

PRPF39-PRTF1 enforces splicing of short and GC-poor introns across metazoans

Simon Bekker-Jensen

sbj@sund.ku.dk

University of Copenhagen <https://orcid.org/0000-0002-7308-4597>

Laura Ryder

University of Copenhagen

Article

Keywords:

Posted Date: July 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9550158/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

PRPF39-PRTF1 enforces splicing of short and GC-poor introns across metazoans

Laura Ryder^{1,*}, José Francisco Martínez^{1,*}, Melanie Blasius^{1,*}, Zhenzhen Wu¹, Anastasia M. Janssen¹, Catarina Pechincha¹, Miruthula Saravanan², Anouk M. Olthof¹, Alexander Jordan¹, Emil Andrup Olsen², Abbas Jafari², Thijn Brummelkamp^{3,4,5,6}, Peter Haahr¹, Hector Herranz², Simon Bekker-Jensen^{1,#}

¹*Center for Gene Expression, Department of Cellular and Molecular Medicine, University of Copenhagen, Blegdamsvej 3, DK-2200 Copenhagen, Denmark;* ²*Department of Cellular and Molecular Medicine, University of Copenhagen, Blegdamsvej 3, DK-2200 Copenhagen, Denmark;* ³*Division of Biochemistry, the Netherlands Cancer Institute, Plesmanlaan 121, 1066CX, Amsterdam, the Netherlands;* ⁴*Onco Institute, Division of Biochemistry, the Netherlands Cancer Institute, Plesmanlaan 121, 1066CX, Amsterdam, Netherlands.* ⁵*CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, 1090 Vienna, Austria.* ⁶*Cancer Genomics Center, Plesmanlaan 121, 1066CX, Amsterdam, Netherlands.*

*Equal contribution

[#]Lead contact. Correspondence and requests for materials should be addressed to S.B.-J.; email: sbj@sund.ku.dk

Abstract

Metazoans predominantly rely on exon-defined splicing due to the presence of long introns and high exon-intron GC differentials. Conversely, short introns embedded in overall GC-rich environments are recognized through intron-defined, intron-spanning spliceosome assembly. Exons followed by short introns within GC-poor environments thus pose significant problems for both exon- and intron-defined splicing. Here we show that a trimeric PRPF39/PRTF1 complex accumulates at unusual short and extremely GC-poor introns to enforce their splicing. Across metazoans, PRPF39/PRTF1 is critical for splicing of a few hundred introns with conserved nucleotide composition and length features. Loss of PRPF39/PRTF1 predominantly results in exon skipping in vertebrates (humans, mice) and intron retention in invertebrates (flies), reflecting their preference for exon-defined and intron-defined splicing, respectively. Failure of this mechanism in apoptosis-deficient *Drosophila* wing discs leads to wing duplication. Together, our data suggest an evolutionarily conserved role for PRPF39/PRTF1 in enforcing excision of seemingly “un-spliceable” introns.

Introduction

The central dogma dictates that a gene encoded by DNA is transcribed into mRNA that is subsequently translated into protein ¹. A unique feature of eukaryotic pre-mRNA transcripts is the presence of non-coding (intron) as well as coding and regulatory (exon) sequences. In the pre-mRNA resulting from transcription, introns must be excised by the spliceosome and exons joined to generate an open reading frame (ORF) that can be translated to protein in ribosomes ². In simple terms, the spliceosome consists of five core particles, each held together by a central non-coding snRNA (small nuclear RNA). In addition to snRNA, these particles consist of around 200 different proteins and are named U1, U2, U4, U5 and U6, respectively. Briefly, U1 and U2 recognize the intronic 5' splice site (SS) and the intron branch point on pre-mRNA, respectively ³. Subsequently, this complex recruits the tri-snRNP (small nuclear ribonuclear protein) complex U4-U5-U6, and in a coordinated reaction, first U1 and subsequently U4 is released from the structure, with U6 replacing U1 in 5' SS recognition. This catalytic spliceosome first allows formation of a lariat between the 5' SS (donor) and the intron branch point, and subsequent cleavage of the 3' SS (acceptor) and joining of the two exon ends ⁴.

Key to achieving correct splicing is the ability to recognize the basic splicing units of pre-mRNA. In budding yeast, less than 300 genes contain an intron, and only 9 genes contain more than one intron ⁵. These introns are all within a narrow length span, and in this organism, splicing is initiated by recognition of introns (intron definition) ⁶. In humans, introns are much more abundant with an average of 8 introns per gene. Furthermore, while human exons fall within a narrow length span, human introns are highly variable in length, ranging from 50 to 300,000 nucleotides (nt) with a median length around 1,500 nt ⁷. In this and most other metazoan organisms, exons are preferred as the splicing unit for initial recognition (exon definition) ^{8,9}. Exon-defined splicing likely evolved to

facilitate splicing of long introns and is dependent on high exon-intron GC differentials ^{10,11}, and strong flanking intronic splice sites.

Though the core splicing reaction is largely conserved from the simplest eukaryotes to humans, the contrasting gene structures of yeast and metazoans have shaped evolution of the functional architecture of the spliceosome. In yeast, where introns are uniformly short and recognized through intron definition, the U1-bound Prp39/Prp42 heterodimer physically bridges U1 and U2 across an intron to stabilize the intron-spanning pre-spliceosome complex (A complex) ^{5,12}. In addition, Nam8 reinforces this intron-defined assembly by contacting the dimer interface of Prp39/Prp42 through its C-terminal RRM, while binding intronic RNA through its two N-terminal RRMs ^{12,13}. As such, Nam8 is suggested to be critical for the splicing of introns with weak 5' SS ¹⁴. Prp39 and Prp42 are essential for survival, whereas Nam8 is non-essential but indispensable for meiosis ¹⁵. In contrast, the abundance of long introns and the reliance on exon definition in vertebrates have coincided with the loss of Prp42 and marked shortening of U1 snRNA ¹⁶, suggesting a reconfiguration of the early spliceosome that no longer depends on an intron-spanning U1-U2 bridge. Instead, human PRPF39 forms a U1-extrinsic homodimer ¹⁶ and has been suggested to recruit U1 to weak 5' SS to support alternative splicing ¹⁷. TIA and TIAL1 are thought to function as the human homologs of Nam8 and work together with PRPF39 in this capacity. Nevertheless, the role of PRPF39 in mRNA splicing remains relatively poorly described and it is unclear how the metazoan spliceosome can recognize and efficiently splice introns that are too short and GC-poor to be handled by either exon or intron definition alone.

Here we show that metazoans have evolved a mechanism to enforce the splicing of short and extremely GC-poor introns. Importantly, a ubiquitous and metazoan-specific trimeric complex

consisting of a dimer of PRPF39 and a monomer of mRNA-binding PRTF1 (TRNAU1AP) proves critical for correct splicing of such introns, and failure of this mechanism leads to pronounced skipping of the upstream exon and/or intron retention. Our results indicate that this U1-extrinsic protein complex evolved to enforce the excision of a class of introns that would otherwise not be amenable to mRNA splicing.

Results

PRPF39 and TRNAU1AP/PRTF1 are required for inclusion of exon 13 in the spliced *ZAK α* transcript

Revisiting historical HAP1-based gene trapping screens (Fig. S1a) that report on *ZAK α* kinase activation¹⁸, we noticed that the genes encoding PRPF39 and TRNAU1AP (henceforth named PRTF1 for PRPF39 RNA Tethering Factor 1) consistently scored as strong positive regulators (Fig. 1a). The DepMap database¹⁹ indicated that these two genes are highly co-dependent for cell growth across a broad panel of human cell lines, suggesting that they function in the same pathway (Fig. S1b). The *MAP3K20/ZAK* gene encodes two p38- and JNK-activating MAP3 kinase splice forms with distinct activation signals. The mechanosensitive kinase encoded by the shorter splice form, *ZAK β* ²⁰, is approximately 10 times more abundant in cells than the ribotoxic stress-activated kinase *ZAK α* ²¹ encoded by the longer splice form (Fig. 1b). Spliced *ZAK α* transcript is produced by skipping of the last *ZAK β* exon in the primary *ZAK* transcript, leading to inclusion of nine more exons. Validation of the genetic screen data highlighted that the *ZAK α* protein was absent from HAP1, U2OS and HeLa cells deleted or depleted for either PRPF39 or PRTF1 while *ZAK β* levels were not affected (Fig. 1c-e; Fig. S1c). PRPF39 is a known, yet poorly understood splicing factor^{16,17}, but we found high quality and publicly accessible genome-wide eCLIP (enhanced crosslinking

and immunoprecipitation) data for its binding to RNAs in RBP-ARK (rbp-ark.com). Notably, PRPF39 showed a single binding peak in the short intron (97 nt) between the *ZAK* α -specific exons 13 and 14 (Fig. 1f), raising the possibility that missplicing of this region was responsible for the loss of *ZAK* α protein. Indeed, RT-PCR (reverse transcription polymerase chain reaction) amplification with primers annealing to the spliced exon 11-12 junction (forward) and exon 14 (reverse) revealed abundant skipping of exon 13 in the absence of PRPF39 or PRTF1 in both HAP1 and U2OS cells (Fig. 1g,h). Because this exon is only 31 nt and not divisible by 3, we hypothesized that skipping of this exon would induce a frameshift and trigger degradation of the transcript by nonsense-mediate decay (NMD). Indeed, cycloheximide treatment, which inhibits NMD, stabilized the exon-skipped transcript, supporting the idea that aberrant splicing underlies the loss of *ZAK* α . Importantly, a region in the shared part of the α/β transcript with similar lengths of exons and introns, but not occupied by PRPF39, was not sensitive to knockout (KO) of these factors (Fig. 1f-h). PRPF39 and PRTF1 are thus required for a single specific splicing event in the *ZAK* α transcript characterized by the (not unique) presence of a short exon followed by a short intron.

PRTF1 mediates inclusion of several auto-regulatory poison exons in *PRPF39* mRNA

In all cellular backgrounds tested (HAP1, U2OS, HeLa), we noticed that PRTF1 KO or depletion by siRNA was accompanied by a dramatic increase in the levels of PRPF39 protein (Fig. 1c,d; Fig. 2a; Fig. S1c,d). PRPF39 was previously demonstrated to negatively regulate its own levels by mediating the inclusion of an NMD-causing alternative poison exon between exons 8 and 9¹⁶. eCLIP data highlighted a prominent double peak of PRPF39 binding to either side of this poison exon, but also strong binding to the intron separating exons 3 and 4 (Fig. 2b). When we subjected both of these regions to analysis of splicing outcomes by RT-PCR, we could first of all confirm the occasional inclusion of the NMD-causing poison exon between exons 8 and 9, and this event was

completely PRTF1- and PRPF39-dependent (Fig. 2c; Fig. S1e,f). Several NMD-targeted *PRPF39* splice products from the region between exons 3 and 4 could also be detected. Two of these were also dependent on PRTF1 (Fig. 2c) and the band sizes were consistent with inclusion of one and even two additional cryptic exons (Fig. S1g). We conclude from these analyses that PRTF1 acts with PRPF39 to buffer the levels of the canonical protein-coding *PRPF39* splice form (Fig. S1h,i). Importantly, the elevated levels of PRPF39 in the absence of PRTF1 did not appear to underlie *ZAKα* mis-splicing (Fig. S1d), suggesting that these proteins are non-functional in the absence of their partner.

PRTF1 engages PRPF39 in a trimeric complex

To determine whether PRPF39 and PRTF1 interact with each other and/or with additional spliceosome factors, we performed AlphaFold3 (AF3) modelling²² for binding of these two proteins to all proteins associated with the term “mRNA splicing” in Reactome (reactome.org - R-HSA-72172). To separate true from false positive interactions, we employed the recently described Structure Prediction and Omics Informed Classifier (SPOC)²³. This analysis returned excellent positive and reciprocal predictions for binding of PRPF39 to both PRPF39 itself and PRTF1, but not to any other splicing-associated proteins (Fig. 2d). Indeed, PRPF39 co-purified with PRTF1 and *vice versa*, both when stably over-expressed (Fig. S1j,k) and at the level of endogenous proteins (Fig. 2e). By contrast, and under the same stringent conditions, endogenous PRPF39 failed to co-purify with protein components of the U1, U2 and U5 snRNPs (Fig. 2e), the branchpoint binding protein SF1 or the splicing factors TIA1 and TIAL1 (Fig. S1l). AF3 predicted a trimeric complex consisting of two copies of PRPF39 and one copy of PRTF1. The resulting predicted aligned error (PAE) matrix (Fig. S2a) and inspection of the proposed structure (Fig. 2f) suggested that the interactions are largely based on the insertion of a short helix in the C-terminus of PRTF1 into a

deep binding pocket formed at the interface of the dimerizing HAT (half a-tetratricopeptide) repeats in PRPF39 (Fig. 2g). Pulldown experiments with gross deletion mutants of PRTF1 confirmed an absolute requirement of the C-terminal region (Δ C-term, last 62 amino acids – mutant Δ 226-287) for interaction (Fig. S2b-d), and a fragment consisting of the last 100 amino acids (186-287, TYR+C-term) efficiently interacted with PRPF39 (Fig. S2e). We also deleted half of the PRPF39-interacting α -helix in PRTF1 (Δ 262-270) and made single amino acid substitutions (E262K or L264R) in the same region with expected deleterious effects on the interaction. Contrary to deletion of the first RRM (Δ RRM1, Δ 3-86), these PRTF1 mutants were severely compromised for binding to PRPF39 (Fig. 2h). Some binding persisted, however, suggesting that additional contacts highlighted by AF3 modelling also contribute to complex formation. We then stably rescued U2OS / Δ PRTF1 cells with RRM1-deleted and helix-mutated versions of PRTF1. Western blotting analysis for the levels of PRPF39 and ZAK α proteins (Fig. S2f) and RT-PCR-based splicing analysis of the underlying exon skipping events (Fig. S2g) indicated that both RRM1- and helix-deficient PRTF1 were hypomorphs but not null mutants - likely due to a contribution to RNA-binding from the second RRM and additional weak contacts between PRTF1 and PRPF39 (Fig. 2f; Fig. S2a). The two proteins thus appear to engage in a trimeric complex outside the core spliceosomal particles with PRTF1 tethering PRPF39 to its target introns (Fig. 2i).

Parallel RNA sequencing experiments highlight a rare exon-skipping phenotype in the absence of PRPF39 and PRTF1

It seemed improbable that inclusion of exon 13 in ZAK α is the only event regulated by the PRPF39-PRTF1 complex. To search for other such events, we carried out parallel RNA sequencing experiments based on *PRPF39*- and *PRTF1*-depleted (siRNA) U2OS cells and PRPF39- and PRTF1-deleted (CRISPR KO) HAP1 cells, and all conditions were analyzed as biological triplicates

with two independent siRNAs for each factor (in the case of U2OS cells) (Fig. 2j). For bioinformatic analysis, we pooled all our data into one big dataset with just two conditions – control and PRPF39/PRTF1-deficient - and performed replicate Multivariate Analysis of Transcript Splicing (rMATS) to calculate Δ PSI (Delta Percent Spliced-in) values²⁴ for all categories of alternative splicing events. Increased exon skipping was the most dominant change (Fig. 2k). A closer inspection of significantly changed cassette exon events highlighted a striking but limited exon skipping phenotype with 321 events in 305 unique genes (Fig. 2l; Table S1). The most significant skipped exons included (next to the ones we demonstrated for *PRPF39* in Fig. 2c) exon 10 in *FANCL* and exon 32 in *RIF1* (Fig. 2l). Using RT-PCR, we readily validated PRPF39/PRTF1-dependent exon skipping for *RIF1*, *FANCL*, *HRNPA2B1*, *PHF14* and *CDC16*. Some of these events have been reported before¹⁷, but not with a mutual dependency on PRPF39 and PRTF1, and unlike *ZAK α* and *PRPF39*, these events did not appear to result in NMD (Fig. 3a; Fig. S3a – cycloheximide condition). The exon skipping phenotype could also be readily observed by inspection of sequencing tracks for *RIF1* exon 32 and *ZAK α* exon 13 (Fig. 3b; Fig. S3b). Given that we do not observe a pronounced intron retention phenotype in the absence of these factors in human cells (Fig. 3a,b), the PRPF39-PRTF1 complex appears to mediate exon-defined rather than intron-defined splicing in human cells. To understand the mechanistic basis for PRPF39/PRTF1-dependent inclusion of these exons, we employed a minigene setup (Fig. 3c). In human U2OS cells, a 3-exon minigene from mouse *Rif1* (spanning exons 31 to 33) was spliced into both a 3- and a 2-exon product, with the former species being dependent on PRTF1 (Fig. 3c). Mutation of splice sites (3' SS of upstream intron and 5' SS of downstream intron, respectively) phenocopied these effects on splicing outcome. PRPF39-PRTF1 thus facilitates spliceosome assembly across poorly defined exons, resulting in their skipping when the complex is absent.

Short downstream intron length is a defining feature of PRPF39/PRTF1-dependent exon inclusion

To identify features associated with PRPF39/PRTF1-dependent splicing, we compared exons that were significantly skipped upon loss of either factor (foreground), to exons showing no detectable change (background). Comparison of these two groups did not reveal any difference in the splice site strength of the flanking introns (Fig. S3c,d). Instead, we found that a short downstream intron (~100 nt) was the most distinguishing characteristic of PRPF39/PRTF1-dependent exon inclusion (Fig. 3d), consistent with our validated examples (Fig. 1f-h; Fig. 2b,c, Fig. 3a,b). This is much shorter than the median length of human introns. However, it is in itself not unusual or predicted to be problematic for the splicing reaction²⁵, suggesting that additional hidden features determine the requirement for PRPF39 and PRTF1.

Integration of eCLIP and RNA sequencing data highlights a PRTF1 consensus binding site in downstream introns

Given that short downstream introns emerged as a defining feature of PRPF39/PRTF1-dependent exons, we asked whether the complex itself preferentially associates with such intronic environments. To this end, we gathered all eCLIP data sets from HepG2 cells deposited in RBP-ARK and plotted the length distribution of bound introns for all RBPs (Fig. 3e). From this analysis, PRPF39 stood out as an RNA binding protein (RBP) with a preference for binding to short introns, again with a median length of around 100 nt. Comparing the foreground and background datasets, PRPF39 eCLIP signal (but not signal for any other RBP) was strongly enriched in the introns downstream of the dependent exons (Fig. 3f; Fig. S3e), again similar to our validated examples (*ZAKα*, *PRPF39* and *RIF1*, Fig. 1f; Fig 2b; Fig. 3b). Motif enrichment analysis revealed the presence of U-rich motifs in the foreground dataset sequences (Fig. 3g), which could constitute

binding platforms for the RRM of PRTF1. U-rich 5-mer binding motifs were also enriched in previously published PRTF1 RNA Bind-N-Seq experiments^{26,27}. We also computationally evaluated the contribution of PRTF1 binding motifs to splicing outcome using the raw rMATS output and the CisBP (Catalog of inferred sequence Binding Preferences)-RNA database²⁸. This analysis highlighted that PRTF1 binding motifs have a positive impact on exon inclusion in our dataset when located between 40 and 100 nucleotides downstream from the 5' SS (Fig. 3h). To confirm that such motifs are responsible for the influence of PRTF1 on individual splicing events, we mutated the most likely binding sequence in a 3-exon mouse *Prpf39* mini-gene (spanning the poison exon and flanking exons 8 and 9) (Fig. 3i). As expected, the WT minigene was spliced into both 3- and 2-exon products in WT cells, but only to the 2-exon product in U2OS / ΔPRTF1 cells. However, the mutated minigene (U->A) was also exclusively spliced into a 2-exon product in both U2OS WT and ΔPRTF1 cells (Fig. 3j). These results highlight that promotion of mRNA splicing by the PRPF39-PRTF1 complex is governed by recognition of U-rich sequence motifs in the intron downstream of the sensitive exon (Fig. 3k).

A feature-based computational model highlights short and GC-poor introns as clients of PRPF39/PRTF1-dependent splicing

In order to determine the importance of individual sequence-based features and feature interactions underlying PRPF39/PRTF1-dependent splicing events, we extracted features from the detected splicing events and trained an Explainable Boosting Machine²⁹ to predict foreground events given those features. We used this type of model for its easily interpretable decision-making (Fig. 4a,b). The individual features with the highest absolute contribution to classification of an exon as PRPF39/PRTF1-dependent for inclusion were short length and low GC content of the downstream intron (Fig. 4c). Other features with positive (but weaker) contribution were short length and low

GC content of the upstream intron, while the contribution from splice site sequence consensus was negligible (Fig. 4c). Among the four feature interactions that the model uncovered was the interaction between short length *and* low GC content of the downstream intron (Fig. 4d).

To match these computational outputs with our data, we proceeded to plot length against GC content for all downstream introns and color them according to presence of the upstream exon in our experimental foreground (PRPF39/PRTF1-dependent for inclusion) and background datasets (Fig. 2l; Fig. 4e). In mammals, these features generally correlate in the way that short introns tend to have a high GC content ³⁰. By contrast, introns following PRPF39/PRTF1-dependent exons were strikingly enriched in a cluster of rare short introns that did not follow this general trend (Fig. 4e). A similar albeit less obvious trend could be observed for the variable exons themselves (Fig. 4f). Metagene plots of areas encompassing both the variable exon and the upstream and downstream introns confirmed a general low GC content across the whole region in our foreground dataset compared to background (Fig. 4g; Fig. S4a-d). A general nucleotide bias towards low G and C and high A and U was also evident in the upstream and downstream exons (Fig. 4h). This bias was strongest in the downstream introns which were also enriched for U, consistent with the presence of U-rich PRTF1-binding motifs. By combining experimental observations and computational modelling, our data highlight a unique requirement for PRPF39 and PRTF1 for successful excision of short and extremely GC-poor introns in the context of an overall GC-poor environment (Fig. 4i).

Artificial intelligence has revolutionized most areas of biology, and recently such tools have been developed to accurately predict splice site strengths and splice site usage. To assess to which extent these models can account for the paradoxical inclusion of PRPF39-PRTF1-dependent exons in mature transcripts despite their unusual sequence context, we computed predicted cassette donor

scores (0-1) for all exons in our experimentally-derived foreground and background dataset with both spliceAI ³¹ and the more recently developed AlphaGenome ³². We plotted these values against observed exon inclusion levels and calculated correlation coefficients (Pearson r) for both datasets and with both prediction models. The performances of both spliceAI and AlphaGenome were clearly superior in correlating predicted splice site strength with observed usage for the background (PRPF39-PRTF1-independent) compared to the foreground dataset (PRPF39-PRTF1-dependent) (Fig. S4e,f). This suggests that these models do not fully appreciate the splicing-supportive effects afforded by PRPF39-PRTF1 in the context of short and extremely GC-poor downstream introns.

Analysis of PRPF39/PRTF1-dependent splicing in mice highlights the relevance of intronic features

To study the evolutionary conservation of PRPF39/PRTF1 function in splicing, we used our computational model, trained on human experimental data, to classify mouse exons according to dependency on PRPF39/PRTF1 for inclusion (Fig. 5a). Among the top predictions were several of the exons that we successfully validated as PRPF39/PRTF1-dependent for inclusion in human cells, such as *MmPrpf39* poison exon and *MmRif1* exon 32 (Fig. 5b, Table S2). Knockdown of *MmPRTF1* in mouse NIH3T3 cells also led to marked upregulation of *MmPRPF39* protein levels, highlighting that the functionality of the complex is conserved in this species (Fig. 5c). The intron downstream of *MmRif1* exon 32 displayed similar short length and low GC content to its human counterpart (Fig. 5d), and also the inclusion of this mouse exon was dependent on PRTF1 and PRPF39, both when assayed at the endogenous transcript level in NIH3T3 cells (Fig. 5e) and within a mini-gene (Fig. 5f). Surprisingly, *ZAKα* exon 13, the exon that initially set us on the course of studying the PRPF39-PRTF1 complex, was not predicted as a PRPF39/PRTF1-dependent exon in mice (Fig. 5b). Interrogation of the individual features from the human vs. mouse predictions

highlighted that this divergence arose from differences in GC content of the downstream intron between the two species (Fig. 5g, arrow). Consistently, sequence alignment highlighted a mouse-specific insertion of a short GC-rich stretch of nucleotides in the intron downstream of *MmZakα* exon 13 (Fig. 5h), explaining why this exon is not classified as PRPF39/PRTF1-dependent in mouse. Indeed, knockdown of *MmPRPF39* and *MmPRTF1* in mouse NIH3T3 cells did not affect ZAKα levels, anisomycin-induced p38 activation or exon 13 inclusion (as it did in human cells) (Fig. 5c,e). Another intron that differs between humans and mice is the one downstream of *FANCL* exon 10. The mouse version displays a somewhat higher GC content than its human counterpart, however, sequence alignment highlighted a likely PRTF1-binding sequence in humans only (Fig. 5i) (compare with logo in Fig. 3g). Consequently, *MmFancl* exon 10 was included in the mature transcript independently of *MmPRPF39* and *MmPRTF1* in mouse NIH3T3 cells. Our results indicate that PRPF39 and PRTF1 function is highly conserved between human and mouse and that downstream intron length, GC content and presence of a U-rich motif defines the dependency of this complex for exon inclusion.

Ubiquitous depletion of *DmPrpf39* and *DmPrtf1* in the *Drosophila melanogaster* larval state results in missplicing of short introns with low GC content

Flies and other invertebrate species generally exhibit an exon-intron structure with much shorter introns and longer exons compared to vertebrates (incl. mammals) and largely rely on intron-defined splicing³³. We used *Drosophila melanogaster* as a divergent model organism to test whether the PRPF39-PRTF1 complex has conserved splicing functions across metazoans in spite of this fundamental difference. Comparative genome analysis identified *Dme1/CG1646* as the PRPF39 orthologue and *Sepe43/CG15440* as the most likely *D. melanogaster* orthologue of PRTF1. AF3 modelling supported this, revealing a C-terminal *DmPrtf1* helix inserted into a *DmPrpf39* dimer

pocket, analogous to the human complex (Fig. 6a). We used the Gal4-UAS-Gal80^{ts} system³⁴ (Fig. S5a,b) to downregulate *DmPrpf39* and *DmPrtf1*. We crossed Actin-Gal4 with UAS-*Prtf1*-RNAi and UAS-*Prpf39*-RNAi lines (two independent sequences per gene) and subjected pools of larvae to RNA sequencing (Fig. 6b). Examination of *Dm-Prpf39* and *Dm-Prtf1* gene tracks confirmed knockdown of these mRNAs (Fig. S5c-f).

Intriguingly, in contrast to humans (Fig. 2k), intron retention was the main contributor to missplicing in *DmPrpf39*- and *DmPrtf1*-deficient fly larvae followed at some distance by exon skipping (Fig. 6c). This species-dependent difference in outcomes likely reflects the consequences of missplicing of a subset of introns in a predominantly exon-defined (humans) vs. intron-defined (flies) splicing context³⁵. We next aimed to determine whether the same sequence features drove the dependency for PRPF39/PRTF1 in both human and fly. To facilitate direct comparison with the human-derived splicing analysis, we focused on the exon skipping events in *DmPrpf39*- and *DmPrtf1*-deficient larvae. Comparative bioinformatic analysis of PRPF39/PRTF1-dependent exons (Table S3) recapitulated key features related to length and GC content (Fig. 6d). These similarities in affected introns between highly divergent species highlights a conserved commitment of the PRPF39-PRTF1 complex to an unusual class of introns.

Knockdown of *DmPrpf39* and *DmPrtf1* in the *D. melanogaster* wing imaginal disc leads to wing duplication

We hypothesized that if PRPF39 and PRTF1 also function as a complex to splice a conserved class of introns in flies, then loss of either factor should lead to similar developmental consequences. To test this, we first crossed UAS-*Prtf1*-RNAi and UAS-*Prpf39*-RNAi lines with the ubiquitous Actin-Gal4 driver. This did not yield any viable progeny, showing that both genes are essential in *D.*

melanogaster, similar to in mice. We therefore next examined the *D. melanogaster* wing imaginal disc, which is the precursor of the adult wing and thorax ³⁶ (Fig. S5a). Immunostaining for Death caspase-1 (Dcp1³⁷) showed increased apoptosis upon *DmPrpf39* or *DmPrtf1* knockdown (Fig. 6e,f). These effects persisted when depletion was restricted to the dorsal wing compartment (Ap-Gal4 driver) (Fig. S5a,b). To circumvent cell death, we blocked apoptosis by co-expressing the baculovirus caspase inhibitor p35 ³⁸ under Ap-Gal4 control. Discs co-expressing p35 with *DmPrpf39*- or *DmPrtf1*-RNAi showed abnormal morphology, suggesting developmental defects. Staining with Nubbin (Nub), a marker of the wing pouch ³⁹ (Fig. S5a), revealed that control discs displayed a single Nub-positive region, whereas apoptosis-blocked discs deficient for *DmPRPF39* or *DmPRTF1* showed two Nub-positive regions, indicative of wing disc duplication (Fig. 6g). This phenotype occurred in 87 % of *DmPrpf39*-depleted (16 discs) and 52 % of *DmPrtf1*-depleted (21 discs) samples. Another wing marker, the morphogen Wingless (Wg) ⁴⁰, also appeared ectopically at duplication sites in *DmPrpf39*- or *DmPrtf1*-deficient discs (Fig. S5g). Only few live adults emerged from these pupae (and only from Ap-Gal4 / p35 / UAS-*Prtf1*-RNAi crosses), but those that did displayed additional underdeveloped ectopic wing structures (Fig. 6h). To confirm that full wing duplication was taking place in *DmPrtf1*-deficient flies, we dissected mature pupae and exposed the emerging flies to an artificial wing inflation treatment, and indeed uncovered flies with three or four wing structures (Fig. 6i). Finally, to visualize shape and localization of ectopic wing structures in intact pupae, we performed μ -CT scanning. Also here, we observed duplicated but hypomorphic additional wing structures that were placed in an anterior-dorsal orientation relative to the main wing structure (Fig. 6j).

The complete overlap of rare phenotypes in *DmPrpf39*-depleted and *DmPrtf1*-depleted flies (Fig. S5h) is remarkable in that it confirms that these two factors are highly co-dependent also in this

divergent organism. Combined with our bioinformatic analysis of fly-specific splicing defects (Fig. 6c,d) it also highlights a striking functional conservation of the PRPF39-PRTF1 complex across metazoan species in supporting excision of an unusual class of introns that are problematic for both exon- and intron-defined splicing.

Discussion

Our data highlight that PRPF39 and PRTF1 work together in a complex that binds to U-rich sequence motifs in short and extremely GC-poor introns. These are embedded in overall GC-poor sequence environments and thus pose a significant challenge for both exon-defined and intron-defined splicing mechanisms. From this location PRPF39/PRTF1 can uniquely enforce the splicing of such atypical introns (Fig. 6k). In organisms that rely mainly on exon-defined splicing (humans, mice), PRPF39/PRTF1 deficiency largely results in skipping of the upstream exon, while intron retention is the dominant outcome in organisms (flies) that mainly rely on intron-defined splicing.

The complexity of mRNA splicing in higher eukaryotes is immense. While many key determinants underlying recognition and inclusion of individual exons into mature mRNAs have been appreciated, splicing outcomes are still notoriously difficult to predict based on features such as sequence and genomic and cellular context. Despite recent advances^{32 41}, the predictive power of holistic and artificial intelligence-based splicing models remain limited. In the present case, even the recently developed and powerful AlphaGenome model fails to appreciate the feasibility of PRPF39-PRTF1-dependent compared to -independent splicing events (Fig. S4f). Fundamental principles regarding the influences of exon/intron length, GC content and predicted SS strength hold up when evaluated at a global scale but often fail to predict and explain individual splicing

outcomes⁴². This suggests that cells have many, still poorly understood mechanisms in place to overrule sequence-based limitations to splicing. The important, supportive role of the PRPF39-PRTF1 complex that we describe here represents a unique example of such mechanisms. Mechanistically, PRPF39-PRTF1 likely facilitates efficient recruitment of U1 to the 5' SS of the bound intron¹⁷, either through direct interaction or by modulating the structure of the short and GC-poor intron. This will then be coupled with co-recruitment of U2 to either the upstream (exon definition, predominant in human and mouse) or downstream (intron definition, predominant in flies) acceptor sites. In the case of exon-defined splicing, this will favor formation of a pre-catalytic splicing complex that connects the donor of the upstream intron with the acceptor of the downstream intron, resulting in exon skipping (Fig. 6k).

Rather than the suggested TIA1 and/or TIAL1, the main metazoan partner of PRPF39 is PRTF1. These two proteins form a stable complex, display remarkable co-dependency across human cell lines, and give rise to near-identical splicing defects and phenotypes when depleted in humans, mice and flies. The dramatic phenocopy of wing duplication in apoptosis-deficient *D. melanogaster* imaginal wing discs underscores that these two proteins are tightly functionally connected across metazoan species. Although not further investigated, the co-dependency of PRPF39 and PRTF1 was noticed before in human cells, with both factors being essential specifically in mismatch repair-deficient cancer cell lines⁴³. PRTF1 has been previously described under the name TRNAU1AP (transfer RNA selenocysteine 1 associated protein 1), arising from its initial mis-annotation as a protein involved in selenocysteine biosynthesis⁴⁴. However, while germline KO of *PRTF1/TRNAU1AP* is lethal in mice, conditional mutation of the gene in liver was not associated with defects in selenocysteine incorporation into proteins⁴⁵. PRTF1 was recently proposed to be a mRNA splicing factor because it scored in a screen for decreased exon inclusion⁴⁶. In that work,

inclusion of the *PRPF39* poison exon was also identified as a prominent PRTF1-regulated event. Nevertheless, these two proteins have not been functionally connected in previous work and their precise roles in mRNA splicing have not been described. TIA1 and TIAL1 display some overlapping features with PRTF1 in splicing regulation⁴⁷. All three accrue at GC-poor introns through binding of U-rich motifs, but while PRTF1 action is uniquely limited to very short introns, TIA1/TIAL1 is not restricted by intron length and regulates alternative splicing through a poorly defined and PRPF39-independent mechanism. In both cases, the RBPs bind to U-rich splicing enhancer sequences in the downstream intron to support inclusion of the upstream exon.

In conclusion, we have uncovered a new mRNA-binding module that is critical for splicing of rare short and extremely GC-poor introns that occur in the context of overall GC-poor sequence environments. The critical supportive role played by this conserved complex expands the repertoire of sequences that are amenable to splicing across metazoan species.

Acknowledgements

We thank Dr. Jesper Q. Svejstrup (University of Copenhagen, Denmark) for critical reading of our manuscript. Work in the Bekker-Jensen lab was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement 863911 - PHYRIST), Independent Research Fund Denmark (grant no. 3101-00344B) and the LEO Foundation (grant no. LF-OC-23-001458). Center for Gene Expression (CGEN) is a Center of Excellence funded by The National Danish Research Foundation (grant no. DNRF166).

Declaration of interests

The authors have no positions, patents or financial interests to declare.

Figure legends

Figure 1.

PRPF39-PRTF1 is required for inclusion of a single exon in *ZAKα*

a. Gene-trap-based genetic screen in haploid human cells for regulators of p38 kinase activation in response to anisomycin (ani). Data re-analysed from ¹⁸ with known ribotoxic stress response pathway hits (red) and *PRPF39-PRTF1* (green) highlighted. **b.** Schematic of protein domains in *ZAKα* and *ZAKβ*. KD, Kinase Domain; LZ, Leucine Zipper; SAM, Sterile Alpha Motif; YLD, YEATS-like Domain; S, Sensor Domain; CTD, C-Terminal Domain; SFBD, Stress Fiber-Binding Domain. **c.** HAP1 cells with the indicated genotypes were treated with ani (1 μM, 1 h). Lysates were analyzed by immunoblotting with the indicated antibodies. **d.** U2OS WT and ΔPRTF1 cells were treated and analyzed as in (c). **e.** U2OS cells were transfected with siRNA targeting PRPF39 for 96 h. Lysates were analyzed as in (c). **f.** Exon-intron structure and eCLIP read tracks for PRPF39 (orange) and SF3A1 (blue) at the *MAP3K20/ZAK* gene locus. Blow-ups highlight the area around PRPF39-PRTF1-dependent exon 13 (right) and -independent exon 6 (left). **g.** RNA extracted from HAP1 cells with the indicated genotypes was reverse-transcribed, and the resulting cDNA was analyzed by PCR with primers annealing to exons upstream and downstream of *ZAKα* exons 13 and 6, respectively. Cells were pre-treated with cycloheximide (100 μg/ml, 2h) when indicated to assay for nonsense mediated decay of exon-skipped splice forms. Reactions were separated and analyzed by agarose gel electrophoresis. **h.** As in (g), except that WT U2OS and ΔPRTF1 cells were used.

Figure 2.

PRPF39 and PRTF1 form a trimeric complex that controls *PRPF39* splicing

a. U2OS cells were transfected with PRTF1-targeting siRNAs (96 h). Lysates were analyzed by immunoblotting with the indicated antibodies. **b.** Exon-intron structure and eCLIP read tracks for PRPF39 at the *PRPF39* gene locus. Blow-ups highlight the areas spanning the known *PRPF39* position exon (right) and another region between annotated exons 3 and 4, both with a pronounced double eCLIP peak for PRPF39. **c.** RNA extracted from U2OS cells with the indicated genotypes was reverse-transcribed, and the resulting cDNA was analyzed by PCR with primers annealing to exons upstream and downstream of PRPF39 eCLIP peaks from (b), respectively. Cells were pre-treated with cycloheximide (100 µg/ml, 2h) when indicated to assay for nonsense mediated decay of exon-skipped splice forms. Reactions were separated and analyzed by agarose gel electrophoresis. **d.** Two-dimensional SPOC-scored AlphaFold3 (AF3) prediction scores (ranging from 0 to 1) for PRPF39 (x-axis) and PRTF1 (y-axis) binding to each of the proteins associated with the term “RNA splicing” in Reactome (reactome.org - R-HSA-72172). **e.** U2OS cell lysates were subjected to immunoprecipitation (IP) with PRPF39 antibodies or IgG control. IP’ed material and whole cell extracts (WCE) were analyzed as in (a). **f.** AF3-predicted structure of a complex consisting of two molecules of PRPF39 and one molecule of PRTF1. Blow-ups of the main interaction pocket (green box) indicate the predicted local distance difference test (pLDDT) score. **g.** Schematic of PRPF39 and PRTF1 protein domains with the interacting helix of PRTF1 highlighted in blue. HAT, half-a-tetratricopeptide domains; RRM, RNA Recognition Motif; TYR, Tyrosine. **h.** HEK293 cells were transfected with the indicated full-length and mutant constructs of strep-HA-PRTF1. Lysates were subjected to strep-purification and pulldown material and WCE were analyzed as in (a). **i.** Model of a PRPF39-PRTF1 complex binding to intronic RNA via the RRM of PRTF1. **j.** Overview of RNA

sequencing experiments performed on human cells. Both HAP1 cells with stable gene deletion and rescue as well as siRNA-treated U2OS cells were used. Two different siRNAs per target were used and all conditions were analyzed as biological triplicates. **k.** Donut-chart of categories of alternative splicing outcomes from a two-way comparison of control conditions and all knock-out/knockdown conditions from (j). All control conditions and all knock-out / siRNA conditions were pooled into two groups for statistical analysis. Number of significant events ($|\Delta\text{PSI}| > 5\%$ and $\text{FDR} < 0.01$) per category is indicated. **l.** A foreground dataset of PRPF39/PRTF1-dependent exon inclusion was defined based on $\Delta\text{PSI} > 5\%$ and $\text{FDR} < 0.01$, and a background dataset was defined based on $|\Delta\text{PSI}| > 5\%$ and $\text{FDR} > 0.95$. SPOC, Structure Prediction and Omics Informed Classifier; AF3, AlphaFold 3.; ΔPSI , Delta Percent Spliced-in.

Figure 3.

PRPF39-PRTF1-dependent exon inclusion is dependent on a U-rich motif in the short downstream intron

a. RNA extracted from HAP1 cells with the indicated genotypes was reverse-transcribed, and the resulting cDNA was analyzed by PCR with primers annealing to the indicated exons. Cells were pre-treated with cycloheximide (100 $\mu\text{g}/\text{ml}$, 2h) when indicated to assay for nonsense mediated decay of exon-skipped splice forms. Reactions were separated and analyzed by agarose gel electrophoresis. **b.** PRPF39 eCLIP and RNA sequencing read tracks from all sequencing conditions across a region encompassing exons 31-33 of the *RIF1* transcript. **c.** Splicing outcomes of 3-exon mouse minigenes containing *Rif1* exons 31-33 transfected into WT or ΔPRTF1 human U2OS cells. Constructs were either WT or mutated in the 3' or 5' SS flanking exon 32. **d.** Log10-transformed length distributions for cassette exons and their surrounding introns and exons in foreground (red) and background (black) datasets (from Fig. 2l). Overrepresentation of short introns directly

downstream of skipped exons is highlighted by an arrow. **e.** Intron length distributions for introns with eCLIP signal enrichment of all RNA binding proteins (RBPs) with datasets available from HepG2 cells in RPB-ARK. eCLIP datasets for two different PRPF39 antibodies are highlighted in red, both indicating preferential binding to short introns. **f.** eCLIP peaks from (e) were mapped to introns and exons surrounding cassette exons from foreground and background datasets (defined in Fig. 2l). RBPs enriched in the foreground by more than 2-fold within distinct elements are highlighted (FDR < 0.01). **g.** Overrepresented RNA sequence motif in introns downstream of the variable exons in foreground relative to background dataset. Sequence windows from -25 nt to +200 nt relative to the 5' splice site (spanning the downstream intron) were analyzed. **h.** Contribution of PRTF1 binding motif from CisBP²⁶ to exon inclusion in foreground dataset according to KATMAP **i.** Partial alignment of the intron downstream of *PRPF39* poison exon (Fig. 2b) from human and mouse genomes. The putative PRTF1-binding motif is highlighted and was mutated in a 3-exon minigene encompassing canonical mouse *Prpf39* exons 8 and 9 ("MUT"). **j.** WT and "MUT" mouse *Prpf39* minigenes were transfected into WT and ΔPRTF1 U2OS cells. 3- and 2-exon splicing products were resolved and analyzed by agarose gel electrophoresis. **k.** Model of PRPF39-PRTF1 binding to short introns with U-rich motifs to mediate inclusion of the upstream exon.

Figure 4.

A feature-based computational splicing model highlights low GC content and short downstream intron length as determinants of PRPF39/PRTF1-dependent exon inclusion

a. Schematic of input features for computational model for prediction of PRPF39/PRTF1-dependent exons. In addition to lengths and GC content of surrounding exons and introns, relative splice site strength based on MaxEntScan score differentials and presence of enhancing (ERE) and inhibiting regulatory elements (IRE) surrounding the cassette exon were used as inputs. The output of the

model is classification of an exon as either PRPF39/PRTF1-dependent or -independent. **b.** Precision-Recall curve for a model trained on 80% of exons in experimental foreground and background (Fig. 21) and tested on the remaining 20%. AUC (Area under the curve) = 0.26 vs. 0.04 for a random classifier. **c.** Contribution towards classification for all 16 single features. Red indicates a positive contribution and blue a negative contribution. **d.** The four most important feature interactions selected by the model, all including aspects of intron length and exon/intron GC content. **e.** Scatter plot (graph) and histograms (axes) of GC content vs length for downstream introns from experimental foreground and background exon-inclusion datasets (Fig. 21). Correlation strength is indicated by the correlation coefficients (Pearson r). **f.** As in (e), except that values for cassette exons were plotted. **g.** Metagene plot highlighting overall low GC content across cassette exons and adjacent introns in foreground compared to background datasets. Dashed lines indicate cassette exon acceptor and donor position, respectively. **h.** Single nucleotide enrichment/depletion in foreground vs. background datasets for cassette exons and adjacent introns and exons. Error bars represent the standard error of the mean (SEM). **i.** Model of PRPF39/PRTF1-mediated inclusion of exons followed by short and extremely GC-poor introns. These are embedded within an overall GC-poor sequence environment and contain PRTF1-accruing U-rich motifs.

Figure 5.

Conservation of PRPF39/PRTF1-dependent exon inclusion events between human and mouse

a. Strategy for computational prediction of PRPF39/PRTF1-dependent exon inclusion events in mouse by using a classifier trained on human data (Fig. 4a-d). Only mouse exons (roughly 6,500) with orthologous exons in humans according to the Vast database (vastdb.crg.eu) were analyzed.

b. Scatter plot of PRPF39-PRTF1-dependence prediction scores (0-1) for orthologous mouse and human exons according to the human-trained classifier. Individual exons are color-coded according

to their presence in experimental human foreground (red) and background datasets (grey) (Fig. 2l).

c. Mouse NIH3T3 cells were transfected with siRNAs targeting *Prpf39* and *Prtfl* as indicated (72 h). Cells were treated with anisomycin (1 μ M, 1 h) and lysates were analyzed by immunoblotting with the indicated antibodies. **d.** Comparison of length and GC content of the intron downstream of *RIF1* exon 32 from humans and mice. **e.** RNA extracted from mouse NIH3T3 and human U2OS cells transfected with the indicated siRNAs were reverse-transcribed, and the resulting cDNA was analyzed by PCR with species-specific primers annealing to the indicated exons. Cells were pre-treated with cycloheximide (100 μ g/ml, 2 h) when indicated to assay for nonsense mediated decay of exon-skipped splice forms. Reactions were separated and analyzed by agarose gel electrophoresis. **f.** Splicing outcomes of 3-exon mouse and human minigenes containing *RIF1* exons 31-33 transfected into WT and Δ PRTF1 U2OS cells. For the human minigene, primer binding sites were mutated to the mouse sequence for comparative PCR analysis. Reactions were analyzed as in (e). **g.** Feature contributions (16 single features and 4 feature interactions) towards overall classification as PRPF39/PRTF1-dependent for *ZAK* exon 13 in humans and mice. The arrow highlights negative contribution from downstream intron GC-content in mouse. **h.** Partial alignment of the intron downstream of *ZAK* exon 13 in humans and mice. A short GC-rich insertion in the mouse genome is indicated. **i.** As in (h) but for the intron downstream of *FANCL* exon 10. Notice the increased GC content in mouse and appearance of a putative PRTF1-binding motif in human.

Figure 6.

Knockdown of *DmPRPF39* and *DmPRTF1* leads to splicing defects and wing disc duplication in fruit flies

a. AF3-predicted structure of a complex consisting of two molecules of the PRPF39 fly homolog (*Dmel/CG1646*) and one molecule of the PRTF1 fly homolog (*Sepc43/CG15440*). Blow-up of the

main interaction pocket (green box) indicates strong similarity to the interaction mode of the human PRPF39-PRTF1 complex (Fig. 2f). **b.** Overview of RNA sequencing experiments in fruit fly larvae ubiquitously expressing control, *DmPrtf1* or *DmPrpf39* RNAi (Actin-Gal4 driver). All conditions were analyzed as biological triplicates in two independent experiments. **c.** Donut-chart of categories of alternative splicing outcomes from a two-way comparison of control conditions and all knockdown conditions from (b). All control conditions and all RNAi conditions were pooled into two groups for statistical analysis. Number of significant events ($|\Delta\text{PSI}| > 5\%$ and $\text{FDR} < 0.01$) per category is indicated with intron retention being most dominant in this organism. **d.** Scatter plot (graph) and histograms (axes) of GC content vs length of downstream introns from fly-derived experimental foreground (Table S3) and background exon inclusion datasets. **e.** Fruit fly larvae from (b) were immunostained for apoptotic cells with antibodies against Dcp1. Images show representative signals from wing imaginal discs. **f.** Quantification of Dcp1-positive areas in discs from (e). Numbers were calculated into fold changes over control RNAi ($n = 9\text{-}13$ discs). Data are plotted as mean and all error bars represent the standard error of the mean (SEM). Statistical significance was defined as **, $p \leq 0.01$; ****, $p \leq 0.0001$ in one-way ANOVA followed by Dunnett post-hoc test for multiple comparisons. **g.** Fly larvae expressing apoptosis-blocking p35 protein and *DmPrtf1* RNAi or *DmPrpf39* RNAi in the dorsal wing imaginal disc compartment (under control of Ap-Gal4 driver - GFP marker) were immunostained with antibodies against the wing pouch marker Nubbin (Nub) and counterstained for DNA content with DAPI. Arrows indicate areas with ectopic Nubbin expression. **h.** Image of an *Ap > p35 + Prtf1i* fly emerging from its pupae with a single ectopic wing structure. **i.** Flies with the indicated genotypes were dissected out of mature pupae and exposed to a chemical wing inflation protocol. Arrows mark wing structures that are resistant to artificial inflation. **j.** μ -CT scanning of mature pupae containing flies with the indicated genotypes shown from the side and from the top. Normal (green) and ectopic (red) wing structures are

highlighted. (e, g) Scale bar = 100 μ m. (h, i, j) Scale bar = 1 mm. **k.** Model of PRPF39-PRTF1 function in metazoan mRNA splicing. PRPF39-PRTF1 accumulates at short and extremely GC-poor introns embedded in overall GC-poor sequence environments through recognition of U-rich motifs. From this location the complex likely facilitates recruitment of U1 to the 5' SS. For exon-defined splicing (predominant in humans and mice), this is coupled to recruitment of U2 to the upstream branch point and failure of this supportive mechanism results in exon skipping. For intron-defined splicing (predominant in flies), it is coupled to recruitment of U2 to the downstream branch point and failure of the mechanism results in intron retention.

Methods

Cell culture and reagents

Male human near haploid cells (HAP1) (RRID: CVCLY019) were cultured in Iscove's Modified Dulbecco's Medium (IMDM) GlutaMAX™ Supplement (Gibco Cat#31980030) supplemented with 10% fetal bovine serum (FBS, Biowest) and penicillin-streptomycin (Gibco Cat #15140122). Female human osteosarcoma cells (U2OS) cells (ATCC, HTB-96, RRID: CVCL0042), female human malignant cervical epithelial cells (HeLa) (ATCC, CCL-2, RRID: CVCL0030), female human embryonic kidney (HEK293) (ATCC, CRL-1573, RRID: CVCL0045) and *Mus musculus* NIH/Swiss embryonic fibroblast (NIH3T3) (ATCC, CRL-1658, RRID: CVCL_0594) were cultured in Dulbecco's Modified Eagle's Medium (DMEM, Biowest Cat# L0104-500) supplemented with 10% FBS and penicillin-streptomycin. All cells were cultured at 37°C in a humidified 5% CO₂ cell incubator.

Cell lines stably expressing S-HA-PRTF1 and GFP-PRPF39, respectively, under DOX-inducible promoters were generated by co-transfecting U2OS cells with either the pcDNA4/TO/Strep-HA-PRTF1 FL or the pcDNA4/TO-GFP-PRPF39 FL construct and the pcDNA6/TR vector (Invitrogen Cat#V102520) in a 1:4 ratio and selecting for 14 days with zeocin (200 µg/ml) and blasticidin (5 µg/ml). Individual clones were picked, and expression analyzed by immunofluorescence and Western blotting. Drugs used in this study were: Doxycycline (DOX) (Merck Cat #D3347, 0.13 µg/ml, overnight), anisomycin (Ani) (Merck Cat #A9789, 0.5 µM/ml, 1h) and cycloheximide (CHX) (Sigma Cat #C7698, 100 µM/ml, 2h).

Plasmids and minigenes

Full-length (FL) strep-HA-PRTF1 (S-HA-PRTF1) and GFP-PRPF39 plasmids were generated by subcloning PRTF1 FL into the mammalian expression vector pcDNA4/TO/Strep-HA and PRPF39 FL into the mammalian expression vector pcDNA4/TO-GFP using NotI restriction sites, under doxycycline (DOX)-inducible promoters. All truncations of PRTF1 were PCR-cloned into the pcDNA4/TO/Strep-HA vector via NotI restriction sites and internal deletions were generated by inverse PCR with phosphorylated primers and subsequent intramolecular ligation. Point mutations in PRTF1 were made by Thermo Fisher Scientific, through their GeneArt Service and cloned into the pcDNA4/TO/Strep-HA vector.

The 3-exon *mRif* minigene construct was generated to comprise the murine genomic region spanning exon 31 to exon 33, including the two intervening introns (intron 31-32 and intron 32-33). To preserve local splicing regulatory elements across the exon-intron junctions, the construct was extended by the addition of a 50-base pair-long upstream and downstream flanking intronic sequence. The 3-exon *mRif* minigene was ordered from IDT in their pEXP-IDT vector. The 2-exon *mRif* minigenes, comprising the murine genomic sequence of exon 32, intron 32-33 and exon 33,

were generated from the 3-exon minigene plasmids using Gibson cloning into the pcDNA3.1+ vector (Thermo Fisher Scientific, #V79020). Splice site point mutations (guanine (G) to adenine (A)) were done by site-directed mutagenesis using the Q5® Site-Directed Mutagenesis Kit (NEB, #E0554S) in accordance with the manufacturer's instructions, primers were designed by using the NEBaseChanger™ (NEB, <https://nebasechanger.neb.com/>) tool. The m*Prpf39* 3-exon minigenes, including the WT and the putative PRTF1-binding motif U>A mutant (Fig. 3i), were generated to comprise the murine genomic region spanning exon 8 to exon 9, including the poison exon (in intron 8-9), the two intervening introns flanking the poison exon, in addition to two long flanking sequences of respectively the 503 bp upstream and 506 bp downstream of the region of interest. Both plasmids were ordered from IDT in their pEXP-IDT vector.

The gRNAs for CRISPR/Cas9-mediated generation of U2OS ΔPRTF1 cells were cloned using the pX459 plasmid (Addgene, #62988). In brief, gRNA DNA oligos were ordered as complimentary sequences with overhangs and mixed at a 1:1 ratio for annealing. pX459 was digested with BbsI-HF (NEB, # R3539L), and the gRNA was introduced using standard ligation reaction. The following gRNA sequences were used: PRTF1-Fw; 5'- CACCGCCATACCCGCGCACCGGGG and PRTF1-Rv; 5'- AAACCCCCGGTGCGCGGGTATGGC. All constructs were verified by sequencing, and all plasmid transfections were done using FUGENE6 (Promega, #E2692) according to the manufacturer's protocol.

Knockdown by siRNA

U2OS and NIH3T3 cells were reversed transfected with siRNA using Lipofectamine™ RNAiMAX Transfection Reagent (Life Technologies Cat#13778030) following the manufacturer's instructions. Briefly, 12.5 nM siRNA was mixed with RNAiMAX in Opti-MEM™ (Gibco Cat#31985062) and incubated at room temperature for 15 minutes. Subsequently, trypsinized cells were added to the

transfection mix and incubated at room temperature for 5 minutes. The siRNA transfected cells were seeded in DMEM supplemented with 10% FBS and 1% penicillin-streptomycin. The medium was changed after 8 hours, and cells were harvested 72-96 hours post-siRNA transfection. siRNAs used in this study includes: *siLuciferase* (*siLuc/mock*): 5'- CGUACGCGGAAUACUUUCGA (dTdT), *siPRTF1*#1 (human): 5'- GUGUAAAGCCAGUGGAAUA (dTdT), *siPRTF1*#4 (human): 5'-GGGUUAUGGUUUUGUGAAA (dTdT), *siPRPF39* #1 (human): 5'- CAGAAGAACCAGAGGAAAA (dTdT), *siPRPF39* #2 (human): 5'- GGGAAAAUGAGCAGGGAAA (dTdT), *siPRPF39* #3 (mouse/human): 5'- AGGCAACAGUUCAGAGGUA (dTdT), *siPRTF1* #m1(mouse): 5'- GGGUAAAGCCAGUGGAGUA (dTdT) and *siPRTF1* #m2 (mouse): 5'- GGGCUAUGGUUUUGUGAAA (dTdT).

Transient plasmid transfection

Cells were transiently transfected using polyethylenimine (PEI). Cells were grown to 40-50% confluency and medium was changed to not contain penicillin-streptomycin prior to transfection. The transfection mix consisting of plasmid and PEI in Opti-MEM™ were incubated for 15 minutes at room temperature, and cells were transfected with a final plasmid concentration of 1.5 µg/ml. 4 hours post-transfection, the medium was changed to DMEM supplemented with 10% FBS and 1% penicillin-streptomycin. Cells were harvested 24 hours after transfection. For minigenes, cells were grown to 70-90% confluency and transfected using Lipofectamine™ 3000 Transfection Reagent (Invitrogen Cat#L3000001) following the manufacture's protocol. Cells were harvested after 48 hours, and RNA isolated using TRI reagent (Thermo Fisher Scientific Cat#AM9738) following the manufacture's instruction. Prior to reverse transcription, the RNA samples were treated with RQ1

DNase (Promega Cat# M6101) treatment using 1 µl RQ1 DNase per 1 µg of RNA in accordance with the manufacturer's instructions.

Splicing event reverse transcription PCR

Cells were grown to confluency and treated with either DMSO (mock) or 100µg/ml cycloheximide (CHX) for 2 hours prior to isolation of total RNA using TRI reagent (Thermo Fisher Scientific Cat#AM9738) following the manufacturer's protocol. Reverse transcription was performed using 1000 ng of the purified RNA per sample combined with 2.5 µM random hexamer primers (Thermo Fisher Scientific), 1 mM dNTPs (Thermo Fisher Scientific), and RevertAid Reverse Transcriptase (Thermo Scientific Cat#EP0442) according to the manufacturer's protocol. The cDNA was diluted 1:4 in UltraPure™ DNase/RNase-Free Distilled Water (Invitrogen Cat# 10977035). Splicing events were studied by gene-specific reverse transcription PCRs, using 1 µL of the diluted synthesized cDNA, 0.2 mM dNTPs, 0.2 µM of both the forward and reverse primer, 2.4% DMSO and 0.625 U of GoTaq® DNA Polymerase (Promega Cat#M7845). The PCR was performed under the following conditions: Initial denaturation at 95 °C for 2 min; 30-40 cycles of denaturation at 95 °C for 30 sec, annealing at a gene-specific temperature for 30 sec, and extension at 72 °C for 1 min; followed by a final extension at 72 °C for 5 min. PCR products were separated on a 3% agarose gel containing SYBR™ safe DNA Gel Stain (Invitrogen, # S33102) and visualized by UV using the ChemiDoc Imaging System (BioRad). PCR primers, annealing temperature and the number of cycles used in the study were: hZAK-exon_11/12-14-fw: 5'- TCCAACACCCCGCTGC and hZAK-exon_11/12-14-rv: 5'-GCCGCTTCCCTGTAATGTTG (ann. temp.: 52 °C, 30 cycles), hZAK-exon_4/5-7-fw: 5'- GTTATAGCTGCTGATGGAGTATTGAAG and hZAK-exon_4/5-7-rv: 5'- CACCATAGGAATATGTGTCACAAGTTTC (ann. temp.: 61 °C, 35 cycles), hPRPF39-exon_6/7-9-fw: 5'-CATAACCGATCCTGCAAAGC and hPRPF39-exon_6/7-9-rv: 5'-

TGCTGAAGACATGCCTCACTC (ann. temp.: 52 °C, 30 cycles), *hPRPF39*-exon_2/3-4-fw: 5'-
 CAGGAGAATCACTTGATGGCTG and *hPRPF39*-exon_2/3-4-rv: 5'-
 GAACAGCATGCTCAAAAGTTCCTCT (ann. temp.: 52 °C, 35 cycles), *hRIF1*-exon_30/31-33-
 fw: 5'-CCTGTTAATAAGGTTGCGCCG and *hRIF1*-exon_30/31-33-rv: 5'-
 TGGTATGGATTCTTTGGCC (ann. temp.: 57 °C, 40 cycles), *hFANCL*-exon_8/9-11-fw: 5'-
 GCACGCAGAATTGCATTAGGTAATAATGTTTCC and *hFANCL*-exon_8/9-11-rv: 5'-
 CAGGATAGCACGAGCTGGAAAATC (ann. temp.: 52 °C, 30 cycles), *hHNRNPA2B* -exon_1-
 2/3-fw: 5'- GCTCCTCGCAGAGTTGTTTC and *hHNRNPA2B* -exon_1-2/3-rv: 5'-
 CAGGATCCCTCATTACCACAC (ann. temp.: 52 °C, 30 cycles), *hCDC16*-exon1-3/4_-fw: 5'-
 GTCCGGCAGTACCTCGAC and *hCDC16*- exon1-3/4_-rv: 5'-
 CGACATGCTTCATACAATTTGTCC (ann. temp.: 51 °C, 30 cycles), *hPHF14*-exon14-16/17_-
 fw: 5'- TGCCACCAGAACCCAAGAAG and *hPHF14*-exon14-16/17_-rv: 5'-
 GTCTGCATTCATCACACCTGAC (ann. temp.: 53 °C, 30 cycles), *hGAPDH*-fw: 5'-
 GTCAGCCGCATCTTCTTTTG and *hGAPDH*-rv: 5'-GCGCCCAATACGACCAAATC (ann.
 temp.: 52 °C, 30 cycles), *mZak*-exon_11/12-14-fw: 5'- ATCCAACACCCCGCTGC and *mZak*-
 exon_11/12-14-rv: 5'- TGTGTCCCTTGGAGACGATG (ann. temp.: 54 °C, 35 cycles), *mRif1*-
 exon_30/31-33-TCTCACCAGTTAACAAGATTAGACGfw: 5'- and *mRif1*-exon_30/31-33-rv: 5'-
 TGTGTTGGGACAGCATGG (ann. temp.: 50 °C, 35 cycles), *mRif1*-exon_31-33-fw: 5'-
 TCATTCTTCAAATAGCTCTCCCATAAT, *mRif1*-exon_32-33-fw: 5'-
 CCACTTCAGCCAAAGGATTTCTG and *mRif1*-exon_31-33-rv: 5'-
 ATGCTGCACAGTTCACCAAA (ann. temp.: 50 °C for 3-exon system and 52 °C for 2-exon
 system and, 40 cycles), *mPrpf39*-exon_8-9-fw: 5'-GAAACCATTTGGAAAAGGCTCAG and
mPrpf39-exon_8-9-fw: 3'-CCATGGGTTTCTTTGGGAGGTGAAC (ann. temp.: 51 °C, 40 cycles),

and *mGapdh*-fw: 5'-ACTCCACTCACGGCAAATTCAACG and *mGapdh*-rv: 5'-AAGACACCAGTAGACTCCACGACA (ann. temp.: 52 °C, 30 cycles).

Western blotting, pull-downs and endogenous PRPF39-immunoprecipitation

Cell lysate was prepared in EBC buffer (50 mM Tris, pH 7.5, 150 mM NaCl, 1 mM EDTA, 0.5% NP-40) supplemented with protease and phosphatase inhibitors. Protein concentrations were normalized between samples within each experiment, and samples were boiled in Laemmli sample buffer. For endogenous PRPF39-immunoprecipitation, samples were prepared from U2OS cells grown to confluency, using one 15 cm tissue culture dish per sample. Protein A/G agarose beads were incubated in DMEM with 1.34 µg of either rabbit IgG (control) or anti-PRPF39 (Abcam Cat# ab97654, RRID:AB 10679976) antibody rotating for 1 hour at 4 °C. The antibody-coated beads were washed three times in EBC buffer prior to 2 hours of incubation with the cell lysate rotating at 4 °C. Next, the beads were washed five times in EBC buffer, and the immunoprecipitated material was eluted by boiling the beads in 1x Laemmli sample buffer. Pull-downs of tagged proteins were done either using Strep-Tactin Sepharose beads (IBA Life Sciences) or magnetic GFP-beads (homemade), incubating the protein lysate with the beads for respectively 2 hours or 1 hour rotating at 4 °C. After incubation, the beads were washed in EBC buffer, and the bound material was eluted by boiling the beads in 1x Laemmli sample buffer. Both total protein lysates and eluted proteins from immunoprecipitated samples or pull-downs were resolved by SDS-PAGE and transferred to a nitrocellulose membrane. Nitrocellulose membranes were blocked in 5% milk in PBS-tween (PBST) and incubated with primary antibody at 4 °C overnight. The next day, membranes were washed five times in PBST and incubated 1 hour with secondary antibody (1:1000 in 5% milk in PBST) at room temperature, using either goat anti-rabbit IgG antibody (H+L) peroxidase (Vector Laboratories Cat# PI-1000, RRID:AB_2336198) or horse anti-mouse IgG antibody (H+L)

peroxidase (Vector Laboratories Cat# PI-2000-1, RRID:AB_2336177). Subsequently, membranes were washed another three times and visualized by chemiluminescence (Clarity Western ECL substrate, Bio-Rad) using the Bio-Rad ChemiDoc imaging system. The antibodies used in this study were: Mouse monoclonal anti-phospho-p38 (Cell Signaling Cat#9216, RRID:AB_331296), rabbit polyclonal anti-p38 (Cell Signaling Cat#9212, RRID:AB_330713), mouse anti-phospho-SAPK/JNK (Cell Signaling Cat#9255, RRID:AB_2307321), rabbit monoclonal anti-SAPK/JNK (Cell Signaling Cat#9258, RRID:AB_2141027), rabbit polyclonal anti-ZAK (Proteintech Cat#14945-1-AP, RRID:AB_10642698), rabbit polyclonal anti-ZAK α (Bethyl Cat#A301-993A, RRID:AB_1576612), mouse monoclonal anti-p150 (BD biosciences Cat#610473, RRID:AB_397845), mouse monoclonal anti- α -tubulin (Sigma Cat# T9026, RRID:AB_477593), rabbit polyclonal anti-TRNAU1AP (Proteintech Cat#15053-1-AP, RRID:AB_10644291), rabbit polyclonal anti-PRPF39 (Abcam Cat#97654, RRID:AB_10679976), rabbit polyclonal anti-SNRNP200 (Bethyl Cat#A303-453A, RRID:AB_10949362), rabbit polyclonal anti-SF3A1 (Invitrogen Cat#PA5-51439, RRID:AB_2647173), mouse monoclonal anti-TIA-1 (Thermo Fisher Cat#MA526474, RRID:AB_2725518), mouse monoclonal anti-TIAR (Proteintech Cat#66907-1-IG, RRID:AB_2882234), rabbit polyclonal anti-SF1 (Boster Biological Cat#PA1864, RRID:AB_3082814) mouse monoclonal anti-U1 snRNP 70 (C3) (Santa Cruz Biotechnology Cat#sc-390899, RRID:AB_2801569), mouse monoclonal anti-HA (Santa Cruz Biotechnology Cat#sc-7392, RRID:AB_627809), mouse monoclonal anti-GFP (Roche Cat#11814460001, RRID:AB_390913) and mouse monoclonal anti-FLAG M2 (Sigma Cat# F1804, RRID:AB_262044).

Quantitative RT-PCR

Total RNA isolation and reverse transcription was performed as described for the splicing event PCR with the two modifications, that only 500 ng of total RNA was used for reverse transcription, and the cDNA was diluted 1:5. qPCRs were performed using SYBR™ Green Universal Master Mix (Applied Biosystems Cat#4309155) following the manufacturer's instructions. Primers used for qPCR in this study were: *PRPF39*-fw1: 5'-CTGTGGACCTTCCAGTGACG, *PRPF39*-rv1: GCAGCCATCAAGTGATTCTCC, *PRPF39*-fw2: 5'- AGTGCTGTGGACCTTCCAGT, *PRPF39*-rv2: GGCAGCCATCAAGTGATTCTCCT and the *hGAPDH* primer pair also used for the splicing event PCRs.

AlphaFold 3 modelling and SPOC interaction scoring

The code for running AlphaFold 3 ²² was obtained from the AlphaFold 3 GitHub repository (www.google-deepmind/alphafold3 - release v3.0.1), while model weights were procured from Google upon request. The model was run locally in a NVIDIA Quadro RTX 8000 GPU with unified memory and memory preallocation enabled.

Screening runs consisted of *PRPF39*:possible interactor or *PRTF1*:possible interactor dimers.

Possible interactors were all canonical protein isoforms, as per Uniprot, in the Reactome pathway R-HSA-72203 ⁴⁸. SPOC ²³ was used to score the dimers, with default parameters.

RNA sequencing

Library preparation and sequencing: For human cell lines, total RNA was isolated from HAP1 cells and siRNA transfected U2OS cells according to the manufacturer's instructions using TRI reagent (Thermo Fisher Scientific, AM9738). All samples were done in biological triplicates. Total RNA samples from HAP1 cells were processed for poly(A)-selected, non-stranded mRNA library preparation, and total RNA samples from U2OS cells were processed by BGI for poly(A)-selected,

stranded mRNA library preparation. Libraries were sequenced using paired-end 100 bp reads sequencing with 50 million paired reads per sample on DNBSEQ sequencer. Whole-body *Drosophila melanogaster* samples were homogenized using a TissueLyser II (Qiagen) in TRI Reagent, and RNA was extracted using TRI Reagent (Thermo Fisher Scientific, AM9738). *Drosophila* RNA samples were sequenced in two batches. In both batches, poly(A)-selected, stranded mRNA libraries were prepared and sequenced using paired-end 150 bp reads.

Alignment to reference genome: Raw fastq files were first trimmed using Trim Galore with default parameters, and then aligned against their corresponding reference genome and GTF annotations (GENCODE v49 GRCh38.p14 for *H. sapiens*, and ENSEMBL BDGP6.54 for *D. melanogaster*) using Spliced Transcripts Alignment to a Reference (STAR v2.7.11b). STAR parameters were set as recommended in Wang et al ⁴⁹. Read coverage tracks were generated as normalized bigWig files using bamCoverage from deeptools (v3.5.5) and visualized in Integrative Genomics Viewer (IGV, version 2.19.2, ⁵⁰).

Splicing analysis

Splicing event detection and quantification: Splicing events were detected with replicate Multivariate Analysis of Transcript Splicing (rMATS-turbo) ^{24,49} v4.3.0 from the alignment files. The *prep* and *post* steps were run separately to allow the comparison of libraries with different strandedness and different read lengths. Before running rMATS, library strandedness was estimated by *infer_experiment* from RSeQC, ⁵¹, and read length as the mode of the length of the first 1000 reads as per pysam (<https://github.com/pysam-developers/pysam>). Then, for each alignment file, we ran `rmats.py --task prep --libType {infer_experiment libType} --readLength {mode length} --mil 30 -t paired --allow-clipping --novelSS --variable-read-length`, using as GTF the corresponding GTF annotation file (GENCODE v49 GRCh38.p14 for *H. sapiens*, and ENSEMBL BDGP6.54 for *D.*

melanogaster). Comparisons for the *post* step were defined as: deficient vs. controls (all PRPF39/PRTF1 or Prp39/Secp43 KO/KD vs. all control conditions, per organism), and deficient vs. control, in which each KD or KO sample was compared to its corresponding control. The *post* step was executed with the same parameters as the *prep* step, except for `--b1 {deficient_bam} --b2 {control_bam} --tmp {prep_output} --od {post_output} --task post`.

Splicing event filtering: The rMATS-sourced splicing events were filtered to obtain a subset of high-confidence, non-overlapping events from which to operate on. First, as recommended in Wang et al ⁴⁹, events with less than mean 10 reads per sample were eliminated. Then, we calculated a p-value describing the confidence for the existence of each detected junction set on each condition, described as:

$$p_{it} = 1 - \text{CDF}_{\text{Poisson}(n_{it})} \left(\sum_j c_{itj} - 1 \right)$$

Where:

- p_{it} is the p-value for condition $i \in \{\text{deficient, control}\}$, junction type $t \in \{\text{skipped, included}\}$
- n_{it} is the number of replicates for junction i , type t
- c_{itj} is the read count for junction i , type t , replicate j
- $\sum_j c_{itj}$ is the total read count across all replicates
- The expected read count under the null hypothesis is 1 read per replicate, modelled as a Poisson distribution of $\mu = n_{it}$

The obtained p-values were Bonferroni-Hochberg corrected to $p_{adj,it}$, and events where

$$\exists i \in \{\text{deficient, control}\}, \forall t \in \{\text{skipped, included}\}: p_{adj,it} < 0.01$$

were kept. Data manipulation for this and subsequent methods was done in python using polars v1.38.0, numpy v2.3.0, and numba v0.63.1.

Removal of overlapping splicing events and defining the foreground/background sets: Two events were considered overlapping if the alternatively spliced exon defining the events overlapped. Events were clustered using polars-bio v0.25.0 so that overlapping events were assigned to the same cluster. Then, we ranked the events within the same cluster using the value

$$p_{summary} = \max_{t \in \{skipped, included\}} \left(\min_{i \in \{perturbation, control\}} p_{adj, it} \right)$$

and kept, for each cluster, only the event with the lowest $p_{summary}$.

For the skipped exon events, we then defined a foreground set (events where $\Delta PSI > 0.05$ and $FDR < 0.01$) and a background set (events where $\Delta PSI < 0.05$ and $FDR > 0.9$).

Bioinformatics

Motif search: Motif search was performed using the MEME suite⁵² v5.5.0 subprogram Xstreme⁵³, with parameters *xstreme --rna --evt 0.05 --minw 4 --maxw 10 --meme-mod zoops --p foreground_sequences.fasta --n background_sequences.fasta*.

All N foreground sequences and $4N$ randomly sampled background sequences were extracted from the relevant reference genome to foreground and background fasta files via pyfaidx v0.9.0.3, respectively using their corresponding coordinates in the rMATS events. Each sequence consisted of the event subsequence ranging from 25nt upstream to 200nt downstream the cassette exon end. The obtained motifs were then visually compared against the registered PRTF1/Secp43 motifs in the CisBP-RNA database²⁸ to check for similarity.

Nucleotide composition metagene profiling: We extracted RNA sequences consisting of the nucleotides in positions -200 to +50 from the cassette exon acceptor site (acceptor site = 0, acceptor sequence), or consisting of the nucleotides -50 to +200 from the cassette exon donor site (donor site

= 0, donor sequence), from the human genome. Using the python packages numpy, polars, numba, and scipy v1.17.1, acceptor and donor sequences were then separately stacked and one-hot encoded, averaged across the stacking dimension and smoothed across the sequence dimension via a 1D Gaussian filter with $\sigma = 2$ and *radius* = 3 to obtain the metagene profiles.

Motif effect estimation: We used KATMAP⁵⁴ to infer the possible impact of known PRTF1 motifs on ΔPSI in our experiments. For this, we used as input the filtered, non-overlapping skipped exon (SE) events described above, the reference human genome, and each of the known PRTF1/TRNAU1AP motifs present in CISPB-RNA¹³. We kept the model for the motif with the least assumed binding changes per exon (11.82 binding changes, motif TRNAU1AP_M00827_2.00).

RNA binding protein enrichment calculations: All RNA binding protein (RBP) CLIP enriched regions from HepG2 cells were directly obtained from the RBP-ARK database⁵⁵. Only the “reproducible enriched windows” were considered.

Foreground and background skipped exon events were divided into 5 non-overlapping segments: upstream exon, upstream intron, cassette exon, downstream intron, and downstream exon. Then, for each CLIP experiment, we counted how many foreground and background events featured at least one enriched window within each of the segments. From those counts, we obtained a foreground-to-background enrichment ratio and an overrepresentation P-value (using Fisher’s exact test), which we adjusted with the Bonferroni-Hochberg method.

Size of RBP-bound introns: For each gene in the GENCODE v49 GRCh38.p14 GFF annotation, we considered only the transcript with the most exons. We defined the introns as the intervals between those exons. Then, we overlapped the enriched window intervals with the intron intervals, keeping at most one overlap per intron per RBP, and excluded RBPs with less than 1000 intron overlaps.

Computational splicing model

Splicing event featurization: We extracted a set of relatively interpretable features for every skipped exon event, which can be summarized as length features, nucleotide composition features, and splice site strength features (Fig. 4a). Length features were defined for each segment composing a skipped exon event as the distance between the start and end of the segment. Similarly, nucleotide composition (as GC content) was calculated for each segment as the GC percentage in it.

We used two sources to quantify species-specific splice site strength: MaxEntScan scores^{56,57} and SMSplice-learned splice regulatory element (SRE) scores⁵⁶. We got all the model weights and parameters from the SMSplice publication⁵⁶ and Github (<https://github.com/kmccue/SMsplice>).

We used the difference in MaxEntScan score between cassette exon splice site and constitutive (surrounding) splice site (that is, cassette exon donor site minus upstream exon donor site, and cassette exon acceptor site minus downstream exon acceptor site) as features. We also calculated SRE score features as the mean SRE scores for the 80nt sequences upstream and downstream the cassette exon donor and acceptor sites, individually.

Splicing event feature selection: After defining foreground and background sets and event featurization, we trained an Explainable Boosting Machine (EBM) to discover which features were predictive of an event belonging to the foreground set. We utilized the ExplainableBoostingClassifier from the interpret v0.7.4 python package²⁹, with parameters `smoothing_rounds=25`, `interactions=4`, `max_bins=64`, `max_interaction_bins=64`, `inner_bags=20`, `learning_rate=0.1`. Foreground to background class imbalance was addressed by setting the weight of the background class events to 1, and the foreground class weight to the ratio of the number of background class to foreground class events. The model was fit with the objective of discerning between foreground (label 1) and background (label 0) observations.

We obtained performance metrics from the trained model using scikit-learn v1.8.0, extracted feature importance from the trained models via a custom python script, and inspected them visually.

Event conservation assessment: Foreground and background human skipped exon events were matched by their cassette exon start and cassette exon end coordinates to those of the skipped exons registered in VASTDB⁵⁸ v11 to associate them with a human VASTDB identifier. We used the human VASTDB identifier to recover the associated ortholog mouse VASTDB event identifier. Events not matching both a human and a mouse identifier were dropped. We then used the genomic coordinates from the identified mouse events, combined with the *Mus musculus* GRCm38 (mm10) reference genome and the mouse MaxEntScan¹⁶ and SRE scores from the SMSplice repository⁵⁶ and Github (<https://github.com/kmccue/SMsplice>), to extract features from them as described for the human SE events. These features were then used as input to score each mouse event according to the EBM trained on human data described in the *Splicing event feature selection* section, thus obtaining a predicted score for each mouse event associated to an already scored ortholog human event.

Analysis of splicing with SpliceAI and AlphaGenome

Splice site scoring from sequence: Skipped exon event sequences were scored using SpliceAI³¹ v1.3.1 locally. For each event, the whole sequence (from the upstream exon start to the downstream exon end coordinates) was extracted from the human reference genome, padded with 5000 'N' at each side, and scored throughout its whole length.

SE events were also scored using Alphagenome³² from the Alphagenome API v0.6.0.

Alphagenome intervals were defined as the genomic intervals from upstream exon start to downstream exon end coordinates, padded to the minimum Alphagenome-compatible length greater than the length of the interval. The instantiated Alphagenome dna_client model was then invoked

for each interval with parameters `ontology_terms=["EFO:0001187"]`,
`requested_outputs=[dna_client.OutputType.SPLICE_SITES,`
`dna_client.OutputType.SPLICE_SITE_USAGE]`, to obtain donor/acceptor scores and splice site
usage predictions across the whole length of each event.

Lastly, for each sequence scoring method (SpliceAI donor/acceptor scoring, Alphagenome
donor/acceptor scoring, and Alphagenome splice site usage scoring), we extracted the score at the
rMATS splice site coordinates upstream exon end, cassette exon start, cassette exon end, and
downstream exon start. The predicted splice site usage was normalized for each event so that

$$\text{Normalized cassette exon donor usage} = \frac{\text{Cassette exon donor usage}}{\text{Upstream exon donor usage}}$$

$$\text{Normalized cassette exon acceptor usage} = \frac{\text{Cassette exon acceptor usage}}{\text{Downstream exon acceptor usage}}$$

Drosophila melanogaster

Strains: The following strains were obtained from the Bloomington *Drosophila* Stock Center (BDSC): *act-Gal4* (RRID:BDSC_4414), *UAS-p35* (RRID:BDSC_5072, insertion in Chr2; ID: 6298, insertion in ChrX), *UAS-mCD8-GFP* (RRID:BDSC_5137, UAS-GFP in the text), and *tub-G80ts* (RRID:BDSC_7018; insertion in Chr3). The following strains were obtained from the Vienna *Drosophila* Resource Center (VDRC): *UAS-TRNAUP RNAi* (GD 19494), and *UAS-Prp39 RNAi* (GD 105999). Other strain used in this article was *ap-Gal4* ⁵⁹.

Cross maintenance: Crosses were kept at 18°C in standard conditions, flipped every 2 or 3 days to allow egg laying, and moved to 29°C to induce Gal4 expression. Vials were transferred to an incubator at 29°C right after flipping.

Immunofluorescence: Primary antibodies used: anti-Wg (Developmental Studies Hybridoma

Bank, ID: 4D4, 1:50); rabbit anti-Dcp1 (Cell Signaling Technology Cat# 9578, RRID:AB_2721060, 1:100), anti-Nubbin³⁹. Secondary antibodies were acquired from Molecular Probes. All of them were generated in goat and used at a 1:400 dilution: anti-mouse IgG-555 (Thermo Fisher Scientific Cat# A-21425, RRID:AB_2535846), anti-rabbit IgG-555 (Thermo Fisher Scientific Cat# A-21430, RRID:AB_2535851). 4',6-diamino-2-phenylindole (DAPI) was obtained from Invitrogen (no. D1306) and used at a 600 nM concentration. Wing imaginal discs from third instar wandering larvae were dissected in cold phosphate-buffered saline (PBS) and subsequently fixed in a 3.7% formaldehyde solution diluted in PBS for 20 minutes at room temperature (RT). The samples were then washed three times for 10 minutes each in 0.2% Triton X-100 in PBS (PBT) and blocked for at least 20 minutes in a solution of 0.5-1% BSA in PBT with 250 mM NaCl (BBT). Following this, the samples were incubated with the primary antibody diluted in BBT at RT overnight or at 4°C for approximately 60 hours. After incubation, the samples were washed three times for 10 minutes each in BBT and then incubated with DAPI and fluorescent secondary antibodies for 1.5 hours at RT. Finally, the samples were washed four times for 15 minutes each in PBT, mounted in 90% glycerol in PBS containing 0.5% propyl gallate, and stored at 4 °C.

Image processing and data analysis: Images were obtained using a Leica SP8 confocal microscope and processed with the Fiji software. The exact procedures for quantification are explained in the sections below. Figures were prepared with Affinity by Canva.

Quantification of apoptotic area: Confocal images of wing imaginal discs were captured using a Leica SP8 confocal microscope. A "median" filter with a value of 1 was applied to the Dcp1 channel, followed by thresholding to isolate the Dcp1-positive signal. The area of the Dcp1 signal within the dorsal compartment was quantified using the "measure" function in the ROI manager. This Dcp1 area was then normalized to the previously measured dorsal compartment area. The

results were presented as the fold change of the Dcp1 area relative to the dorsal area in the control sample.

Pupal extraction: Pharate adult *Drosophila* were first carefully extracted from their pupal cases. Pupae were positioned ventral-side up under a stereomicroscope, and the operculum and anterior portion of the puparium were carefully opened using fine forceps. The pharate adult was gently extracted to expose the developing wings while preserving their attachment to the thorax.

KOH treatment: Extracted flies were transferred to a glass micro-well containing 10% potassium hydroxide (KOH) prepared in distilled water. Samples were incubated for 10 minutes at 110 °C by placing the glass well directly onto a pre-heated dry heat block. This step softens residual cuticle and facilitates subsequent wing expansion.

Osmotic wing inflation: Immediately after KOH exposure, samples were moved to distilled water, where the wings expanded by osmotic swelling. This passive inflation assists in fully unfolding the compact pupal wing structure. When wings did not fully extend after the first cycle, the KOH incubation followed by osmotic inflation was repeated once under identical conditions.

Micro-CT analysis of *Drosophila* pupae: Micro-computed tomography (μ -CT) was performed on phosphotungstic acid (PTA)-stained *drosophila* pupae using a Bruker Skyscan 1272 (CMOS edition; control software v1.23.0.2+). Scans were acquired at 50 kV and 200 μ A with frame averaging 2, 2 $\times$ 2 binning, and a pixel size of 4 μ m. Prior to scanning, pupae were incubated in 1 mL 0.5% Tween 20 in PBS (PBST) at 100°C for 20 s, then cooled at room temperature for 5 min. Bouin's solution (5% acetic acid, 9% formaldehyde, 0.9% picric acid) was added for 2 h, after which the anterior pupal case was punctured with a microdissection needle. Samples were then fixed in fresh Bouin's solution for 24 h, washed 3 $\times$ 30 min in micro-CT wash buffer (0.1 M Na₂HPO₄, 1.8% sucrose) and stained with 0.5% PTA (w/v) in distilled water for 5 days.

Reconstruction was performed using NRecon (v2.2.0.6), and volumes were visualized in Bruker CTVox (v3.3.1).

Data representation and statistical analysis

Data visualization and statistical analyses were performed using R programming language (R Core Team, 2025) within RStudio (Version 2025.09.1+401; Posit Team, 2025) using the Tidyverse suite of packages. Figure 6f was generated using GraphPad Prism (version 10.4.2).

Linear relationship between two continuous variables were assessed using the Pearson correlation coefficient (r). Depending on data distribution, statistical comparisons were performed using either the Kruskal–Wallis test followed by Dunn’s post hoc test with Bonferroni correction or a one-way ANOVA followed by Dunnett post hoc test.

Statistical significance was defined as ns (not significant), $p \leq 0.05$ (*), $p \leq 0.01$ (**), $p \leq 0.001$ (***) and $p \leq 0.0001$ (****).

Data and code availability

Demultiplexed FASTQ files containing RNA-seq data and processed BigWig files have been deposited in the Gene Expression Omnibus (GEO) with the accession number **GSE324756** and reviewer token **kbxyocxfgvjyx**. Code is available at the github repo <https://github.com/BekkerJensenLab/Splicing>. It is private until publication but can be accessed with the personal access token:

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. I.e. to clone the repo run “*git clone*

https://user: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@github.com/BekkerJensenLab/Splicing.git

Supplementary tables

Table S1.

Human foreground exon skipping data set – Δ PSI and FDR

Table S2.

Human vs. mouse prediction scores for orthologous exons

Table S3.

Fly foreground exon skipping data set – Δ PSI and FDR

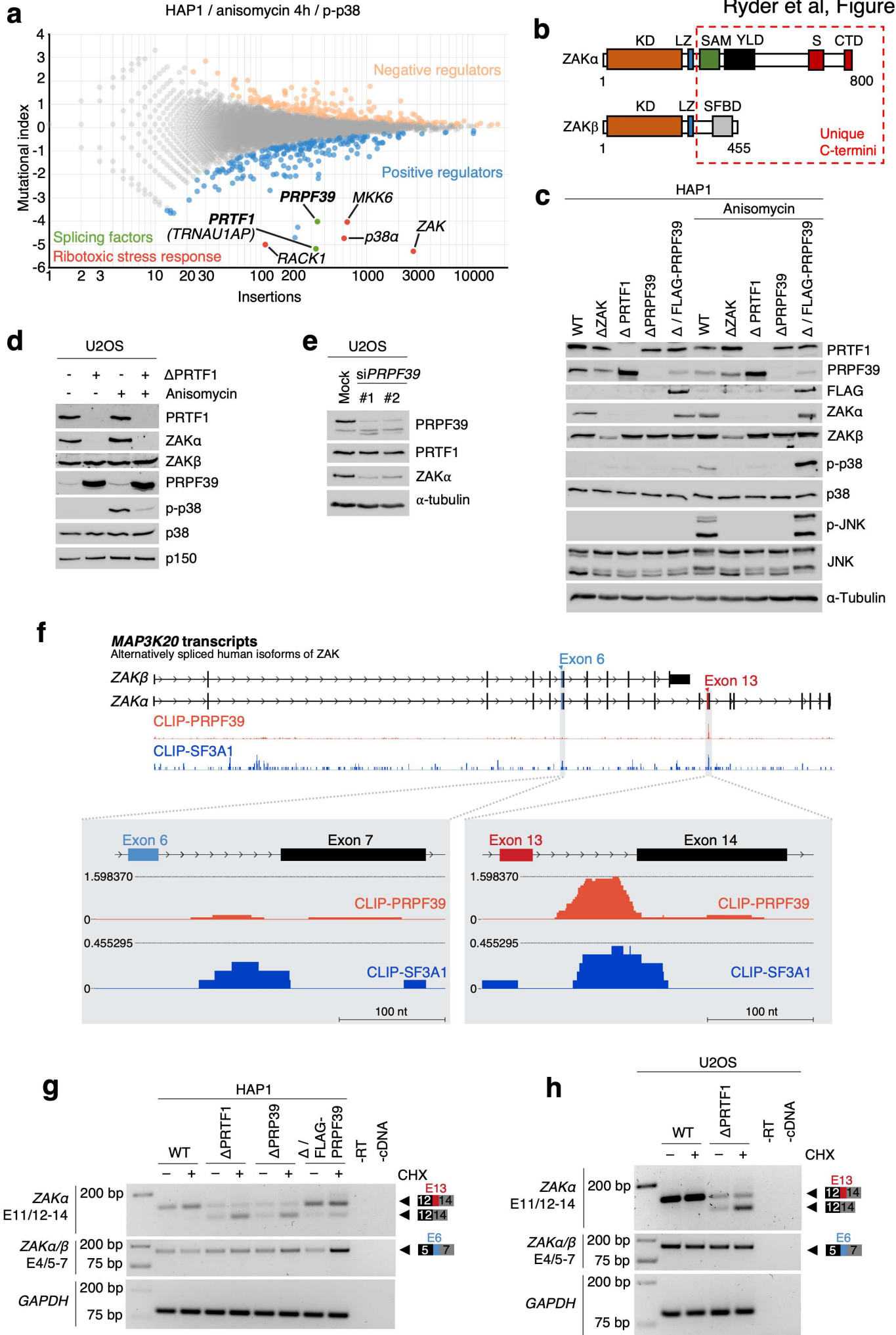
References

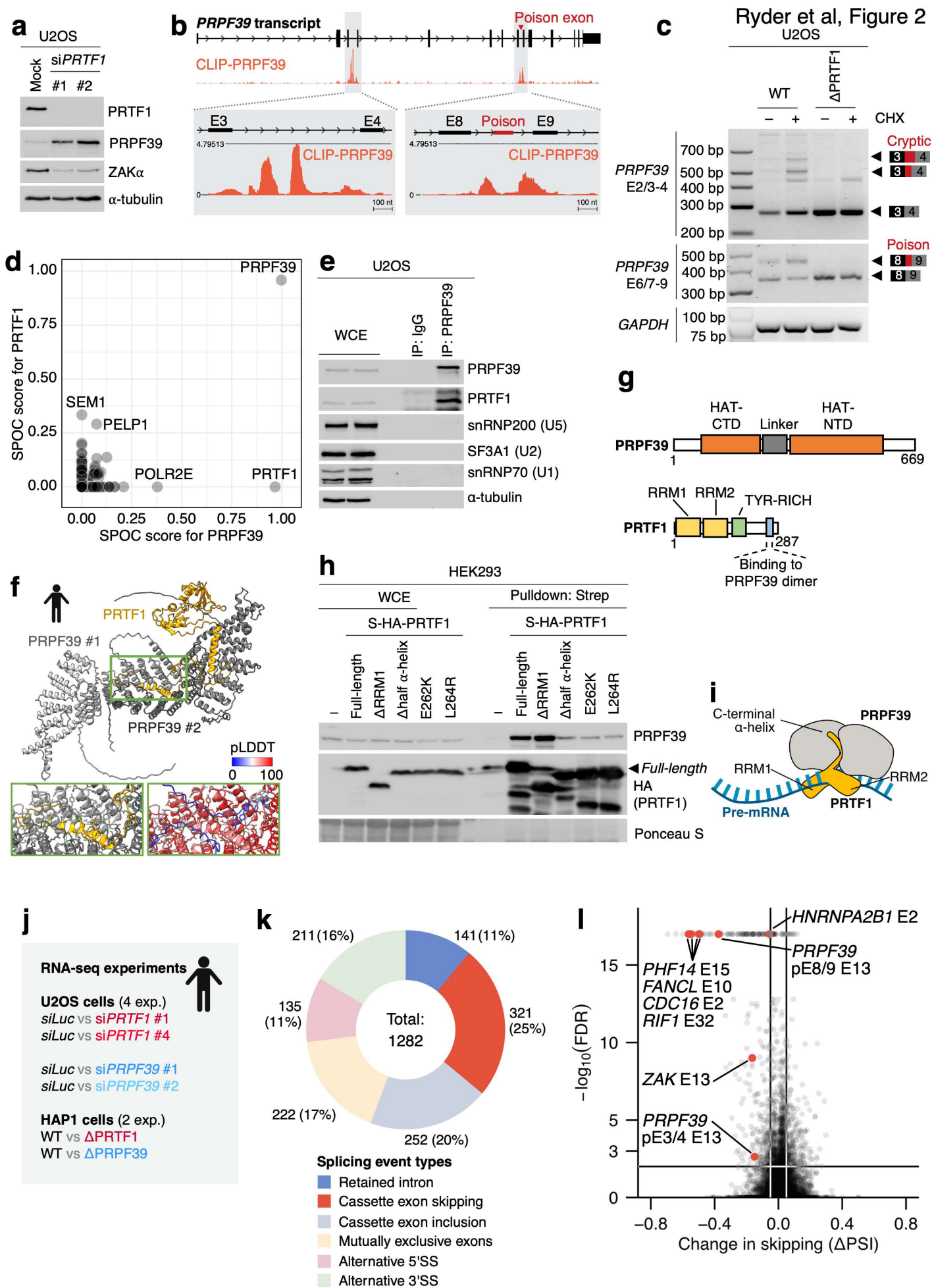
- 1 Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).
<https://doi.org/10.1038/227561a0>
- 2 Shine, M. *et al.* Co-transcriptional gene regulation in eukaryotes and prokaryotes. *Nat Rev Mol Cell Biol* **25**, 534–554 (2024). <https://doi.org/10.1038/s41580-024-00706-2>
- 3 Wilkinson, M. E., Charenton, C. & Nagai, K. RNA Splicing by the Spliceosome. *Annu Rev Biochem* **89**, 359–388 (2020). <https://doi.org/10.1146/annurev-biochem-091719-064225>
- 4 Rogalska, M. E., Vivori, C. & Valcarcel, J. Regulation of pre-mRNA splicing: roles in physiology and disease, and therapeutic prospects. *Nat Rev Genet* **24**, 251–269 (2023).
<https://doi.org/10.1038/s41576-022-00556-8>
- 5 Plaschka, C., Newman, A. J. & Nagai, K. Structural Basis of Nuclear pre-mRNA Splicing: Lessons from Yeast. *Cold Spring Harb Perspect Biol* **11** (2019).
<https://doi.org/10.1101/cshperspect.a032391>
- 6 De Conti, L., Baralle, M. & Buratti, E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA* **4**, 49–60 (2013). <https://doi.org/10.1002/wrna.1140>
- 7 Parenteau, J. & Abou Elela, S. Introns: Good Day Junk Is Bad Day Treasure. *Trends Genet* **35**, 923–934 (2019). <https://doi.org/10.1016/j.tig.2019.09.010>
- 8 Will, C. L. & Luhrmann, R. Spliceosome structure and function. *Cold Spring Harb Perspect Biol* **3** (2011). <https://doi.org/10.1101/cshperspect.a003707>
- 9 Berget, S. M. Exon recognition in vertebrate splicing. *J Biol Chem* **270**, 2411–2414 (1995).
<https://doi.org/10.1074/jbc.270.6.2411>
- 10 Tammer, L. *et al.* Gene architecture directs splicing outcome in separate nuclear spatial regions. *Mol Cell* **82**, 1021–1034 e1028 (2022).
<https://doi.org/10.1016/j.molcel.2022.02.001>
- 11 Amit, M. *et al.* Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. *Cell Rep* **1**, 543–556 (2012).
<https://doi.org/10.1016/j.celrep.2012.03.013>
- 12 Plaschka, C., Lin, P. C. & Nagai, K. Structure of a pre-catalytic spliceosome. *Nature* **546**, 617–621 (2017). <https://doi.org/10.1038/nature22799>
- 13 Li, X. *et al.* CryoEM structure of *Saccharomyces cerevisiae* U1 snRNP offers insight into alternative splicing. *Nat Commun* **8**, 1035 (2017). <https://doi.org/10.1038/s41467-017-01241-9>
- 14 Qiu, Z. R., Schwer, B. & Shuman, S. Determinants of Nam8-dependent splicing of meiotic pre-mRNAs. *Nucleic Acids Res* **39**, 3427–3445 (2011). <https://doi.org/10.1093/nar/gkq1328>
- 15 Puig, O., Gottschalk, A., Fabrizio, P. & Seraphin, B. Interaction of the U1 snRNP with nonconserved intronic sequences affects 5' splice site selection. *Genes Dev* **13**, 569–580 (1999). <https://doi.org/10.1101/gad.13.5.569>
- 16 De Bortoli, F. *et al.* Increased versatility despite reduced molecular complexity: evolution, structure and function of metazoan splicing factor PRPF39. *Nucleic Acids Res* **47**, 5867–5879 (2019). <https://doi.org/10.1093/nar/gkz243>
- 17 Espinosa, S. *et al.* Human PRPF39 is an alternative splicing factor recruiting U1 snRNP to weak 5' splice sites. *RNA* **29**, 97–110 (2022). <https://doi.org/10.1261/rna.079320.122>

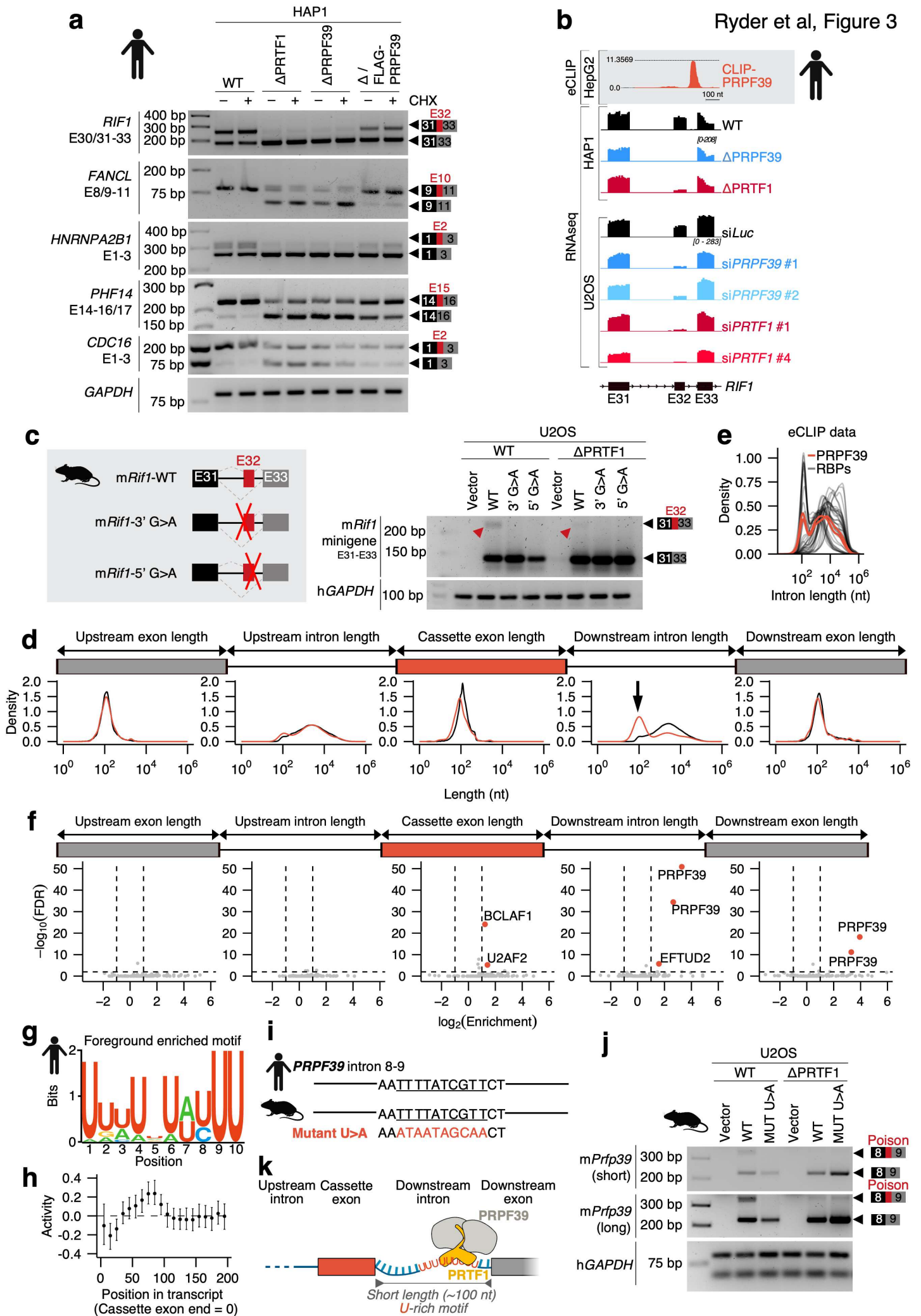
- 18 Brockmann, M. *et al.* Genetic wiring maps of single-cell protein states reveal an off-switch for GPCR signalling. *Nature* **546**, 307–311 (2017). <https://doi.org/10.1038/nature22376>
- 19 Arafeh, R., Shibue, T., Dempster, J. M., Hahn, W. C. & Vazquez, F. The present and future of the Cancer Dependency Map. *Nat Rev Cancer* **25**, 59–73 (2025). <https://doi.org/10.1038/s41568-024-00763-x>
- 20 Nordgaard, C. *et al.* ZAKbeta is activated by cellular compression and mediates contraction-induced MAP kinase signaling in skeletal muscle. *EMBO J* **41**, e111650 (2022). <https://doi.org/10.15252/emboj.2022111650>
- 21 Vind, A. C., Zhong, F. L. & Bekker-Jensen, S. Death by ribosome. *Trends Cell Biol* **35**, 615–626 (2025). <https://doi.org/10.1016/j.tcb.2024.10.013>
- 22 Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024). <https://doi.org/10.1038/s41586-024-07487-w>
- 23 Schmid, E. W. & Walter, J. C. Predictomes, a classifier-curated database of AlphaFold-modeled protein-protein interactions. *Mol Cell* **85**, 1216–1232 e1215 (2025). <https://doi.org/10.1016/j.molcel.2025.01.034>
- 24 Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A* **111**, E5593–5601 (2014). <https://doi.org/10.1073/pnas.1419161111>
- 25 Yu, J. *et al.* Minimal introns are not "junk". *Genome Res* **12**, 1185–1189 (2002). <https://doi.org/10.1101/gr.224602>
- 26 Dominguez, D. *et al.* Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell* **70**, 854–867 e859 (2018). <https://doi.org/10.1016/j.molcel.2018.05.001>
- 27 Van Nostrand, E. L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020). <https://doi.org/10.1038/s41586-020-2077-3>
- 28 Rosado-Tristani, D. A. *et al.* CisBP-RNA: a web resource for eukaryotic RNA-binding proteins and their motifs. *Nucleic Acids Res* **54**, D98–D105 (2026). <https://doi.org/10.1093/nar/gkaf1081>
- 29 Nori, H., Jenkins, S., Koch, P. & Caruana, R. InterpretML: A Unified Framework for Machine Learning Interpretability. arXiv:1909.09223 (2019). <https://ui.adsabs.harvard.edu/abs/2019arXiv190909223N>.
- 30 Gazave, E., Marques-Bonet, T., Fernando, O., Charlesworth, B. & Navarro, A. Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biol* **8**, R21 (2007). <https://doi.org/10.1186/gb-2007-8-2-r21>
- 31 Jaganathan, K. *et al.* Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548 e524 (2019). <https://doi.org/10.1016/j.cell.2018.12.015>
- 32 Avsec, Z. *et al.* Advancing regulatory variant effect prediction with AlphaGenome. *Nature* **649**, 1206–1218 (2026). <https://doi.org/10.1038/s41586-025-10014-0>
- 33 Talerico, M. & Berget, S. M. Intron definition in splicing of small Drosophila introns. *Mol Cell Biol* **14**, 3434–3445 (1994). <https://doi.org/10.1128/mcb.14.5.3434-3445.1994>
- 34 McGuire, S. E., Le, P. T., Osborn, A. J., Matsumoto, K. & Davis, R. L. Spatiotemporal rescue of memory dysfunction in Drosophila. *Science* **302**, 1765–1768 (2003). <https://doi.org/10.1126/science.1089035>
- 35 Li, X. *et al.* A unified mechanism for intron and exon definition and back-splicing. *Nature* **573**, 375–380 (2019). <https://doi.org/10.1038/s41586-019-1523-6>

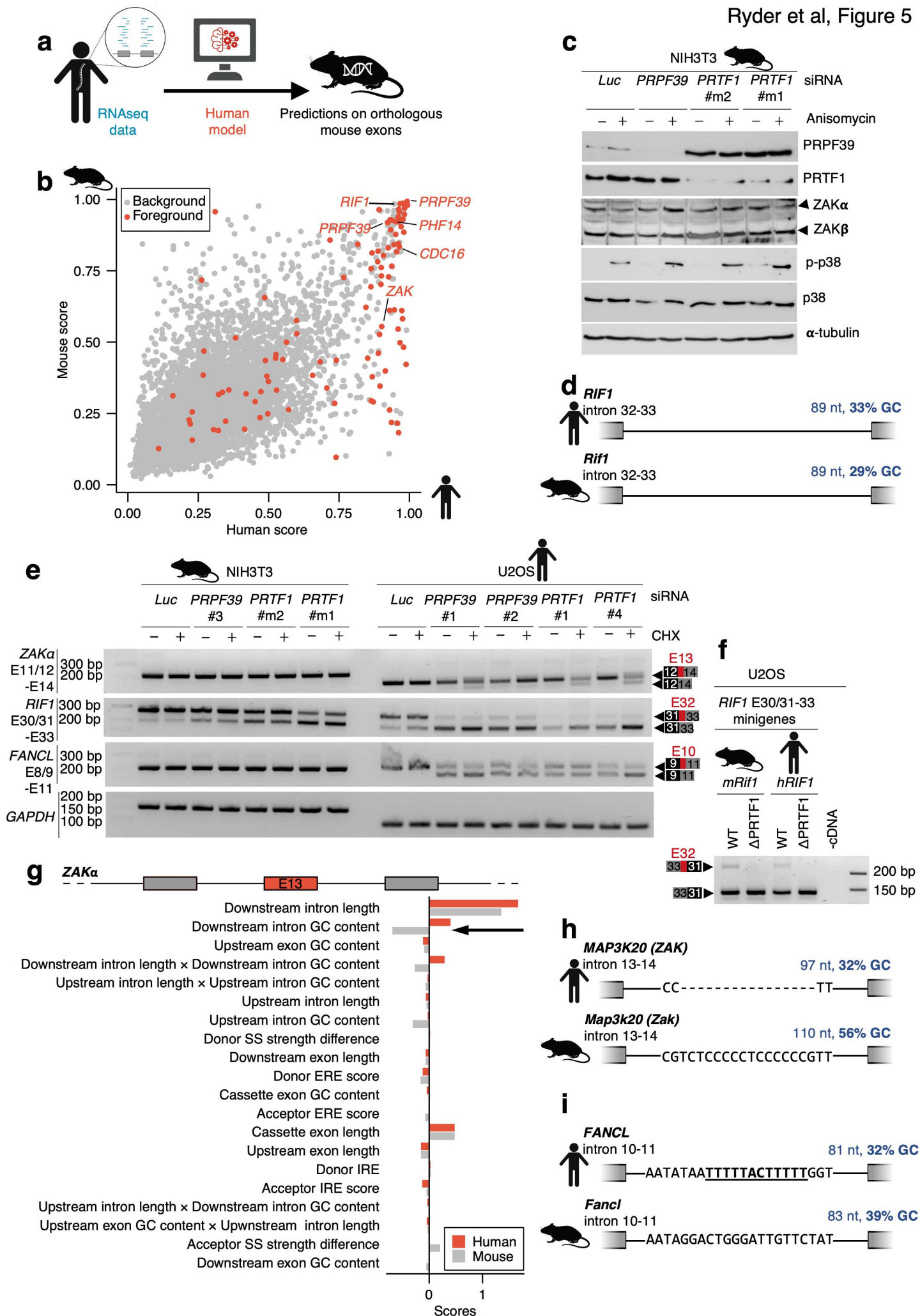
- 36 Tripathi, B. K. & Irvine, K. D. The wing imaginal disc. *Genetics* **220** (2022).
<https://doi.org/10.1093/genetics/iyac020>
- 37 Tenev, T., Zachariou, A., Wilson, R., Ditzel, M. & Meier, P. IAPs are functionally non-equivalent and regulate effector caspases through distinct mechanisms. *Nat Cell Biol* **7**, 70–77 (2005). <https://doi.org/10.1038/ncb1204>
- 38 Hay, B. A., Wolff, T. & Rubin, G. M. Expression of baculovirus P35 prevents cell death in *Drosophila*. *Development* **120**, 2121–2129 (1994). <https://doi.org/10.1242/dev.120.8.2121>
- 39 Neumann, C. J. & Cohen, S. M. Boundary formation in *Drosophila* wing: Notch activity attenuated by the POU protein Nubbin. *Science* **281**, 409–413 (1998).
<https://doi.org/10.1126/science.281.5375.409>
- 40 Ng, M., Diaz-Benjumea, F. J., Vincent, J. P., Wu, J. & Cohen, S. M. Specification of the wing by localized expression of wingless protein. *Nature* **381**, 316–318 (1996).
<https://doi.org/10.1038/381316a0>
- 41 Callaway, E. DeepMind's new AlphaGenome AI tackles the 'dark matter' in our DNA. *Nature* **643**, 17–18 (2025). <https://doi.org/10.1038/d41586-025-01998-w>
- 42 Capitanchik, C., Wilkins, O. G., Wagner, N., Gagneur, J. & Ule, J. From computational models of the splicing code to regulatory mechanisms and therapeutic implications. *Nat Rev Genet* **26**, 171–190 (2025). <https://doi.org/10.1038/s41576-024-00774-2>
- 43 Chen, X. *et al.* Context-defined cancer co-dependency mapping identifies a functional interplay between PRC2 and MLL-MEN1 complex in lymphoma. *Nat Commun* **14**, 4259 (2023). <https://doi.org/10.1038/s41467-023-39990-5>
- 44 Ding, F. & Grabowski, P. J. Identification of a protein component of a mammalian tRNA(Sec) complex implicated in the decoding of UGA as selenocysteine. *RNA* **5**, 1561–1569 (1999). <https://doi.org/10.1017/s1355838299991598>
- 45 Mahdi, Y. *et al.* Expression of Selenoproteins Is Maintained in Mice Carrying Mutations in SECp43, the tRNA Selenocysteine 1 Associated Protein (Trna^{u1ap}). *PLoS One* **10**, e0127349 (2015). <https://doi.org/10.1371/journal.pone.0127349>
- 46 Schmok, J. C. *et al.* Large-scale evaluation of the ability of RNA-binding proteins to activate exon inclusion. *Nat Biotechnol* **42**, 1429–1441 (2024). <https://doi.org/10.1038/s41587-023-02014-0>
- 47 Meyer, C. *et al.* The TIA1 RNA-Binding Protein Family Regulates EIF2AK2-Mediated Stress Response and Cell Cycle Progression. *Mol Cell* **69**, 622–635 e626 (2018).
<https://doi.org/10.1016/j.molcel.2018.01.011>
- 48 Milacic, M. *et al.* The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res* **52**, D672–D678 (2024). <https://doi.org/10.1093/nar/gkad1025>
- 49 Wang, Y. *et al.* rMATS-turbo: an efficient and flexible computational tool for alternative splicing analysis of large-scale RNA-seq data. *Nat Protoc* **19**, 1083–1104 (2024).
<https://doi.org/10.1038/s41596-023-00944-2>
- 50 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26 (2011).
<https://doi.org/10.1038/nbt.1754>
- 51 Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012). <https://doi.org/10.1093/bioinformatics/bts356>
- 52 Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res* **43**, W39–49 (2015). <https://doi.org/10.1093/nar/gkv416>

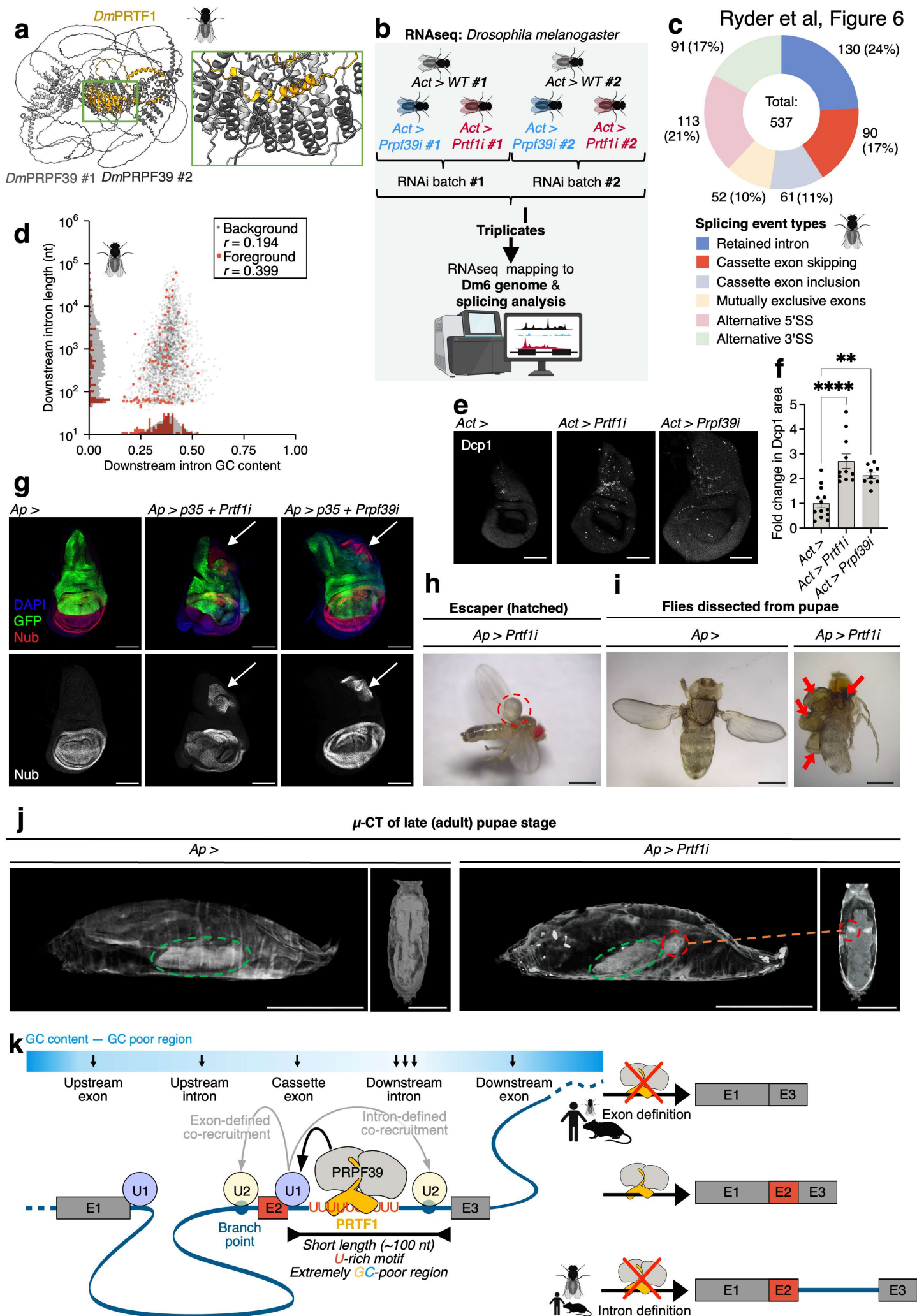
- 53 Grant, C. E. & Bailey, T. L. XSTREME: Comprehensive motif analysis of biological sequence datasets. *bioRxiv*, 2021.2009.2002.458722 (2021).
<https://doi.org/10.1101/2021.09.02.458722>
- 54 McGurk, M. P., McWatters, D. C. & Burge, C. B. KATMAP infers splicing factor activity and regulatory targets from knockdown data. *Nat Biotechnol* (2025).
<https://doi.org/10.1038/s41587-025-02881-9>
- 55 Boyle, E. A. *et al.* Skipper analysis of eCLIP datasets enables sensitive detection of constrained translation factor binding sites. *Cell Genom* **3**, 100317 (2023).
<https://doi.org/10.1016/j.xgen.2023.100317>
- 56 McCue, K. & Burge, C. B. An interpretable model of pre-mRNA splicing for animal and plant genes. *Sci Adv* **10**, eadn1547 (2024). <https://doi.org/10.1126/sciadv.adn1547>
- 57 Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**, 377–394 (2004).
<https://doi.org/10.1089/1066527041410418>
- 58 Tapial, J. *et al.* An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Res* **27**, 1759–1768 (2017). <https://doi.org/10.1101/gr.220962.117>
- 59 Calleja, M., Moreno, E., Pelaz, S. & Morata, G. Visualization of gene expression in living adult *Drosophila*. *Science* **274**, 252–255 (1996).
<https://doi.org/10.1126/science.274.5285.252>











Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementaryfiguresandlegends.pdf](#)