Human cells contain myriad excised linear intron RNAs with potential functions in gene regulation and utility as disease biomarkers

Jun Yao[1], Hengyi Xu[1], Douglas C. Wu[1†], Manuel Ares, Jr.[2], and Alan M. Lambowitz[1]*

[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

†Present address: QIAGEN

*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 end-to-end sequence reads of highly structured RNAs, to identify >8,000 short (≤300 nt) full-length excised intron (FLEXI) RNAs in human cells. Most FLEXI RNAs are predicted to have stable secondary structures, making them difficult to detect by other RNA-seq methods. Some FLEXI RNAs correspond to annotated mirtron pre-miRNAs (introns that are processed into functional miRNAs) or agotrons (introns that bind AGO2 and function in a miRNA-like manner), but the vast majority had not been identified or characterized previously. FLEXI RNA profiles are cell-type specific, reflecting differences in host gene transcription, alternative splicing, or 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 of the FLEXI RNAs contained one or more experimentally identified binding sites for a spliceosomal protein, AGO1-4, DICER, or a number of different regulatory proteins, suggesting multiple ways in which FLEXIs could contribute to the regulation of gene expression. As FLEXI RNAs are linked to the expression of thousands of protein-coding and lncRNA genes, they potentially constitute a new class of broadly applicable, highly discriminatory biomarkers for human diseases.

cancer / diagnostics / liquid biopsy / miRNA / RNA-binding protein / RNA sequencing / RNA splicing

Introduction

Most protein-coding genes in eukaryotes consist of coding regions (exons) separated by introns, which must be removed by RNA splicing to produce a functional mRNA. RNA splicing is performed by a large ribonucleoprotein complex, the spliceosome, which catalyzes transesterification reactions yielding ligated exons and a distinctive excised intron lariat RNA whose 5' end is linked to a branch-point nucleotide near its 3' end by a 2',5'-phosphodiester bond (1). In most cases, excised intron lariat RNAs are acted upon by a dedicated debranching enzyme (DBR1) to produce a linear intron RNA, which is rapidly degraded by cellular ribonucleases (2). In a few cases, excised intron RNAs have been found to persist after excision, either as branched circular RNAs (lariats whose tails have been removed) or as unbranched linear RNAs, with some contributing to cell or viral regulatory processes (3-11). The latter include a group of yeast introns that are rapidly degraded in log phase cells but are debranched and accumulate as linear RNAs in stationary phase where they may contribute to a generalized cellular stress response (10, 12). Other examples of excised intron RNAs with cellular functions are mirtrons, structured pre-miRNA introns that are debranched by DBR1 and processed by DICER into functional miRNAs (13-15), and agotrons, structured excised linear intron RNAs that bind AGO2 and function directly to repress target mRNAs in a miRNA-like manner (16).

Recently, while analyzing human plasma RNAs by using Thermostable Group II Intron Reverse Transcriptase sequencing (TGIRT-seq), a method that gives full-length, end-to-end sequence reads of highly structured RNAs, we identified ~50 short (≤300 nt), full-length excised intron (FLEXI) RNAs, subsets of which corresponded to annotated mirtrons or agotrons (17). Here, we used TGIRT-seq to systematically search for FLEXI RNA in human cell lines and tissues. We thus identified >8,000 short (≤300 nt) FLEXI RNAs, the vast majority of which had not been characterized previously. FLEXI RNAs could potentially function in gene regulatory pathways and constitute a new class of broadly applicable, highly discriminatory RNA biomarkers for human diseases.

Results and Discussion

A search of Ensembl GRCh38 Release 93 annotations (http://www.ensembl.org) revealed 51,645 short introns (≤300 nt) in 12,020 different genes that could potentially give rise to FLEXI RNAs. To investigate which of these introns might produce FLEXI RNAs in biological samples, we obtained TGIRT-seq datasets of intact (non-fragmented) human cellular RNAs (Universal Human Reference RNA (UHRR) and RNAs from HEK-293T, K-562, and HeLa S3 cells) and reanalyzed the plasma RNA datasets by mapping to the hg38 human genome reference sequence (Table 1). The searches were done using combined datasets obtained from multiple replicate libraries totaling 666-768 million mapped reads for each of the cellular RNA samples (Table 1). We thus identified 8,144 different FLEXI RNAs originating from 3,743 protein-coding genes, lncRNA genes, or pseudogenes (collectively denoted FLEXI host genes) represented by at least one read in any of the sample types (Fig. 1*A*). Most FLEXI RNAs were present in relatively low abundance in whole-cell RNA preparations, with density plots showing sharp peaks for the UHRR, K-562, and HEK-293T samples at 0.003-0.01 reads per million (RPM) and broader peaks for the HeLa S3 and plasma samples at 0.02 RPM (Fig. S1). The most abundant FLEXIs in different sample types were present at 1.3-6.9 RPM. Pairwise scatterplots showed that FLEXI RNA profiles are cell-type specific with many FLEXIs differentially expressed in each sample type (Fig. 1*B*).

The FLEXI RNAs identified by TGIRT-seq appear to be linear RNAs that extend from the 5'- to 3'-splice site with no evidence of base substitutions or impediments that might indicate the presence of a branch-point nucleotide (Fig. 2*A*). In addition to host gene transcription, the cell-type specific detection of FLEXI RNAs reflects differences in alternative splicing and stability of the excised intron RNAs (examples shown in Fig. 2*B*). Most of the detected FLEXI RNAs have sequence characteristics of major (U2-type) spliceosomal introns, with (98.7%) have canonical GU-AG ends and 1.3% having GC-AG or other non-canonical 5' and 3' end (18), and only 36 having sequence characteristics expected for minor (U12-type) spliceosomal introns (19) (Fig. 3*A* and Fig. S2). The identified FLEXI RNAs have a canonical branch-point (BP) consensus

sequence (Fig. 3*A*) (20), suggesting that most if not all were excised as lariat RNAs and debranched by DBR1, as found for mirtron pre-miRNAs (14, 15).

Density distribution plots showed that the FLEXI RNAs detected in plasma are a relatively homogenous subset with peaks at 90-nt length, 70% GC, and -40 kcal/mole minimum free energy (MFE; ΔG) for the most stable RNA secondary predicted by RNAfold (Fig. 1*C*-*E*). By comparison, the FLEXI RNAs detected in cells were more heterogenous, with similar peaks but larger shoulders extending to longer lengths, lower GC contents, and less stable predicted secondary structures (≥-25 kcal/mole; Fig. 1*C*-*E*). Similar trends were seen in density distribution plots for FLEXI RNAs present at higher abundance (≥0.01 and ≥0.02 RPM; Fig. S3*A* and *B*). The more homogenous subset of FLEXI RNAs found in plasma could reflect preferential export or greater resistance to plasma RNases of shorter, more stably structured FLEXI RNAs. Consistent with these possibilities, 23% of the FLEXI RNAs found in plasma but only 1.7-2.6% of those found in cells corresponded to annotated mirtron pre-miRNAs or agotrons, whose biological functions require a stable stem-loop structure (Fig. 3*B* and *C*). A small number of FLEXI RNAs found in cells but not plasma (43, 0.5% of the total) contained embedded snoRNAs (Fig. 3*C*).

About half (4,505; 55%) of the detected FLEXI RNAs contained an experimentally identified binding site for one or more of 126 different RNA-binding proteins (RBPs) in human eCLIP (21), DICER PAR-CLIP (22), or AGO1-4 PAR-CLIP (23) datasets (Fig. 4*A* and *B*). Compared to all annotated introns or protein-binding sites in the above datasets, the detected FLEXI RNAs were enriched in binding sites for spliceosomal proteins (*e.g.*, PRPF8, SF3B4, AQR, EFTUD2, BUD13), as well as AGO1-4 (250 FLEXIs, including 23 annotated agotrons (16)) and DICER (308 FLEXIs, including 44 annotated mirtrons (24)). Sixty-six FLEXIs contained binding sites for both AGO1-4 and DICER (including five annotated as both agotrons and mirtrons). Notably, FLEXI RNAs were also enriched in binding sites for several regulatory proteins, including PPIG (protein folding), GRWD1 (ribosome biogenesis and histone methylation), UCHL5 (protease), and ZNF622 (transcription regulation).

Most the identified FLEXI RNAs (7,775, 95%), including those corresponding to mirtron pre-miRNAs or agotrons, had low PhastCons scores (<0.5 calculated for 27 primates including humans plus mouse, dog, and armadillo), with those encoding snoRNAs having somewhat higher PhastCons scores (four at ≥0.5; Fig. 1*F*). GO enrichment analysis using biological process annotations (25, 26) showed that the more evolutionarily conserved FLEXIs (PhastCons scores ≥0.50) were enriched in genes involved in regulation of RNA metabolic processes and gene expression, while those with lower PhastCons scores (<0.5) were enriched in genes involved in other biological processes (Fig. 5). Among the most highly conserved FLEXIs (PhastCons score ≥0.99; n=44), several are multiples of three nucleotides long and would encode additional in-frame protein sequence if retained in the mRNA (examples in *HNRNPL*, *HNRNPM*, and *FXR1*). Another FLEXI (*EIF1*, chr17:41,690,819-41,690,902) appears to be unique in humans and arose from a point mutation in the 3' untranslated region (UTR) that generates a new 3'-splice site, resulting in a novel, human-specific EIF1 isoform.

The cell- and tissue-specific patterns of FLEXI RNA expression suggested that they might be useful as biomarkers that distinguish normal and abnormal cellular states. To test this idea, we compared FLEXI RNAs and FLEXI host genes in commercial matched tumor and neighboring healthy tissue from two breast cancer patients (patients A and B) 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. 6*A* and Fig. S4 for FLEXIs detected at ≥0.01 RPM). The discriminatory power of FLEXIs is also evident in scatter plots comparing 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 comparing all transcripts from the same FLEXI host genes (Fig. 6*B*). The scatter plots also 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 each cancer patient, but not detected at all in the matched healthy tissue (highlighted in red in Fig. 6*B*). GO enrichment analysis of FLEXI RNA host genes detected in the four cancer samples but

not healthy tissues showed significant enrichment ($p \leq 0.05$) in hallmark gene sets (27) that are perturbed in many cancers (Fig. 6*C*, 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 V1/2, estrogen response early/late, androgen response, oxidative phosphorylation, mTORC1 signaling, apical junction, and cholesterol homeostasis (Fig. 6*C*, pathway names in orange).

The short introns that give rise to FLEXI RNAs could have originated by splice-site acquisition, as found for the *EIF1* intron described above, or by an active intron transposition process (28-30), with their retention in the human genome reflecting that they acquired a cellular function or were not sufficiently deleterious for loss by purifying selection. 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 found previously for one of the plasma FLEXI RNAs (17). However, intron transposition followed by sequence divergence cannot be excluded.

Based on literature precedents (3, 4, 8, 10, 12), FLEXIs RNAs could function in cellular regulation by base pairing to a complementary target RNA or by sequestering proteins that function in RNA splicing or other cellular processes, the latter possibility supported by the identification of multiple FLEXI RNAs corresponding to binding sites for regulatory proteins (Fig. 4*B*). The additional FLEXI RNAs identified here as being DICER or AGO1-4 binding sites could be previously unannotated mirtrons or agotrons or could function to sequester these proteins, thereby impacting miRNA biogenesis or function. Additionally, FLEXI RNAs that have stable stem-loop structures could be acted upon by a double-stranded RNA-specific endonuclease, like DICER or DROSHA, to generate discrete RNA fragments that could function as miRNAs or other short regulatory RNAs. More generally, the stable predicted secondary structure found for many FLEXI RNAs may be a common feature of introns that have acquired a secondary function, as such structures could facilitate splicing by bringing splice sites closer together, contribute to 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 origin or function, FLEXI RNAs potentially provide a large new class of RNA biomarkers that are linked to the expression of thousands of different protein-coding and lncRNA genes, with particular utility as biomarkers in bodily fluids such as plasma where their stable secondary structures and/or bound proteins may protect them from RNases (17). Although FLEXI RNAs are not abundant in whole-cell RNA preparations, their relative abundance for diagnostics applications could be increased readily by incorporating a size-selection step that retains RNAs ≤300 nt prior to RNA-seq or other detection methods. As many FLEXIs are structured RNAs, their initial detection and characterization is best done by TGIRT-seq to obtain full-length, end-to-end sequence reads. Once identified, targeted assays for optimal FLEXI RNAs or combinations therefore could use different types of read outs, such as RT-qPCR, microarrays, other hybridization-based assays, or targeted RNA-seq. Targeted RNA panels of FLEXI RNAs by themselves or combined with other analytes could provide a rapid cost-effective method for the diagnosis and routine monitoring of progression and response to treatment for a wide variety of human diseases.

Materials and Methods

DNA and RNA oligonucleotides. The DNA and RNA oligonucleotides used for TGIRT-seq on the Illumina sequencing platform are listed in Table 2. 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) + 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, L-glutamine, and 110 mg/L 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 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 (31), followed by clean-up using a Zymo RNA Clean & Concentrator kit (8X ethanol protocol). The RNA fragment size range was confirmed and the RNA concentration determined by using an Agilent 2100 Bioanalyzer with a 6000 RNA Pico chip, and the RNA was aliquoted into 4 ng portions for storage in -80 °C.

TGIRT-seq. TGIRT-seq libraries were prepared as described (31) 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 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 purified from commercial plasma pooled from multiple healthy individuals was described previously (17), and the resulting datasets, which were reused in this study, 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 (Table 1). Illumina TruSeq adapters and PCR primer sequences were trimmed from the reads with Cutadapt v2.8 (32) (sequencing quality score cut-off at 20; p-value <0.01) 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 (33) 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 sncRNA and rRNA reference sequences, including human miRNA, tRNA, Y RNA, Vault RNA, 7SL RNA and 7SK RNA, 5S rRNA and 45S rRNA genes

including the 2.2-kb 5S rRNA repeats from the 5S rRNA cluster on chromosome 1 (1q42, GeneBank: X12811) and the 43-kb 45S rRNA repeats that contained 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 (34) 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 (35) 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 (36). To avoid miscounting reads with embedded sncRNAs that were not filtered out in Pass2 (*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). 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 calculate the coverage for 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, and Venn diagram were plotted by using the VennDiagram package v1.6.20 in R. 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.

5'- and 3'-splice-sites and branch-point consensus sequences of human U2- and U12-type spliceosomal introns were obtained from previous publications (19, 20). 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 (19) using FIMO (37) with the following settings: FIMO --text --norc <GU_AG_U12_5SS motif file> <sequence file>. The branch-point (BP) 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 (38) 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.

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 (24), agotrons (16), 150 RBPs (eCLIP, GENCODE, annotations with irreproducible discovery rate analysis) (21), DICER PAR-CLIP (22), and Ago1-4 PAR-CLIP (23) datasets by using BEDTools.

FLEXI RNAs containing embedded snoRNAs were identified by intersecting the FLEXI RNA coordinates with the coordinates of annotated snoRNA and scaRNA from the Ensembl GRCh38 annotations.

GO enrichment analysis was done by using ShinyGO (26) with the GO term of Biological Process (25) and hallmark gene sets from the Molecular Signatures Database (MSigDB) (27).

ACKNOWLEDGEMENTS. We thank Marta Mastroianni and Ryan Nottingham for cell culture and RNA preparations, and Blerta Xhemalce (University of Texas at Austin) for a fresh stock of MDA-MB-231 cells. 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.

Conflict-of-interest statement: Thermostable group II intron reverse transcriptase (TGIRT) enzymes and methods for their use are the subject of patents and patent applications that have been licensed by the University of Texas and East Tennessee State University to InGex, LLC.

A.M.L., some former and present members of the Lambowitz laboratory, and the University of Texas are minority equity holders in InGex, LLC 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. M.A. has no competing interests.

Data availability: TGIRT-seq datasets have been deposited in the Sequence Read Archive (SRA) under accession numbers PRJNA648481 and PRJNA640428.

Code availability: 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.

References

1. M. E. Wilkinson, C. Charenton, K. Nagai, RNA splicing by the spliceosome. *Annu. Rev. Biochem.* **89**, null (2020).
2. K. B. Chapman, J. D. Boeke, Isolation and characterization of the gene encoding yeast debranching enzyme. *Cell* **65**, 483-492 (1991).
3. M. J. Farrell, A. T. Dobson, L. T. Feldman, Herpes simplex virus latency-associated transcript is a stable intron. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 790-794 (1991).
4. C. A. Kulesza, T. Shenk, Murine cytomegalovirus encodes a stable intron that facilitates persistent replication in the mouse. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 18302-18307 (2006).
5. E. J. Gardner, Z. F. Nizami, C. C. Talbot, J. G. Gall, Stable intronic sequence RNA (sisRNA), a new class of noncoding RNA from the oocyte nucleus of *Xenopus tropicalis*. *Genes Dev.* **26**, 2550-2559 (2012).

6. W. N. Moss, J. A. Steitz, Genome-wide analyses of Epstein-Barr virus reveal conserved RNA structures and a novel stable intronic sequence RNA. *BMC Genomics* **14**, 543 (2013).
7. Y. Zhang *et al.*, Circular intronic long noncoding RNAs. *Mol. Cell* **51**, 792-806 (2013).
8. J. W. Pek, I. Osman, M. L.-I. Tay, R. T. Zheng, Stable intronic sequence RNAs have possible regulatory roles in *Drosophila melanogaster*. *J. Cell Biol.* **211**, 243-251 (2015).
9. G. J. S. Talhouarne, J. G. Gall, Lariat intronic RNAs in the cytoplasm of vertebrate cells. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E7970-E7977 (2018).
10. J. T. Morgan, G. R. Fink, D. P. Bartel, Excised linear introns regulate growth in yeast. *Nature* **565**, 606-611 (2019).
11. H. Saini, A. A. Bicknell, S. R. Eddy, M. J. Moore, Free circular introns with an unusual branchpoint in neuronal projections. *Elife* **8**, e47809 (2019).
12. J. Parenteau *et al.*, Introns are mediators of cell response to starvation. *Nature* **565**, 612-617 (2019).
13. E. Berezikov, W.-J. Chung, J. Willis, E. Cuppen, E. C. Lai, Mammalian mirtron genes. *Mol. Cell* **28**, 328-336 (2007).
14. K. Okamura, J. W. Hagen, H. Duan, D. M. Tyler, E. C. Lai, The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell* **130**, 89-100 (2007).
15. J. G. Ruby, C. H. Jan, D. P. Bartel, Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**, 83-86 (2007).
16. T. B. Hansen *et al.*, Argonaute-associated short introns are a novel class of gene regulators. *Nat Commun* **7**, 11538 (2016).
17. J. Yao, D. C. Wu, R. M. Nottingham, A. M. Lambowitz, Identification of protein-protected mRNA fragments and structured excised intron RNAs in human plasma by TGIRT-seq peak calling. *eLife* **9**, e60743 (2020).
18. M. Burset, I. A. Seledtsov, V. V. Solovyev, Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* **28**, 4364-4375 (2000).

19. N. Sheth *et al.*, Comprehensive splice-site analysis using comparative genomics. *Nucleic Acids Res.* **34**, 3955-3967 (2006).

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

21. E. L. Van Nostrand *et al.*, Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508-514 (2016).

22. A. Rybak-Wolf *et al.*, A variety of Dicer substrates in human and *C. elegans*. *Cell* **159**, 1153-1167 (2014).

23. M. Hafner *et al.*, Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129-141 (2010).

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

25. The Gene Ontology Consortium, The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330-D338 (2018).

26. S. X. Ge, D. Jung, R. Yao, ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**, 2628-2629 (2019).

27. A. Liberzon *et al.*, The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417-425 (2015).

28. S. Lee, S. W. Stevens, Spliceosomal intronogenesis. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 6514-6519 (2016).

29. M. P. Simmons *et al.*, Intron invasions trace algal speciation and reveal nearly identical arctic and antarctic micromonas populations. *Mol. Biol. Evol.* **32**, 2219-2235 (2015).

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

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

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

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

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

35. H. Li *et al.*, The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

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

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

38. T. L. Bailey *et al.*, MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202-W208 (2009).

Fig. 1. Characteristics of FLEXI RNAs in human cells and plasma. (*A*) UpSet plots of FLEXI RNAs (excised intron RNAs ≤300 nt with 5' and 3' ends within 3 nt of annotated splice sites) and host genes encoding FLEXI RNAs detected at ≥1 read in unfragmented RNA preparations from the indicated sample type. (*B*) Scatter plots comparing $\log_2$-transformed RPM of FLEXI RNAs in different samples. *r* and $r_{s,}$ are Pearson and Spearman correlation coefficients, respectively. (*C*-*F*) Density distribution plots of characteristics of FLEXI RNAs detected in different sample types (color coded as indicated in the Figure) compared to all 51,645 annotated introns ≤300 nt in the human genome (GRCh38; black).

Fig. 2. IGV screenshots showing read alignments for FLEXI RNAs. Gene names are at the top with the arrow indicating the 5’ to 3’ orientation of the encoded RNA and tracts below showing the gene annotation (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/or 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 are shown at the bottom. Mismatched boxed red or green nucleotides at the 5' end of RNA sequences are non-templated nucleotides added by TGIRT-III during TGIRT-seq library preparation. Some *MAZ* FLEXIs have a non-coded 3' A or U tail.

Fig. 3. FLEXI RNA splice site and branch-point consensus sequences and numbers and percentage of FLEXI RNAs corresponding to annotated mirtrons and/or agotrons or encoding an embedded snoRNAs in different samples types. (*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 end sequences. (*B*) Venn diagrams showing the relationships between FLEXI RNAs corresponding to annotated mirtrons (left) or agotrons (right) detected in different sample types. FLEXI RNAs annotated as both a mirtron and agotron are included in both Venn diagrams. (*C*) Number and percentages of FLEXI RNAs corresponding to annotated agotrons and mirtrons or containing an embedded snoRNA in different sample types compared to all short introns (≤300 nt) in the human genome (GRCh38). "Agotron and Mirtron" indicates the number of FLEXI RNAs detected in the indicated sample type that are annotated as both a mirtron and an agotron. "Agotron or Mirtron" indicates the number and percentage of FLEXIs detected in the indicated sample type that correspond to an annotated mirtron and/or an annotated agotron. Embedded snoRNAs indicates the number and percentage of FLEXI RNAs detected in the indicated sample type that contain an embedded snoRNA and (/) the number of those snoRNAs that are small Cajal body-specific snoRNAs (scaRNAs).

Fig. 4. Protein-binding sites in FLEXI RNAs. (*A* and *B*) Percentages of different categories of intron RNAs corresponding to experimentally identified protein-binding sites grouped by function or individual RNA-binding proteins (RBPs), respectively for: all FLEXI RNAs (≥1 read) in the UHHR, HEK-293T, HeLa S3, K-562 and plasma samples (black); all annotated short introns (≤300 nt) in the human genome (gray); all annotated introns in the human genome (blue); and all annotated binding sites for 150 different RBPs in the searched datasets (eCLIP RNA-binding proteins (GENCODE), AGO1-4 PAR-CLIP, and DICER PAR-CLIP; orange). In panel A, the apparent enrichment of FLEXI RNAs in the snoRNA/snRNA/telomerase group is due to the presence of spliceosomal protein AQR in that group. In panel B, the names of proteins that are spliceosome components or function in splicing regulation are in red, and those involved in miRNA processing or function are in orange.

Fig. 5. GO analysis for genes encoding more and less highly conserved FLEXI RNAs. The 8,144 FLEXI RNAs identified in UHHR, K-562 cells, HEK-293T cells, HeLa cells, and plasma were divided into groups comprised of 369 FLEXI RNAs with PhastCons scores ≥0.5 and 7,775 FLEXI RNAs with PhastCons cores <0.5, and GO enrichment analysis for the term Biological Process (25) was done using ShinyGO (26). The numbers in the parentheses for each biological process indicate the number of FLEXI host genes and (/) the total number of genes annotated for that process. Adjusted p values for the enrichment of FLEXI host genes in different biological processes are color coded as indicated in the Figure.

Fig. 6. FLEXI RNAs in breast cancer tumors and cell lines. (*A*) UpSet plots of FLEXI RNAs and FLEXI host genes detected (≥1 read) 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. The most abundant FLEXI RNAs and host genes with the highest numbers of aggregated FLEXI RNAs are listed below some sample groups. (*B*) Scatter plots comparing $\log_2$ transformed RPM of FLEXI RNAs and all transcripts from FLEXI host genes in unfragmented and chemically fragmented RNAs from cancer and healthy breast tissue 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 are in red. (*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. Names of pathways significantly enriched (p≤0.05) in all or at least one cancer sample are in red and orange, respectively.

TABLE 1. 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) (17).
# 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.

Table 2. Oligonucleotides used in TGIRT-seq library construction

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

Fig. 1

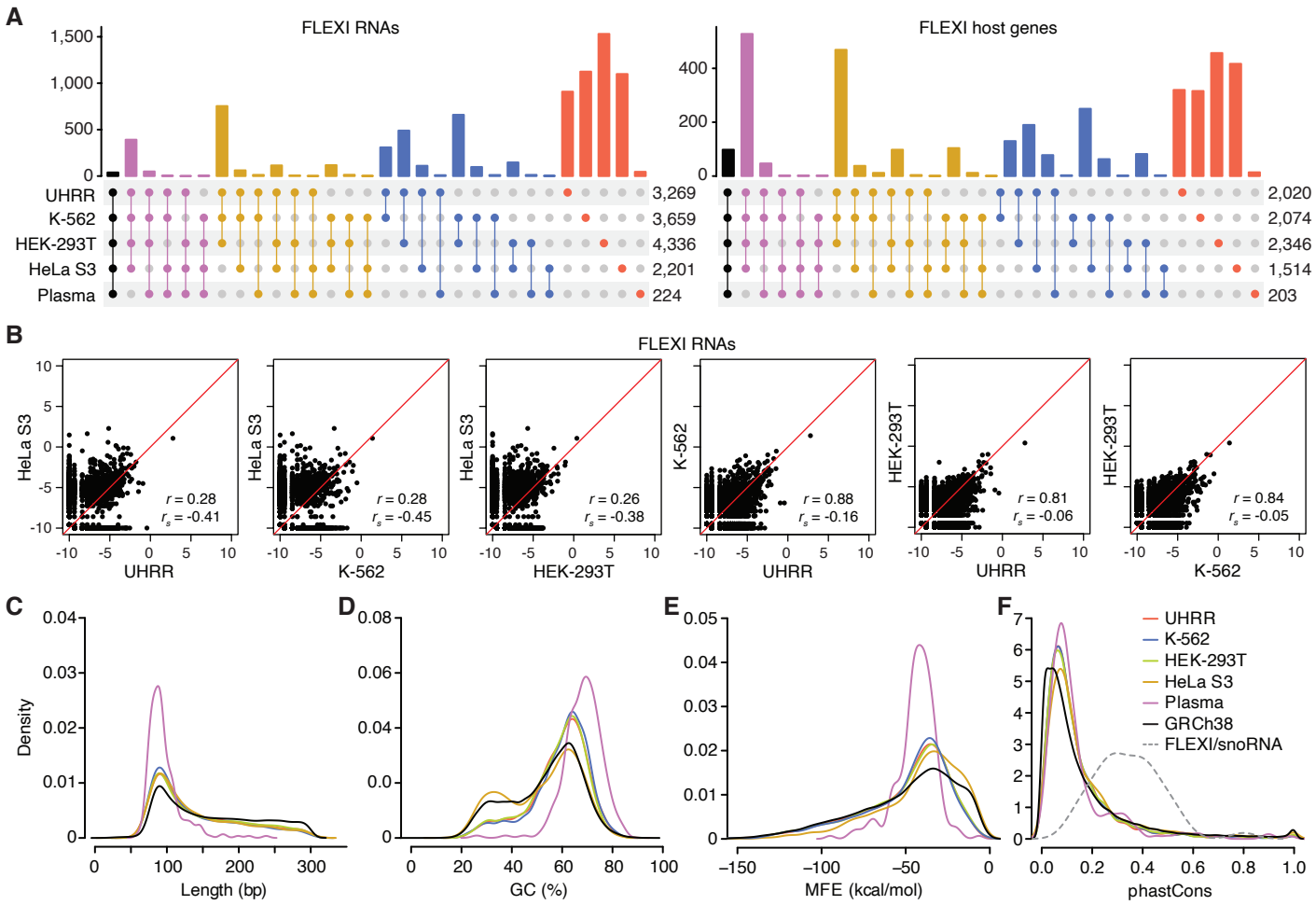


Fig. 2

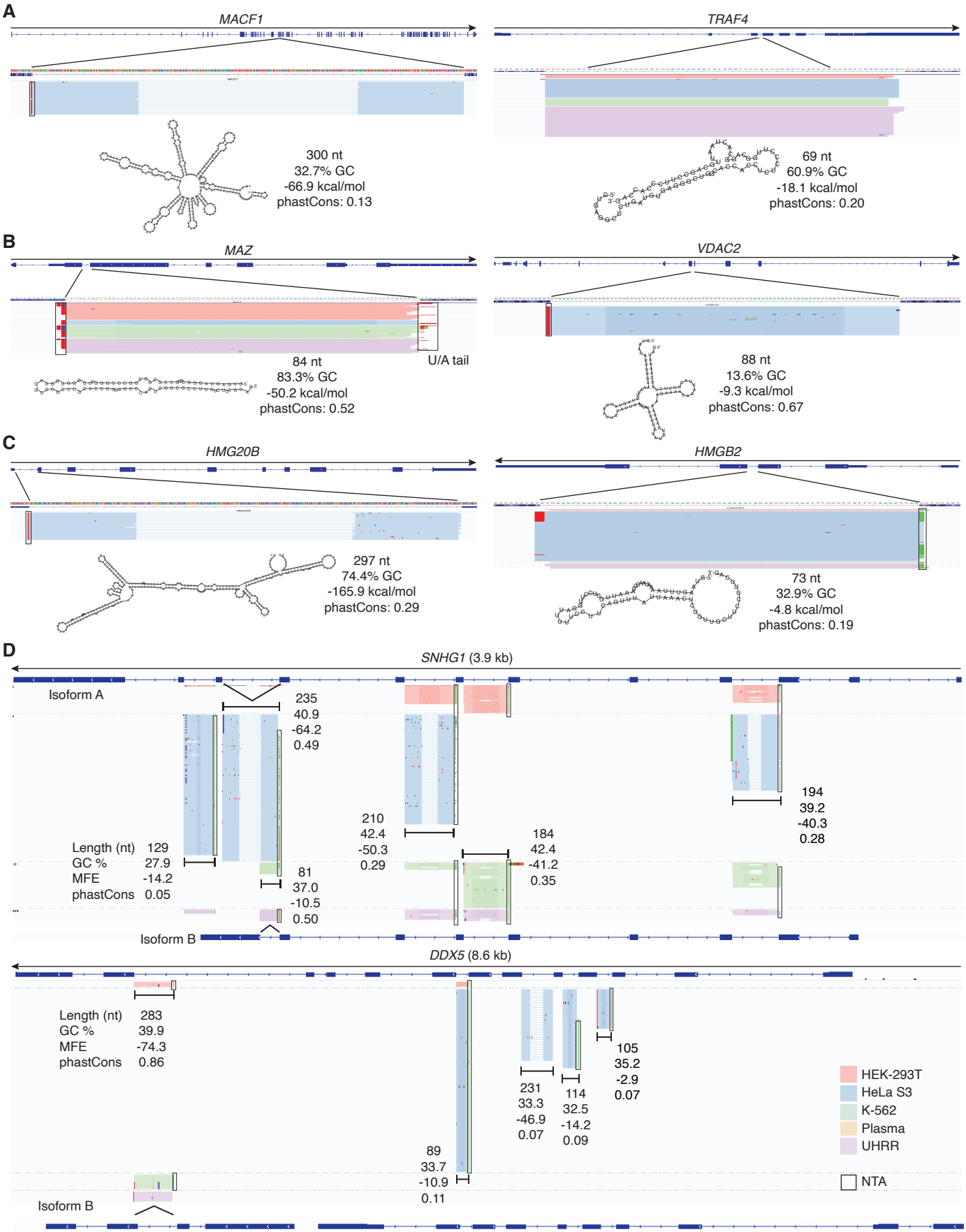
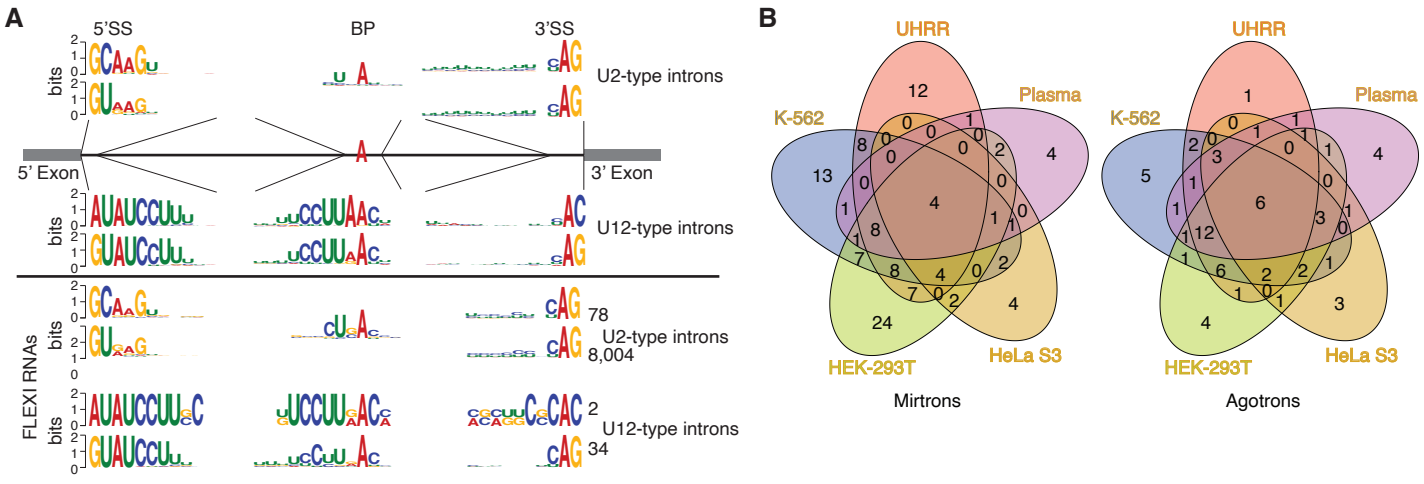


Fig. 3



C

Dataset	FLEXI RNAs	Agotrons	Mirtrons	Agotron and Mirtron	Agotron or Mirtron	Embedded snoRNAs
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

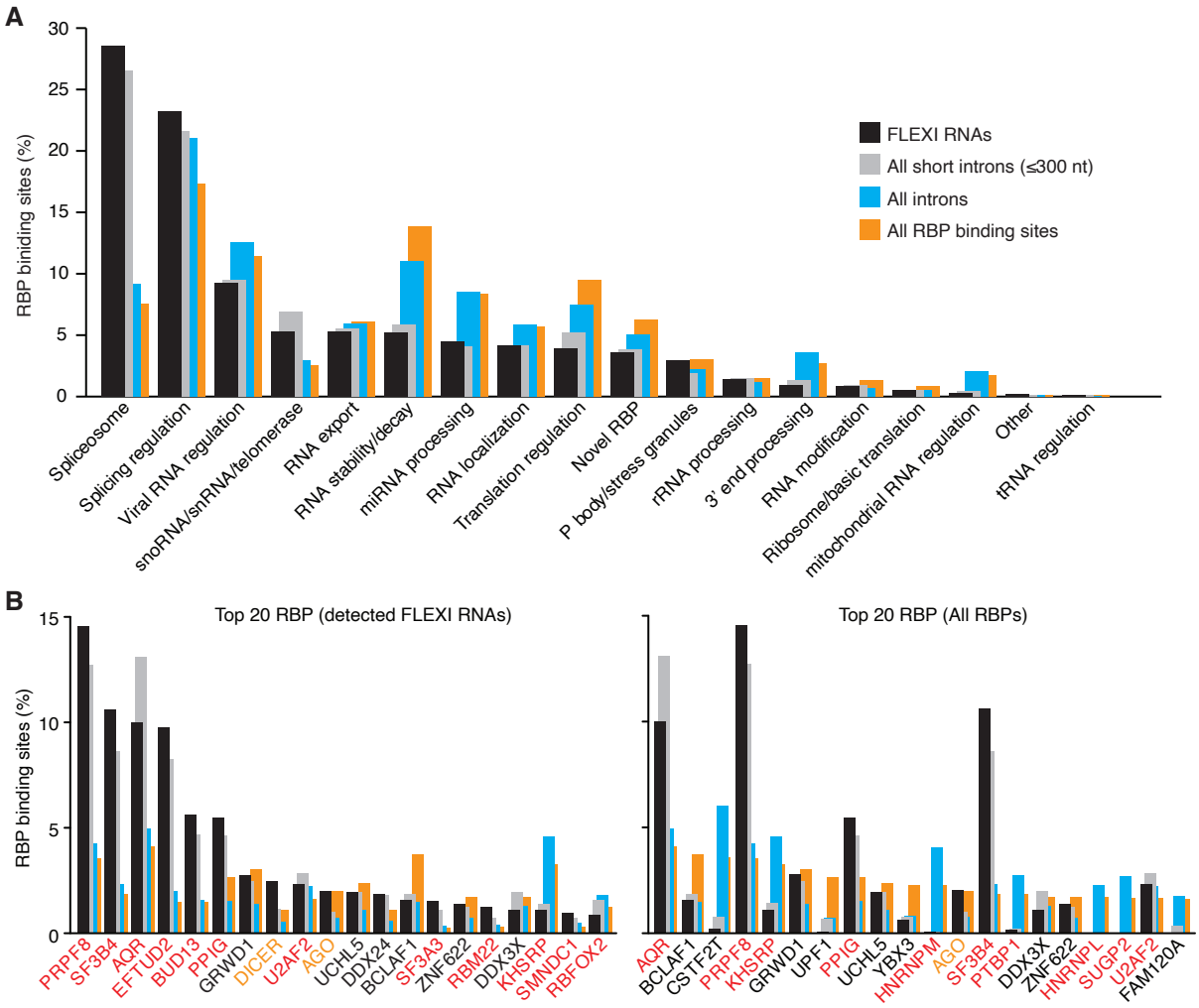


Fig. 5

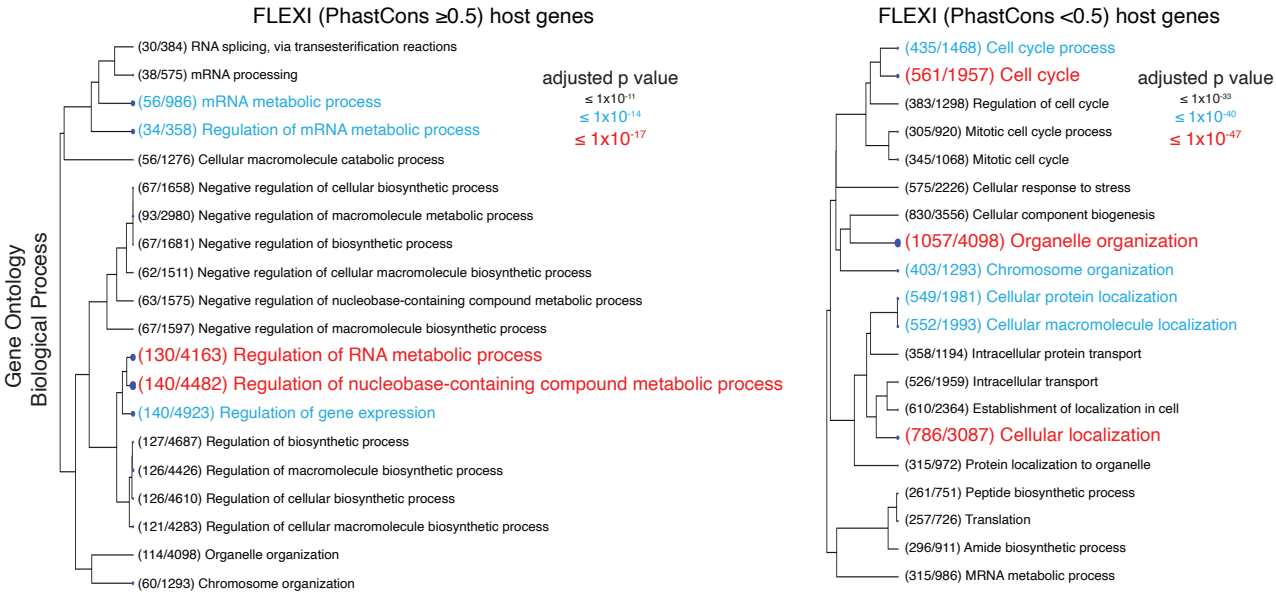


Fig. 6