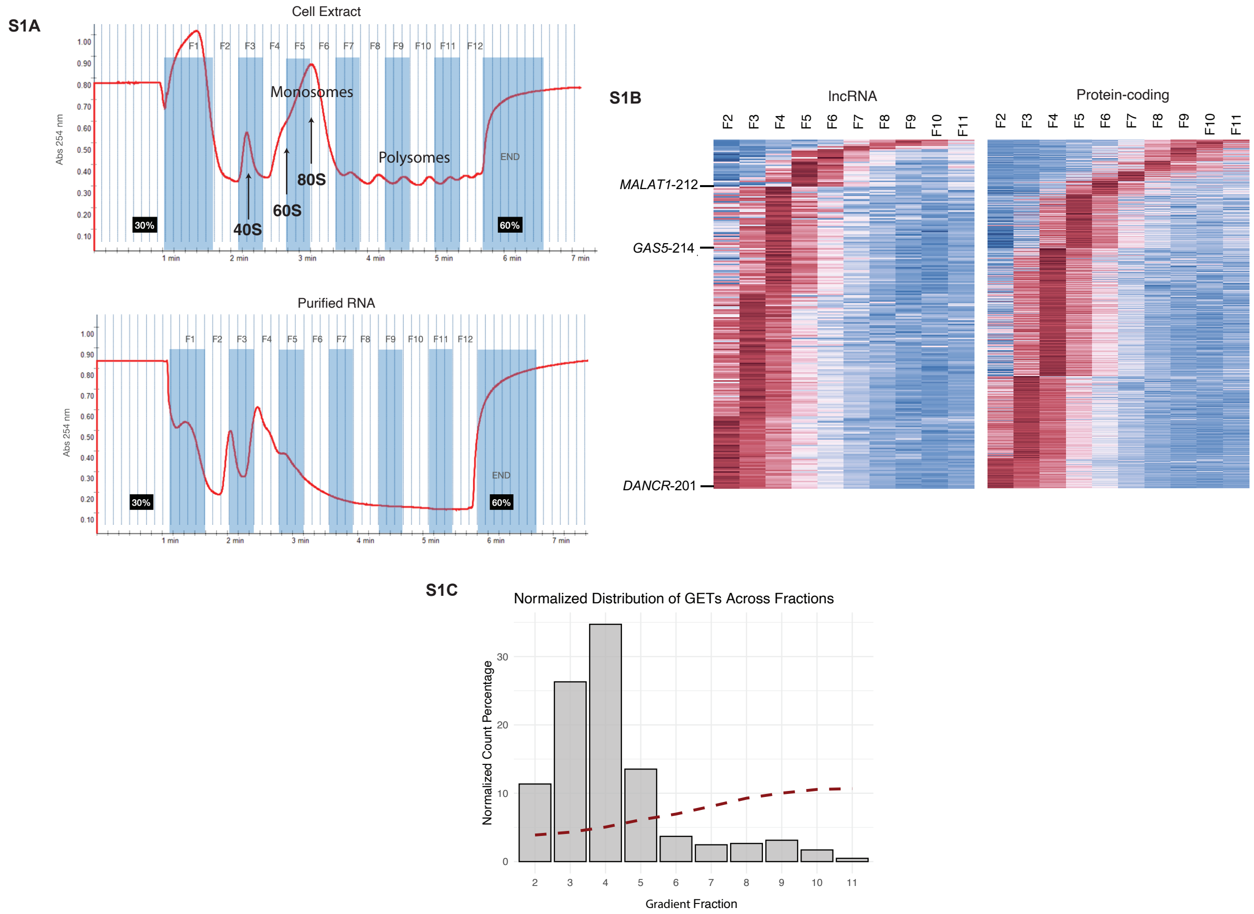
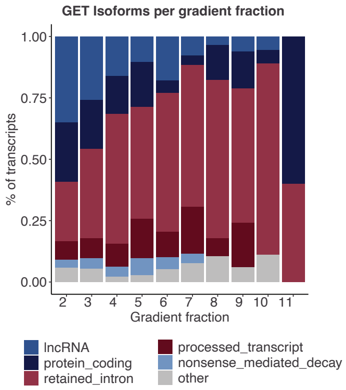
Supplemental material – Pereira et al.

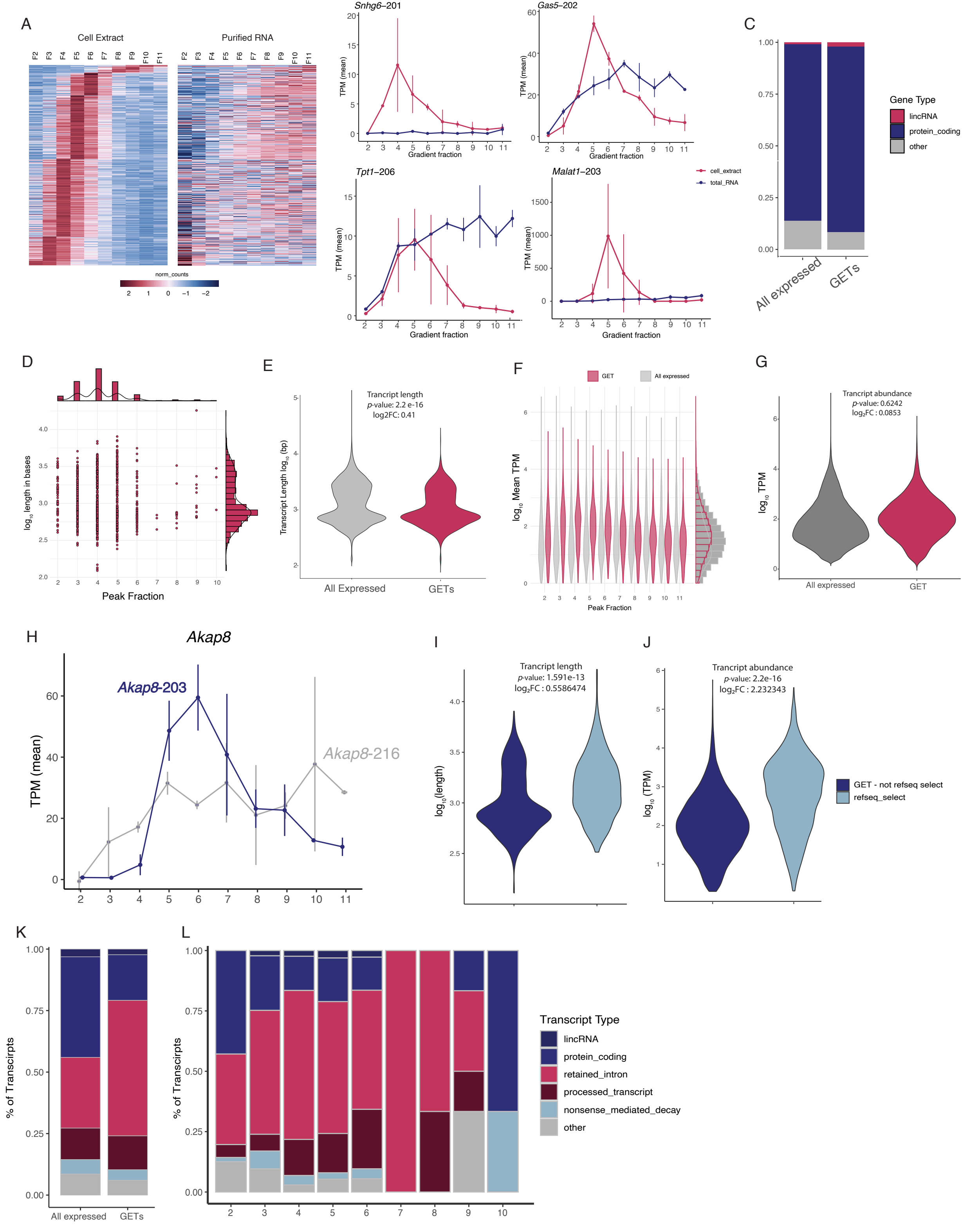


Supplemental Fig S1. (A) Absorbance (254 nm) traces for cell extract and purified RNA gradients and indication of how they were fractionated. Ribosome monosome and polysome peaks can be seen in the cell extract traces but not in the deproteinized RNA. 40, 60 and 80S are indicated. (B) Heatmap representing the abundance (normalized counts) of lncRNA and protein coding gene transcripts in gradient fraction of cell extracts. Columns are gradient fractions; rows are RNA transcripts. (C) Normalized distribution of GETs across gradient fractions. Dashed line represents the standard deviation of the distribution curve (width) of abundance for each GET in the indicated fraction.



Supplemental Fig S2. Relative amount of GET isoform classes based on GENCODE/Ensembl annotation per gradient fraction.

Supplementary Figure S3



Supplemental Fig S3: Identification of Gradient Enriched Transcripts (GETs) in mESCs.

(A) Heatmap representing the abundance (normalized counts) of RNA transcripts in each gradient fraction of cell extracts and purified RNA gradients. Columns are gradient fractions; rows are RNA transcripts. (B) Examples of GETs and their RNA abundance throughout the gradient. (C) Bar plot of percentage of lncRNAs and protein coding genes in All-expressed and GETs. (D) GET length in nucleotides (log10) in each gradient fraction that GETs showed maximum abundance (peak fraction). On the top and side, bar plots represent the number of events. (E) Distribution of GET length in nucleotides (log10) compared to the other expressed transcripts in the cell (All-expressed). Log2FC, fold change (log2) of median values. *P-value*, two-sided t-test. (F) GET abundance in TPM (log10) in peak gradient fraction compared to other expressed transcripts in the cell (All-expressed). On the side, bar plots representing the number of events. (G) Distribution of GET abundance in TPM (log10) compared to the other expressed transcripts in the cell (All-expressed). Log2FC, fold change (log2) of median values. *P-value*, two-sided t-test. (H) RNA abundance throughout the gradient for *Akap8* isoforms. (I) Distribution of GET length in bases and (J) abundance in TPM (log10) compared to RefSeq-select isoforms from the same gene. Log2FC, fold change (log2); *p-value*, Two-sided T-test. (K) Relative amount of transcript isoform classes based on GENCODE/Ensembl annotation in All expressed transcripts and in GETs. (L) Relative amount of GET isoform classes based on Gencode/Ensembl annotation per gradient fraction.

Supplementary Table S1: Results of t-tests and Log2FoldChange analyses comparing the intron properties of GET-RI and validated GET-RI against all expressed RI events. The *Cond* column indicates the specific intron characteristic measured, while the *Group* column specifies whether the value corresponds to the GET-RI or validated GET-RI group.

cond	**log2FC**	**pvalue (t.test)**	**pvalue (wilcoxon)**	**group**
PolypyrimidineTract	-0.2294818	9.80E-01	4.56E-02	GET-RI
IntronLength	-1.5208615	1.88E-66	1.33E-08	GET-RI
PercentageCG	0.00362879	0.136844807	0.1352	GET-RI
IntronDepth	3.58665861	9.70E-08	2.20E-16	GET-RI
IRratio	3.42092282	1.87E-22	2.20E-16	GET-RI
PresenceofMobileElements	-0.3143859	6.79E-01	6.95E-01	GET-RI
Stability	0.05306716	1.23368191766559e-319	2.20E-16	GET-RI
PolypyrimidineTract	0	7.30E-02	4.77E-02	GET-RI validated
IntronLength	-1.0689097	1.51E-14	2.20E-16	GET-RI validated
PercentageCG	0.08755743	3.95E-18	2.20E-16	GET-RI validated
IntronDepth	0.9011979	5.33E-09	2.20E-16	GET-RI validated
IRratio	0.87209278	1.12E-12	2.20E-16	GET-RI validated
Stability	1.36798438	4.87E-62	2.20E-16	GET-RI validated
PresenceofMobileElements	0.53067156	6.38E-01	5.87E-01	GET-RI validated

Supplementary Table S2: Different cutoffs tested to define GETs. The first combination colored in red was used in this study.

Parameters			Number of GETs
Gaussian fit	FoldChange	Correlation to purified RNA (pval)	
s *p*-value < 1e-2 m *p*-value < 1e-6 n *p*-value < 1e-6	10	0.001	1057
s *p*-value < 1e-2 m *p*-value < 1e-4 n *p*-value < 1e-4	10	0.001	7051

s *p*-value < 1e-2 m *p*-value < 1e-2 n *p*-value < 1e-2	10	0.001	14789
s *p*-value < 1e-2 m *p*-value < 1e-6 n *p*-value < 1e-6	4	0.001	2129
s *p*-value < 1e-2 m *p*-value < 1e-4 n *p*-value < 1e-4	4	0.001	11313
s *p*-value < 1e-2 m *p*-value < 1e-2 n *p*-value < 1e-2	4	0.001	23136
s *p*-value < 1e-2 m *p*-value < 1e-6 n *p*-value < 1e-6	10	0.01	904
s *p*-value < 1e-2 m *p*-value < 1e-4 n *p*-value < 1e-4	10	0.01	5668
s *p*-value < 1e-2 m *p*-value < 1e-2 n *p*-value < 1e-2	10	0.01	12166
s *p*-value < 1e-2 m *p*-value < 1e-6 n *p*-value < 1e-6	4	0.01	1867
s *p*-value < 1e-2 m *p*-value < 1e-4 n *p*-value < 1e-4	4	0.01	9405
s *p*-value < 1e-2 m *p*-value < 1e-2 n *p*-value < 1e-2	4	0.01	19297

Supplementary Table S3: List of genes with possible unannotated intron retentions found using IRFinder and manual curation.

Gene
ITFG2
RBIS

CGGBP1
MRLP20
PRDX6
CCDC130
CLDN12
MAT2A
NDUFB1
ORMDL1
MTL12A
PROSER3
PIM2
UMPS
TRAPPC2
SCAMP4
SNPRNP35
UPF3A
SNX5
ZBED5
ZNF224

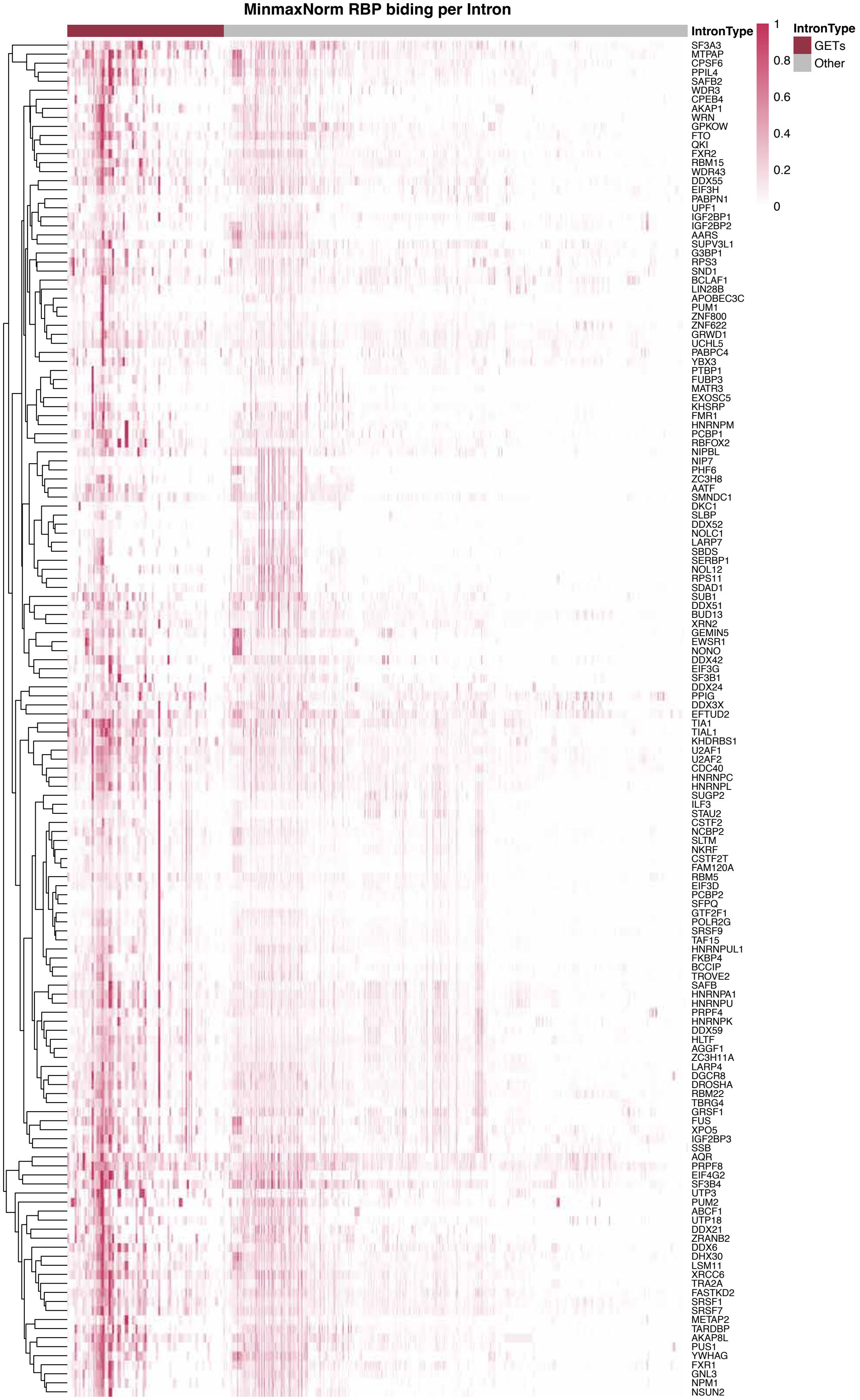

MinmaxNorm RBP biding per Intron
IntronType
1
0.8
0.6
0.4
0.2
0
IntronType
GETs
Other
SF3A3
MTPAP
CPSF6
PPIL4
SAFB2
WDR3
CPEB4
AKAP1
WRN
GPKOW
FTO
QKI
FXR2
RBM15
WDR43
DDX55
EIF3H
PABPN1
UPF1
IGF2BP1
IGF2BP2
AARS
SUPV3L1
G3BP1
RPS3
SND1
BCLAF1
LIN28B
APOBEC3C
PUM1
ZNF800
ZNF622
GRWD1
UCHL5
PABPC4
YBX3
PTBP1
FUBP3
MATR3
EXOSC5
KHSRP
FMR1
HNRNPM
PCBP1
RBFOX2
NIPBL
NIP7
PHF6
ZC3H8
AATF
SMNDC1
DKC1
SLBP
DDX52
NOLC1
LARP7
SBDS
SERBP1
NOL12
RPS11
SDAD1
SUB1
DDX51
BUD13
XRN2
GEMIN5
EWSR1
NONO
DDX42
EIF3G
SF3B1
DDX24
PPIG
DDX3X
EFTUD2
TIA1
TIAL1
KHDRBS1
U2AF1
U2AF2
CDC40
HNRNPC
HNRNPL
SUGP2
ILF3
STAU2
CSTF2
NCBP2
SLTM
NKRF
CSTF2T
FAM120A
RBM5
EIF3D
PCBP2
SFPQ
GTF2F1
POLR2G
SRSF9
TAF15
HNRNPUL1
FKBP4
BCCIP
TROVE2
SAFB
HNRNPA1
HNRNPU
PRPF4
HNRNPK
DDX59
HLTF
AGGF1
ZC3H11A
LARP4
DGCR8
DROSHA
RBM22
TBRG4
GRSF1
FUS
XPO5
IGF2BP3
SSB
AQR
PRPF8
EIF4G2
SF3B4
UTP3
PUM2
ABCF1
UTP18
DDX21
ZRANB2
DDX6
DHX30
LSM11
XRCC6
TRA2A
FASTKD2
SRSF1
SRSF7
METAP2
TARDBP
AKAP8L
PUS1
YWHAG
FXR1
GNL3
NPM1
NSUN2

Figure S4. **Normalized RNA-binding protein (RBP) occupancy across intron types.** The heatmap displays the normalized binding intensities (Min-Max scaled) of RBPs across two intron categories: GET introns and other introns of the same genes. Each row corresponds to an individual RBP, and each column represents a distinct intron. Color intensity reflects the relative binding score, with darker shades indicating higher binding affinity.

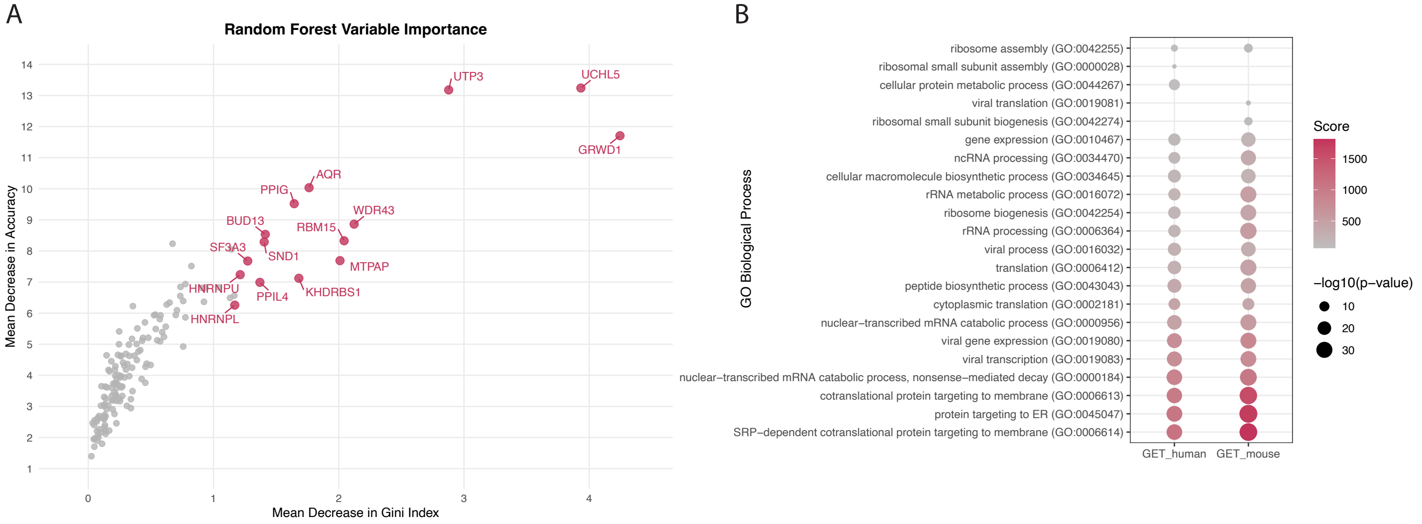


Figure S5: Classification of RBP biding using random forest and functional enrichment of GET transcripts. S5A: Scatter plot of the importance of each RBP in classifying between GET introns and all other introns. Each point represents an individual RBP, with the x-axis indicating the *Mean Decrease in Gini Index* and the y-axis showing the *Mean Decrease in Accuracy*, two metrics reflecting each variable's contribution to model performance. The top 15 RBPs, ranked by Gini importance, are highlighted in red and labeled. S5B: Dotplot of the top 15 pathways, ranked Score, that appeared in the functional enrichment of GET transcripts in human and mouse. On the y axis the GO biological pathways and on the x axis human and mouse results. The radius of each dot represents the -$\log_{10}$ of the *p*-value and the color intensity the score.