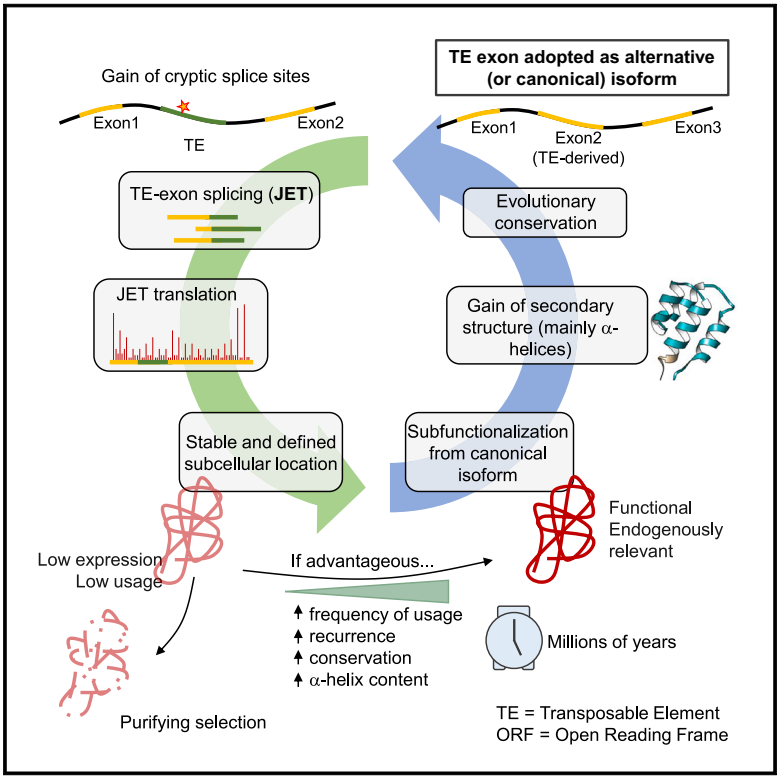
Transposable element exonization generates a reservoir of evolving and functional protein isoforms

Yago A Arribas, Blandine Baudon, Maxime Rotival, Guadalupe Suárez, Pierre Emmanuel Bonté, Vanessa Casas, Apollinaire Roubert, Paul Klein, Elisa Bonnin, Basma Mchich, et al.

Cell

Transposable element exonization generates a reservoir of evolving and functional protein isoforms

Graphical abstract



Highlights

- Transposable element (TE) exonization expands the diversity of the human proteome
- TE-spliced isoforms can have distinct functions compared with canonical isoforms
- Exonized TEs can be evolutionarily conserved and add young sequences to ancient genes
- Exonized TEs contribute to the overall secondary structure of proteins

Authors

Yago A. Arribas, Blandine Baudon, Maxime Rotival, ..., Marianne Burbage, Antonela Merlotti, Sebastian Amigorena

Correspondence

sebastian.amigorena@curie.fr

In brief

Transposable element exonization by unannotated splicing events produces stable protein isoforms with acquired functions that are subject to evolutionary selection.

Arribas et al., 2024, Cell *187*, 7603–7620
December 26, 2024
https://doi.org/10.1016/j.cell.2024.11.011

CellPress

Cell

CellPress
OPEN ACCESS

Article

Transposable element exonization generates a reservoir of evolving and functional protein isoforms

Yago A. Arribas,[1] Blandine Baudon,[1] Maxime Rotival,[2] Guadalupe Suárez,[1] Pierre-Emmanuel Bonté,[1] Vanessa Casas,[3] Apollinaire Roubert,[1] Paul Klein,[4,5] Elisa Bonnin,[1] Basma Mchich,[6] Patricia Legoix,[7] Sylvain Baulande,[7] Benjamin Sadacca,[1,4,5] Julien Diharce,[6] Joshua J. Waterfall,[4,5] Catherine Etchebest,[6] Montserrat Carrascal,[3] Christel Goudot,[1] Lluís Quintana-Murci,[2,8] Marianne Burbage,[1,9] Antonela Merlotti,[1,9] and Sebastian Amigorena[1,9,10,*]

[1]Institut Curie, PSL University, Inserm U932, Immunity and Cancer, 75005 Paris, France
[2]Institut Pasteur, Université Paris Cité, CNRS UMR2000, Human Evolutionary Genetics Unit, 75015 Paris, France
[3]Biological and Environmental Proteomics, Institut d'Investigacions Biomèdiques de Barcelona-CSIC, IDIBAPS, Roselló 161, 6a planta, 08036 Barcelona, Spain
[4]INSERM U830, PSL Research University, Institute Curie Research Center, Paris, France
[5]Department of Translational Research, PSL Research University, Institut Curie Research Center, Paris, France
[6]Université Paris Cité and Université de la Réunion and Université des Antilles, INSERM, BIGR, DSIMB UMR_S1134, 74014 Paris, France
[7]Institut Curie, Centre de Recherche, Genomics of Excellence Platform, PSL Research University, Paris Cedex 05, France
[8]Chair Human Genomics and Evolution, Collège de France, 75005 Paris, France
[9]These authors contributed equally
[10]Lead contact
*Correspondence: sebastian.amigorena@curie.fr
https://doi.org/10.1016/j.cell.2024.11.011

SUMMARY

Alternative splicing enhances protein diversity in different ways, including through exonization of transposable elements (TEs). Recent transcriptomic analyses identified thousands of unannotated spliced transcripts with exonizing TEs, but their contribution to the proteome and biological relevance remains unclear. Here, we use transcriptome assembly, ribosome profiling, and proteomics to describe a population of 1,227 unannotated TE exonizing isoforms generated by mRNA splicing and recurrent in human populations. Despite being shorter and lowly expressed, these isoforms are shared between individuals and efficiently translated. Functional analyses show stable expression, specific cellular localization, and, in some cases, modified functions. Exonized TEs are rich in ancient genes, whereas the involved splice sites are recent and can be evolutionarily conserved. In addition, exonized TEs contribute to the secondary structure of the emerging isoforms, supporting their functional relevance. We conclude that TE-spliced isoforms represent a diversity reservoir of functional proteins on which natural selection can act.

INTRODUCTION

Advances in proteogenomics led to the discovery of previously unannotated proteins,[1–3] generated by different mechanisms. Open reading frames (ORFs) from non-genic regions encode microproteins or short ORFs.[4,5] Variants of existing proteins can be generated through splicing or gene duplication,[6] while protecting the functions of the main isoform and probing emerging variants. Cryptic splice sites emerge in the vicinity of genes through random mutations, providing opportunities for the splicing machinery to generate evolutionarily recent exons and transcripts.[7–9] If they are evolutionarily advantageous, such emerged splicing sites can be fixed in the population and become alternative, and even constitutive, spliced isoforms.

A significant proportion of these recently acquired splice signals are located within intronic transposable elements (TEs).[10,11] TEs, which cover around 45% of the genome, are mobile entities[12] that represent a reservoir of polymorphic sites upon which evolution can act.[13,14] Mammalian TEs are divided into four main classes, including DNA transposons, short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), and long terminal repeats (LTRs). TE contribution to the human proteome has been a matter of debate.[15–17] *De novo* TE exonization has been studied in cancer and healthy tissue samples, based on the detection of unannotated splicing junctions between exons and TEs (JETs).[18–22] Most JETs are sample specific, whereas a few percent are shared across individuals. Some JETs are only detected in tumors and rarely observed across healthy tissue samples. We have recently shown that recurrent, tumor-associated JET-derived peptides are presented by major histocompatibility complex class I (MHC

class I) molecules to T lymphocytes, in both humans and mice.[18,23]

Although unannotated spliced transcripts are, in part, recurrent between individuals and produce peptides for MHC class I presentation, they could still represent "random mistakes" that lead to rapidly degraded proteins.[15] Alternatively, unannotated spliced isoforms could represent a "diversity reservoir" for natural selection.[9,24] If they do, these spliced transcripts must encode stable protein isoforms that can be assayed for cellular functions. To test this hypothesis, we coupled ribosome profiling (Ribo-seq) and proteomics to investigate if JETs that are non-annotated as protein coding in Ensembl (i.e., unannotated JETs) can generate functional protein isoforms. We show that some of these JET-derived ORFs (JET-ORFs) are stable, localize to defined subcellular compartments, and have modified functions compared with canonical (CAN) isoforms. Evolutionary and structural analyses show that JET-ORFs derive preferentially from evolutionarily ancient genes, whereas the exonized TE sequences are relatively younger and can adopt secondary structures enriched in α helices, contributing to the secondary structure of the emerging isoforms. We conclude that TE-spliced isoforms represent a reservoir of evolving proteins under ongoing selective pressures.

RESULTS

JETs can be recurrent and present tissue-dependent expression

We identified splicing JETs in The Cancer Genome Atlas (TCGA, n = 9,191, Figure 1A) as in Merlotti et al.[18] 22,947 JETs are detected in at least 2 patients (n = 6–512, average = 236, Figure S1A). Although most JETs are not recurrent, 2,506 JETs are shared by at least 1% of the samples and 467 by at least 10% (Figure 1B). JET recurrence distribution is similar in all TCGA cancer indications (examples in Figure S1B). JETs expressed in over 10% can be preferentially expressed in certain cancer types or expressed across all indications (Figure 1C). Uniform manifold approximation and projection (UMAP) visualization based on JET expression clusters TCGA samples according to the tissue of origin (Figure 1D). Recurrent JETs are also present in tumor-adjacent normal tissues from TCGA (n = 679, Figure S1C). JET recurrence in tumors is correlated with recurrence in normal tissues (R^2 = 0.88, Figure S1D) and most JETs are found in both tumor and healthy tissues. To validate recurrent JETs in an independent cohort, we used the Cancer Cell Line Encyclopedia (CCLE), which contains RNA sequencing (RNA-seq) data from 1,019 cell lines. The overlap between datasets increases proportionally with JET recurrence (Figure S1E). To confirm that JETs derive from TE exonization rather than genomic rearrangements, we validated the presence of 17 recurrent JETs by PCR at the RNA level, which are not identified using genomic DNA (Figure S1F). We conclude that a subpopulation of JETs that is recurrent across individuals and independent sample cohorts presents tissue-dependent expression profiles.

Recurrent JETs have, on average, 10-fold lower expression levels than CAN (according to Gencode annotation) junctions (Figure 1E), and the two are not correlated (Figure S1G). The ratio between each JET and the corresponding CAN junction is variable and can be conserved or not between tissues (see two examples in Figure 1F). For example, the chrX:54469833 splice site in *TSR2* has two junctions: a major exon-exon junction and a JET, which corresponds to 1% of all reads mapping to the splice site (each dot represents a TCGA indication). For the chr1:217915317 splice site in *SPATA17*, the JET proportion is variable across tissues.

The proportion of exon-exon junctions and recurrent JETs among all reads of a given splicing breakpoint display a bimodal and symmetric distribution (Figure 1G, top). By contrast, most JETs are found at low proportions (Figure 1G, bottom). A total of 572 JETs, however, contribute to over 50% of all splicing junctions for the corresponding exon in at least one TCGA indication (examples in Figures S1H and S1I). Expression levels are higher for JETs with a higher proportion among all splicing reads (Figure S1J). No correlation, by contrast, is observed between the proportion of JETs among all splicing reads and their recurrence in TCGA samples (Figure 1H). JETs that represent a few percent of all splicing events for a given exon can be as recurrent as JETs representing much higher proportions. This analysis revealed a population of low abundance, unannotated splicing junctions containing exonized TEs that can be tissue associated and recurrent across individuals.

JET splice sites can be conserved and are a preferential source of alternative splicing

To address the evolutionary implications of TE exonization, we investigated the age and conservation of the TE splice sites involved in JETs. We used age (million years old [mya]) and conservation values of a curated collection of 782,112 splice sites from Rotival et al.[25] The splice sites were dated mapping the first ancestor with an essential AG/GT sequence. In parallel, conservation is defined by the difference between the observed rate of nucleotide substitutions across mammals and the expected rate under neutral evolution (genomic evolutionary rate profiling rejection substitution [GerpRS]). From the 22,947 JETs identified in TCGA, 8,534 overlap with a splice motif of the dataset (37%). Overall, as expected, there is a correlation between conservation and age (Figure 2A, middle): old splice motifs are generally conserved (e.g., in CDC-like kinase 4 [CLK4], the TE and splice sites were acquired simultaneously 105 mya, Figure 2A, left), whereas recent splice sites tend to be non-conserved and evolve under neutrality (e.g., SZRD1, the TE is 140 mya with a recently acquired splice motif 8 mya). Some splice sites, however, are old but non-conserved, or recent and conserved (examples from the 4 categories are shown in Figure 2A). A JET in CPSF1, for example, bears a splice motif less than 5 mya that is highly conserved (GerpRS = 3.7), suggesting a recent repurposing of the function of the associated JET (Figure 2A, right).

To characterize the evolutionary features of JET splice sites, all splice sites were categorized based on their overlap or not with a TE (i.e., TE and noTE splice site), and on the transcript type annotation from Gencode (i.e., protein-coding transcript [PCod], not protein-coding transcript [notPCod], or not annotated [unAnnot], Figure 2B). JET splice sites were annotated in an independent category (TE/JET). Most splice sites do not overlap TEs (94.75%), and only 6.13% (5.95% noTE + 0.18% TE) correspond to already annotated protein-coding regions (Figure 2B, left).

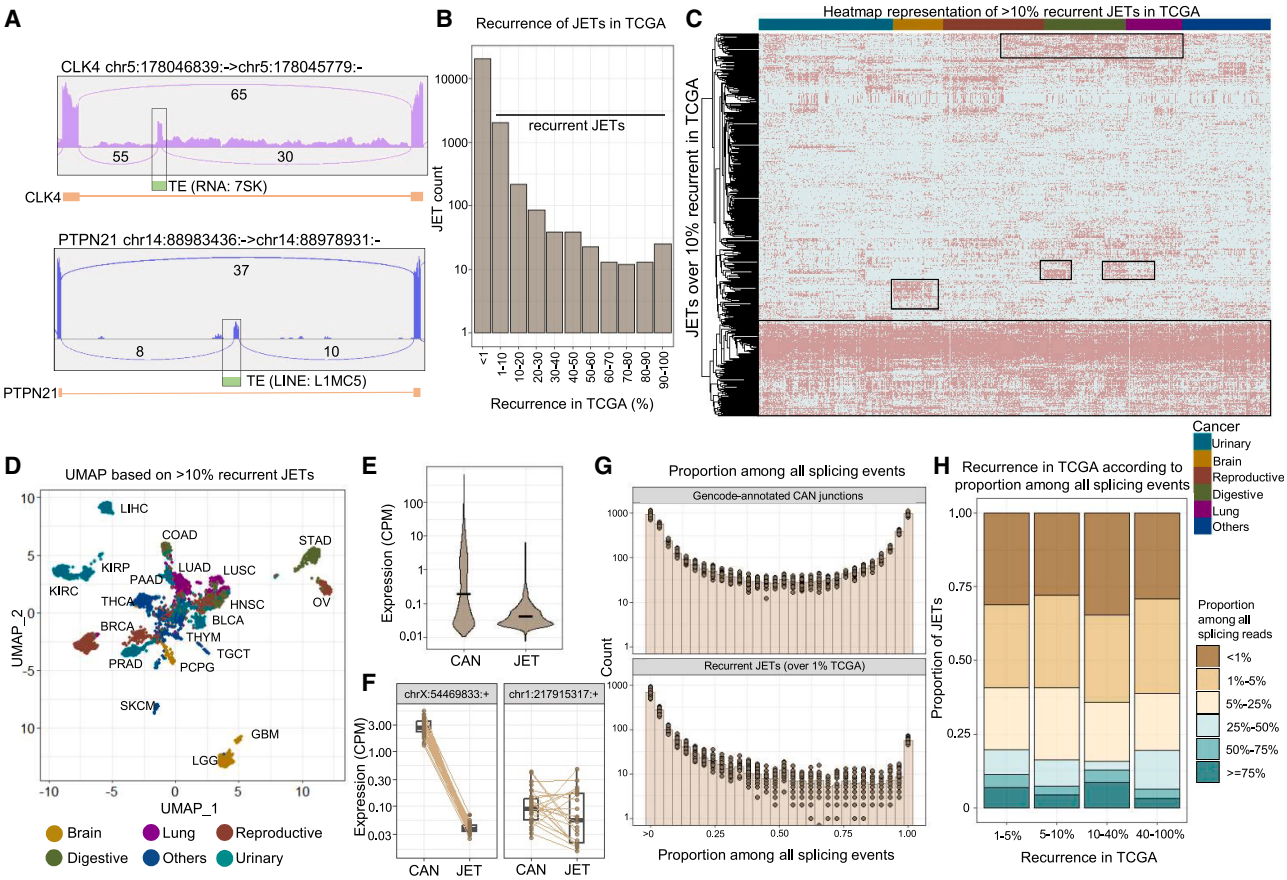


Figure 1. JETs can be recurrent and present tissue-dependent expression

(A) Splicing junctions between exons (orange) and TE (green).

(B) JET recurrence across TCGA tumors. Only JETs present in at least 2 samples are included.

(C) Heatmap of JETs recurrent in more than 10% of TCGA tumors. Squares highlight tissue-associated or ubiquitously expressed JETs.

(D) UMAP representation based on the expression of >10% recurrent JETs.

(E) Expression of canonical junctions and JETs. The median is indicated.

(F) Expression of canonical junctions and JETs sharing the exon splice site (paired according to the TCGA indication). Dots indicate the average expression per indication.

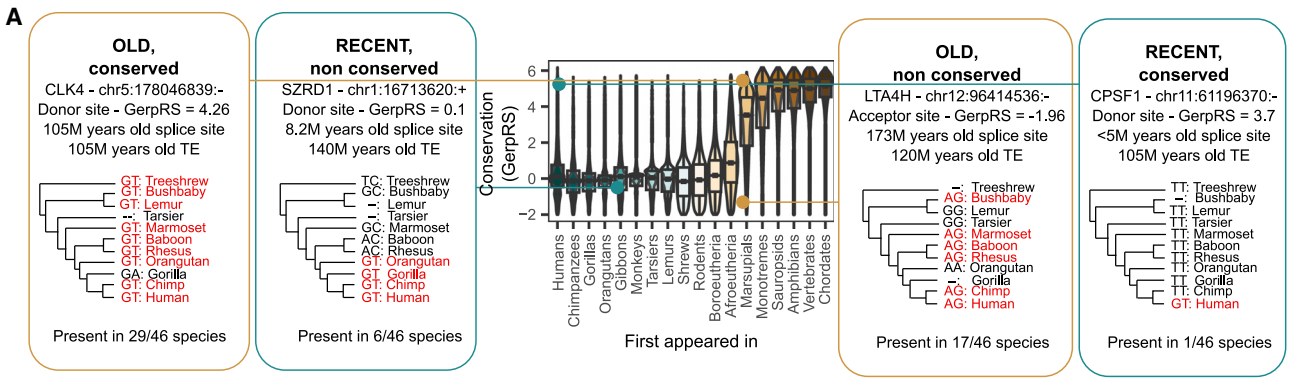
(G) Average proportion of canonical (top) or JET (bottom) junctions among all splicing reads overlapping the same breakpoint per TCGA indication (dots).

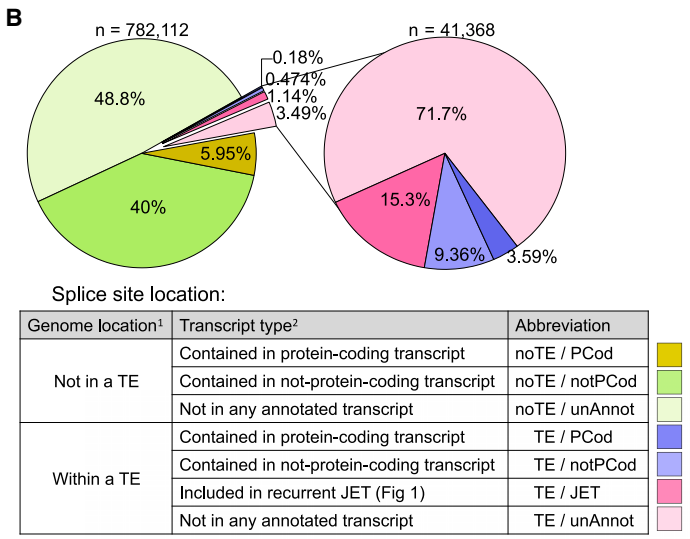
(H) JET proportion among all splicing reads grouped according to their recurrence.

See also Figure S1.

Among the 41,368 splice sites that do overlap a TE (Figure 2B, right), 71.7% are not annotated (TE/unAnnot.), 15.3% correspond to TE/JETs, and 9.36% are present in non-coding annotated transcripts (TE/notPCod). The remaining 1,485 (3.59%) TE splice sites are present in known and annotated protein-coding transcripts (TE/PCod), probably corresponding to already fixed, exonized TEs in the CAN proteome. TE splice sites are overall younger than non-TE splice sites in all categories (Figure 2C). TE/PCod splice sites, however, are older and more conserved than the other TE splice sites (Figures 2C and S2A, respectively). TE splice sites are also overall less conserved (Figure S2A), consistent with their lower conservation being driven by their young age. Nevertheless, the conservation of TE/JET splice sites increases proportionally with their recurrence in TCGA (Figure 2D). The percentage of conserved splice sites of JETs recurrent in over 40% of TCGA samples is closer to the proportion of conserved TE/PCod splice sites (11.11% versus 17.14%, respectively). Splice sites of more recurrent JETs (>40% TCGA) are also slightly older than less recurrent JET splice sites (Figure S2B). We conclude that splice sites in exonized TEs, both annotated and JETs, have been acquired relatively recently in evolution and that recurrently expressed JETs are more evolutionary conserved than non-recurrent JETs.

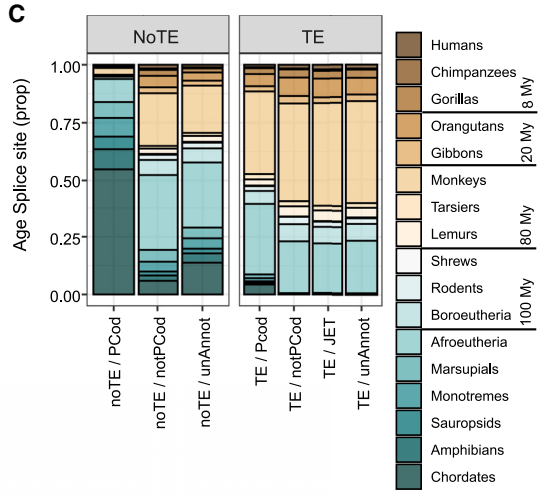
The density distribution of splice sites in the Rotival et al. dataset[25] has a bimodal usage distribution, with a clear population of splice motifs that are highly used (i.e., representing over 50% of splicing events, Figure S2C). When categorizing the splice sites based on their overlap in TEs and transcriptomic annotation (as in Figure 2B), TE/PCod and TE/notPCod splice sites display higher frequency of usage (close to 100%) compared with other





Splice site location:

Genome location[1]	Transcript type[2]	Abbreviation
Not in a TE	Contained in protein-coding transcript	noTE / PCod
	Contained in not-protein-coding transcript	noTE / notPCod
	Not in any annotated transcript	noTE / unAnnot
Within a TE	Contained in protein-coding transcript	TE / PCod
	Contained in not-protein-coding transcript	TE / notPCod
	Included in recurrent JET (Fig 1)	TE / JET
	Not in any annotated transcript	TE / unAnnot



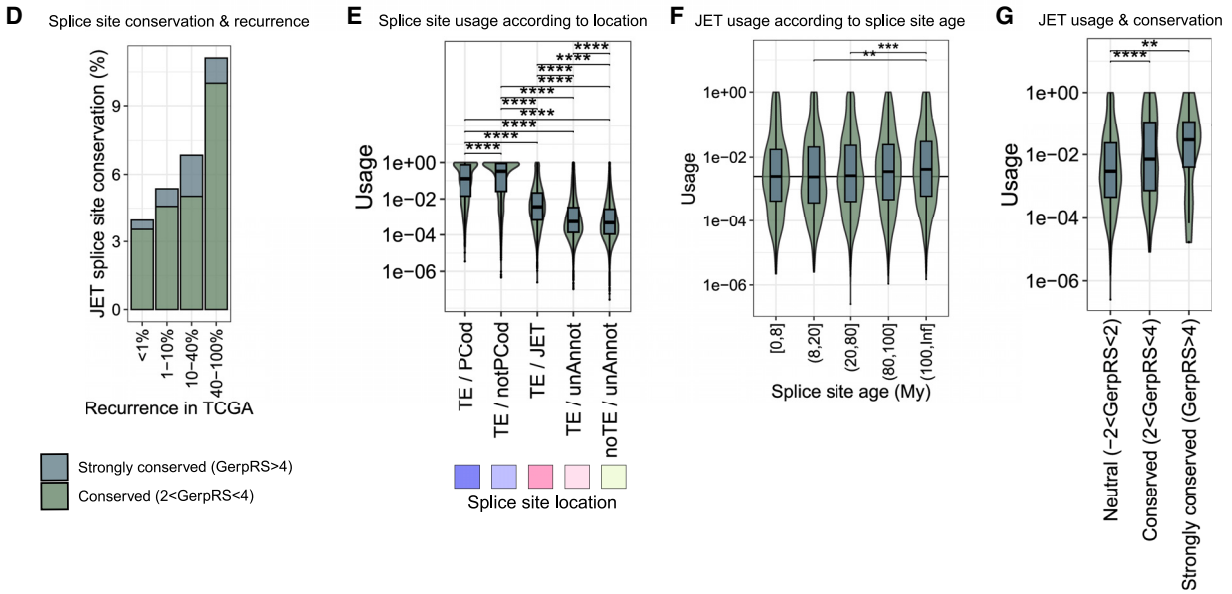


Figure 2. JET splice sites can be conserved and are a preferential source of alternative splicing

(A) Splice site age and conservation (GerpRS).

(B) Splice site classification considering their overlap with an annotated TE in genome and their inclusion in any annotated transcript. JET splice sites are classified separately.

(C) Splice site age based on the previous classification.

(legend continued on next page)

splice sites in TEs, probably corresponding to constitutive splicing isoforms (Figure 2E). Moreover, JET splice sites are used at intermediate frequencies between the annotated (i.e., TE/PCod and TE/notPCod) and not annotated splice sites (both TE/unAnnot and noTE/unAnnot). JET usage increases slightly, but consistently, with the age of the splice sites (Figure 2F), indicating that JET usage rises during evolution (which could correspond to ongoing evolutionary pressures). Furthermore, conserved and strongly conserved JET splice sites (GerpRS > 2) are more frequently used than neutrally evolving JET splice sites (0 < GerpRS < 2, Figure 2G). Annotated TE/PCod splice sites also have increasing frequencies of usage according to the age and conservation of the splice site (Figures S2D and S2E, respectively). These results suggest that more recurrent and more used JETs are subjected to ongoing selective pressures.

Recurrent JETs can be translated and encode unannotated protein isoforms

To investigate if recurrent JETs are translated into unannotated protein isoforms, we combined transcriptome assembly, Ribo-seq, and mass spectrometry (MS)-based proteomics (Figure 3A). Among the 12,953 JETs that expressed at least in 1% of TCGA and/or CCLE samples, 3,801 were assembled into transcripts using RNA-seq datasets from 5 human cell lines (Table S1). These transcripts were interrogated in Ribo-seq datasets from 5 cell lines (Table S1), and translated ORFs were blasted against RefSeq (a curated protein database). Only ORFs with less than 95% similarity and/or with a gain of at least 5 amino acids were selected. A total of 1,227 unannotated JET-ORFs were identified, corresponding to 807 unique translated JETs (one JET can be present in more than one ORF, Table S2). Each cell line expresses between 296 and 644 JET-ORFs (Figures S3A and S3B). Examples of uniquely mapping P-site profiles are shown in Figure S3C. Exonized TEs were categorized according to their position within the ORF (start, internal, and end, Figure 3B). Although TEs located in the start (5′ end) and internal exons of the ORF represent 17.29% and 17.54% of the JET-ORFs, respectively, over 65% of the identified ORFs (n = 810) contain the TE at the 3′ end (and introduce a stop codon, Figure 3C). JET-ORFs are overall shorter than the corresponding CAN-ORFs (Figure 3D). Although JET-induced exons are globally also shorter than CAN exons, internal JET-induced exons are longer than those that either start or end the JET-ORF (Figure 3E). We conclude that recurrent JETs can be translated into unannotated ORFs typically leading to shorter protein isoforms.

To comprehend the proportion of translated recurrent JET-ORFs compared with all JETs expressed in a certain cell line (recurrent and sample specific), we used RNA-seq and Ribo-seq from the H1650 cell line (Table S3). In total, 2,783 JETs were identified in H1650, from which 2,175 (78.15%) were recurrent in more than 1% of samples TCGA and/or CCLE (as defined in Figure 3A). Among all expressed JETs, 815 were assembled in at least one transcript, and 96% of the assembled JETs were recurrent. These JET transcripts were interrogated in the H1650 Ribo-seq to identify 191 unique translated JETs (all of them recurrent). Therefore, JETs can be consistently identified across different technologies in a cell line-based approach, and recurrent JETs preferentially generate unannotated isoforms, compared with sample-specific JETs.

To investigate if JET-ORFs are efficiently translated, we compared the expression levels of translated JETs and the corresponding CAN junctions at the RNA-seq and Ribo-seq levels in H1395 and H1650 cell lines. Translated JETs had, on average, lower expression levels than CAN junctions in both RNA-seq and Ribo-seq (Figure 3F). Some JETs, however, were better detected by Ribo-seq than by RNA-seq, suggesting differences in the translation efficiency (examples shown in Figure 3G). Overall, JETs and CAN junctions had similar translation efficiencies (Figure S3D). These results indicate that most identified JET-ORFs are efficiently translated and generate unannotated protein isoforms.

To explore translated JET-ORFs at the protein level, we used MS-based proteomics on six cell lines[26] generated using deep fractionation and six proteases (trypsin, LysC, LysN, AspN, GluC, and chymotrypsin), achieving high coverage of the proteoform diversity. On average, 17,677 total proteins (CAN + JET-ORFs) were identified by each digestion protocol (Figure S3E). The identified peptides were filtered with two human proteome databases, Swissprot and Refseq Curated (Figure 3A). A total of 153 peptides map uniquely to a JET-ORF (and not to the corresponding CAN-ORF), derived from 112 unique JET-ORFs (Figure 3H). Among these, 27 JET-ORFs (24%) are identified with at least two JET-ORF-specific peptides (Figure S3F). The remaining 76% of JET-ORFs are only defined by one peptide, indicating that JET-ORF identification highly relies on enzymatic digestions (Figure S3G). MS-identified peptides mapping to huntingtin (HTT) and death-associated protein 1 (DAP) JET-ORFs are shown in Figure 3I. Similar proportions of the TE location within the ORF (start, internal, and end) are observed in MS-identified JET-ORFs (Figure S3H) compared with the whole JET-ORF population (Figure 3C). Despite the limitations of shotgun proteomics for protein isoform identification, JET-ORFs can be detected by MS-based proteomics. These results provide additional evidence that unannotated isoforms derived from TE exonization contribute to the cell proteome.

JET-ORFs can be stable and localize to distinct subcellular compartments

To evaluate JET-ORF stability, we predicted their instability index (based on the primary amino acid sequence[27]). JET-ORFs have a slightly increased predicted instability compared with CAN-ORFs (Figure 4A, 50.1 versus 48.8, p value < 0.05). Nevertheless, 99.4% of JET-ORFs fall within the ranges of the

(D) Percentage of conserved (2 < GerpRS < 4) and strongly conserved (GerpRS > 4) JET splice sites according to their recurrence.
(E) Usage (frequency) among different types of splice sites.
(F and G) JET splice site usage according to their age (F) and conservation (G). *p < 0.05, **p < 0.01, ***p < 0.001, and ****p < 0.0001 (Mann-Whitney test with Bonferroni adjustment).
See also Figure S2.

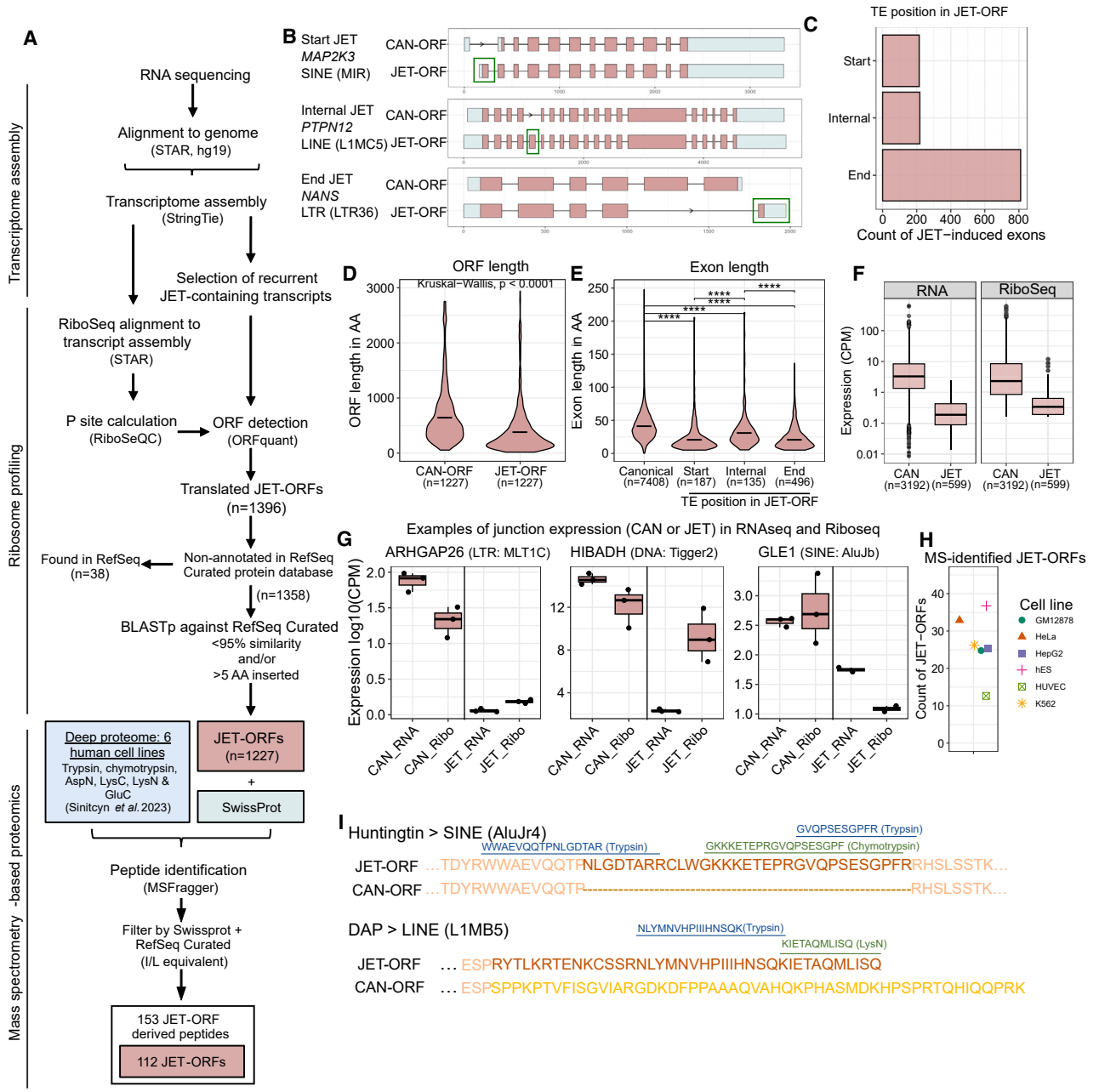


Figure 3. Recurrent JETs can be translated and encode unannotated protein isoforms

(A) JET-ORF identification workflow.

(B) JET-ORF transcripts and the corresponding RefSeq-annotated CAN-ORF. The color of the exon indicates whether it is included in the ORF (pink) or not translated (gray). JET-induced exons are highlighted in green.

(C) Count of JET-ORFs according to JET position within the ORF.

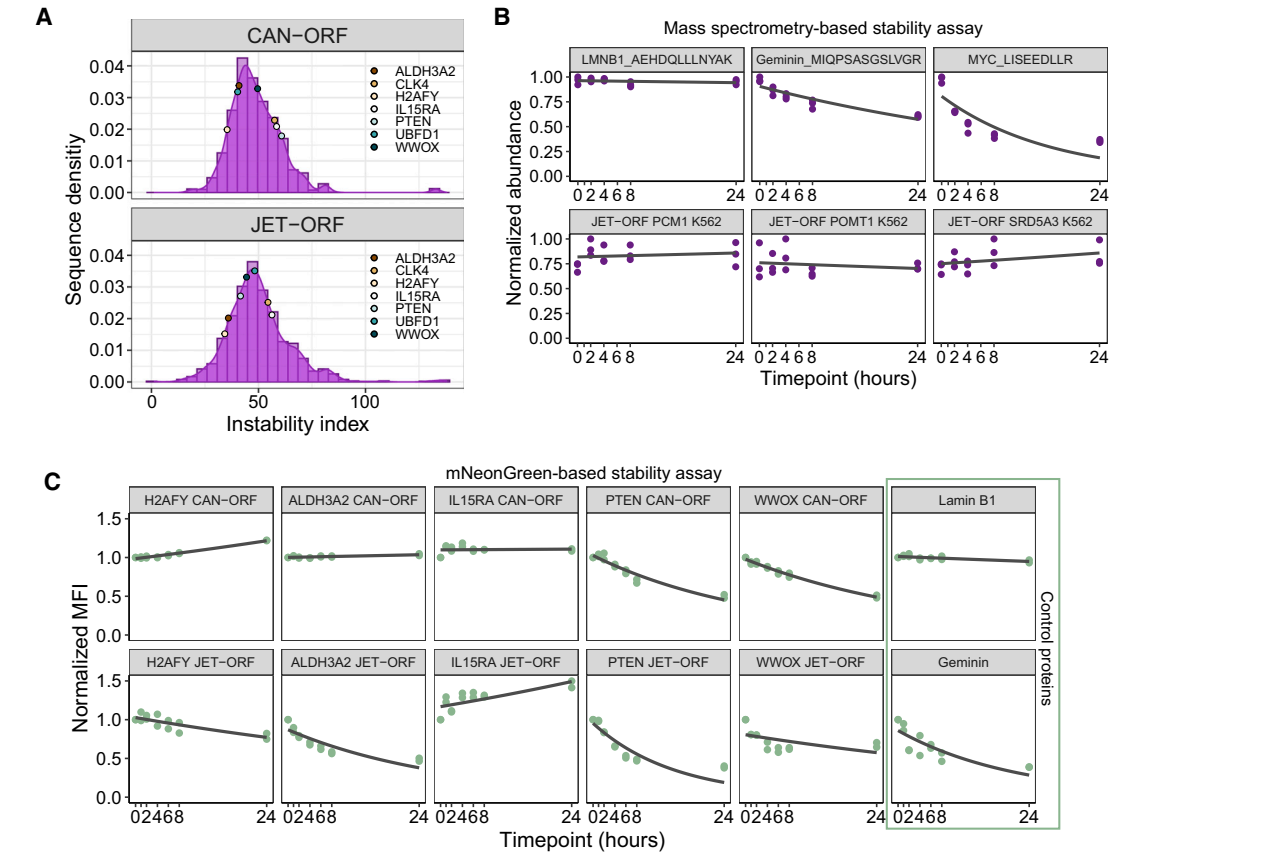
(D and E) Length of JET-ORFs and the RefSeq-annotated CAN-ORFs (D) and according to their position within the ORF (E). ****$p < 0.0001$ (Mann-Whitney test with Bonferroni adjustment).

(F and G) Expression of JETs and canonical junctions in RNA-seq and Ribo-seq (F). Individual examples are shown in (G).

(H) Count of MS-identified JET-ORFs per cell line.

(I) MS-identified peptides mapping to Huntingtin and DAP JET-ORFs.

See also Figure S3.



(legend on next page)

predicted instability for CAN-ORFs (16.89–132.87). To experimentally validate the overall stability of JET-ORFs, we measured protein half-life by coupling quantitative MS and timed translation arrest (adapted from Li et al.[28]). We combined the in-house data on K562 cell line with the publicly available dataset on HCT116, HEK293, RPE1, and U2OS cell lines.[28] To validate the assay, we quantified the stability of three short-lived control proteins: geminin, PRELID3B, and MYC (Figures 4B, top, and S4A).[28–31] As control for stable proteins, lamin B1 is highly stable across the evaluated time points[28] (Figures 4B, top, and S4A). Among the MS-identified JET-ORFs, seven were additionally quantified, and their normalized abundances indicate little or no protein degradation (Figures 4B, bottom, and S4A), supporting that JET-ORFs can be stable.

To directly address the stability of JET-ORFs, we selected 7 JET-ORFs based on their JET recurrence and functions of the CAN isoforms: *H2AFY*, *ALDH3A2*, *IL15RA*, *CLK4*, *UBFD1*, *PTEN*, and *WWOX* genes (Data S1). Except for the interleukin (IL)-15RA and ubiquitin family domain containing 1 (UBFD1) JET-ORFs, in which TEs are internal, the other JET-ORFs contain TE insertions at the end of the ORF (Figure S4B). TE insertions in CLK4, phosphatase and tensin homolog (PTEN), and WW domain-containing oxidoreductase (WWOX) lead to truncated proteins; in H2AFY and ALDH3A2, the TE substitutes the last region of the CAN-ORF, resulting in an ORF of similar length. The endogenous expression of these JETs was quantified by RNA-seq and/or RT-qPCR in two cell lines and compared with the CAN exon-exon junctions (Figures S4C and S4D). To experimentally validate the stability of the selected candidates, they were ectopically expressed using mNeonGreen (mNG) split fluorescence system[32] (Figure S4E; Table S4). JET-ORFs were tagged with the mNG11 fragment, and complementation with the remaining mNG1-10 fragments constitutively expressed by host cells (HeLa mNG1-10 and K562 mNG1-10) was detected by flow cytometry at baseline (Figure S4F) or after cycloheximide blocking of translation at different time points (Figures 4C and S4G). The experimental model was validated with known stable (lamin B1-mNG11, Figure 4C) and short-lived (geminin-mNG11, Figure 4C, and PRELID3B-mNG11, Figure S4G) control proteins. mNG11-tagged IL-15RA, UBFD1, CLK4, PTEN, and WWOX JET-ORFs display similar degradation profiles compared with the corresponding CAN-ORFs (Figure 4C). By contrast, H2AFY and ALDH3A2 JET-ORFs are slightly less stable than the CAN-ORFs but still within the ranges of stability of CAN control proteins. In conclusion, most JET-ORFs are stable at levels similar to CAN proteins.

H2AFY histone JET-ORF, similar to CAN-ORF, is found in the nucleus and colocalizes with chromatin in both interphase and mitosis (Figures 4D and S4H).[33] ALDH3A2 JET-ORF localized to the endoplasmic reticulum, as observed for CAN-ORF (Figure 4E).[34] TE exonization does not impair the ALDH3A2 transmembrane domain, according to TMHMM predictor (Figure S4I). Additionally, membrane isolation confirmed the transmembrane insertion of ALDH3A2 JET-ORF (Figure S4J). IL-15RA JET-ORF also preserves the transmembrane domain and is expressed at the plasma membrane (Figure 4F).[35,36] By contrast, JET-ORFs from the two tumor suppressors, PTEN and WWOX, display different subcellular locations compared with the CAN-ORFs (Figures 4G and 4H, respectively). Although PTEN CAN-ORF is mainly cytosolic, and WWOX CAN-ORF localizes to the Golgi, both JET-ORFs are enriched in the nucleus. Quantification of the mean fluorescence intensity (MFI) ratio between nuclei and total cell confirms the nuclear enrichment of PTEN and WWOX JET-ORFs (Figure 4I). Additional examples include CLK4, on which both JET-ORF and CAN-ORF localize to the nucleus (Figure S4K), and the UBFD1, on which the JET-ORF is restricted to the nucleus, whereas CAN-ORF has ubiquitous localization (Figure S4L). We conclude that JET-ORFs can be stable and localize to specific subcellular locations that can be similar to or different from the localizations of the corresponding CAN-ORFs.

JET-ORFs are functional and can have divergent functions compared with CAN isoforms

To be subject to natural selection, JET-ORFs should have modified biological functions compared to the corresponding CAN-ORFs. To investigate acquired functions (neo-functionalization), we selected four JET-ORFs: WWOX, UBDF1, HTT, and PTEN JET-ORFs.

WWOX JET-ORF conserves the WW1 and WW2 domains and the nuclear translocation signal (Figure 5A). WW-domains interact with TP73, TP53, and other transcription factors (e.g., ELF5 and KLF5)[37–40] and also degrade HIF1α.[41–43] Overexpressing WWOX JET-ORF and CAN-ORF in HeLa cells (Figure S5A) reduces HIF1α protein levels compared with control HeLa mNG1-10 cells (Figures S5B and S5B). No HIF1α mRNA expression differences between cell lines are observed (Figure S5C), consistent with the reported post-translational regulation of HIF1α by WWOX.[41,43] RNA-seq from HeLa cells overexpressing the WWOX JET-ORF or CAN-ORF and control HeLa mNG1-10 cells shows 219, 589, and 118 genes differentially expressed in JET-ORF, CAN-ORF, or in both, respectively, compared with control cells (Figure S5D). In line with the nuclear localization of WWOX JET-ORF (Figure 4H), we observed 5 erythroblast transformation specific (ETS) transcription factors differentially expressed in WWOX JET-ORF-expressing cells, compared with WWOX CAN-ORF or control HeLa mNG1-10

Figure 4. JET-ORFs can be stable and localize to distinct subcellular compartments

(A) Instability index for JET-ORFs (bottom) and CAN-ORFs (top). Colored dots indicate the indexes of the ectopically expressed candidates.

(B and C) Stability curves in K562 cells using quantitative proteomics (B) or mNeonGreen-based assay (C). Lamin-B1 (stable) and geminin and MYC (short-lived) proteins are used as controls. In (B), JET-ORF normalized abundances along time points are shown. In (C), normalized MFI along time points are shown for each CAN-ORF (top) and JET-ORF (bottom).

(D–H) Confocal microscopy images of reexpressed JET-ORFs and the corresponding CAN-ORFs (green). sirDNA (pink) indicates the nucleus.

(I) Quantification of the nuclear enrichment (MFI nucleus/MFI total cell) of PTEN JET-ORF and WWOX JET-ORFs compared with the CAN-ORFs. ***$p = 0.0005$; ****$p < 0.0001$ by unpaired t test.

See also Figure S4.

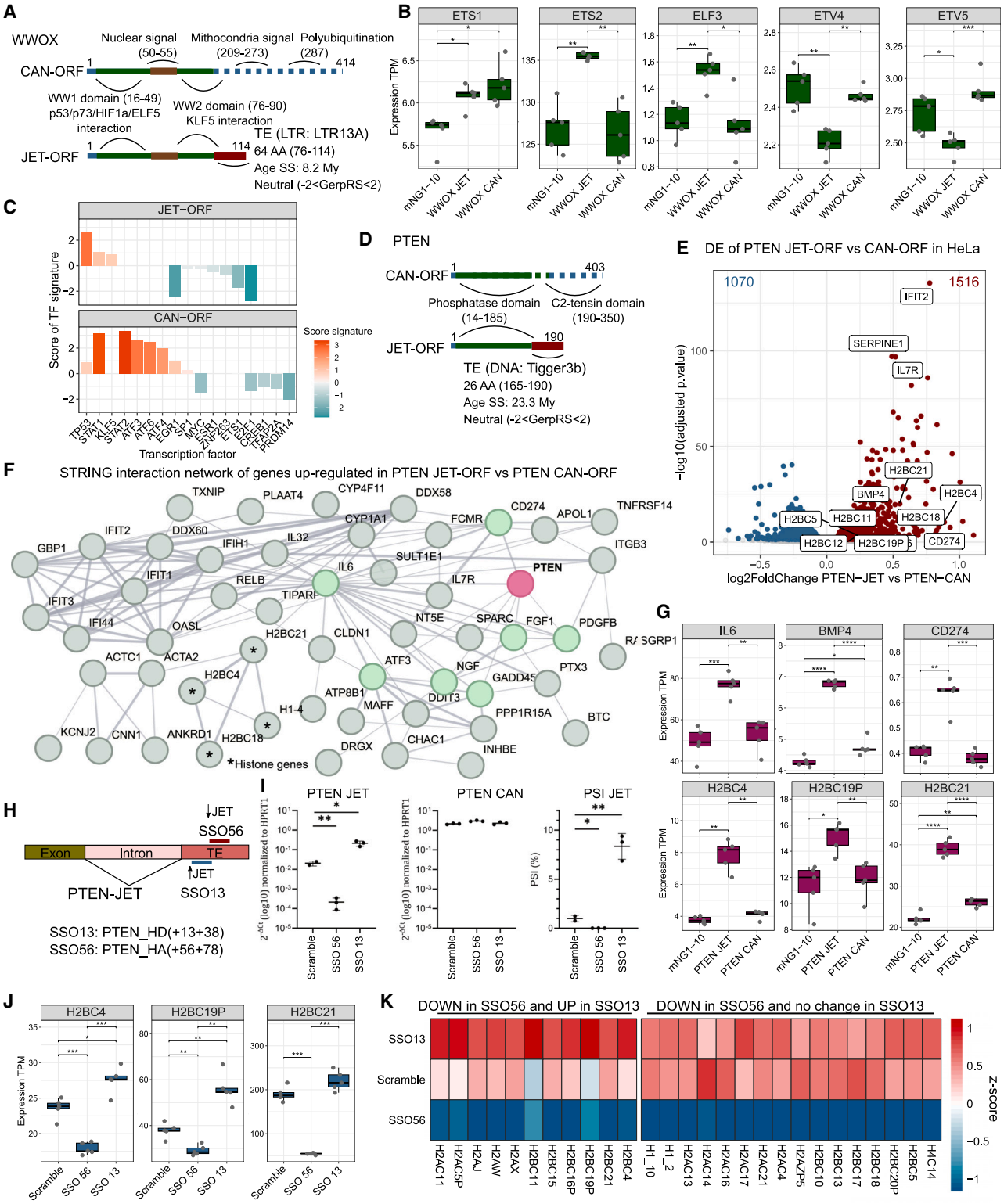


Figure 5. JET-ORFs are functional and can have divergent functions compared with CAN isoforms

(A) Scheme of WWOX JET-ORF and CAN-ORF.

(B) Expression levels of ETS transcription factors differentially expressed in WWOX JET-ORF overexpressing cells.

(C) Transcription factor inference in WWOX JET-ORF and WWOX CAN-ORF overexpressing cells compared to HeLa mNG1-10 cells (negative control).

(legend continued on next page)

(Figure 5B). Transcription factor inference from the RNA-seq data shows that, although WWOX CAN-ORF induces *ATF6*, *ATF4*, *ATF3*, and *STAT1/2* responses compared with negative control cells, WWOX JET-ORF is predicted to activate *KLF5* and *TP53* (Figure 5C). *KLF5* and *TP53* expression levels are regulated by WWOX translocation to the nucleus.[37,39,40] *KLF5* regulates other transcription factors, such as *ELF3*, which is increased in WWOX JET-ORF-overexpressing cells (Figure 5B). The overexpression results suggest that its nuclear localization allows WWOX JET-ORF to display, at steady state, the tumor suppressor functions performed by WWOX CAN-ORF following activation and nuclear translocation.

To examine the functions of endogenous JET-ORFs (which are generally expressed at low levels), we generated single-cell clones from HEK293FT cells after CRISPR-Cas9 depletion of the exonized TE in HTT and UBFD1 JET-ORFs (Figure S5E). For the TEs involved in both HTT and UBFD1 JETs (Figure S5F), we obtained several KO clones that expressed the CAN isoforms at normal levels (Figure S5G). In the UBFD1 JET-ORF, an internal TE exonization creates a frameshift on the downstream CAN exon that leads to a shorter isoform (Figure S5F) with a change in subcellular location compared with the CAN-ORF (Figure S4L). Differential gene expression analysis uncovered 21 differentially expressed genes between UBFD1 JET-KO and mock clones (Figure S5H), including the upregulation of the transcription factor *MAFB* and *PTCHD1*. HTT is a large protein that serves as scaffold for several biological processes, such as microtubule-mediated transport[44] and cell response to calcium.[44,45] HTT JET-ORF consists of a 29-amino-acid-long TE insertion in the middle of the protein (Figure S5F). RNA-seq from HTT-JET-ORF KO clones and mock clones revealed strong downregulation of one single gene, *CACNA1H*, after HTT-JET depletion (Figure S5I). The slight modification of the protein sequence after TE exonization compared with the CAN-ORF aligns with the low number of differentially expressed genes and suggests complementarity between the two isoforms. The expression of *HLA-A*, *SDHA*, and *ACTB* (housekeeping genes) remains unchanged between mock and KO clones, which reinforces the robustness and relevance of the differences in *CACNA1H* expression (Figure S5J). To compare the phenotypic changes with CAN *HTT* depletion, we used publicly available RNA-seq data (GEO: GSE178467) from two clones with full *HTT* depletion in the SY5Y neuroblastoma cell line compared with wild-type (WT) cells.[46] Full HTT depletion leads to the downregulation of several calcium channels (i.e., *CACNA1B*, *CACNG4*, and *CACNAD2*, Figure S5K). *CACNA1H* expression after HTT CAN-ORF depletion could not be evaluated because the gene is not expressed in SY5Y cells. Because full HTT depletion affects proliferation,[46] we measured cell proliferation in HTT-JET KO clones using non-invasive electrical impedance monitoring (Figure S5L, left). Real-time measurements showed decreased proliferation of HTT-JET-KO clones compared with mock clones (Figure S5L, right). Analysis of UBFD1-JET and HTT-JET KO clones suggests that JET-ORFs can be functional at endogenous levels of expression.

The PTEN JET-ORF codes for a truncated isoform that loses the C2-tensin domain but retains the phosphatase domain of the CAN-ORF (Figure 5D). RNA-seq from HeLa mNG1-10 control cells or cells overexpressing similar levels of either isoform (Figure S5M) shows differential expressions of 1,070 and 1,516, respectively (Figure 5E). Functional association network analysis of the genes upregulated by the JET-ORF underscored a gene network related to *IL6* (e.g., *CD274* and *FGF1*, Figures 5F and 5G). PTEN JET-ORF overexpression also increases the expression of seven H2BC histones compared with control cells (Figures 5G and S5N). This finding is consistent with the previous observation that PTEN translocation to the nucleus impacts chromatin organization.[47,48] Overexpression of PTEN JET-ORF or PTEN CAN-ORF in the H1650 cell line reproduced the effects of the JET-ORF on histone upregulation (Figure S5O). Therefore, although PTEN JET-ORF and CAN-ORF share structural domains essential for their functions, the results suggest that JET-ORF acquires distinct functions compared with the CAN-ORF.

To evaluate if endogenous PTEN JET-ORF has similar functions, we modulated its splicing using splice-switching oligonucleotides (SSOs). SSOs are designed to bind splice site flanking RNA sequences, impairing the binding of splicing regulators and thereby increasing or decreasing splicing levels[49] (Figure 5H). Although the PTEN JET has an endogenous usage or percent spliced in (PSI) of 1% in H1650 cell line, SSO56 diminishes PTEN JET to almost undetectable levels, and SSO13 increases PTEN JET PSI to around 8% (Figure 5I). H1650 cells transfected with SSO13, SSO56, or scramble control SSO were analyzed by RNA-seq. Inhibition of the JET splicing by SSO56 had opposite effect than the overexpression of JET-ORF on H2BC histones and *IL6*, *BMP4*, and *CD274* expression (Figures 5J and S5P). In addition, 11 histones are significantly upregulated in SSO13 and downregulated in SSO56, compared with control cells (Figure 5K). Moreover, the expression of 16 additional histones is reduced in SSO56. The modulation of the endogenous expression of PTEN JET suggests that PTEN JET-ORF is physiologically functional and is involved on the regulation of histone expression.

(D) Scheme of PTEN JET-ORF and CAN-ORF.
(E) Differentially expressed genes between PTEN JET-ORF (red) and CAN-ORF (blue) expressing HeLa cells.
(F) STRING interaction network of genes enriched in PTEN JET-ORF-overexpressing cells and with a foldchange $\geq$ 1.5. Only genes directly (green) or indirectly (gray) connected to *PTEN* (red) are represented.
(G) Expression levels of representative genes in PTEN JET-ORF- and PTEN CAN-ORF-expressing cells and HeLa mNG1-10 control cells.
(H) SSO-based modulation of PTEN-JET expression.
(I) RT-qPCR quantification of PTEN JET (left) and PTEN-CAN (middle) expression in SSO56, SSO13, or scramble-treated cells. Percent spliced in (PSI) changes across conditions are also represented (right).
(J) Expression of *H2BC4*, *H2BC19P*, and *H2BC21* genes in H1650 cells treated with the scramble control, SSO56, or SSO13.
(K) Expression changes of 27 histone genes in SSO13-, SSO56-, and scramble-treated cells. Color gradient represents the *Z* score normalization based on TPM expression. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$ (t test with Bonferroni adjustment).
See also Figure S5.

Altogether, these results indicate that endogenous JET-ORFs, even if expressed at low levels, can have biological functions. These results are consistent with JET-ORFs representing a diversity reservoir for natural selection.

JET-ORFs show intronic LINE enrichment in ancient genes

If JET-ORFs have acquired functions, it is most likely that the exonized TEs contribute to these functions. The 1,227 JET-ORFs defined in Figure 3, include LINEs (n = 264), SINEs (n = 256), LTRs (138), and DNA elements (122) (Figure 6A). In comparison with JET transcripts (i.e., assembled JET-containing transcripts), JET-ORFs (i.e., detected by Ribo-seq) are enriched in LINEs (adj. p value < 0.0001, Figure 6B). TE classes vary according to the position of the JET within the ORF (Figure 6C). SINEs are enriched in JETs at start positions, whereas LTRs are enriched in JETs at the end of the ORF (Figure 6D). No significant changes in the translation efficiency are observed between TE classes (Figure S6A). TEs involved in JET-ORFs are mostly proximal to genes (Figure S6B) and intronic (Figure 6E). A total of 21% of LTRs in JET-ORFs are intergenic, whereas the proportion of intergenic elements for LINEs, SINEs, and DNAs is lower (2.5%). Intergenic LTRs are numerous in JET-ORFs with the TE at the end of the ORFs (Figure S6C), suggesting that LTRs are more prone to abort translation.

The evolutionary age of LINEs and SINEs involved in JET transcripts and JET-ORFs is similar to the ages of all copies in the genome (Figure S6D). By contrast, LTR and DNA elements in JET transcripts and JET-ORFs are overall younger than genomic copies (Figure S6D). No significant differences are observed between the age of TEs in JET transcripts and JET-ORFs, suggesting that the results are not driven by translation but by the splicing event. We conclude that LINEs are the most enriched TEs in JET-ORFs and that exonized LTRs and DNA elements are younger than genomic TE copies from the same classes, informing the selectivity of the TE exonization process.

Previous studies showed that co-evolution of the transcriptional machinery with ancestral genes has optimized their expression efficiency[50,51] and that evolutionarily older genes are more prone to use alternative splicing to gain variability.[52] To investigate if TEs are preferentially exonized in ancient genes, we classified JET-ORFs according to their phylostrata (i.e., time period when a gene first appeared in evolution) of the involved genes.[53] Genes common to all living organisms (phylsotratum 1) represent 25% of all human genes and the 42% of genes with at least one JET-ORF (Figure 6F). By contrast, the proportion of youngest genes (phylostratum $\geq$ 12) decreases from 40% in the genome to around 20% in JET-containing genes (both JET transcripts and JET-ORFs). JET-ORFs with LTRs involve younger genes compared with JET-ORFs involving other TE classes (Figure 6G). The enrichment of JET transcripts and JET-ORFs in older genes could either be due to increased detection power by transcriptomics or to a genuine biological enrichment of JETs in older genes. The latter scenario could result from both a higher number of TE insertions and a longer time to exonize these TEs during evolution. Indeed, younger genes (phylostratum $\geq$ 12) contain fewer intronic TEs than older genes (Figure 6H). Although the intronic TE density remains stable across phylostrata (Figure S6E), older genes are characterized by a higher number of introns (Figure S6F), which results in more intronic TEs per gene (Figure S6G). The proportions of the TE classes, however, do not vary between phylostrata (Figure S6H). Furthermore, TE exonization preferentially occurs in genes with more intronic TEs, as shown by the higher number of intronic TEs in JET-containing genes, compared with genes without JETs (Figure 6I). These results suggest that older genes bearing more introns, and therefore more intronic TEs, have greater chances to exonize TEs.

Because ancient genes generate more JETs, we investigated the conservation across vertebrates of the exonized TE sequences using phastCons.[54] JET-induced exons are less conserved than all exons in the transcriptome (Figure 6J). Furthermore, translated JET-exons (in JET-ORFs) are slightly more conserved than non-translated JET-exons (JETs found in transcriptome but not in translatome, Figure S6I). Among TE classes, LINE and DNA-derived exons are evolutionary older than SINE and LTR (Figure S6J). We conclude that JETs are more frequent in ancient genes, which allows acquisition of younger and less conserved TE-derived exons during evolution.

Translated JETs can increase the α helix content of the host protein

To gain insights into the structural domains acquired in JET-ORFs after TE exonization, we used ColabFold to predict the three-dimensional structures of ALDH3A2, H2AFY, WWOX, IL-15RA, and PTEN JET-ORFs (from which we have shown functional data, Figures 4 and 5). Structure comparison of the ALDH3A2 JET-ORF and CAN-ORF notes that the exonized TE (in green) elongates the host isoform and is exposed to the cytosol by crossing the membrane (Figure S7A). In H2AFY JET-ORF, the TE exonization leads to changes in the Macro domain, compared with the CAN-ORF (Figure S7B). Similarly, the WW2 domain of the WWOX JET-ORF retains two β strands (instead of three in the CAN-ORF), which also changes their orientation (Figure S7C). These changes are driven by the acquired α helix of the exonized TE (Figure S7D). Interestingly, the WW2 domain interacts with KLF5,[39] which is among the most activated transcription factors after WWOX JET-ORF overexpression (Figure 5C). Concerning IL-15RA JET-ORF, the TE inserts two α helices between the signal peptide and the Sushi domain involved in IL-15 interaction (Figure S7E). Finally, the exonized TE shortens PTEN (scheme in Figure 5D; Figure S7F) and modifies the original α helix (CAN-ORF) to become a β sheet formed by 2 β strands (Figure S7G). These results indicate that exonized TE sequences are not random and can have secondary structures that modify the overall structure of the host isoforms.

We next predicted the three-dimensional structures of the JET-ORFs with internal TE exonization, together with the structures of the corresponding CAN-ORFs (Figure 7A). We obtained reliable predictions for 36 JET-ORFs with the paired CAN-ORFs, corresponding to 25 unique exonized TE sequences (examples shown in Figures 7B–7D). Using Chimera (v1.16),[55] we annotated the secondary structures adopted by the TE exons. Twenty-four out of 36 structures contain at least 1 α helix on the TE region (Figure 7E). Only 3 of them bear β strands, whereas no secondary structure was predicted in 9 exonized TE sequences. Although

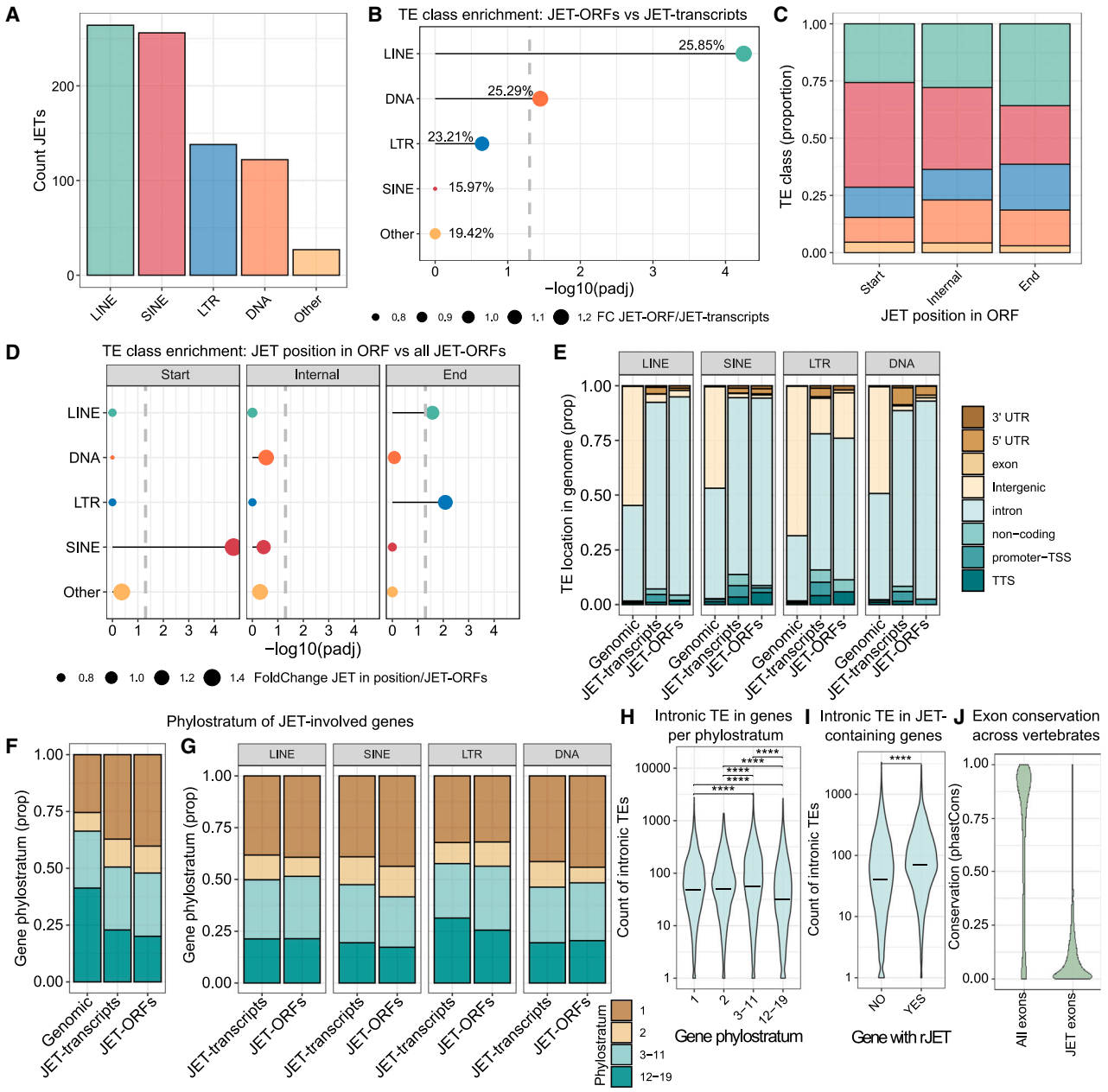


Figure 6. JET-ORFs show intronic LINE enrichment in ancient genes

(A) JET count per TE class.

(B) TE class enrichment in JET-ORFs versus all JET-containing transcripts. Dot size represents the ratio of translated JETs versus all transcribed JETs.

(C) TE class proportions depending on the JET position within the ORF.

(D) TE class enrichment based on the JET position within the ORF versus all JET-ORFs.

(E) Genomic TE location (in proportion) for all genomic TEs, TEs involved in JET-transcripts, and TEs involved in JET-ORFs.

(F and G) Gene age (phylostratum, in proportion) of all genes in genome, genes in JET-transcripts, and genes in JET-ORFs. In (G), gene ages are represented according to the involved TE class.

(H and I) Count of intronic TEs per gene according to (H) gene age or (I) whether the gene contains a recurrent JET. The median is indicated (black line). ****$p <$ 0.0001 by Mann-Whitney test with Bonferroni adjustment.

(J) Conservation across vertebrates (phastCons) of all transcribed exons and JET-induced exons.

See also Figure S6.

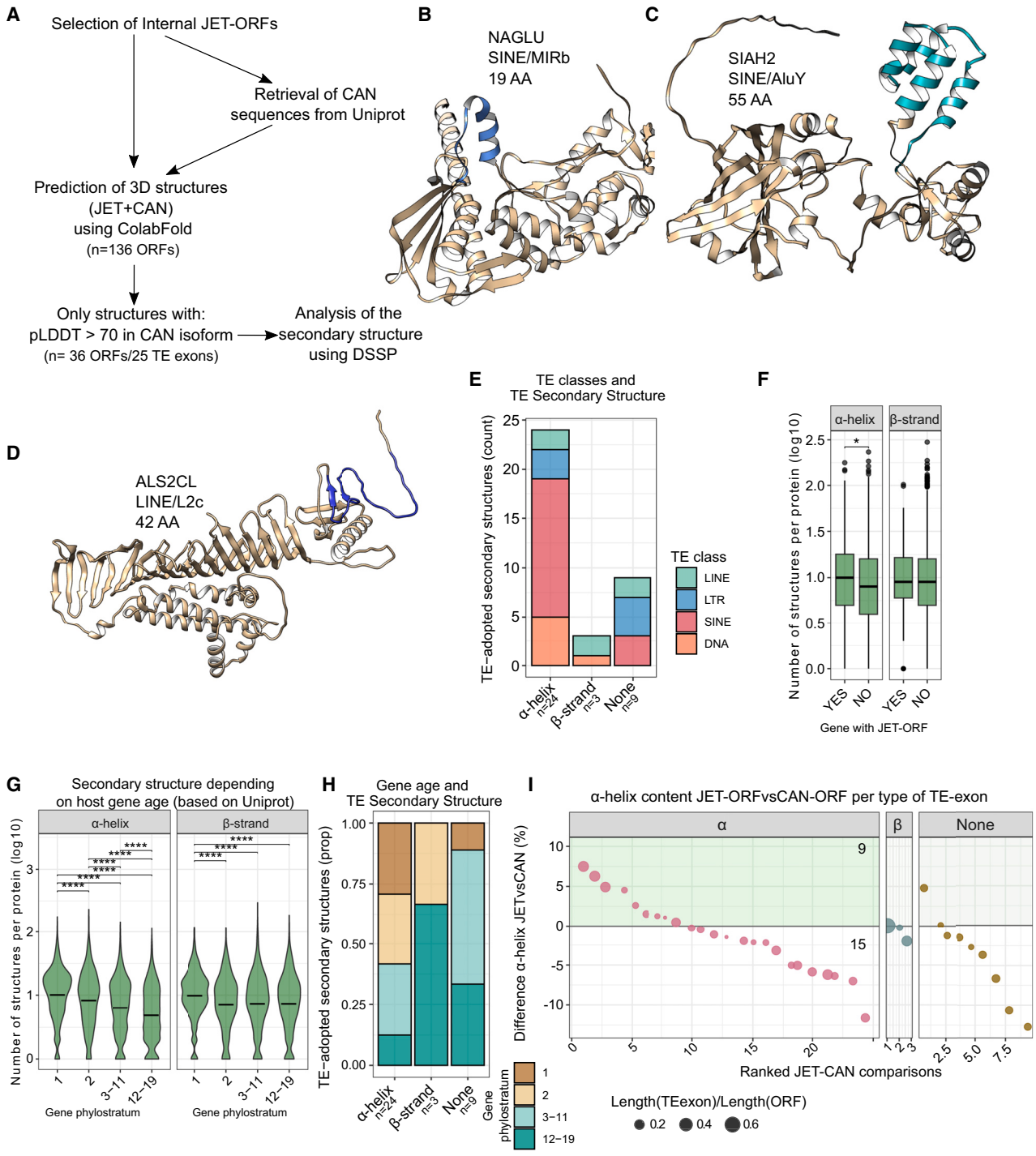


Figure 7. JET-ORFs shift the secondary structure content of the host protein toward α helices

(A) Structure prediction workflow for the selected JET-ORFs.

(B–D) Examples of JET-ORF 3D structures. Exonized TE sequence is colored in blue.

(E) Count of TE exons that adopt an α helix, β strand, or no structure (*n* JET-ORFs = 36). The color indicates the TE class.

(F and G) Count of α helix or β strand annotated in the canonical proteins of genes containing or not containing any identified JET-ORF (F) and depending on the host gene age (G). $^{*}p < 0.05$; $^{****}p < 0.0001$ by Mann-Whitney test with Bonferroni adjustment.

(H) Gene age (proportion) depending on the secondary structure of the TE exon.

(I) Difference of the overall α helix content (%) in JET-ORFs compared with CAN-ORFs, depending on the structure adopted by the exonized TE.

See also Figure S7.

SINEs provide 14 of the 24 exonized α helices, 2 of the 3 β strands come from LINEs. On average, around 30% of the amino acids supplied by the TE exon are included in the acquired α helices (Figure S7H), a proportion similar to that of α helix-adopting amino acids in the CAN proteome.[56] To understand if the TE preference to fold in α helices is explained by their amino acid composition, we compared amino acid proportions between the TE exons and the whole JET-ORFs among the 24 structures on which the TE exon folds in α helix (Figure S7I, left). Although the amino acids with major preference to fold in α helix (i.e., L, E, and A[57]) are reduced in the TE exon compared with the JET-ORF, we also observe a reduction of G (major blocker of α helices) and an increase of K and Q, which are a second layer of amino acids that favor α helix conformations. We also observe a reduction of V (which favors β strands). This trend was confirmed by an extended analysis using all JET-ORFs identified in Figure 3 (and with a TE exon longer than 20AA, Figure S7I, right). Other amino acids considered formers of α helices (F and H) are increased in TE exons compared with the proportions in JET-ORFs (Figure S7I, right). In conclusion, JETs provide sequences with a divergent amino acid composition compared with the CAN sequence, and the exonized TEs preferentially adopt α helix conformations.

Because JET-ORFs containing TE exons that fold in α helices have a larger overall content of secondary structures than JET-ORFs with TEs without structure (Figure S7J), we hypothesized that the TE exonization could impact the overall structural content of the host proteins and be involved in evolution of protein structures. We first counted the number of secondary structures (α helices or β strands) per CAN protein among the 8,196 entries with annotated structure information in Uniprot. CAN proteins encoded by JET-ORF-containing genes harbor more α helices, compared with genes without any JET-ORF, whereas no changes are perceived in β strand content (Figure 7F). When categorizing proteins depending on their gene phylostrata, we observe that proteins encoded by the oldest genes have 10 α helix on average, and this number diminishes together with the gene age (Figure 7G). Although CAN proteins encoded by genes from phylostratum 1 also contain more β strand than the other groups, no differences are observed from phylostrata 2 to 19 (Figure 7G). In line with these results, JET-ORFs with TE exons bearing α helices are preferentially found in genes from older phylostrata (Figure 7H). TE sequences fold in α helix and increase the overall α helix content in 9 out of 24 JET-ORFs (Figure 7I). By contrast, no overall increase in α helix is driven by TEs without structure or adopting β strands. Despite the TE sequence in POLRJ2 JET-ORF adopting three small α helix, which also contains a large unstructured region, the overall percentage of α helix in the JET-ORF decreased, compared with the CAN-ORF (Figure S7K). In GOPC JET-ORF (Figure S7L) and in SIAH2 JET-ORF (Figure S7M), the exonized TE sequences increase up to 6.3% and 4.9% the α helix content compared with the CAN-ORF, respectively. Out of the 36 JET-ORFs, only one increased more than 1% the overall β strand content in the host isoform (Figure S7N), suggesting an α helix selectivity of the process. Overall, we conclude that exonized TEs can adopt secondary structures and can contribute to increasing the overall α helix content during protein evolution.

DISCUSSION

Current protein evolutionary models suggest that alternative splicing represents a source of diversity upon which natural selection can target emerging protein variants, whereas the existing functions are preserved by the main isoforms.[9] The biological relevance of TE exonization in this context is increasingly well established.[11,17,58] In direct support of a model in which TE exonization plays a role in protein evolution, our results indicate that JET-ORFs are not merely transcriptional noise but rather represent a fixed image of a reservoir of evolving proteins under ongoing natural selection.

If JET-ORFs are subjected to natural selection, the emerging isoforms, or a fraction of them, should be functional. Functional analyses of PTEN and WWOX tumor suppressor JET isoforms showed modified subcellular localization and functions, compared with the CAN isoforms (Figures 5 and S5). SSO-mediated depletion of endogenous PTEN JET expression impacts the expression of several histones (Figure 5K). JET depletion in *HTT* and *UBFD1* genes also showed mild but detectable endogenous phenotypic changes. These experiments show that at least a proportion of JET-ORFs can be translated into functional protein isoforms.

Can we exclude, however, that this population of proteins, rather than being under natural selection, have stable biological roles at low expressed levels? Even if the two interpretations are not mutually exclusive, several of our findings support the selective hypothesis. First, we showed that JET usage increased proportionally to the age of the splice site (Figure 2F), with some JETs being already the major isoform in certain tissues (Figures S1H and S1I). Second, JET-ORFs have been fixed recently because the splice sites of recurrent JETs are evolutionary young (Figure 2C) but also more conserved than splice sites of non-recurrent JETs (Figure 2D). These results are all consistent with exonization of intronic regions being a multistep process, in which series of random mutations give birth to new exons by increasing the strength of splice motifs. Quantifying the proportion of JET-ORFs that will be gradually targeted by natural selection and attain physiological relevance poses an important challenge. Around 6% of the 1,227 identified JET-ORFs are evolutionarily conserved (increasing up to 12% among more recurrent JETs). Supporting this interpretation, exonized LINEs are translated preferentially over SINE-containing transcripts, supplying young sequences primarily to ancient genes. Additionally, exonized TE sequences favorably adopt α helix conformations and can increase the content of secondary structures in the host protein (Figures 7E and 7I). Because JET-ORFs are preferentially found in genes from older phylostrata (Figure 6F) and because genes from older phylostrata have more α helices (Figure 7G), our results support TE exonization as a mechanism to increase the overall secondary structure in proteins during evolution.

Regardless of the evolutionary interpretation of our results, this study identifies previously unannotated protein isoforms and describes an approach to discover more in other tissues and organisms. In addition, although random mutation-based generation of isoforms allows functional adaptations, it also produces a large pool of unstable proteins that are a productive

source of MHC class I-presented peptides.[18,23] The analysis of this low abundance but highly variable proteome may shed light on emerging cell functions in physiological and pathological contexts.

Limitations of the study

Our study has limitations that should be acknowledged. Assembly of the JET transcriptome can lead to incomplete annotations. Although these issues are minimized at the ORF level, they impact the interpretation of our results. Second, JET-ORF functions are probably context dependent. Our observations were made under steady-state conditions in a specific cell line, and it remains unclear how these processes might vary under different stimuli. For example, although HTT-JET-KO cells grow slower, only one gene was differentially expressed compared with WT cells. Cell cycle synchronization would probably reveal other transcriptomic differences related to proliferation. Tissue-dependent expression of JETs suggests that the observed sub-functionalization is environment-dependent. Variability between cell lines and single-cell clones can also impact the results. Future studies incorporating long-read sequencing and a variety of biological contexts (reflecting tissue environment) will be required to fully understand the functional aspects of JET-ORF diversity and selection pressures.

RESOURCE AVAILABILITY

Lead contact

Requests for information should be directed to and will be fulfilled by the lead contact, Sebastian Amigorena (sebastian.amigorena@curie.fr).

Material availability

Cell lines or other resources generated in this study are available from the lead contact with a completed materials transfer agreement.

Data and code availability

Sequencing data are deposited in NCBI's GEO at GEO: GSE234223 and are publicly available as of the date of publication. Proteomic data on K562 cells are available in PRIDE: PXD054305 and are publicly available as of the date of publication. This paper does not report original code. Any additional information required is available from the lead contact upon reasonable request.

ACKNOWLEDGMENTS

We used data produced by TCGA, managed by the NCI and NHGRI (http://cancergenome.nih.gov). We thank the ICGex NGS, Bioinformatics, and the flow cytometry platforms from the Institut Curie; M.G. Delgado, M. Maurin, V. Fraisier, and the CurieCoreTech cell and tissue imaging (PICT) for the support in microscopy experiments; Mnemo Therapeutics for providing Ribo-seq samples; M. Robert de Massy for helping in the Ribo-seq analysis; and O. Lantz, A. Gros, N. Manel, X. Lahaye, E. Zueva, S. López-Cobo, J. Fuentealba, J. Tosello, and members of the Amigorena team for scientific discussions. The ICGex NGS platform is funded by ANR-10-EQPX-03 (Equipex) and ANR-10-INBS-09-08 (France Génomique Consortium) from the Agence Nationale de la Recherche, by the ITMO-Cancer Aviesan (Plan Cancer III), and by the SiRIC-Curie program (SiRIC Grant INCa-DGOS-465 and INCa-DGOS-Inserm_12554). S.A. is supported by ANR-10-IDEX-0001-02 PSL and ANR-11-LABX-0043 CIC IGR-Curie 1428 (Center for Clinical Investigation), as well as ANR-16CE1500180, ANR 16CE18002003, 2017-1-PL BIO-03-ICR-1, 2017-1-PL BIO-05-ICR-1, ANR-22-CE44-0014-03, Foundation ARC, and Mnemo Therapeutics. Y.A.A. received PhD fellowships from Ligue contre le Cancer and Foundation ARC. B.B. was supported by the Fondation pour la Recherche Médicale (FDT202304016610). J.J.W. is supported by the SiRIC-Curie program (INCa-DGOS-4654). M.B. received fellowships from Foundation ARC. This work has received support under the program France 2030 launched by the French government.

AUTHOR CONTRIBUTIONS

Y.A.A. conceived the project, generated the data, prepared figures, and wrote the manuscript. S.A. conceptualized and supervised the project, acquired funding, and wrote the manuscript. A.M. provided critical help during the project and edited the manuscript. B.B. designed and performed experiments. M.R. designed and performed the studies about evolution and edited the manuscript. C.E., B.M., and J.D. performed JET structure analyses. P.-E.B., A.R., P.K., and B.S. assisted with bioinformatics. V.C., G.S., and E.B. assisted with experiments. P.L. and S.B. performed the sequencing. C.G., M.C., M.B., J.J.W., and L.Q.-M. provided key insights in project development.

DECLARATION OF INTERESTS

Y.A.A., S.A., A.M., J.J.W., M.B., B.S., and C.G. have filed patent applications for the therapeutic use of JET-derived peptides (WO 2018/234367, WO 2022/189626, and WO 2022/189639). Y.A.A., M.B., C.G., and P.-E.B. are advisors in Mnemo Therapeutics. S.A. and J.J.W. are advisors and shareholders in Mnemo Therapeutics. A.M. and B.S. are currently employed by Mnemo Therapeutics.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Bacteria
  - Cell lines
- METHOD DETAILS
  - JET identification in TCGA and CCLE
  - Dimension reduction and unsupervised clustering
  - Age and conservation of JET splice sites
  - DNA/RNA extraction and RNA sequencing
  - PCR amplification
  - Genome-guided transcriptome assembly
  - Ribosome profiling analysis
  - Translation efficiency calculation
  - Deep proteome mass spectrometry analysis
  - Instability index calculation
  - Proteomics for half-life measurement
  - JET-ORF reexpression using mNG
  - *In vitro* mNG-based protein stability assay
  - Confocal microscopy
  - RT-qPCR
  - Membrane enrichment for ALDH3A2 JET-ORF
  - Western blot
  - JET KO and single-cell clone generation
  - Differential expression analyses
  - Structure analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.11.011.

Received: June 7, 2023
Revised: May 26, 2024
Accepted: November 11, 2024
Published: December 11, 2024

REFERENCES

1. Chen, J., Brunner, A.D., Cogan, J.Z., Nuñez, J.K., Fields, A.P., Adamson, B., Itzhak, D.N., Li, J.Y., Mann, M., Leonetti, M.D., et al. (2020). Pervasive functional translation of noncanonical human open reading frames. Science *367*, 1140–1146. https://doi.org/10.1126/science.aav5912.
2. Calviello, L., Hirsekorn, A., and Ohler, U. (2020). Quantification of translation uncovers the functions of the alternative transcriptome. Nat. Struct. Mol. Biol. *27*, 717–725. https://doi.org/10.1038/s41594-020-0450-4.
3. Ruiz Cuevas, M.V., Hardy, M.P., Hollý, J., Bonneil, É., Durette, C., Courcelles, M., Lanoix, J., Côté, C., Staudt, L.M., Lemieux, S., et al. (2021). Most non-canonical proteins uniquely populate the proteome or immunopeptidome. Cell Rep. *34*, 108815. https://doi.org/10.1016/J.CELREP.2021.108815.
4. Basrai, M.A., Hieter, P., and Boeke, J.D. (1997). Small open reading frames: beautiful needles in the haystack. Genome Res. *7*, 768–771. https://doi.org/10.1101/GR.7.8.768.
5. Mudge, J.M., Ruiz-Orera, J., Prensner, J.R., Brunet, M.A., Calvet, F., Jungreis, I., Gonzalez, J.M., Magrane, M., Martinez, T.F., Schulz, J.F., et al. (2022). Standardized annotation of translated open reading frames. Nat. Biotechnol. *40*, 994–999. https://doi.org/10.1038/S41587-022-01369-0.
6. Kopelman, N.M., Lancet, D., and Yanai, I. (2005). Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. Nat. Genet. *37*, 588–589. https://doi.org/10.1038/ng1575.
7. Mantica, F., and Irimia, M. (2022). The 3D-Evo Space: Evolution of Gene Expression and Alternative Splicing Regulation. Annu. Rev. Genet. *56*, 315–337. https://doi.org/10.1146/annurev-genet-071719-020653.
8. Sibley, C.R., Blazquez, L., and Ule, J. (2016). Lessons from non-canonical splicing. Nat. Rev. Genet. *17*, 407–421. https://doi.org/10.1038/nrg.2016.46.
9. Ast, G. (2004). How did alternative splicing evolve? Nat. Rev. Genet. *5*, 773–782. https://doi.org/10.1038/nrg1451.
10. Zhang, F., Raabe, C.A., Cardoso-Moreira, M., Brosius, J., Kaessmann, H., and Schmitz, J. (2022). ExoPLOT: Representation of alternative splicing in human tissues and developmental stages with transposed element (TE) involvement. Genomics *114*, 110434. https://doi.org/10.1016/J.YGENO.2022.110434.
11. Schmitz, J., and Brosius, J. (2011). Exonization of transposed elements: A challenge and opportunity for evolution. Biochimie *93*, 1928–1934. https://doi.org/10.1016/J.BIOCHI.2011.07.014.
12. Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., et al. (2018). Ten things you should know about transposable elements. Genome Biol. *19*, 199. https://doi.org/10.1186/s13059-018-1577-z.
13. Trizzino, M., Kapusta, A., and Brown, C.D. (2018). Transposable elements generate regulatory novelty in a tissue-specific fashion. BMC Genomics *19*, 468. https://doi.org/10.1186/s12864-018-4850-3.
14. Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. *10*, 691–703. https://doi.org/10.1038/nrg2640.
15. Gotea, V., and Makałowski, W. (2006). Do transposable elements really contribute to proteomes? Trends Genet. *22*, 260–267. https://doi.org/10.1016/J.TIG.2006.03.006.
16. Makałowski, W., Kischka, T., and Makałowska, I. (2017). Contribution of Transposable Elements to Human Proteins. In eLS (Wiley) https://doi.org/10.1002/9780470015902.a0020793.pub2.
17. Lin, L., Jiang, P., Park, J.W., Wang, J., Lu, Z.X., Lam, M.P.Y., Ping, P., and Xing, Y. (2016). The contribution of Alu exons to the human proteome. Genome Biol. *17*, 15. https://doi.org/10.1186/s13059-016-0876-5.
18. Merlotti, A., Sadacca, B., Arribas, Y.A., Ngoma, M., Burbage, M., Goudot, C., Houy, A., Rocañín-Arjó, A., Lalanne, A., Seguin-Givelet, A., et al. (2023). Noncanonical splicing junctions between exons and transposable elements represent a source of immunogenic recurrent neo-antigens in patients with lung cancer. Sci. Immunol. *8*, eabm6359. https://doi.org/10.1126/sciimmunol.abm6359.
19. Attig, J., Young, G.R., Hosie, L., Perkins, D., Encheva-Yokoya, V., Stoye, J.P., Snijders, A.P., Ternette, N., and Kassiotis, G. (2019). LTR retroelement expansion of the human cancer transcriptome and immunopeptidome revealed by de novo transcript assembly. Genome Res. *29*, 1578–1590. https://doi.org/10.1101/gr.248922.119.
20. Ng, K.W., Attig, J., Young, G.R., Ottina, E., Papamichos, S.I., Kotsianidis, I., and Kassiotis, G. (2019). Soluble PD-L1 generated by endogenous retroelement exaptation is a receptor antagonist. eLife *8*, e50256. https://doi.org/10.7554/eLife.50256.
21. Shah, N.M., Jang, H.J., Liang, Y., Maeng, J.H., Tzeng, S.-C., Wu, A., Basri, N.L., Qu, X., Fan, C., Li, A., et al. (2023). Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. Nat. Genet. *55*, 631–639. https://doi.org/10.1038/S41588-023-01349-3.
22. Clayton, E.A., Rishishwar, L., Huang, T.C., Gulati, S., Ban, D., McDonald, J.F., and Jordan, I.K. (2020). An atlas of transposable element-derived alternative splicing in cancer. Philos. Trans. R. Soc. Lond. B Biol. Sci. *375*, 20190342. https://doi.org/10.1098/rstb.2019.0342.
23. Burbage, M., Rocañín-Arjó, A., Baudon, B., Arribas, Y.A., Merlotti, A., Rookhuizen, D.C., Heurtebise-Chrétien, S., Ye, M., Houy, A., Burgdorf, N., et al. (2023). Epigenetically controlled tumor antigens derived from splice junctions between exons and transposable elements. Sci. Immunol. *8*, eabm6360. https://doi.org/10.1126/SCIIMMUNOL.ABM6360.
24. Singh, P., and Ahi, E.P. (2022). The importance of alternative splicing in adaptive evolution. Mol. Ecol. *31*, 1928–1938. https://doi.org/10.1111/MEC.16377.
25. Rotival, M., Quach, H., and Quintana-Murci, L. (2019). Defining the genetic and evolutionary architecture of alternative splicing in response to infection. Nat. Commun. *10*, 1671. https://doi.org/10.1038/s41467-019-09689-7.
26. Sinitcyn, P., Richards, A.L., Weatheritt, R.J., Brademan, D.R., Marx, H., Shishkova, E., Meyer, J.G., Hebert, A.S., Westphall, M.S., Blencowe, B.J., et al. (2023). Global detection of human variants and isoforms by deep proteome sequencing. Nat. Biotechnol. *41*, 1776–1786. https://doi.org/10.1038/s41587-023-01714-x.
27. Guruprasad, K., Reddy, B.V.B., and Pandit, M.W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng. *4*, 155–161. https://doi.org/10.1093/PROTEIN/4.2.155.
28. Li, J., Cai, Z., Vaites, L.P., Shen, N., Mitchell, D.C., Huttlin, E.L., Paulo, J.A., Harry, B.L., and Gygi, S.P. (2021). Proteome-wide mapping of short-lived proteins in human cells. Mol. Cell *81*, 4722–4735.e5. https://doi.org/10.1016/J.MOLCEL.2021.09.015.
29. McGarry, T.J., and Kirschner, M.W. (1998). Geminin, an inhibitor of DNA replication, is degraded during mitosis. Cell *93*, 1043–1053. https://doi.org/10.1016/S0092-8674(00)81209-X.
30. Hann, S.R., and Eisenman, R.N. (1984). Proteins encoded by the human c-myc oncogene: differential expression in neoplastic cells. Mol. Cell. Biol. *4*, 2486–2497. https://doi.org/10.1128/MCB.4.11.2486-2497.1984.
31. Farrell, A.S., and Sears, R.C. (2014). MYC Degradation. Cold Spring Harb. Perspect. Med. *4*, a014365. https://doi.org/10.1101/CSHPERSPECT.A014365.
32. Feng, S., Sekine, S., Pessino, V., Li, H., Leonetti, M.D., and Huang, B. (2017). Improved split fluorescent proteins for endogenous protein labeling. Nat. Commun. *8*, 370. https://doi.org/10.1038/S41467-017-00494-8.
33. Hernández-Muñoz, I., Lund, A.H., Van Der Stoop, P., Boutsma, E., Muijrers, I., Verhoeven, E., Nusinow, D.A., Panning, B., Marahrens, Y., and Van Lohuizen, M. (2005). Stable X chromosome inactivation involves the PRC1 Polycomb complex and requires histone MACROH2A1 and the CULLIN3/SPOP ubiquitin E3 ligase. Proc. Natl. Acad. Sci. USA *102*, 7635–7640. https://doi.org/10.1073/PNAS.0408918102.

34. Kelson, T.L., Secor McVoy, J.R., and Rizzo, W.B. (1997). Human liver fatty aldehyde dehydrogenase: microsomal localization, purification, and biochemical characterization. Biochim. Biophys. Acta *1335*, 99–110. https://doi.org/10.1016/S0304-4165(96)00126-2.

35. Ratthé, C., and Girard, D. (2004). Interleukin-15 enhances human neutrophil phagocytosis by a Syk-dependent mechanism: importance of the IL-15Ralpha chain. J. Leukoc. Biol. *76*, 162–168. https://doi.org/10.1189/JLB.0605298.

36. Dubois, S., Magrangeas, F., Lehours, P., Raher, S., Bernard, J., Boisteau, O., Leroy, S., Minvielle, S., Godard, A., and Jacques, Y. (1999). Natural splicing of exon 2 of human interleukin-15 receptor alpha-chain mRNA results in a shortened form with a distinct pattern of expression. J. Biol. Chem. *274*, 26978–26984. https://doi.org/10.1074/JBC.274.38.26978.

37. Lo, J.Y., Chou, Y.T., Lai, F.J., and Hsu, L.J. (2015). Regulation of cell signaling and apoptosis by tumor suppressor WWOX. Exp. Biol. Med. (Maywood) *240*, 383–391. https://doi.org/10.1177/1535370214566747.

38. Aqeilan, R.I., Pekarsky, Y., Herrero, J.J., Palamarchuk, A., Letofsky, J., Druck, T., Trapasso, F., Han, S.Y., Melino, G., Huebner, K., et al. (2004). Functional association between Wwox tumor suppressor protein and p73, a p53 homolog. Proc. Natl. Acad. Sci. USA *101*, 4401–4406. https://doi.org/10.1073/PNAS.0400805101.

39. Ge, F., Chen, W., Yang, R., Zhou, Z., Chang, N., Chen, C., Zou, T., Liu, R., Tan, J., and Ren, G. (2014). WWOX suppresses KLF5 expression and breast cancer cell growth. Chin. J. Cancer Res. *26*, 511–516. https://doi.org/10.3978/J.ISSN.1000-9604.2014.09.03.

40. Xu, Y., Yan, Y.C., Hu, Y.K., Fang, L.S., Li, Q., Xu, J., and Yan, H.C. (2020). WWOX regulates the Elf5/Snail1 pathway to affect epithelial-mesenchymal transition of ovarian carcinoma cells in vitro. Eur. Rev. Med. Pharmacol. Sci. *24*, 1041–1053. https://doi.org/10.26355/EURREV_202002_20154.

41. Abu-Remaileh, M., and Aqeilan, R.I. (2014). Tumor suppressor WWOX regulates glucose metabolism via HIF1α modulation. Cell Death Differ. *21*, 1805–1814. https://doi.org/10.1038/cdd.2014.95.

42. Abu-Remaileh, M., Khalaileh, A., Pikarsky, E., and Aqeilan, R.I. (2018). WWOX controls hepatic HIF1α to suppress hepatocyte proliferation and neoplasia. Cell Death Dis. *9*, 511. https://doi.org/10.1038/s41419-018-0510-4.

43. Baryła, I., Styczeń-Binkowska, E., Płuciennik, E., Kośla, K., and Bednarek, A.K. (2022). The WWOX/HIF1A Axis Downregulation Alters Glucose Metabolism and Predispose to Metabolic Disorders. Int. J. Mol. Sci. *23*, 3326. https://doi.org/10.3390/IJMS23063326.

44. Saudou, F., and Humbert, S. (2016). The Biology of Huntingtin. Neuron *89*, 910–926. https://doi.org/10.1016/j.neuron.2016.02.003.

45. Rockabrand, E., Slepko, N., Pantalone, A., Nukala, V.N., Kazantsev, A., Marsh, J.L., Sullivan, P.G., Steffan, J.S., Sensi, S.L., and Thompson, L.M. (2007). The first 17 amino acids of Huntingtin modulate its sub-cellular localization, aggregation and effects on calcium homeostasis. Hum. Mol. Genet. *16*, 61–77. https://doi.org/10.1093/HMG/DDL440.

46. Bensalel, J., Xu, H., Lu, M.L., Capobianco, E., and Wei, J. (2021). RNA-seq analysis reveals significant transcriptome changes in huntingtin-null human neuroblastoma cells. BMC Med. Genomics *14*, 176. https://doi.org/10.1186/S12920-021-01022-W.

47. Chen, Z.H., Zhu, M., Yang, J., Liang, H., He, J., He, S., Wang, P., Kang, X., McNutt, M.A., Yin, Y., et al. (2014). PTEN Interacts with Histone H1 and Controls Chromatin Condensation. Cell Rep. *8*, 2003–2014. https://doi.org/10.1016/J.CELREP.2014.08.008.

48. Gong, L., Govan, J.M., Evans, E.B., Dai, H., Wang, E., Lee, S.W., Lin, H.K., Lazar, A.J., Mills, G.B., and Lin, S.Y. (2015). Nuclear PTEN tumor-suppressor functions through maintaining heterochromatin structure. Cell Cycle *14*, 2323–2332. https://doi.org/10.1080/15384101.2015.1044174.

49. Bauman, J.A., Li, S.D., Yang, A., Huang, L., and Kole, R. (2010). Anti-tumor activity of splice-switching oligonucleotides. Nucleic Acids Res. *38*, 8348–8356. https://doi.org/10.1093/NAR/GKQ731.

50. Nagaraj, S.H., Ingham, A., and Reverter, A. (2010). The interplay between evolution, regulation and tissue specificity in the Human Hereditary Diseasome. BMC Genomics *11*, S23. https://doi.org/10.1186/1471-2164-11-S4-S23.

51. Zhang, J.Y., and Zhou, Q. (2019). On the Regulatory Evolution of New Genes Throughout Their Life History. Mol. Biol. Evol. *36*, 15–27. https://doi.org/10.1093/MOLBEV/MSY206.

52. Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nat. Rev. Genet. *11*, 345–355. https://doi.org/10.1038/nrg2776.

53. Litman, T., and Stein, W.D. (2019). Obtaining estimates for the ages of all the protein-coding genes and most of the ontology-identified noncoding genes of the human genome, assigned to 19 phylostrata. Semin. Oncol. *46*, 3–9. https://doi.org/10.1053/j.seminoncol.2018.11.002.

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

55. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612. https://doi.org/10.1002/JCC.20084.

56. de Brevern, A.G. (2023). An agnostic analysis of the human AlphaFold2 proteome using local protein conformations. Biochimie *207*, 11–19. https://doi.org/10.1016/J.BIOCHI.2022.11.009.

57. Chou, P.Y., and Fasman, G.D. (1974). Prediction of protein conformation. Biochemistry *13*, 222–245. https://doi.org/10.1021/BI00699A002.

58. Jang, H.S., Shah, N.M., Du, A.Y., Dailey, Z.Z., Pehrsson, E.C., Godoy, P.M., Zhang, D., Li, D., Xing, X., Kim, S., et al. (2019). Transposable elements drive widespread expression of oncogenes in human cancers. Nat. Genet. *51*, 611–617. https://doi.org/10.1038/s41588-019-0373-3.

59. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607. https://doi.org/10.1038/NATURE11003.

60. Park, J.E., Yi, H., Kim, Y., Chang, H., and Kim, V.N. (2016). Regulation of Poly(A) Tail and Translation during the Somatic Cell Cycle. Mol. Cell *62*, 462–471. https://doi.org/10.1016/J.MOLCEL.2016.04.007.

61. Clamer, M., Tebaldi, T., Lauria, F., Bernabò, P., Gómez-Biagi, R.F., Marchioretto, M., Kandala, D.T., Minati, L., Perenthaler, E., Gubert, D., et al. (2018). Active Ribosome Profiling with RiboLace. Cell Rep. *25*, 1097–1108.e5. https://doi.org/10.1016/J.CELREP.2018.09.084.

62. Sun, Z., Xue, S., Xu, H., Hu, X., Chen, S., Yang, Z., Yang, Y., Ouyang, J., and Cui, H. (2019). Effects of NSUN2 deficiency on the mRNA 5-methylcytosine modification and gene expression profile in HEK293 cells. Epigenomics *11*, 439–453. https://doi.org/10.2217/EPI-2018-0169.

63. Martinez, T.F., Chu, Q., Donaldson, C., Tan, D., Shokhirev, M.N., and Saghatelian, A. (2020). Accurate annotation of human protein-coding small open reading frames. Nat. Chem. Biol. *16*, 458–468. https://doi.org/10.1038/s41589-019-0425-0.

64. Calviello, L., Mukherjee, N., Wyler, E., Zauber, H., Hirsekorn, A., Selbach, M., Landthaler, M., Obermayer, B., and Ohler, U. (2016). Detecting actively translated open reading frames in ribosome profiling data. Nat. Methods *13*, 165–170. https://doi.org/10.1038/nmeth.3688.

65. Alonso, R., Flament, H., Lemoine, S., Sedlik, C., Bottasso, E., Péguillet, I., Prémel, V., Denizeau, J., Salou, M., Darbois, A., et al. (2018). Induction of anergic or regulatory tumor-specific CD4+ T cells in the tumor-draining lymph node. Nat. Commun. *9*, 2113. https://doi.org/10.1038/S41467-018-04524-X.

66. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast

universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/BIOINFORMATICS/BTS635.

67. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at arXiv.
68. Davis, M.W., and Jorgensen, E.M. (2022). ApE, A Plasmid Editor: a Freely Available DNA Manipulation and Visualization Program. Front. Bioinform. *2*, 818619. https://doi.org/10.3389/FBINF.2022.818619.
69. Shumate, A., Wong, B., Pertea, G., and Pertea, M. (2022). Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLoS Comput. Biol. *18*, e1009730. https://doi.org/10.1371/JOURNAL.PCBI.1009730.
70. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12. https://doi.org/10.14806/EJ.17.1.200.
71. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.
72. Gustavsson, E.K., Zhang, D., Reynolds, R.H., Garcia-Ruiz, S., and Ryten, M. (2022). ggtranscript: an R package for the visualization and interpretation of transcript isoforms using ggplot2. Bioinformatics *38*, 3844–3846. https://doi.org/10.1093/BIOINFORMATICS/BTAC409.
73. Osorio, D., Rondón-Villarreal, P., and Torres, R. (2015). Peptides: A package for data mining of antimicrobial peptides. R J. *7*, 4–14. https://doi.org/10.32614/RJ-2015-001.
74. Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods *9*, 671–675. https://doi.org/10.1038/nmeth.2089.
75. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/BIOINFORMATICS/BTT656.
76. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97. https://doi.org/10.1093/NAR/GKW377.
77. P.Badia, I.-Mompel, Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., Taus, P., Dugourd, A., Holland, C.H., Ramirez Flores, R.O., et al. (2022). decoupleR: ensemble of computational methods to infer biological activities from omics data. Bioinform. Adv. *2*, vbac016. https://doi.org/10.1093/BIOADV/VBAC016.
78. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682. https://doi.org/10.1038/S41592-022-01488-1.
79. Touw, W.G., Baakman, C., Black, J., Te Beek, T.A.H., Krieger, E., Joosten, R.P., and Vriend, G. (2015). A series of PDB-related databanks for everyday needs. Nucleic Acids Res. *43*, D364–D368. https://doi.org/10.1093/NAR/GKU1028.
80. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput. Biol. *6*, e1001025. https://doi.org/10.1371/JOURNAL.PCBI.1001025.
81. Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295. https://doi.org/10.1038/nbt.3122.
82. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/S13059-014-0550-8/FIGURES/9.
83. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613. https://doi.org/10.1093/NAR/GKY1131.
84. Choudhary, M.N., Friedman, R.Z., Wang, J.T., Jang, H.S., Zhuo, X., and Wang, T. (2020). Co-opted transposons help perpetuate conserved higher-order chromosomal structures. Genome Biol. *21*, 16. https://doi.org/10.1186/s13059-019-1916-8.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
anti-FLAG (FG4R clone)	Fisher Scientific	MA1-91878, RRID: AB_1957945
anti-HIF1α (polyclonal)	Proteintech	20960-1-AP, RRID: AB_10732601
anti-PTEN (polyclonal)	Proteintech	22034-1-AP, RRID: AB_2878977
anti-Actin (C4 clone)	MerkMillipore	MAB1501, RRID: AB_2223041
anti-mouse HRP-conjugated	Jackson ImmunoResearch	111-035-003
anti-rabbit HRP-conjugated	Jackson ImmunoResearch	111-035-144
Bacterial and virus strains		
NEB® 10-beta Competent E. coli	NewEngland Biolabs	C3019H
Chemicals, peptides, and recombinant proteins		
Difco LB Broth, Miller Luria-Bertani medium	BD Biosciences	244620
Ampicillin	Euromedex	EU0400-A
Cycloheximide	Sigma-Aldrich	C4859-1ML
sirDNA	Spirochrome	SC007
β-mercaptoethanol	Sigma-Aldrich	M3148
SYBR Safe Dye	Invitrogen	S33102
Triethylammonium bicarbonate buffer	Sigma	T7408
Tris-HCl	FisherScientific	15893661
SDS	Sigma-Aldrich	71736
Acetonitrile	ThermoFisher	10001334
Methanol	Millipore-Merck	1.06018.2500
Trifluoroacetic Acid, Optima™ LC/MS Grade	FisherScientific	10125637
Opti-MEM I Reduced Serum Medium	Gibco	31985070
TransIT-293 Reagent	Mirus Bio	MIR 2706
FluoroBrite DMEM medium	Gibco	A1896701
cOmplete protease inhibitor	Roche	11836170001
NaCl solution	ThermoFisher	AM9759
EDTA	ThermoFisher	15575-038
n-dodecyl-bêta-maltoside	ThermoFisher	89902
Laemmli sample buffer	BioRad	1610747
Tween20	BioRad	1706531
DMEM	Gibco	31966047
RPMI 1640	Gibco	21875034
Normocin	Invivogen	ant-nr-05
Penicillin/Streptomycin	Life Technologies	15140122
Lipofectamine™ 3000 transfection reagent	ThermoFisher	L3000015
Trypsine-EDTA	Gibco	25300054
Critical commercial assays		
SuperScript III Reverse transcriptase	ThermoFisher	18080093
RNeasy Mini Kit	Qiagen	74104
QIAwave DNA Blood Tissue kit	Qiagen	69554
Bioanalyzer RNA 6000 nano assay	Agilent	5067-1511
GoTaq G2 Hot Start Polymerase	Promega	M7401
QIAquick Gel Extraction Kit	Qiagen	28704

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
microBCA method	ThermoFisher	23235
10 kDa Amicon centrifugal filters	Millipore	UFC501096
TMT 16-plex isobaric label reagents	ThermoFisher	A44521 Lot: YJ374827
NucleoBond Xtra Midi Plus EF kit	Macherey-Nagel	740422-50
Pierce BCA Protein Assay kit	ThermoFisher	23225
SF Cell Line Nucleofector® Solution	Lonza	V4XC-2032
Mini-Protean TGX Stain-Free gels 4-15%	BioRad	4568084
ImmunoBlot PVDF membranes	BioRad	1620174
Clarity Western ECL	BioRad	1705060
Deposited data		
TCGA repository	NCI and NHGRI	http://cancergenome.nih.govhttp://cancergenome.nih.gov
CCLE	Barretina et al.[59]	https://sites.broadinstitute.org/ccle/
In-house sequencing data on cell lines	This Study	GEO: GSE234223
RNAseq for transcriptome assembly	Park et al.[60]	GEO: GSE79664
RNAseq for transcriptome assembly	Clamer et al.[61]	GEO: GSE112295
RNAseq for transcriptome assembly	Sun et al.[62]	GEO: GSE122425
RNAseq for transcriptome assembly	Martínez et al.[63]	GEO: GSE125218
RiboSeq data	Calviello et al.[64]	GEO: GSE73136
RiboSeq data	Clamer et al.[61]	GEO: GSE112353
RiboSeq data	Park et al.[60]	GEO: GSE79664
RiboSeq data	Martínez et al.[63]	GEO: GSE125218
RiboSeq data	Calviello et al.[2]	GEO: GSE129061
RNAseq on full HTT-KO clones	Bensalel et al.[46]	GEO: GSE178467
Proteomics on cycloheximide-treated K562 cells	This Study	PRIDE: PXD054305
Proteomics on cycloheximide-treated cell lines	Li et al.[28]	PRIDE: PXD024513
Deep proteome for isoform identification	Sinitcyn et al.[26]	Massive: MSV000086944
Experimental models: Cell lines		
HeLa mNG1-10	This Study	N/A
HEK293 mNG1-10	This Study	N/A
HEK293T	ATCC	CRL-3216
H1650	ATCC	CRL-5883
K562	ATCC	CCL-243
H1395	ATCC	CRL-5868
HEK293T-Lenti-X cells	Takara	632180
Oligonucleotides		
See Table S5 for PCR primer sequences	This Study	N/A
See Table S6 for CRIPSR/Cas9 guides	This Study	N/A
Recombinant DNA		
pTwist Lenti SFFV Puro WPRE	TwistBioscience	N/A
pCMV-VSVG	Alonso et al.[65]	N/A
psPAX2	Alonso et al.[65]	N/A
Software and algorithms		
STAR (v2.5.3a)	Dobin et al.[66]	https://github.com/alexdobin/STAR
Umap R package	McInnes et al.[67]	https://github.com/tkonopka/umap
Ape, A Plasmid Editor	Davis and Jorgensen[68]	https://jorgensen.biology.utah.edu/wayned/ape/
StringTie v2.1.4	Shumate et al.[69]	https://github.com/mpertea/stringtie2-initial-release
cutadapt v1.8	Martin[70]	https://github.com/marcelm/cutadapt

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
bowtie2 v2.2.5	Langmead and Salzberg[71]	https://github.com/BenLangmead/bowtie2
RiboseQC v1.1 R package	Calviello et al.[2]	https://github.com/ohlerlab/RiboseQC
ORFquant v1.02.0 R package	Calviello et al.[2]	https://github.com/lcalviell/ORFquant
ggtranscript R package	Gustavsson et al.[72]	https://github.com/dzhang32/ggtranscript
Peptides R package	Osorio et al.[73]	https://github.com/dosorio/Peptides
ProteomeDiscoverer 3.0	ThermoFisher	https://cran.r-project.org/bin/windows/base/old/4.2.2/
FlowJo v10.6.1	BD Biosciences	https://www.flowjo.com/solutions/flowjo/downloads
Metamorph	Gataca Systems	https://github.com/Nesvilab/FragPipe
Image J v1.53	Schneider et al.[74]	https://imagej.net/
AlleleID	Premier Biosoft	https://www.premierbiosoft.com
featureCounts	Liao et al.[75]	https://github.com/ShiLab-Bioinformatics/subread
enrichr R package	Kuleshov et al.[76]	https://github.com/wjawaid/enrichR
decoupleR	Badia-I-Mompel et al.[77]	https://github.com/saezlab/decoupleR
ColabFold (version 1.5. 1)	Mirdita et al.[78]	https://github.com/sokrypton/ColabFold
DSSP (version 4.4)	Touw et al.[79]	https://swift.cmbi.umcn.nl/gv/dssp/index.html
Chimera 1.16	Pettersen et al.[55]	https://www.cgl.ucsf.edu/chimera/download.html
Prism 9.0	GraphPad	https://www.graphpad.com/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

All human cell lines and bacteria strains used in this study are listed in the key resources table.

Bacteria

NEB® 10-beta Competent E. coli (NewEngland Biolabs: C3019H) were cultured in Difco LB Broth, Miller Luria-Bertani medium (BD, 244620) at 37°C in agitation.

Cell lines

HeLa mNG1-10 (female), HEK293T mNG1-10 (female), HEK293T-Lenti-X (female, Takara: 632180) and HEK293T (female, ATCC: CRL-3216) were cultured in DMEM 10% fetal bovine serum (FBS) 1% penicillin/streptomycin (P/S) at 37°C and 5% CO_2. K562 (female, ATCC: CCL-243), H1395 (female, ATCC: CRL-5868), and H1650 (male, ATCC: CRL-5883) cells were cultured in RPMI 10% FBS 1% P/S at 37°C and 5% CO_2.

METHOD DETAILS

JET identification in TCGA and CCLE

RNA-sequencing FASTQ files from TCGA and CCLE were retrieved from gdc-legacy portal. FASTQ files were aligned using STAR (v2.5.3a) two-pass mode was used to align the reads against the hg19 human reference genome (Ensembl) and to annotate junctions from SJ.out and Chimeric.out.junction files. The following STAR parameters were used: –sjdbOverhang 100, –outSAMtype BAM SortedByCoordinate, –outSAMunmapped Within, –bamRemoveDuplicatesType UniqueIdentical, –outMultimapperOrder Random, –outFilterMismatchNmax 6, –alignSJDBoverhangMin 1, –outFilterType BySJout, –alignSJoverhangMin 5, –outSAMattributes All, –quantMode GeneCounts, –outFilterMismatchNoverLmax 0.04, –outFilterMatchNminOverLread 0.33, –outFilterScoreMinOverLread 0.33, –outFilterMultimapNmax 1000, –winAnchorMultimapNmax 1000, –chimOutType WithinBAM, –chimSegmentMin 10, –chimJunctionOverhangMin 10.

JET is defined as a junction between a canonical CDS exon and repeated (transposable) element as defined by ENSEMBL and RepeatMasker databases, respectively. After normalizing the number of splicing reads by the number of unique mapped reads, we kept all JETs with a level of expression over $2*10^{-7}$ and present in at least 2 samples. Recurrent JETs were defined as JETs present in more than 1% of tumor TCGA and/or CCLE.

To calculate the proportion of a junction among all overlapping splicing events (Figure 1G), the counts-per-million (CPM) expression value was divided against the CPM values of all junctions involving the same breakpoint of the canonical exon. All junctions were considered; and no thresholds of expression were used.

Dimension reduction and unsupervised clustering

For heatmap representations, hierarchical clustering of JETs (rows) was performed based on Euclidean distance. Pheatmap R package was used for representation. The presence/absence across samples of the 467 JETs recurrent in more than 10% of tumors in TCGA was used for Figure 1C. UMAP visualization (Figure 1D) was based on JET expression across samples in TCGA. Umap R package was used for projection and plotting. Only JETs recurrent in more than 10% of the samples were selected.

Age and conservation of JET splice sites

To characterize the evolutionary history of splice sites underlying JETs, we focused on a set of 782,112 high quality splice sites detected in myeloid cells that were previously described in Rotival et al.[25] To measure the conservation of splice sites across mammals, we retrieved pre-computed base-wise GerpRS scores[80] for hg19 from the website of the Sidow Lab (http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html). Positions that (i) had a null GerpRS score, (ii) were absent from the UCSC MULTIZ-46way alignments or (iii) had a negative score in the MULTIZ-46way alignments were excluded from our analyses. We then considered, for each splice site, the mean GerpRS score over the 2 bp that constitute the essential splice site (AG/GT sequence) as a measure of splice site conservation. Splice site with a mean GerpRS>2 were considered as conserved. Phylogenetic ages of splice sites were estimated from MULTIZ-46way alignments by extracting orthologous sequences in a 100 bp window around each splice site reconstructing the ancestral sequence by maximum likelihood with the *ancestral_inference* script from the *treetime* software (https://github.com/neherlab/treetime) and dating the first occurrence of the essential splice site (AG/GT sequence) based on the ancestors of modern humans, as described in Rotival et al.[25]

DNA/RNA extraction and RNA sequencing

For DNA extraction, cell pellets were washed, and DNA was extracted using QIAwave DNA Blood Tissue kit (Qiagen) following manufacturer instructions. DNA was measured by Nanodrop, and 1 μg was diluted in 200 μL of nuclease-free water for further PCR applications (see below). For RNA extraction, 1M cells were collected with 750 μL of RLT lysis buffer supplemented with 1% β-mercaptoethanol. Total RNA was extracted from cell lines using RNeasy Mini Kit (Qiagen) following manufacturer instructions. 1 μg of RNA was retrotranscribed using SuperScript III Reverse transcriptase (ThermoFisher), with oligo dT primers, following manufacturer instructions. cDNA was diluted (1:5) in nuclease-free distilled H_2O for further use.

For sequencing purposes, RNA interrogation number (RIN) was measured using a Bioanalyzer RNA 6000 nano assay (Agilent) to verify a RIN >7. Libraries were generated from 500 ng or 1000 ng of total RNA using the Illumina TruSeq Stranded mRNA Library Preparation Kit. PolyA enrichment was done using magnetic beads and then, RNA was fragmented. cDNA was synthetized from the resulting fragments, and after dA-tailing, they were ligated to TruSeq indexed adapters, and amplified by PCR (12-15 cycles). Sequencing was performed (paired-end reads, 100 nucleotide length) on a Novaseq 6000 system with a demanded depth of 200 million reads.

PCR amplification

PCR primers were designed in ApE software, ordered for synthesis to Eurogenentec, and resuspended with H_2O at a final concentration of 0.5 μM (sequences indicated in Table S5). For each JET primer combination, 5 ng of cDNA or genomic DNA was amplified using GoTaq G2 Hot Start Polymerase (Promega) following manufacturer instructions. Reactions were carried out in a Veriti 96-Well Thermal Cycler (Thermo Fisher Scientific). PCR products were run in a 2% agarose gel SYBR Safe Dye (1:10,000, Invitrogen), and bands at expected length were purified using the QIAquick Gel Extraction Kit (Qiagen), sequenced by EuroFins Scientific, and compared with JET sequence using ApE software.

Genome-guided transcriptome assembly

RNA-sequencing data generated in-house were subjected to the standard quality controls of the Curie sequencing platform. Additional FASTQ files from cell lines of interest were retrieved from publicly available repositories[60–63] (specified in Table S1). Raw RNA-seq files were aligned using STAR (v2.5.3a) single-pass and two-pass modes.[66] When single-pass mode was used, –FileChrStartEnd option was enabled including the list of recurrent JET coordinates. The other STAR parameters were the following: –sjdbOverhang 100, –outSAMtype BAM SortedByCoordinate, –outSAMunmapped Within, –bamRemoveDuplicatesType UniqueIdentical, –outMultimapperOrder Random, –outFilterMismatchNmax 6, –alignSJDBoverhangMin 1, –outFilterType BySJout, –alignSJoverhangMin 5, –outSAMattributes All, –quantMode GeneCounts, –outFilterMismatchNoverLmax 0.04, –outFilterMatchNminOverLread 0.33, –outFilterScoreMinOverLread 0.33, –outFilterMultimapNmax 1000, –winAnchorMultimapNmax 1000, –chimOutType WithinBAM, –chimSegmentMin 10, –chimJunctionOverhangMin 10.

The outputs of both modes were processed in downstream processes in parallel. Bam files for each cell line were processed by StringTie v2.1.4.[81] A consensus gtf file was then generated using the StringTie-merge option, and it was concatenated to the hg19 human reference genome (Ensembl).

Ribosome profiling analysis

RiboSeq data from H1650 and H1395 cell lines (cell expansion, sample preparation, and sequencing) was performed by EIRNA BIO. Briefly, H1650 (ATCC, CRL-5883) and H1395 (ATCC, CRL-5868) were cultured up to 70-80% confluence. Cells were detached and

washed with PBS/Cycloheximide (CHX) prior to lysis in polysome lysis buffer. Lysate was RNase-treated before ribosome enrichment. Then, ribosome protected mRNA fragments were isolated and used to generate RiboSeq libraries. RiboSeq libraries were sequenced on a depth of 100M reads/sample in Illumina HiSeqX. The following read structure was used: QQQ - sequence-of-interest- NNNNN - BBBBB – adaptor; where Q corresponds to untemplated additions, N to UMIs, B to demultiplexing barcodes, and the following adaptor sequence: AGATCGGAAGAGCACACGTCTGAA.

Additional RiboSeq publicly available data (Table S1) was used from the following publications.[2,60,61,63,64]

Adaptors from RiboSeq FASTQ were removed using cutadapt v1.8[70] with the parameters specified by the publication of origin, or–u 3 –a AGATCGGAAGAGCACACGTCTGAA for the *in-house* generated data. Then, rRNA contaminants were discarded by aligning using bowtie2 v2.2.5.[71] Unaligned reads were then mapped against the assembled transcriptome using STAR with the following parameters: –sjdbOverhang 29, –outSAMstrandField intronMotif, –outSAMtype BAM SortedByCoordinate, –outFilterMismatchNmax 2, –outFilterMultimapNmax 20, –outFilterType BySJout, –outSAMattributes All, –quantMode GeneCounts, –outWigType bedGraph, –outWigNorm RPM, –outMultimapperOrder Random, –outFilterMismatchNoverLmax 0.04.

Generated bam files were processed using RiboseQC v1.1. Briefly, genome annotation from StringTie assembled transcriptome was generated with prepare_annotation_files function from RiboseQC R package. The generated.gtf_Rannot file was used as reference to deduce P-sites from RiboSeq Bam file using RiboseQC_analysis function from RiboseQC R package. The output was interrogated using ORFquant v1.02.0[2] to obtain the fasta sequences and gtf coordinates of the translated JET-containing transcripts (i.e., JET-ORFs). The called ORFs were blasted against RefSeq Curated protein databases (retrieved on December 2022). Only ORFs with less than 95% similarity and/or with the insertion of 5 amino acids were considered. Only ORFs overlapping regions with annotated CDS were considered as JET-derived isoforms. Custom scripts in R were used to locate the JET-induced exon within the ORF (Start, Internal or End). Transcripts and ORF schemes were generated using ggtranscript package using the coordinates obtained from ORFquant generated GTF files.

Translation efficiency calculation

Translation efficiency was calculated from matched RNAseq and RiboSeq data from H1395 and H1650 cell lines. Cell lines were used in triplicates. Translation efficiency was defined by the ratio between RiboSeq CPM junction expression and RNAseq CPM junction expression. Read counts were extracted from SJ.out.tab file generated by STAR.

Deep proteome mass spectrometry analysis

Raw files were downloaded from MASSIVE repository (MSV000086944) and processed using MSFragger v3.7 in FragPipe v19.1 environment with the following parameters: precursor mass tolerance 10ppm and fragment mass tolerance 0.02 Da. Methionine oxidation (+15.995Da), N-acetylation (+42.011Da) were enabled as dynamic modifications. Carbamidomethylation (+57.021Da) was considered as fixed modification. Enzymatic digestion was selected accordingly (trypsin, chymotrypsin, AspN, GluC, LysC and LysN). MSBooster rescoring was enabled and Percolator was used to filter at a false discovery rate (FDR) of 1% at peptide level. No FDR was used at protein level. MS/MS spectra were searched against the human proteome from Uniprot/SwissProt with isoforms (updated 06.03.2020) and concatenated with the JET-ORFs. Identified peptides were filtered by the human proteome (SwissProt + RefSeq) considering L and I as equivalent. Only peptides uniquely mapping the JET-ORF were considered.

Instability index calculation

The instability index of JET-ORF was calculated using the instaIndex function from the Peptides R package.[73]

Proteomics for half-life measurement

Cycloheximide treatment on K562 cells

K562 cells (ATCC: CCL-243) were cultured in RPMI medium supplemented with 10% FBS in 24-well plates at a concentration of 5 million cells per well. 10 μL of cycloheximide (Sigma C4859-1ML) at 10mg/mL were added to each well (excepting in 0h timepoint) to achieve a final concentration of 50 μg/mL. Cycloheximide was homogenized by resuspending 6 times each well. Cells were incubated at 37°C and collected at the following time points (0h, 1.5h, 4h, 8h, 24h) in 15 mL tubes, washed with PBS and snap-frozen.

Cell lysis

K562 cell pellets were disrupted in 200 μL lysis buffer (2.2 % SDS, 0.1 M Tris-HCl, pH 7.5) with an ultrasonic processor VibraCellTM VCX 130PB (30 % amplitude; 10 x 1 s pulse; tubes kept on ice). Cell lysates were then heated at 95° C and 300 rpm for 5 min. After being tempered, lysates were sonicated again (5 x 1 s pulse) and centrifuged at 16000 g at 13°C for 20 min. These two steps were repeated to ensure a good cell disruption and to obtain clarified lysates to be used in further analysis. Aliquots from the supernatants were taken from each cell lysate and kept at -80° C until use. Protein abundance in the cell lysates was quantified by the microBCA method (ThermoFisher, #: 23235).

Protein digestion and TMT mass tag labeling

Fifty μg of each sample were digested with trypsin by the Filter-Aided Sample Preparation (FASP) method using 10 kDa Amicon centrifugal filters (Millipore; # UFC501096). An additional sample of 50 μg was created from the mixture of the three replicates from sample 24 hours. Tryptic peptides were evaporated in speed-vac before being tagged for quantitative mass spectrometry analysis. Each peptide sample was reconstituted with 100 μL 100 mM TEAB (Triethylammonium bicarbonate buffer; Sigma, # T7408) for 10 min at

700 rpm at 20° C. TMT 16-plex isobaric label reagents (ThermoFisher, # A44521 Lot: YJ374827) were reconstituted with 20 μL acetonitrile (Fisher Scientific, # 10001334) and labelling was conducted following the manufacturer's instructions. Briefly, samples were incubated with TMT mass tags for 1h at room temperature (RT) as indicated in:

126	127N	127C	128N	128C	129N	129C	130N	130C	131N	131C	132N	132C	133N	133C	134N
24_1	4_2	24_3	4_3	0_3	8_1	4_1	2_2	2_1	8_2	24_2	0_1	24_mix	8_3	2_3	0_2

Each individual labeling reaction was quenched with 5 μL of 5% hydroxylamine for 15 min before being mixed.

High-pH–Reverse phase liquid chromatography

The TMT mix aliquot corresponding to the 10% of the total sample (80ug) was reconstituted with 200 μL 5mM ammonium formate, pH 10, 2% ACN before fractionation in a HPLC 1100 UV-Vis (Agilent), with a XBridge Peptide BEH C18, 130Å, 5 μm, 2.1mm x 100mm column (Waters, #186003575).

The sample was loaded into the chromatographic system. The High-pH–Reverse phase chromatographic gradient was conducted to separate labelled peptides at a 200 μL/min flow for a total run time of 81 min, using solvent A (5mM ammonium formate, pH 10, 2% ACN) and B (5mM ammonium formate, pH 10, 90% ACN). Fractions were collected every minute between 0 and 70 minutes, 200 μL per fraction. All the fractions obtained were stored at -80°C. In accordance with the chromatographic profiles obtained, fractions from 2 to 66 min were considered for further analysis. The fractions were evaporated until dryness and after that reconstituted with 5% MeOH, 0.5% TFA. The fractions were mixed two by two, a total of 34 mixes of fractions were obtained.

LC-MS/MS quantitative analysis

High-resolution LC-MS/MS was conducted for the 34 mixes, the percentage analyzed corresponds to the 12.5% of each individual fraction. The MS system used was an Orbitrap Exploris 480 (ThermoFisher) equipped with a nanoESI ion source. The samples were loaded into the chromatographic system consisting in a C18 nanoEase M/Z Symmetry, 180 μm x 20mm Trap Column (Waters) connected to a nanoEase M/Z HSS C18 T3, 75 μm x 100 mm column (Waters). The separation was done at 0.5 μL/min in a 60 min acetonitrile gradient from 2 to 30% (solvent A: 0.1% formic acid, solvent B: acetonitrile 0.1% formic acid). The HPLC system used was an ACQUITY UPLC M-Class. The Orbitrap Exploris 480 was operated in the positive ion mode with a spray voltage of 1.8 kV. The spectrometric analysis was performed in a data dependent mode, acquiring a full scan followed by 10 MS/MS scans of the 10 most intense signals detected in the MS scan from the global list. The full MS (range 400-1600) was acquired in the Orbitrap with a resolution of 60.000. The MS/MS spectra were also done in the Orbitrap with a resolution of 15000 (TurboTMT: TMTpro reagent). The collision energy was 35% for MS2 HCD.

Bioinformatic analysis

For the bioinformatics analysis, we combined the *in-house* generated data with a publicly available dataset.[28] Raw files were processed using Sequest and Comet in ProteomeDiscoverer 3.0 with the following parameters: precursor mass tolerance 10ppm and fragment mass tolerance 0.02 Da (for public data: 50 ppm and 0.6 Da, respectively). Carbamidomethylation (+57.021 Da), and TMTpro (+304.3127 Da) at N-terminus and K were considered as fixed modifications. INFERYS rescoring was enabled and Percolator was used to filter at a FDR of 1% at peptide level. No FDR was used at protein level. Reporter ions were quantified in MS2 (*in-house*) and MS3 (public). MS/MS spectra were searched against the human proteome from Uniprot/SwissProt with isoforms (updated 06.03.2020) and concatenated with the JET-ORFs. Identified peptides were filtered by human proteome (SwissProt + RefSeq) considering L and I as equivalent. Only peptides uniquely mapping the JET-ORF were considered.

JET-ORF reexpression using mNG

Plasmid amplification

Expression plasmids encoding mNG11 tagged ORFs under a SFFV promoter and both ampicilin and puromycin resistance (pTwist Lenti SFFV Puro WPRE) were ordered at TwistBioscience. Expression plasmids were resuspended with nuclease-free H2O at a final concentration of 100 μg/mL and amplified using NEB® 10-beta Competent E. coli (NewEngland Biolabs: C3019H). Briefly, 100 ng of plasmid DNA were added to 25 μL of bacteria and placed on ice for 30 minutes. Heat shock was performed for 30 seconds at 42°C and then, placed again on ice for 5 minutes. 100 μL of 10-beta/Stable Outgrowth Medium was added into the mixture and incubated for 90 minutes at 37°C and 180 rpm of agitation. Later, the mixture was spread into pre-warmed bacteria culture plates with ampicillin selection and incubated overnight at 37°C. Individual colonies were then picked and pre-amplified in 2 mL of Difco LB Broth, Miller Luria-Bertani medium (BD, 244620) for 4-6h in the presence of ampicillin (1mg/mL, Euromedex, EU0400-A) at 37°C and 180 rpm of agitation. Then, transformed bacteria were amplified overnight in 180 mL. Bacteria were harvested by centrifugation and DNA was extracted using NucleoBond Xtra Midi Plus EF kit (Macherey-Nagel, 740422-50) following manufacturer indications. DNA was eluted in nuclease-free water and was quantified by Nanodrop.

Lentivirus production

First, 19 μg of plasmid containing the mNG11 tagged JET-ORF, 5.64 μg of envelope (pVSVG), and 13.5 μg of packaging (psPAX2) plasmids were mixed in 3.8 mL of Opti-MEM I Reduced Serum Medium. 114 μL of TransIT-293 Reagent was added to the DNA mixture and incubated at RT for 15-30 minutes. The TransIT:DNA complexes were added drop-wise to HEK293T-Lenti-X cells (Takara) plated at concentration of 9 M cells per T175 flask the day before. After 60 h after transfection, supernatant was collected,

and lentiviral particles were ultracentrifugated (31000 g 90min 4°C) in a 20% sucrose gradient. Lentivirus preparations were aliquoted and stocked at -80°C.

Lentivirus transduction

For transduction in HeLa and HEK293, 50 μL of lentivirus suspension was transferred to 0.25M cells constitutively expressing mNeonGreen 1-10 (mNG1-10) fragments in 6-well plates cultured with DMEM 10% FBS 4 μg/mL polybrene. For transduction in K562, 50 μL of lentivirus suspension was transferred to 0.5M K562-mNG1-10 cells in 24-well plates in the presence of RPMI 10% FBS 4 μg/mL polybrene. Then, cells were spinoculated during 90min at 800g and 32°C. Ectopic expression was evaluated at least after 48h by flow cytometry and confocal microscopy.

In vitro mNG-based protein stability assay

The following mNG11-tagged JET-ORFs, CAN-ORFs, and control proteins were used: PTEN JET/CAN, WWOX JET/CAN, ALDH3A2 JET/CAN, H2AFY JET/CAN, IL15RA JET/CAN, CLK4 JET-/CAN, UBFD1 JET/CAN, Laminin B1, Geminin, and PRELIB3. At day 6 post-transduction (as specified above), mNG11-tagged JET-ORF expressing K562-mNG1-10 cells were counted and 60,000 cells/well were seeded in 96-well plates in RPMI 10% FBS. Cycloheximide (Sigma C4859-1ML) was added to each well (excepting in 0h timepoint) to achieve a final concentration of 50 μg/mL. Cycloheximide was homogenized by resuspending 6 times each well. Cells were incubated at 37°C and collected at the following time points (0h, 1h, 2h, 4h, 6h, 8h, 24h) in 96-well plates, washed, and resuspended in 100 μL PBS 1% FBS 1/1000 DAPI. Cells were acquired in BD FACSVerse™ Cell Analyzer and FCS files were analysed using FlowJo v10.6.1.

Confocal microscopy

Target cells were plated in 10mm petri dishes at a concentration of 50,000 cells/mL in DMEM 10% FBS. After overnight incubation, the medium was replaced with culture medium containing 1/5000 dilution of sirDNA (Spirochrome), and it was incubated for 30-60 minutes. Then, medium was replaced with FluoroBrite DMEM medium (Gibco) and live imaging was performed using Inverted Eclipse Ti-E (Nikon) and Spinning disk CSU-X1miscroscope (Yokogawa) integrated in Metamorph software (Gataca Systems). Fluorescence intensity was quantified using Image J.[74]

RT-qPCR

JET-specific TaqMan probes were designed using the AlleleID software (Premier Biosoft) using default parameters (or slight modifications in the probe or primer lengths). Custom TaqMan probes and primers were ordered in Eurogentec. Concentrated TaqMan assays (20×) were prepared by mixing 5 μl of TaqMan probe, 18 μl of each primer, and 59 μl of ddH2O. Glyceraldehyde-3-phosphate dehydrogenase (*GAPDH*), Beta-2-Microglobulin (*B2M*), and hypoxanthine-guanine phosphoribosyl-transferase (*HPRT1*) TaqMan assays (Fisher Scientific) were used as housekeeping gene controls. RT-qPCR was performed in 384-well plates, using dTTTP Master-Mix (Eurogentec), according to the manufacturer's instructions (for each well, 0.5 μl of TaqMan assay, 5 μl of Mastermix, and 4.5 μl of cDNA). The amplification was performed as recommended by the manufacturer, using a Roche LightCycler 480, and the relative quantification tool from the analysis software was applied. The detection threshold was placed manually in the first one-third of the linear amplification phase.

For the qPCR-based Percent Spliced In (PSI) calculation, the expression of each JET and associated canonical junction were quantified by qPCR with primers and Taqman probes. The PSI was calculated using the following formula:

$$PSI = \frac{2^{-\Delta Ct}(\text{JET})}{2^{-\Delta Ct}(\text{JET}) + 2^{-\Delta Ct}(\text{canonical})} \times 100$$

Membrane enrichment for ALDH3A2 JET-ORF

15 million of HeLa cells expressing ALDH3A2 JET-ORF tagged with FLAG were collected and washed twice with PBS. Untransduced HeLa cells were used in parallel as negative control. Cells were lysed with 1 mL of lysis buffer (150 mM NaCl, 50 mM Tris-HCl, 5 mM EDTA, and cOmplete protease inhibitor (Roche) at pH8) and without detergent. After sonication, cell lysates were centrifuged at 5,000g to remove big organelles and cell debris. Supernatant was collected and centrifuged at 100,000g during 1h. Cell membranes in the pellet were solubilized with 400 μL of lysis buffer with 1% n-dodecyl-bêta-maltoside (FisherScientific). Finally, non-solublised membranes were removed by centrifugation at 100,000g during 1h. Protein fractions (i.e., total, cytosol, and membrane) were quantified using Pierce BCA Protein Assay kit (ThermoFisher) and analyzed by western blot.

Western blot

Cell lysates (HIF1α levels – 500,000 initial cells) or solubilized membranes (ALDH3A2 experiment – 2 μg total protein) were mixed with Laemmli sample buffer (Bio-Rad) with 10% β-mercaptoethanol and boiled 10min at 95°C. Samples were run in Mini-Protean TGX Stain-Free gels 4-15% gradient (Bio-Rad) during 30min at 200V and gel protein content was transferred to ImmunoBlot PVDF membranes (Bio-Rad). Membranes were blocked with 3% milk in TBS 0.1% Tween20 during 1h before being incubated with the primary antibody overnight at 4°C and with the secondary HRP-conjugated antibodies for 1h at RT. Membranes were developed using Clarity

Western ECL solutions (BioRad). The primary antibodies were used in TBS 1% Tween20 3% milk. The following primary antibodies were used: anti-FLAG (1/1000 dilution, FG4R clone, MA1-91878, FisherScientific, RRID: AB_1957945), anti-HIF1α (1/5000 dilution, polyclonal, 20960-1-AP, Proteintech, RRID: AB_10732601), anti-PTEN (1/1000 dilution, polyclonal, 22034-1-AP, Proteintech, RRID: AB_2878977), and anti-actin (1/1000 dilution, clone C4, MAB1501, MerkMillipore, RRID: AB_2223041). The anti-mouse HRP-conjugated (1/10000, Jackson ImmunoResearch) and anti-rabbit HRP-conjugated (1/10000, Jackson ImmunoResearch) were used as secondary antibodies. Band intensity was calculated using Image J.

JET KO and single-cell clone generation

Guide design

For the crRNA guide design, the exonized TE sequences together with the flanking 400 nucleotides were submitted to Alt-R Custom Cas9 crRNA Design Tool from IDT. Three crRNA guides were selected targeting the upstream TE sequence and three crRNA guides for the downstream sequence. The 9 possible dual combination (1 upstream + 1 downstream) were tested experimentally (as explained below) and the best combination was used for the single cell clone generation. CRISPR/Cas9 guides are indicated in Table S6.

rRNA:Cas9 complex nucleofection

crRNAs were chemically synthesized by IDT, and resuspended in IDT duplex buffer at a concentration of 100 μM. rRNA:Cas9 preparation was performed in parallel for the 2 guides targeting the upstream and downstream region of the exonized TE. Briefly, 3.5 μL of crRNA were mixed with 3.5 μL of tracrRNA and annealed by heating at 95 °C for 5 min. Complexes were slowly cooled to RT by ramp down. Then, 1.91 μL of Cas9 were added and incubated at RT for 15 minutes. HEK293FT cells, already in culture for 2 passages using DMEM 10% FBS 1% penicillin/streptomycin, were prepared at a concentration of 0.2 million cells per reaction. After 2 washes with PBS, the supernatant was removed. Before nucleofection, 2.5 μL of each gRNA:Cas9 (5 μL in total) were added to 20 μL of SF Cell Line Nucleofector® Solution:Supplement 1 (V4XC-2032, Lonza). Cells were resuspended with the total 25 μL solution and transferred to a certified cuvette. Cells were electroporated using CM-130 program in a 4-D Nucleofector (Lonza). Immediately after electroporation, cells were rescued by adding 75 μL of pre-warmed DMEM 10% FBS directly into the electroporation well. After 10min at 32 °C, nucleofected cells were transferred to 48-well plates with 500 μL of pre-warmed medium. Cells were incubated at 32 °C overnight, and then moved to 37 °C. To evaluate CRISPR/Cas9 KO efficiency, DNA was extracted at day 4-5 (as mentioned before) and PCR followed by gel electrophoresis was performed. Mock control cells were performed in parallel using exactly the same conditions.

Single cell clone generation

Nucleofected cells (KO and mock) were collected at day 5 post-transfection. In general, 3 reactions were performed in parallel to increase the cell number. Cells were then single cell sorted in 96 well-plates containing 150 μL of DMEM 10% FBS 1% penicillin/streptomycin 100 μg/ml Normocin (ant-nr-05, Invivogen) using a Sony SH800 sorter. Single cells were cultured for 2-3 weeks at 37 °C. Wells with amplified cells were trypsinized and amplified to 48 well-plates. DNA was extracted for each cell clone (as specified above) and the TE deletion was verified by PCR and gel electrophoresis. Clones with homozygous deletion of the exonized TE were selected, amplified and cryopreserved. The absence of the JET expression was evaluated by PCR from extracted RNA for the selected KO clones (as specified above).

Differential expression analyses

WWOX and PTEN overexpression experiments

HeLa mNG1-10 cells expressing WWOX JET-ORF or CAN-ORF, PTEN JET-ORF or CAN-ORF were plated in 6-well plates (0.5M cells/well) in DMEM 10% FBS. HeLa cells expressing only mNG1-10 were used as negative control. Five replicates were used per condition. Expression levels of the ectopic expressed ORFs were previously evaluated using flow cytometry. Cells were incubated overnight at 37 °C before lysing for RNA extraction and RNA sequencing (following the specifications mentioned above).

In an independent experiment, PTEN JET-ORF and CAN-ORF were overexpressed in H1650 cell line. Transduced H1650 were plated in 6-well plates (0.5M cells/well) in RPMI 10% FBS. Untransduced H1650 cells were used as negative control. Five replicates were used per condition.

PTEN-JET SSO experiments

SSO probes were reconstituted in sterile RNAse-free water at a 100 μM concentration. H1650 cells were transfected with SSOs using Lipofectamine™ 3000 transfection reagent (ThermoFisher, L3000015) according to the manufacturer's instructions. Briefly, cells were seeded in RPMI 1640 media (Gibco, 21875034) without antibiotics in 12-well plates (10^5 cells/well) and allowed to reach 70-80% confluence prior to transfection. Lipofectamine reagent was diluted in Opti-MEM® Reduced Serum Medium, and SSOs were diluted in Opti-MEM® media, containing the P3000 reagent. An incubation of 10-15 minutes at RT was performed to allow complex formation. The transfection mix was added dropwise to the cells, and incubated for 48 hours at 37°C with 5% CO2. Six hours after the transfection, the media was removed and replaced by RPMI 1640 media containing 10% FBS and 1% P/S. Twenty-four hours after the first transfection, cells were detached with Trypsine-EDTA (Gibco, 25300054) and each well was divided in 2 wells of new 12-well plates in 2ml of RPMI 10% FBS 1% P/S. After 24h, a second transfection was done following the same parameters, and cells were incubated for 24 hours before lysis.

Single cell HTT- and UBFD1-JET KO clones

For HTT-JET, 3 KO and 4 mock clones were used. For UBFD1-JET, 5 KO and 5 mock clones were used. Each clone was considered as a biological replicate. Cell clones were plated in 6-well plates (0.5M cells/well) in DMEM 10% FBS 1% P/S. After overnight incubation at 37 °C, the medium was removed carefully and RLT buffer (+ 1% β-mercaptoethanol) was used to lyse cells and proceed to RNA extraction (as described above).

RNA sequencing and bioinformatics analysis

Sequencing parameters and RNA-sequencing data alignment were performed as previously described in this manuscript. Bam files were processed by featureCounts to quantify gene expression.[75] Gene read counts were uploaded to R and converted to Transcript per million (TPM) using tpm function from drfun R package. Only raw gene counts from genes with TPM >0.5 were submitted to differential expression analysis using DESeq2 R package.[82]

Differentially expressed genes were analyzed by GeneOntology using enrichr R package,[76] and their functional association was evaluated in STRING-DB.[83] Transcription factor activity was inferred from WWOX JET-ORF, CAN-ORF or HeLa mNG1-10 using decoupleR.[77] Only genes with an absolute log2foldChange greater than 0.15 were used.

Gene phylostratum and conservation

Gene age information was retrieved from Litman and Stein[53] and conservation across vertebrates was calculated using phastCons.[54]

Structure analysis

Gene phylostratum and secondary structures

The number of secondary structures per protein was retrieved from Uniprot (Swissprot without isoforms, 11.04.2024). Only protein entries with an associated PDB file were considered (8,251 Uniprot entries).

Three-dimensional modelling

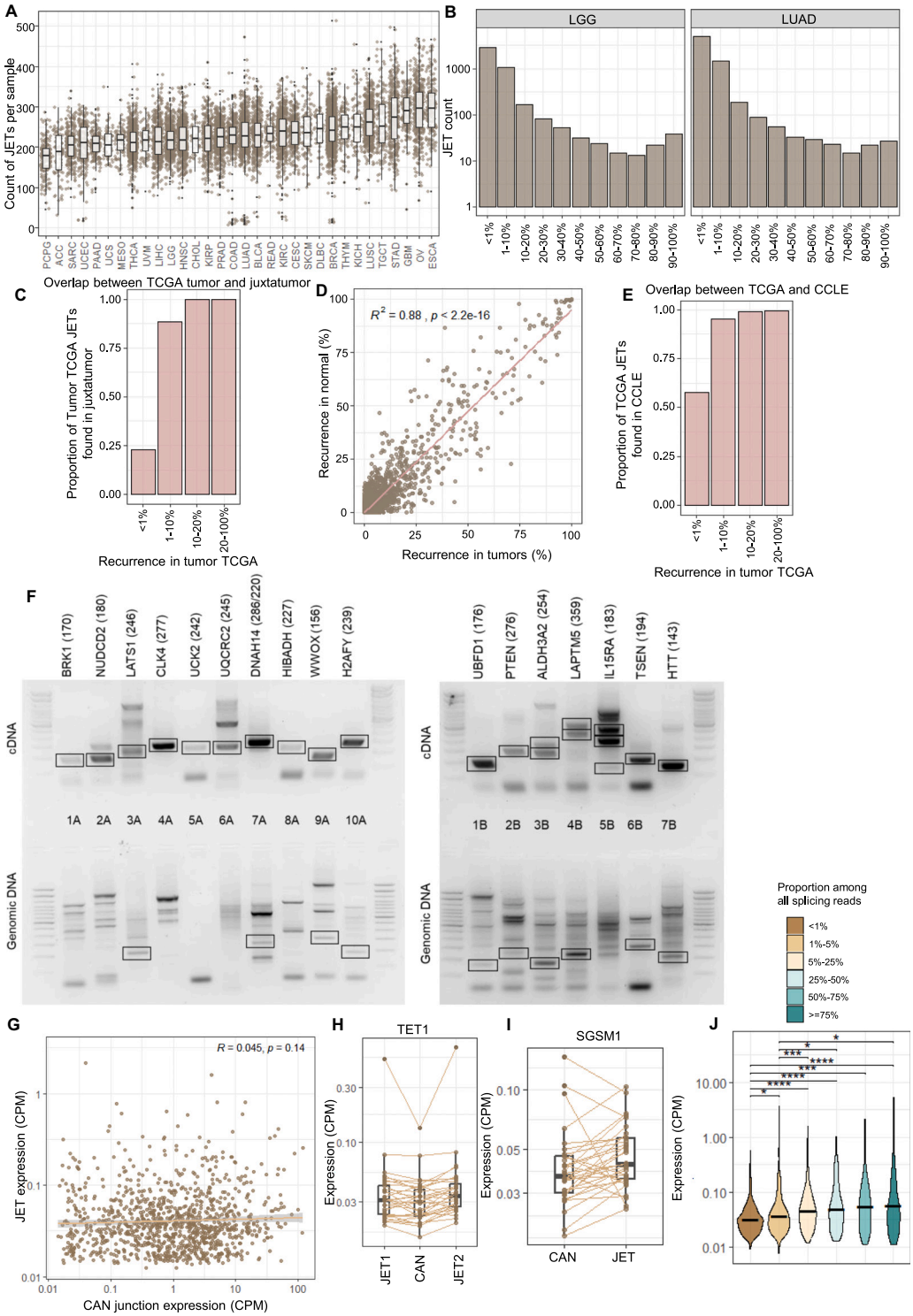
A set of 135 TE sequences with an internal TE insertion and additional 5 JET-ORFs (i.e., PTEN, WWOX, IL15RA, H2AFY, and ALDH3A2) were selected. The 3D structures of the JET-ORF and the corresponding CAN-ORF were predicted using a local installation of ColabFold (version 1.5. 1). The prediction used templates and was performed with 3 recycles. The structures were relaxed with amber option. All the other parameters were fixed as their default values. The best ranked relaxed model was kept for the analysis. Only the structures with a predicted local distance difference test (pLDDT) average score greater than 70, which is the threshold for good quality models, were kept for further analyses. The secondary structure of these 3D models was analyzed with a local installation of DSSP (version 4.4). Among the assigned states (H, G, I, E, B, S, T, P), the five first ones were considered as secondary structures whereas the three last ones plus the non-assigned state (blank state) were grouped in the coil structure. Calculating the % Sec. Structures for Cano & JET. The canonical sequences with a content of secondary structure > 40% were selected. The differences in secondary structures content between canonical sequences and TE sequences were calculated.

QUANTIFICATION AND STATISTICAL ANALYSIS

The number of replicates and independent experiments are indicated in the figure and/or figure legends. The statistical tests were performed by ggpubr in R: unpaired t-test, Kruskal Wallis (for non-parametric comparisons with 2 groups), and Mann-Whitney test (for multiple comparisons in non-parametric data). If the opposite is not specified in the figure legend, asterisks denote Mann-Whitney or t test p-values. P-values are indicated in the figures (*p<0.05; **p<0.01; ***p<0.001 and ****p<0.0001).

TE class enrichment tests were performed with Hypergeometric tests. Multiple testing corrections were done by calculating the FDR-adjusted p-value for each class. An FDR-adjusted p-value <0.05 was considered significant. TE age was retrieved from Choudhary et al.[84]

Supplemental figures



(legend on next page)

Figure S1. JET expression and recurrence in TCGA and cell lines, related to Figure 1

(A) Boxplot of the number of JETs (y axis) per sample (dots) in each TCGA tumor indication (x axis).

(B) Histogram of JET recurrence across tumor samples from low-grade glioma (LGG,left) and lung adenocarcinoma (LUAD, right) TCGA projects.

(C) Bar plot indicating the overlap of JETs (in %) between TCGA tumors and tumor-adjacent normal tissues (i.e., juxtatumor), according to the JET recurrence in TCGA tumors (x axis).

(D) Scatterplot of JET recurrence in tumor versus tumor-adjacent normal TCGA samples.

(E) Bar plot indicating the overlap of JETs (in %) between TCGA tumors and CCLE samples, according to the recurrence in TCGA tumors.

(F) Gel electrophoresis of the PCR amplification of 17 representative JETs using both RNA (top) and genomic DNA (bottom) from HEK293FT cell line. Bands highlighted with a square were purified and send to sequencing. The size of the expected band is indicated in brackets.

(G) Scatterplot correlating JET expression (CPM) versus the expression of the corresponding Gencode-annotated canonical junction.

(H and I) Boxplots of the CPM expression of JETs in *TET1* (H) and *SGSM1* (I) and the corresponding canonical exon-exon junction sharing the same coding exon splice site. Each dot represents the average expression in one TCGA indication. Canonical junctions and JETs are paired according to the tissue of origin. Sashimi plots illustrating TET1 of splicing junctions between exons (orange) and TE (green).

(J) Violin plot of JET expression (in CPM, y axis) according to their proportion among all splicing events (%, x axis). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$ (Mann-Whitney test with Bonferroni adjustment).

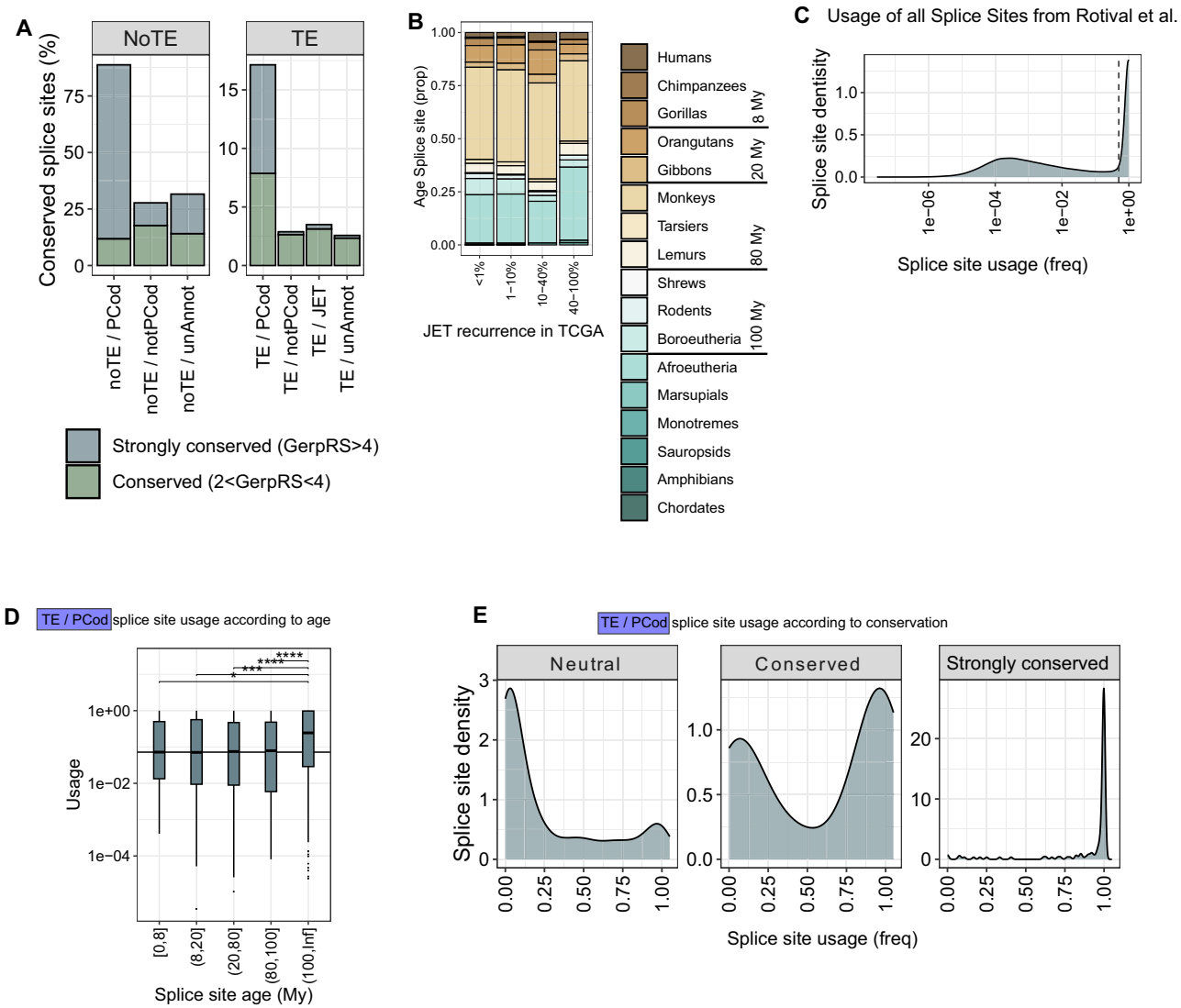


Figure S2. Age and conservation of TE splice sites, related to Figure 2

(A) Bar plot indicating the percentage of conserved (2 < GerpRS < 4, green) and strongly conserved (GerpRS > 4, blue) splice sites (y axis) according to their classification (x axis).

(B) Bar plot showing the age (color gradient) of JET splice sites (in proportion, y axis) according to their recurrence in TCGA (%, x axis).

(C) Density histogram of the frequency of usage (x axis) of the splice sites from the Rotival et al. dataset. The dashed line indicates 50% of usage.

(D) Boxplot of the frequency of usage (y axis) of TE/PCod splice sites according to the age of the splice site (x axis, My, million years). *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$ (Mann-Whitney test with Bonferroni adjustment).

(E) Density histogram of the frequency of usage (x axis) of the TE/PCod splice sites according to their evolutionarily conservation (i.e., neutral, conserved, or strongly conserved).

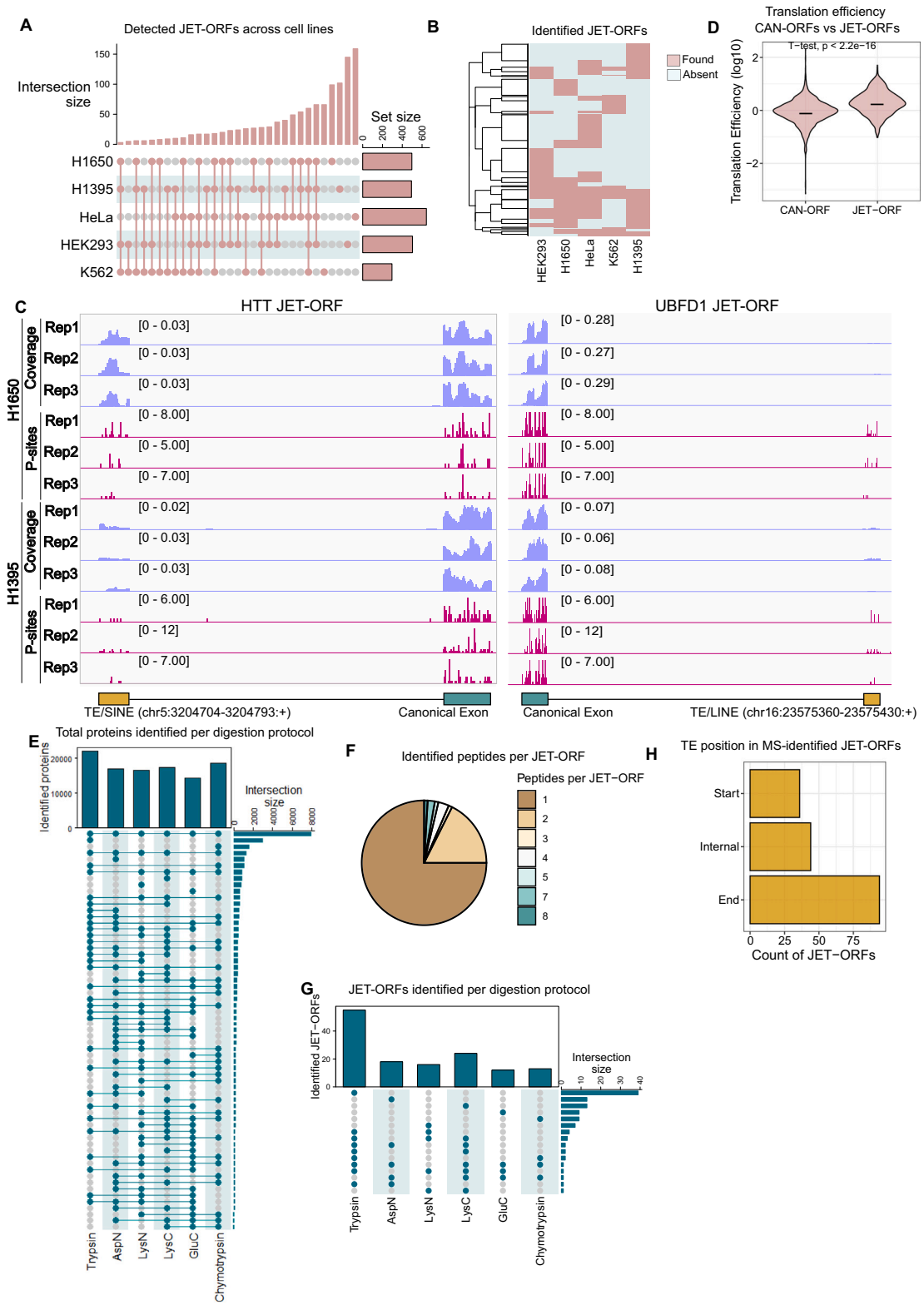


Figure S3. Identification of JET-ORFs using Ribo-seq and MS-based proteomics, related to Figure 3

(A) UpSet diagram summarizing the overlap of Ribo-seq-detected JET-ORFs between cell lines. The top bar plot shows the intersection size, and the bar plot in the right indicates the number of JET-ORFs identified per cell line.

(legend continued on next page)

(B) Heatmap representation based of the identified JET-ORFs (rows) in each cell line (columns). Euclidean clustering has been used (row order).
(C) Representative examples of the Ribo-seq coverage and uniquely mapping P-sites of HTT JET-ORF (left) and UBFD1 JET-ORF (right) identified in H1650 and H1395 cell lines.
(D) Violin plot summarizing the translation efficiency (log_{10}, y axis) of canonical junctions and JETs.
(E) Upset diagram of the total number of proteins identified by each digestion protocol.
(F) Pie chart of the number of peptides identified by mass spectrometry per JET-ORF.
(G) Upset diagram of the JET-ORFs identified by each digestion protocol.
(H) Bar plot indicating the count of JET-ORFs containing the JET-induced exon in start, internal, or end position.

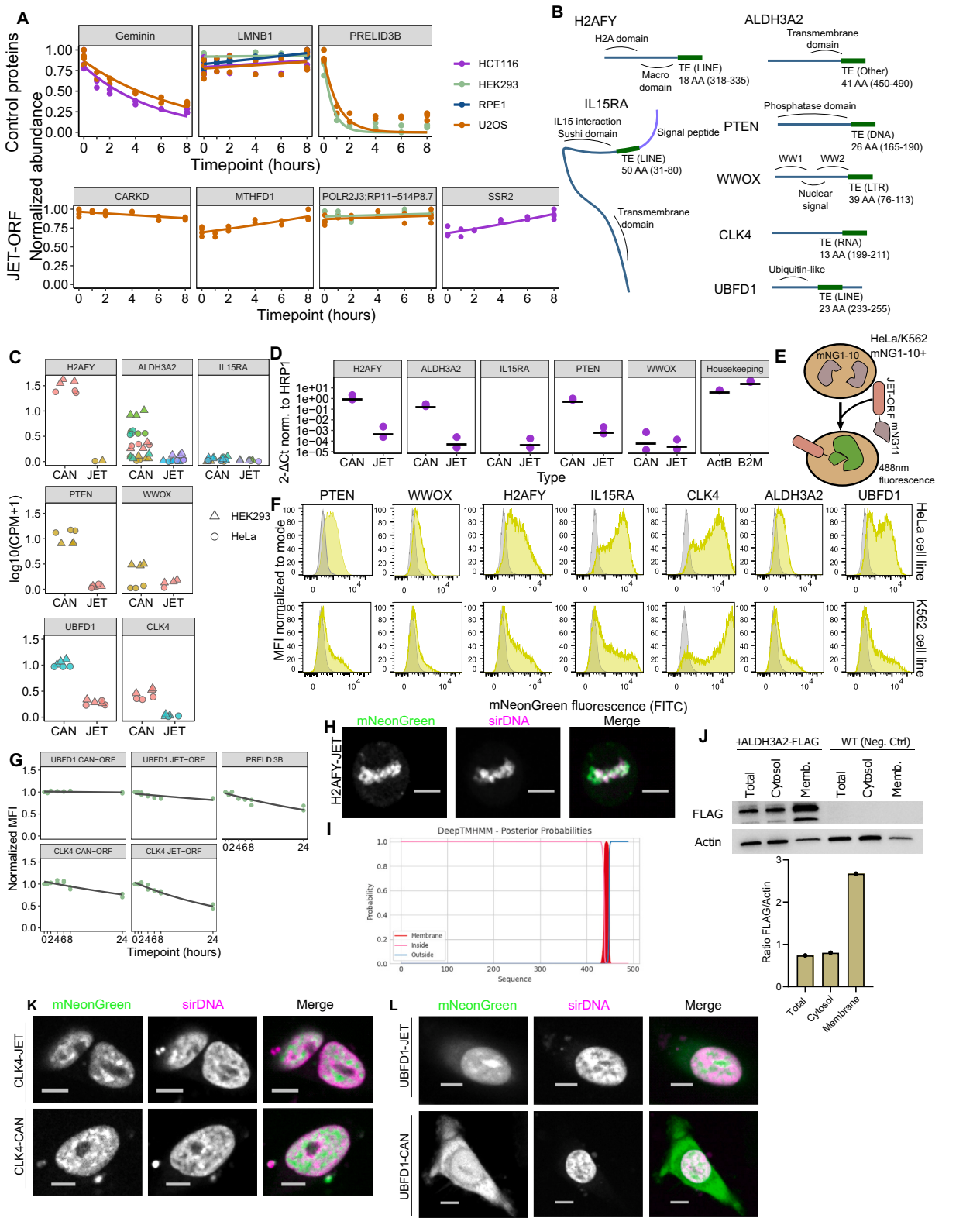
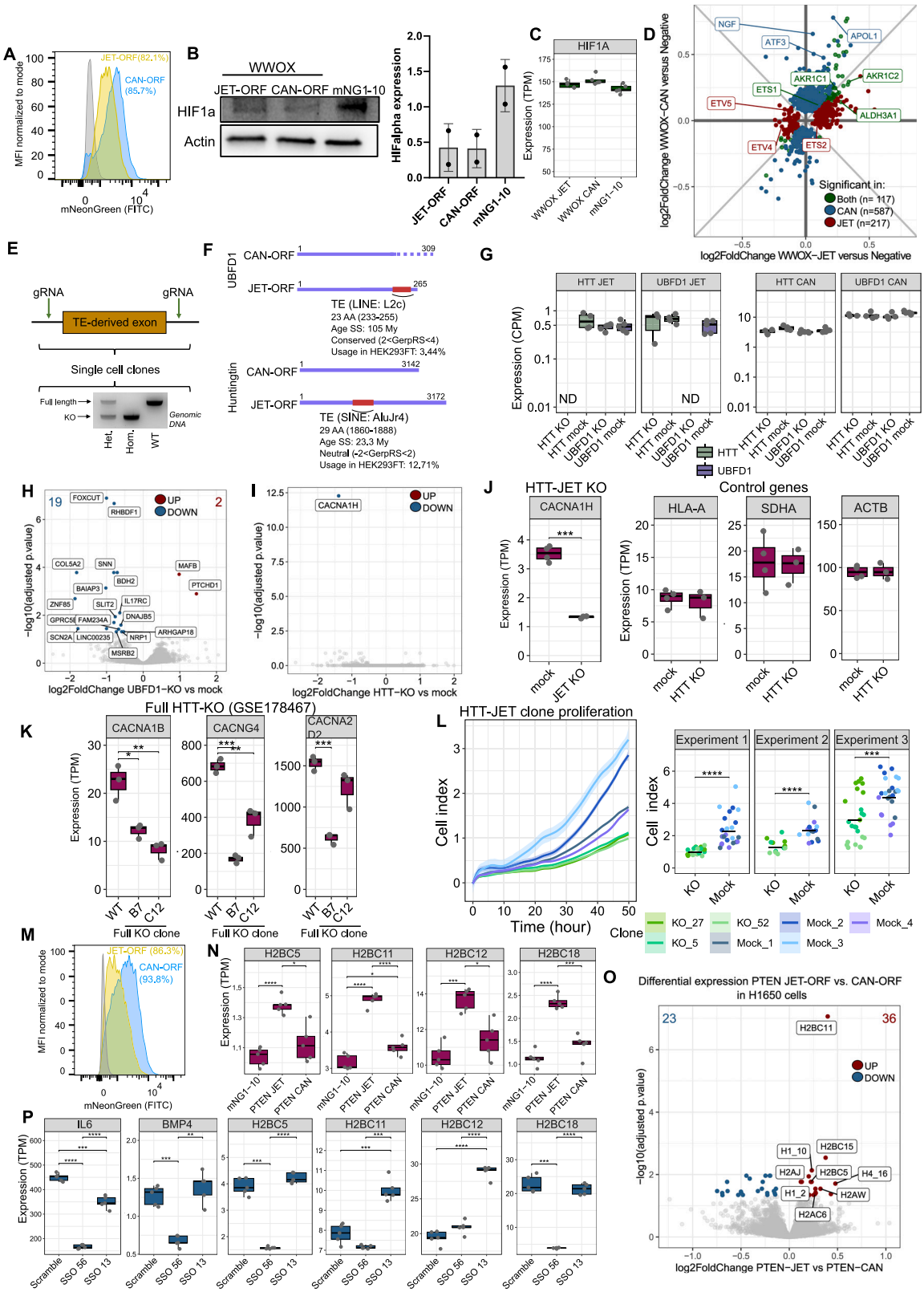


Figure S4. JET-ORF expression, protein stability, and subcellular localization, related to Figure 4

(A) Protein degradation curves of the quantitative proteomics-based stability assay in HCT116 (purple), HEK293 (green), RPE1 (blue), and U2OS (orange) cell lines. Normalized abundance of JET-ORFs across different time points (hours, x axis) of cycloheximide-based translation inhibition is shown (bottom). Geminin and PRELID3B (limited stability) and lamin-B1 (highly stable) proteins are represented as controls (top).

(legend continued on next page)

(B) Schematic representation of the seven overexpressed JET-ORFs.
(C) Expression in $\log_{10}$(CPM + 1) of the annotated exon-exon junctions and JETs in *H2AFY*, *ALDH3A2*, *IL15RA*, *PTEN*, *WWOX*, *UBFD1*, and *CLK4* based on RNA-seq from HeLa (circle) and HEK293FT (triangle) cells.
(D) Expression quantification by RT-qPCR of JETs and their corresponding annotated exon-exon junctions in HeLa cells for *H2AFY*, *ALDH3A2*, *IL15RA*, *PTEN*, and *WWOX* genes.
(E) mNeonGreen (mNG) split fluorescence system. Complementation of the mNG1-10 (expressed in target cells) and the mNG11 (linked to JET-ORF) emits fluorescence at 488 nm.
(F) Flow cytometry histograms of the ectopically reexpressed JET-ORFs (green) in the HeLa mNG1-10 (top) and K562 mNG1-10 (bottom) cell lines. HeLa or K562 mNG1-10 cells without any lentivirus transduction were used as negative control (gray).
(G) Protein degradation curves of the mNeonGreen-based stability assay in the K562-mNG1-10 cell line. The normalized MFI across different time points (hours, x axis) of treatment are shown for each CAN-ORFs (top) the corresponding JET-ORFs (bottom). PRELID3B is used as control.
(H) Confocal microscopy images of cells ectopically expressing H2AFY JET-ORF.
(I) DeepTMHMM predicted transmembrane topology of the ALDH3A2 JET-ORF.
(J) Western blot of ALDH3A2-JET-FLAG levels in total lysate, cytosol, and membrane-enriched fraction (top) and quantification shown as FLAG/actin ratio (bottom).
(K and L) Confocal microscopy images of cells ectopically expressing CLK4 JET-ORF and CAN-ORF (K) and UBFD1 JET-ORF and CAN-ORF (L).



(legend on next page)

Figure S5. Functional characterization of WWOX, UBDF1, HTT, and PTEN JET-ORFs, related to Figure 5

(A) Flow cytometry histograms of the ectopically expression levels of WWOX JET-ORF (green) and CAN-ORF (blue) in HeLa mNG1-10 cell line. HeLa mNG1-10 cells were used as negative control (gray).

(B) Representative western blot image of HIF1α protein levels after overexpression of WWOX-JET, WWOX-CAN, or control HeLa mNG1-10 cells (n = 2). Quantification is shown in the right.

(C) Quantification of HIF1α expression after overexpression of WWOX-JET, WWOX-CAN, or control HeLa mNG1-10 cells at RNA level (using RNA-seq).

(D) Plot representing the $\log_2$foldchange of gene expression in WWOX CAN-ORF (y axis) or JET-ORF (x axis) expressing cells versus the negative control (HeLa mNG1-10). Dots are colored according if they are significantly expressed in cells transduced with either WWOX CAN-ORF (blue) or WWOX JET-ORF (red) or in both cells (green).

(E) Scheme representation of the CRISPR-Cas9 strategy for TE-derived exon depletion.

(F) Schematic representation of UBFD1 and HTT JET-ORFs and the corresponding CAN-ORFs.

(G) Boxplots of the expression (in CPM, y axis) of the HTT and UBFD1 JET-ORF or CAN-ORF in KO cells or mock cells (x axis). Each dot indicates a single-cell clone. ND, not detected.

(H) Volcano plots of the expression $\log_2$foldchange (x axis) between UBFD1-JET depleted clones (red) and mock clones (blue). Colored dots indicate significant p values.

(I) Volcano plots of the expression $\log_2$foldchange (x axis) between HTT-JET depleted clones (red) and mock clones (blue). Colored dots indicate significant p values.

(J) Boxplot of expression (TPM, y axis) of *CACNA1H*, *HLA-A*, *SDHA*, and *ACTB* genes in mock and HTT-JET KO clones. Each dot corresponds to a clone.

(K) Boxplot of expression (TPM, y axis) of three calcium channels (*CACNA1B*, *CACNG4*, and *CACNA2D2*) differentially expressed between WT cells and full HTT KO clones.

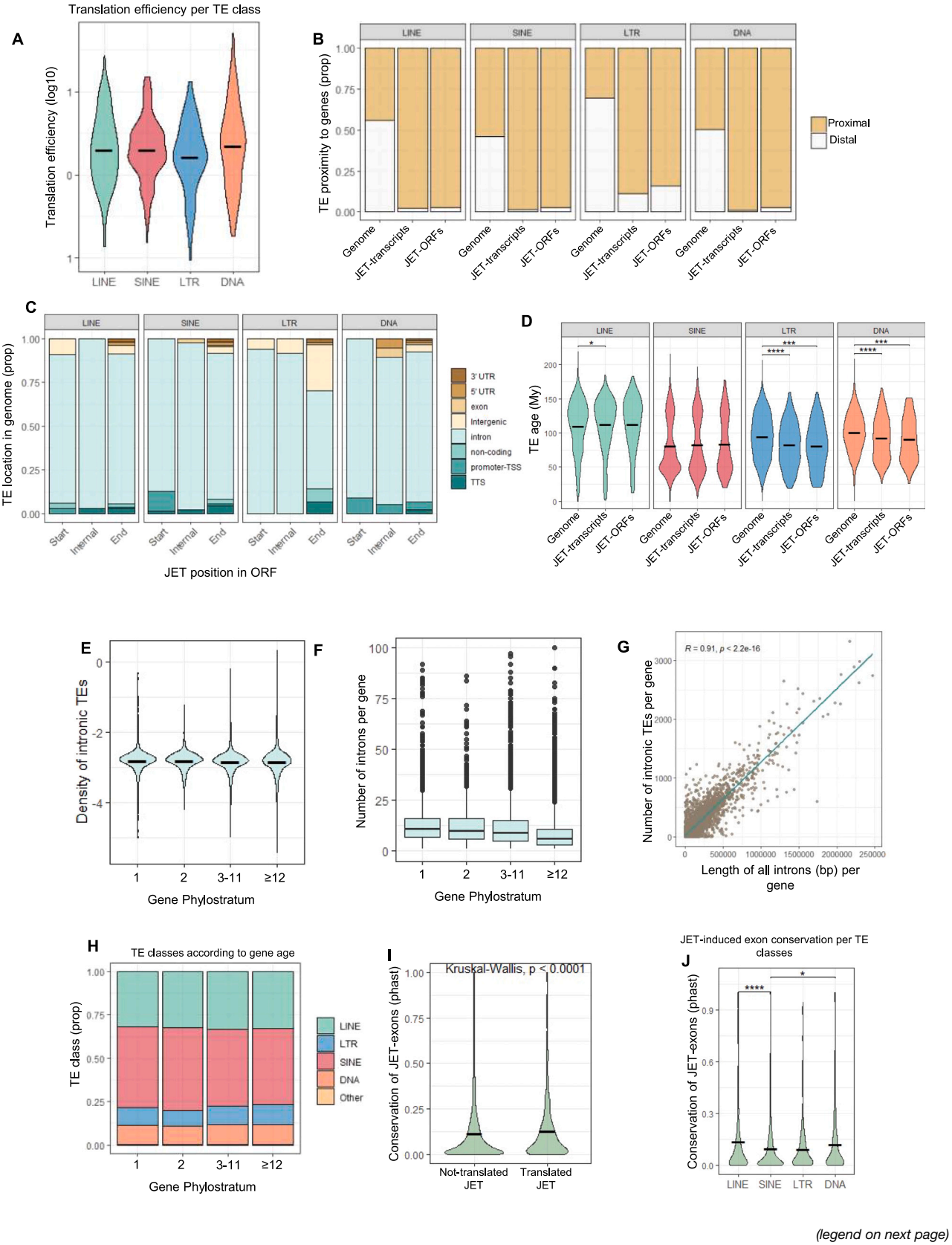
(L) In the left, real-time proliferation based on cell-impedance (cell index) of HTT-JET depleted and mock cells (x axis). In the right, jitter plots of the cell index (y axis) of HTT-JET and mock clones at 50 h (experiment 1) or 72 h (experiments 2 and 3). Colored dots indicate cell clones.

(M) Flow cytometry histograms of the ectopic expression levels of PTEN JET-ORF (green) and CAN-ORF (blue) in HeLa mNG1-10 cell line. HeLa mNG1-10 cells were used as negative control (gray).

(N) Boxplots of TPM expression of *H2BC5*, *H2BC11*, *H2BC12*, and *H2BC18* in cells overexpressing PTEN JET-ORF or CAN-ORF and HeLa mNG1-10 cells (negative control).

(O) Volcano plot of the differentially expressed genes between PTEN JET-ORF (red) and CAN-ORF (blue) expressing H1650 cells. Colored dots indicate significant p values.

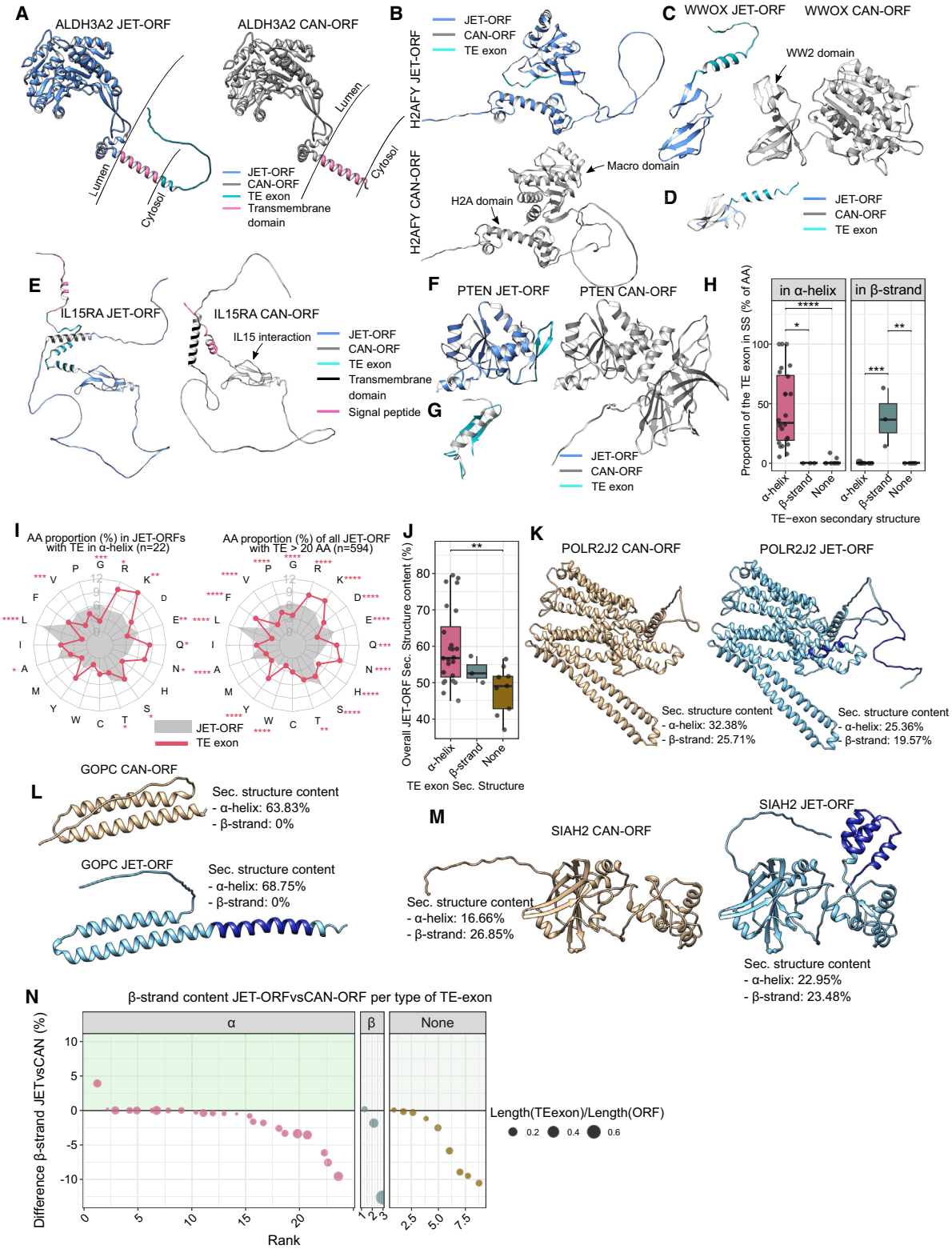
(P) Boxplots of the TPM expression of selected differentially expressed genes between SSO13- and/or SSO56- treated cells versus Scramble control cells. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$ (t test with Bonferroni adjustment).



(legend on next page)

Figure S6. Characterization of TE involved in JET-ORFs, related to Figure 6

(A) Violin plots representing the translation efficiency (log_{10}) of JETs depending on the TE class (x axis). The median is indicated with the black line.
(B) Bar plot indicating the proportion of TEs proximal or distal to genes in the genome, genes involved in JETs in transcriptome, and genes involved in JETs in translatome.
(C) Bar plot indicating the proportion of the TE genomic location depending on the JET position within the ORF (Start, Internal, or End).
(D) Violin plots of the age of TEs in genome, involved in transcribed JETs, and in translated JETs.
(E) Violin plots of the density of intronic TEs per gene according to gene age (phylostratum).
(F) Boxplots of the number of introns per gene according to gene age (phylostratum).
(G) Scatter plot correlating the number of intronic TEs per gene with the length (in bp) of all introns in the gene.
(H) Bar plot indicating the proportion of each TE class in introns of genes grouped according to their age (phylostratum).
(I) Violin plot indicating the conservation (phastCons) of translated versus non-translated JET-induced exons.
(J) Violin plot indicating the conservation (phastCons) of JET-induced exons according to the involved TE class.



(legend on next page)

Figure S7. Structure characterization of JET-ORFs, related to Figure 7

(A–C) Three-dimensional structure of JET-ORF (blue) and CAN-ORF (gray) of ALDH3A2 (A), H2AFY (B), and WWOX (C). The TE exon is highlighted in green. The transmembrane domains of ALDH3A2 are indicated in pink. pLDDT values for ALDH3A2 and H2AFY predictions are higher than 70, pLDDT value in WWOX is lower than 70.

(A) Structure superposition highlighting the changes of the WW2 domain of WWOX JET-ORF (blue) and CAN-ORF (gray), with the TE exon displayed in green.

(B) Three-dimensional structure of IL-15RA JET-ORF (blue) and CAN-ORF (gray), including the TE exon (green), the transmembrane domains (black), and the signal peptide (pink). The pLDDT value for IL-15RA JET-ORF is below 70.

(C) Three-dimensional structure of PTEN JET-ORF (blue) and CAN-ORF (gray), with the TE exon in green. pLDDT values for PTEN predictions are above 70.

(D) Zoomed view of the structure adopted by the TE exon in PTEN JET-ORF (green) and the corresponding sequence in PTEN CAN-ORF (gray).

(E) Boxplots representing the percentage of amino acids in the TE exon that participate in the adopted α helix (left) or β strand (right), depending on the structure adopted by the TE exon itself (i.e., α helix, β strand, or none; x axis).

(F) Radar plots representing the proportion (in %) of each amino acid in the TE exon (red) or the JET-ORF (gray). In the left, the 22 JET-ORFs with internal exonized TEs folded in α helix and longer than 20 amino acids have been used. In the right, the 594 JET-ORFs with an exonized TE longer than 20 amino acids have been selected.

(G) Boxplot indicating the overall percentage of secondary structures of the JET-ORF (y axis) depending on the structure adopted by the TE exon (x axis).

(K–M) Three-dimensional structure of the POLR2J2 (E), SIAH2 (F), and GOPC (G) JET-ORFs (cyan) and the corresponding CAN-ORFs (beige). The exonized TE sequence is highlighted in dark blue.

(N) Difference of the overall β strand content in the JET-ORF compared with the corresponding CAN-ORF (%, y axis) depending on the structure (α helix, β strand, or none) adopted by the exonized TE. Proteins are ranked from higher to lower differences in the x axis.