HNRNPA1-2I PPP2R1A-5I FASN-8I ACTN4-20I HNRNPA1-3I IKBKE-14I INPPL1-23I EEF1D-1I RING1-6I
SGK1-8I MUC1-6I SEPT9-9I FASN-12I FASN-37I FASN-30I FASN-27I KIT-11I PIM1-2I CUL7-18I
FASN-39I JUP-4I PTMA-5I PPP1R10-19I CBFB-1I
(0.002-0.084) (0.003-0.020) (0.005-0.072) (0.004-0.112) (0.004-0.462)
B
Oncogene FLEXIs (Down)
Oncogene FLEXIs
29 3 2 1 8 7 1 1 18 6 1 4
Patient A 61
Patient B 13
MDA–MB–231 39
MCF7 47
ATF3-3I BCL9L-4I BIRC5-1I BRD4-2I ERBB3-12I FASN-13I FASN-4I FASN-25I FDPS-5I FGFR3-12I
FASN-8I PIM1-3I TAF15-3I
ERBB2-21I STAT3-5I
FASN-26I
SREBF1-16I CXCL2-2I STRA6-4I CXCL2-3I
(0) (0-0.004) (0) (0) (0-0.006)
C
TSG FLEXIs (Up)
TSG FLEXIs
75 24 21 18 20 16 11 5 3 1 9 7 5 4
Patient A 51
Patient B 147
MDA–MB–231 69
MCF7 62
ESRP1-5I STUB1-4I TXNIP-4I SIK1-1I PLK2-8I PTPN6-6I THBS1-4I CIZ1-9I TCF3-8I IRF1-1I
KLF4-4I CASP2-4I ARF1-4I GADD45G-3I FLNA-19I ITGA5-18I MYH9-5I TNK1-7I BECN1-2I FLNA-33I
THBS1-18I THBS1-15I MAT2A-6I TRIM3-4I GBP1-2I FLNA-3I GADD45B-1I FLNA-4I MUS81-11I CCAR2-19I
DUSP1-3I KRT19-2I PKD1-1I PTPN23-15I ZYX-7I CIC-5I MUS81-7I AMH-4I CNOT3-14I FANCG-5I
JUP-4I FLNA-41I DNMT1-21I GADD45B-3I
(0.002-0.080) (0.003-0.014) (0.005-0.120) (0.004-0.179) (0.006-0.111)
D
TSG FLEXIs (Down)
TSG FLEXIs
25 4 3 9 6 3 3 1 17 13 1 2
Patient A 50
Patient B 24
MDA–MB–231 50
MCF7 53
ERRFI1-1I KAT5-12I ATF3-3I CCDC154-15I CCNDBP1-1I CIC-8I DNMT1-26I FLNA-12I H2AFX-1I HSP90B1-10I
FLNA-20I ITGA5-15I RBM5-13I TXNIP-4I
KRT19-2I CNOT3-8I STAT3-5I
SIK1-4I TAGLN-3I
(0-0.005) (0-0.001) (0-0.005) (0-0.004)
E
MCF7
MDA-MB-231
All cancer
Oncogene FLEXIs (% RBP sites)
All FLEXIs (% RBP sites)
PRPF4
GTF2F1
AQR
PRPF8
FUS
CSTF2T
HNRNPUL1
NCBP2
TBRG4
KHSRP
U2AF2
RBM22
TIA1
CDC40
PTBP1
BUD13
NCBP2
RBM15
RBM22
GPKOW
XPO5
FTO

Fig. S1

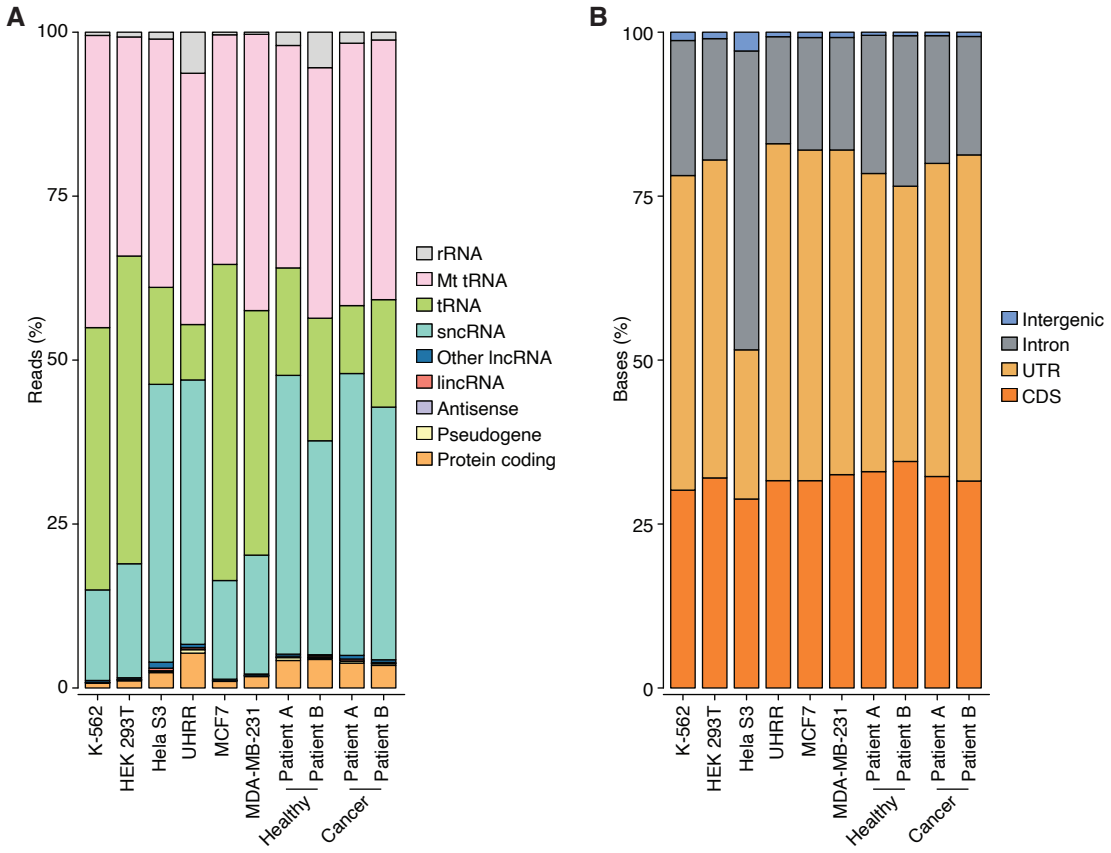

A
100
75
50
25
0
Reads (%)
rRNA
Mt tRNA
tRNA
sncRNA
Other lncRNA
lincRNA
Antisense
Pseudogene
Protein coding
K-562
HEK 293T
Hela S3
UHRR
MCF7
MDA-MB-231
Patient A
Patient B
Patient A
Patient B
Healthy
Cancer
B
100
75
50
25
0
Bases (%)
Intergenic
Intron
UTR
CDS
K-562
HEK 293T
Hela S3
UHRR
MCF7
MDA-MB-231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

Fig. S2

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

Fig. S3

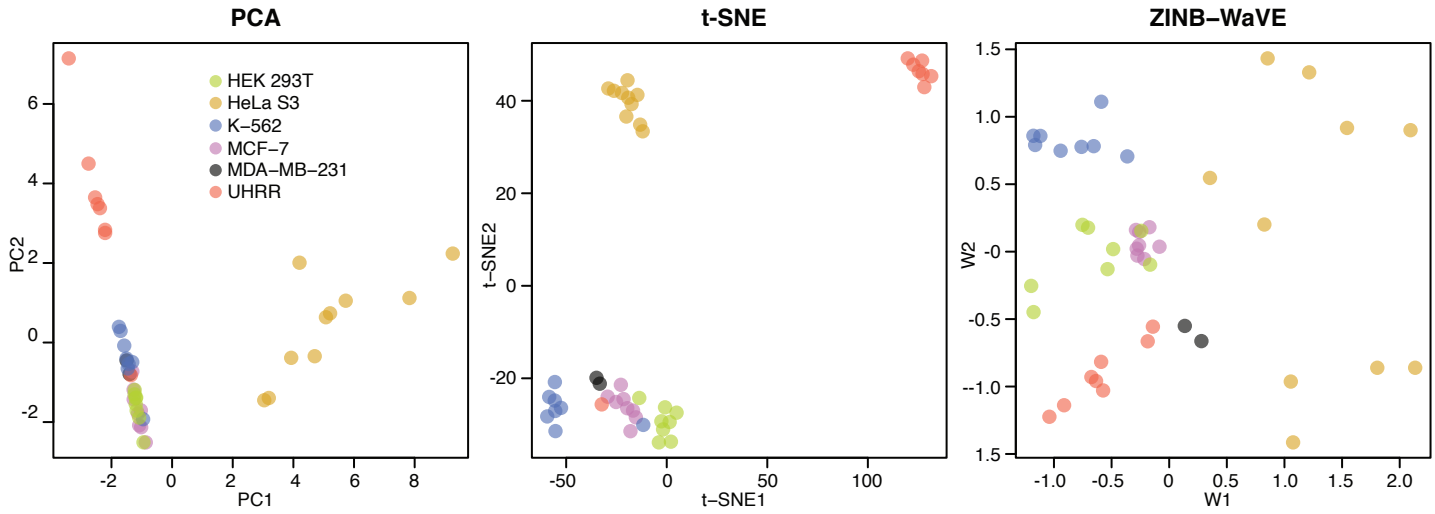

PCA
t-SNE
ZINB−WaVE
HEK 293T
HeLa S3
K−562
MCF−7
MDA−MB−231
UHRR
PC1
PC2
t−SNE1
t−SNE2
W1
W2

Fig. S4

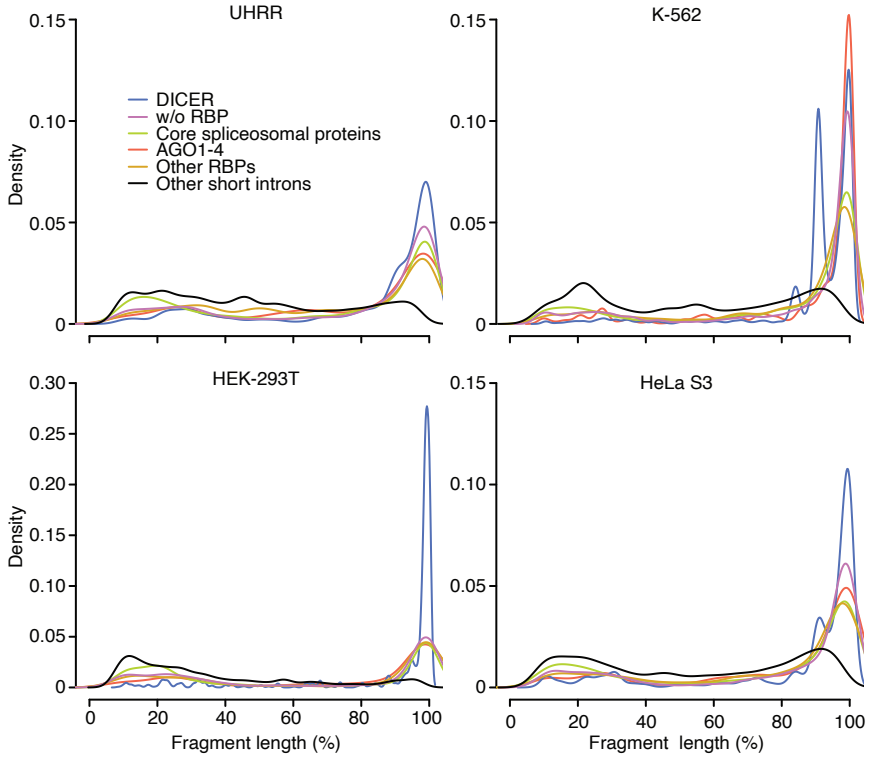

UHRR
K-562
HEK-293T
HeLa S3
DICER
w/o RBP
Core spliceosomal proteins
AGO1-4
Other RBPs
Other short introns
Density
Fragment length (%)
0.30
0.25
0.20
0.15
0.10
0.05
0
20
40
60
80
100

Fig. S5

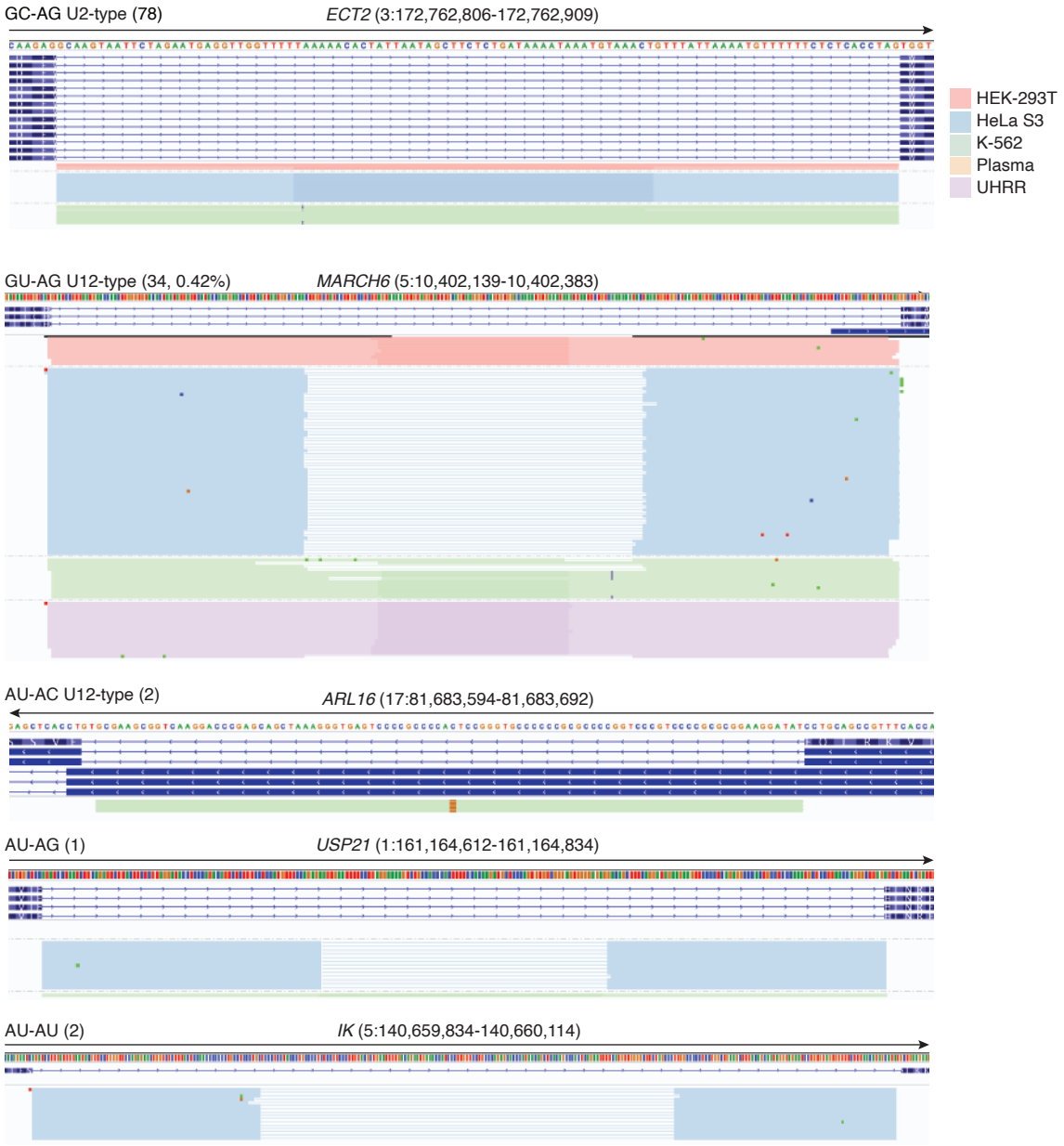

GC-AG U2-type (78)
ECT2 (3:172,762,806-172,762,909)
HEK-293T
HeLa S3
K-562
Plasma
UHRR
GU-AG U12-type (34, 0.42%)
MARCH6 (5:10,402,139-10,402,383)
AU-AC U12-type (2)
ARL16 (17:81,683,594-81,683,692)
AU-AG (1)
USP21 (1:161,164,612-161,164,834)
AU-AU (2)
IK (5:140,659,834-140,660,114)

Fig. S6

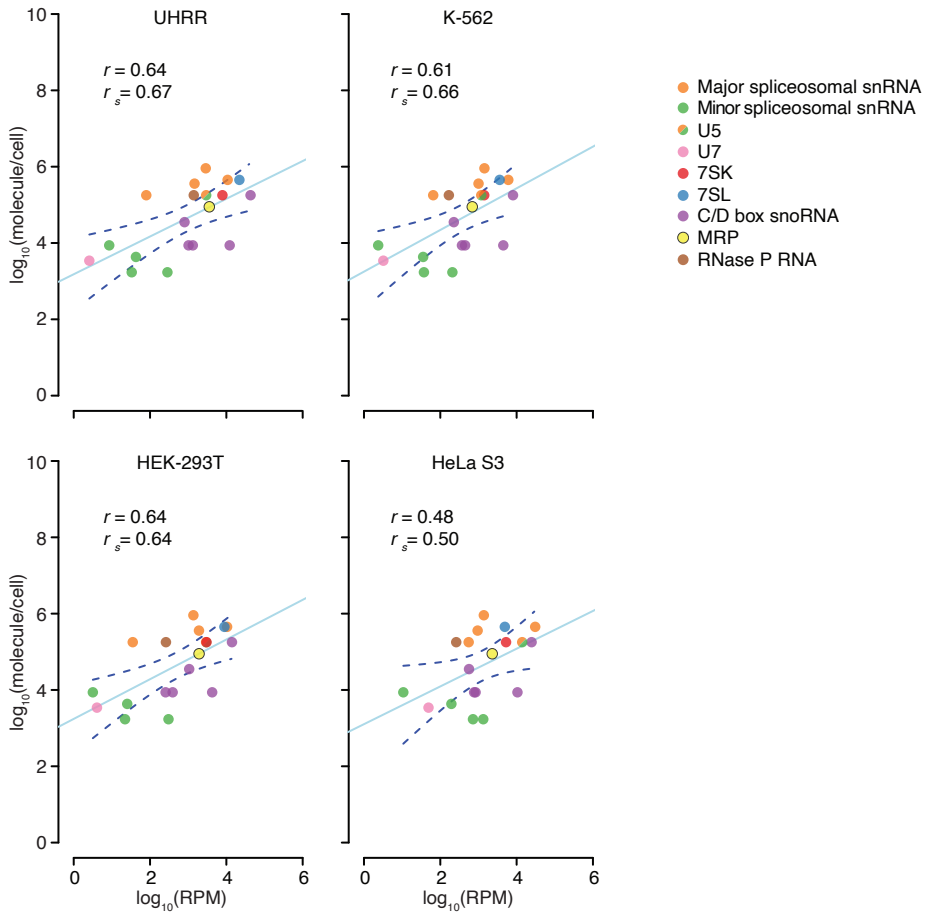
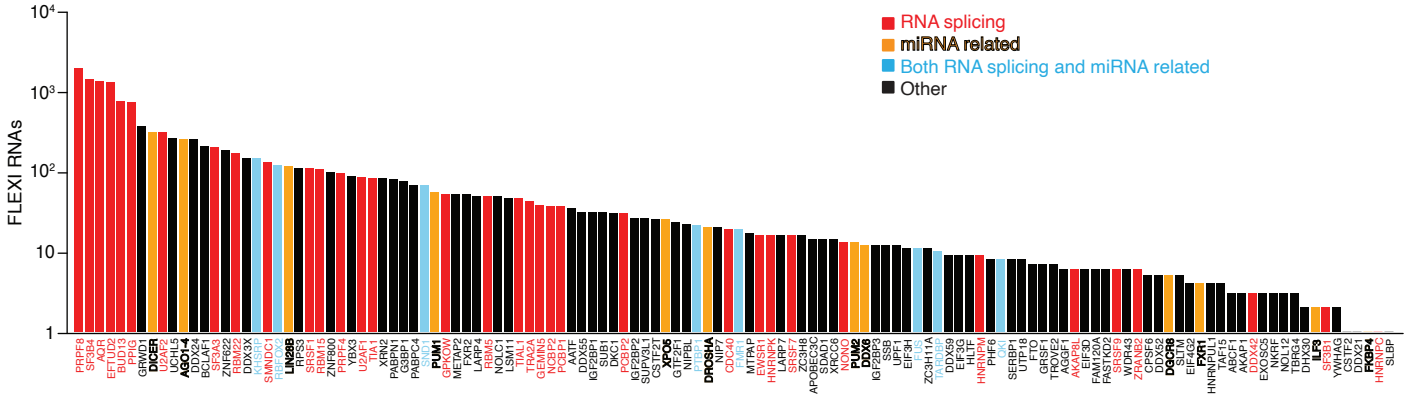

UHRR
r = 0.64
r_s = 0.67
K-562
r = 0.61
r_s = 0.66
HEK-293T
r = 0.64
r_s = 0.64
HeLa S3
r = 0.48
r_s = 0.50
log10(molecule/cell)
log10(RPM)
0
2
4
6
8
10
Major spliceosomal snRNA
Minor spliceosomal snRNA
U5
U7
7SK
7SL
C/D box snoRNA
MRP
RNase P RNA

Fig. S7



RNA splicing
miRNA related
Both RNA splicing and miRNA related
Other
FLEXI RNAs
10^4
10^3
10^2
10
1
PRPF8 SF3B4 AQR EFTUD2 BUD13 PPIG GRWD1 DICER U2AF2 UCHL5 AGO1-4 DDX24 BCLAF1 SF3A3 ZNF622 RBM22 DDX3X KHSRP SMNDC1 RBFOX2 LIN28B RPS3 SRSF1 RBM15 ZNF800 PRPF4 YBX3 U2AF1 TIA1 XRN2 PABPN1 G3BP1 PABPC4 SND1 PUM1 GPKOW METAP2 FXR2 LARP4 RBM5 NOLC1 LSM11 TIAL1 TRA2A GEMIN5 NCBP2 PCBP1 AATF DDX55 IGF2BP1 SUB1 DKC1 PCBP2 IGF2BP2 SUPV3L1 CSTF2T XPO6 GTF2F1 NIPBL PTBP1 DROSHA NUP7 CDC40 FMR1 MTRAP EWSR1 HNRNPK LARP7 SRSF7 ZC3H8 APOBEC3C SDAD1 XRCC6 NOP10 PUM2 DDX6 IGF2BP3 SSB UPF1 EIF3H FUS ZC3H11A TAFDP DDX59 EIF3G HLTF HNRNPM PHF6 DKC1 SERBP1 UTP18 FTO GRSF1 TROVE2 AGGF1 AKAP8L EIF3D FAM120A FASTKD2 SRSF9 WDR43 ZRANB2 CPSF6 DDX52 DGCR8 SLTM EIF4G2 FXR1 HNRNPUL1 TAF15 ABCF1 AKAP1 DDX42 EXOSC5 NKRF NOL12 TBRG4 DHX30 ILF3 SF3B1 YWHAG CSTF2 DDX21 FKBP4 HNRNPC SLBP

Fig. S8

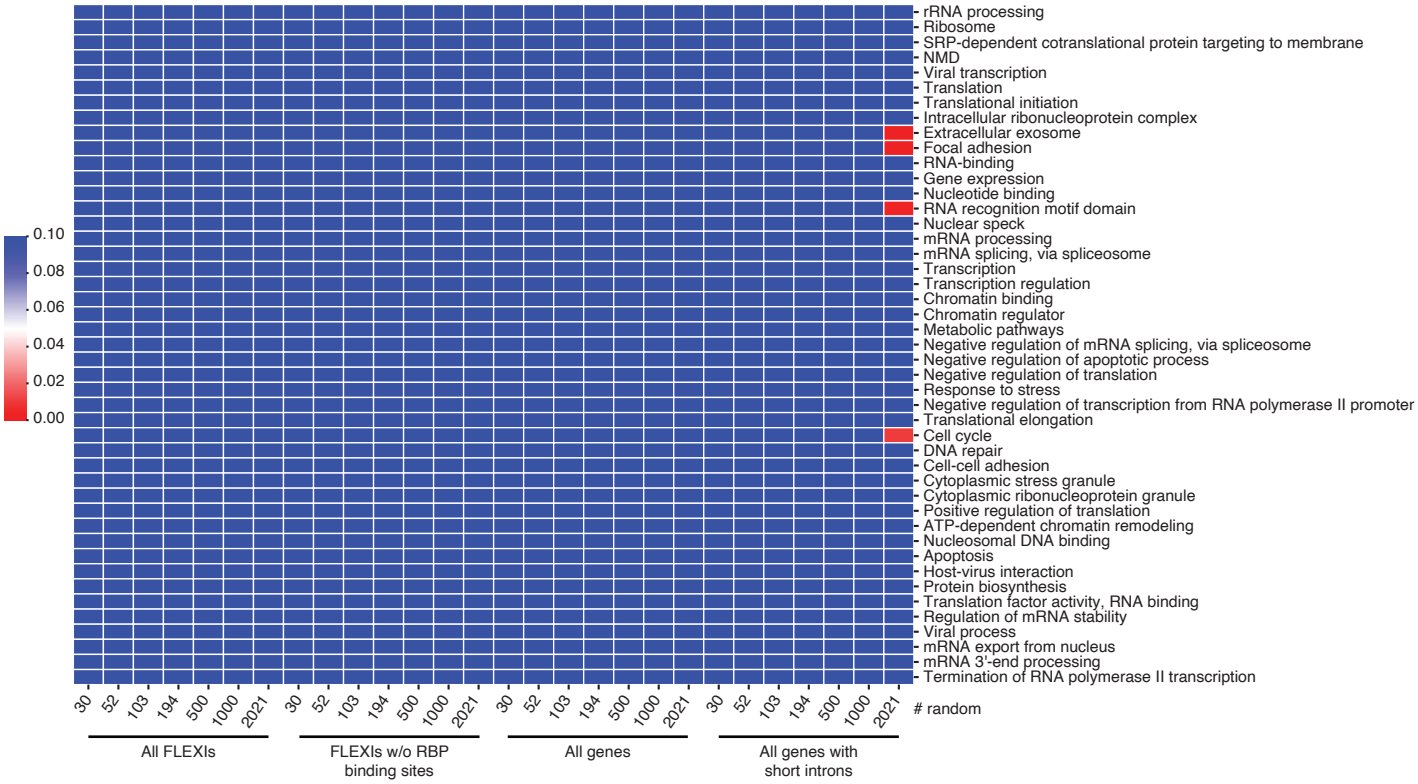

rRNA processing
Ribosome
SRP-dependent cotranslational protein targeting to membrane
NMD
Viral transcription
Translation
Translational initiation
Intracellular ribonucleoprotein complex
Extracellular exosome
Focal adhesion
RNA-binding
Gene expression
Nucleotide binding
RNA recognition motif domain
Nuclear speck
mRNA processing
mRNA splicing, via spliceosome
Transcription
Transcription regulation
Chromatin binding
Chromatin regulator
Metabolic pathways
Negative regulation of mRNA splicing, via spliceosome
Negative regulation of apoptotic process
Negative regulation of translation
Response to stress
Negative regulation of transcription from RNA polymerase II promoter
Translational elongation
Cell cycle
DNA repair
Cell-cell adhesion
Cytoplasmic stress granule
Cytoplasmic ribonucleoprotein granule
Positive regulation of translation
ATP-dependent chromatin remodeling
Nucleosomal DNA binding
Apoptosis
Host-virus interaction
Protein biosynthesis
Translation factor activity, RNA binding
Regulation of mRNA stability
Viral process
mRNA export from nucleus
mRNA 3'-end processing
Termination of RNA polymerase II transcription
0.10
0.08
0.06
0.04
0.02
0.00
30
52
103
194
500
1000
2021
random
All FLEXIs
FLEXIs w/o RBP binding sites
All genes
All genes with short introns

Fig. S9

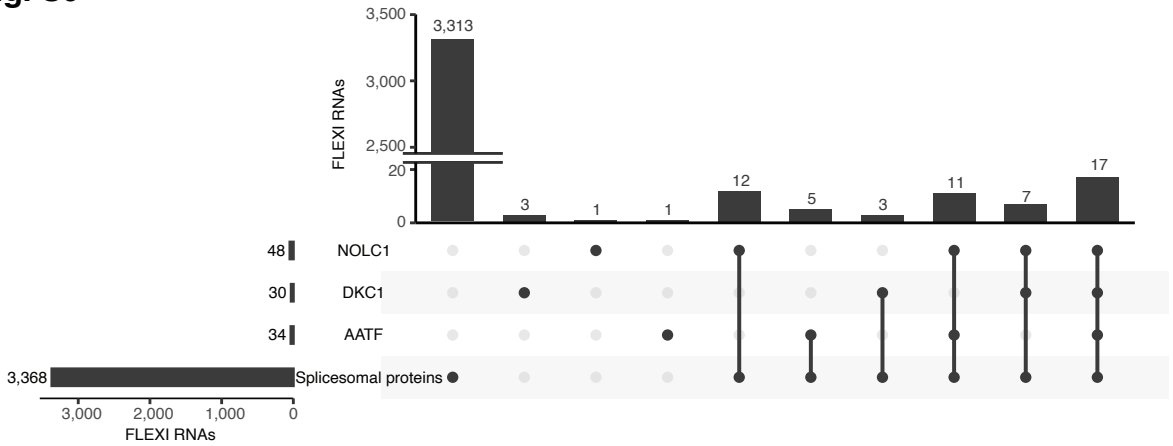
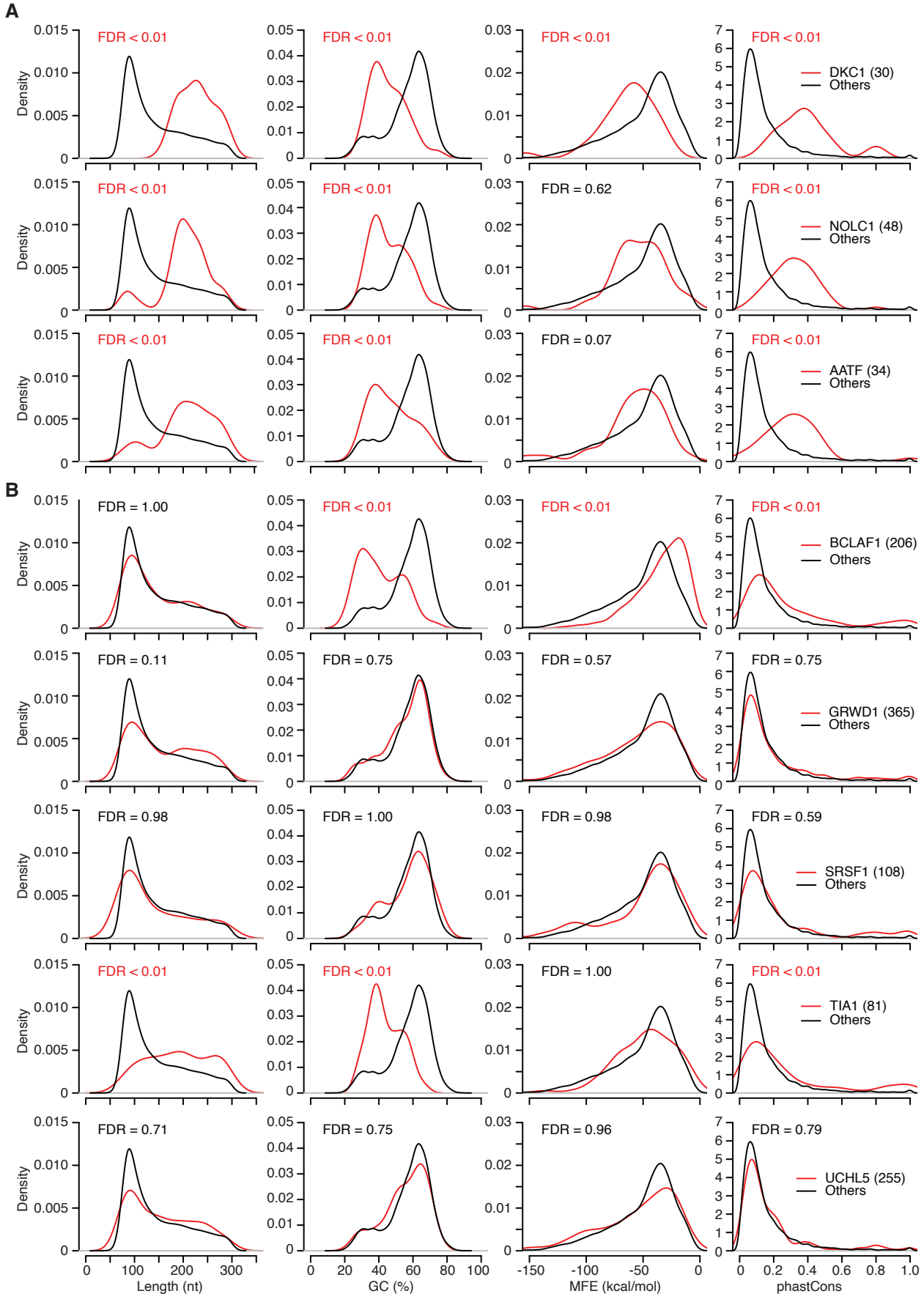

3,500
3,000
2,500
20
0
FLEXI RNAs
3,313
3
1
1
12
5
3
11
7
17
48
NOLC1
30
DKC1
34
AATF
3,368
Splicesomal proteins
3,000
2,000
1,000
0
FLEXI RNAs

Fig. S10


A
B
Density
FDR < 0.01
FDR = 0.62
FDR = 0.07
FDR = 1.00
FDR = 0.11
FDR = 0.75
FDR = 0.57
FDR = 0.98
FDR = 0.59
FDR = 0.71
FDR = 0.96
FDR = 0.79
DKC1 (30)
NOLC1 (48)
AATF (34)
BCLAF1 (206)
GRWD1 (365)
SRSF1 (108)
TIA1 (81)
UCHL5 (255)
Others
Length (nt)
GC (%)
MFE (kcal/mol)
phastCons

B (Cont.)

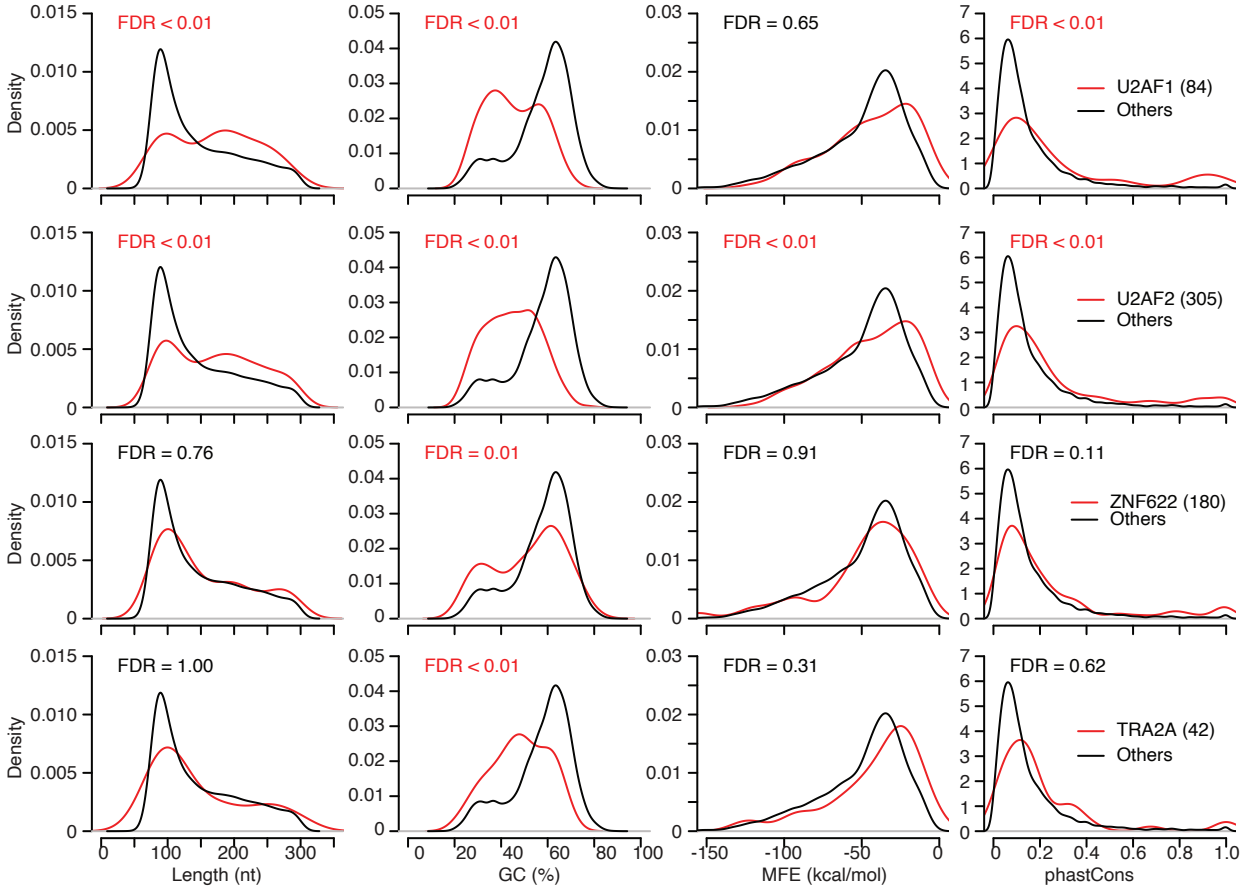
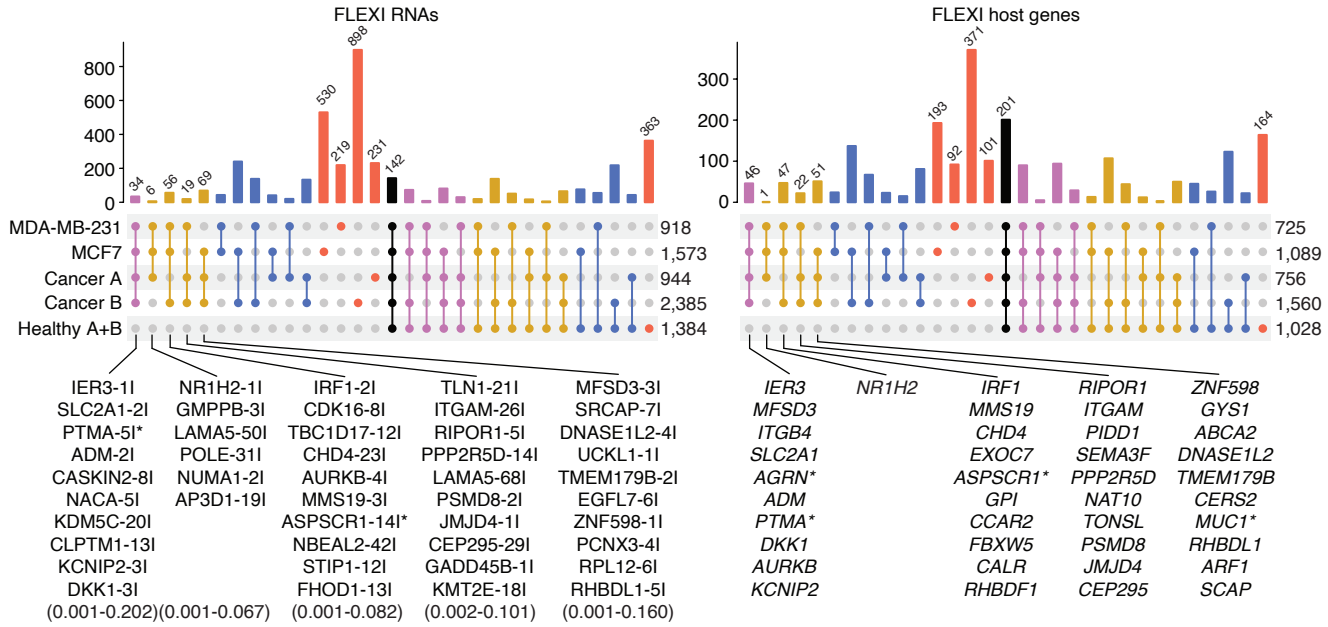

FDR < 0.01
FDR = 0.65
U2AF1 (84)
Others
U2AF2 (305)
FDR = 0.76
FDR = 0.01
FDR = 0.91
FDR = 0.11
ZNF622 (180)
FDR = 1.00
FDR = 0.31
FDR = 0.62
TRA2A (42)
Density
Length (nt)
GC (%)
MFE (kcal/mol)
phastCons

Fig. S11


FLEXI RNAs
FLEXI host genes
MDA-MB-231
MCF7
Cancer A
Cancer B
Healthy A+B
918
1,573
944
2,385
1,384
725
1,089
756
1,560
1,028
IER3-1I
SLC2A1-2I
PTMA-5I*
ADM-2I
CASKIN2-8I
NACA-5I
KDM5C-20I
CLPTM1-13I
KCNIP2-3I
DKK1-3I
(0.001-0.202)
NR1H2-1I
GMPPB-3I
LAMA5-50I
POLE-31I
NUMA1-2I
AP3D1-19I
(0.001-0.067)
IRF1-2I
CDK16-8I
TBC1D17-12I
CHD4-23I
AURKB-4I
MMS19-3I
ASPSCR1-14I*
NBEAL2-42I
STIP1-12I
FHOD1-13I
(0.001-0.082)
TLN1-21I
ITGAM-26I
RIPOR1-5I
PPP2R5D-14I
LAMA5-68I
PSMD8-2I
JMJD4-1I
CEP295-29I
GADD45B-1I
KMT2E-18I
(0.002-0.101)
MFSD3-3I
SRCAP-7I
DNASE1L2-4I
UCKL1-1I
TMEM179B-2I
EGFL7-6I
ZNF598-1I
PCNX3-4I
RPL12-6I
RHBDL1-5I
(0.001-0.160)
IER3
MFSD3
ITGB4
SLC2A1
AGRN*
ADM
PTMA*
DKK1
AURKB
KCNIP2
NR1H2
IRF1
MMS19
CHD4
EXOC7
ASPSCR1*
GPI
CCAR2
FBXW5
CALR
RHBDF1
RIPOR1
ITGAM
PIDD1
SEMA3F
PPP2R5D
NAT10
TONSL
PSMD8
JMJD4
CEP295
ZNF598
GYS1
ABCA2
DNASE1L2
TMEM179B
CERS2
MUC1*
RHBDL1
ARF1
SCAP