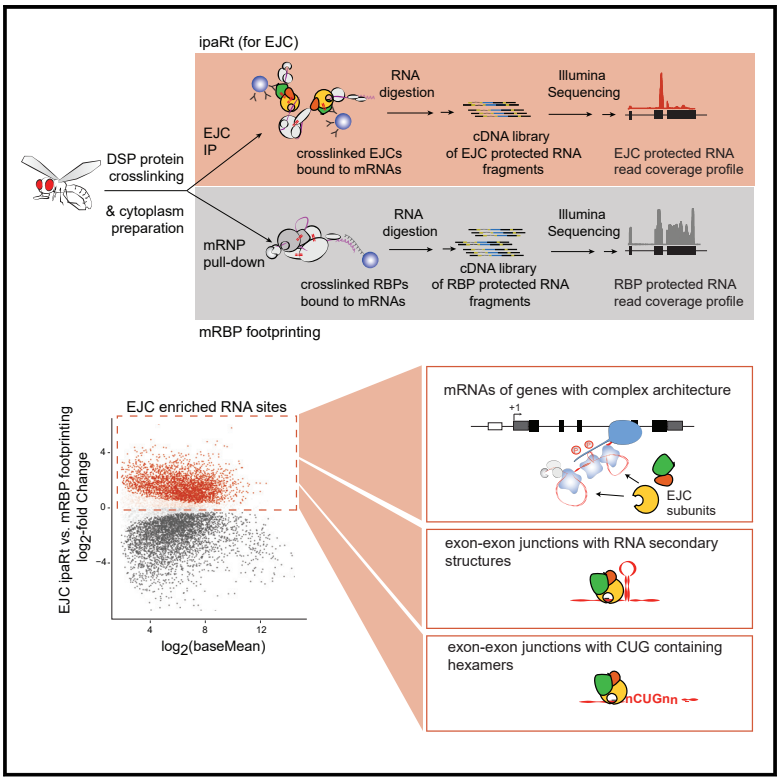
Article

Cell Reports

The Transcriptome-wide Landscape and Modalities of EJC Binding in Adult *Drosophila*

Graphical Abstract



Highlights

- ipaRt: a method for protein-RNA-binding landscape definition in cells and organisms
- *Drosophila* EJC-bound mRNAs are biased toward differentiation and development
- *Drosophila* EJC assembly on mRNAs favors genes with complex gene architecture
- Splice site strength, RNA structure, and CG-rich hexamers enhance EJC binding in flies

Authors

Ales Obrdlik, Gen Lin, Nejc Haberman, Jernej Ule, Anne Ephrussi

Correspondence

obrdlik@embl.de (A.O.), ephrussi@embl.de (A.E.)

In Brief

Obrdlik et al. present ipaRt, an approach for definition of the EJC-RNA-binding landscape in adult *Drosophila melanogaster*. Their study uncovers the impact of gene architecture, splice site strength, RNA structures, and CUG hexamers on EJC binding and provides insights into the evolution of EJC functions.

Obrdlik et al., 2019, Cell Reports *28*, 1219–1236
July 30, 2019
https://doi.org/10.1016/j.celrep.2019.06.088

CellPress

Cell Reports
Article

The Transcriptome-wide Landscape and Modalities of EJC Binding in Adult *Drosophila*

Ales Obrdlik,[1,4,*] Gen Lin,[1,4] Nejc Haberman,[2] Jernej Ule,[2,3] and Anne Ephrussi[1,5,*]
[1]European Molecular Biology Laboratory, 69117 Heidelberg, Germany
[2]Department for Neuromuscular Diseases, UCL Institute of Neurology, London WC1N 3BG, UK
[3]The Francis Crick Institute, London NW1 1AT, UK
[4]These authors contributed equally
[5]Lead Contact
*Correspondence: obrdlik@embl.de (A.O.), ephrussi@embl.de (A.E.)
https://doi.org/10.1016/j.celrep.2019.06.088

SUMMARY

Exon junction complex (EJC) assembles after splicing at specific positions upstream of exon-exon junctions in mRNAs of all higher eukaryotes, affecting major regulatory events. In mammalian cell cytoplasm, EJC is essential for efficient RNA surveillance, while in *Drosophila*, EJC is essential for localization of *oskar* mRNA. Here we developed a method for isolation of protein complexes and associated RNA targets (ipaRt) to explore the EJC RNA-binding landscape in a transcriptome-wide manner in adult *Drosophila*. We find the EJC at canonical positions, preferably on mRNAs from genes comprising multiple splice sites and long introns. Moreover, EJC occupancy is highest at junctions adjacent to strong splice sites, CG-rich hexamers, and RNA structures. Highly occupied mRNAs tend to be maternally localized and derive from genes involved in differentiation or development. These modalities, which have not been reported in mammals, specify EJC assembly on a biologically coherent set of transcripts in *Drosophila*.

INTRODUCTION

The exon junction complex (EJC) consists of a heterotetramer core composed of eIF4AIII, Mago, Y14, and Barentsz (Btz) (Bono et al., 2006; Stroupe et al., 2006) and auxiliary factors that form the EJC periphery (Tange et al., 2005). The complex assembles on mRNAs during splicing, −20 to −24 nt upstream of exon-exon junctions (Le Hir et al., 2000). EJC assembly is a multi-step process that begins with CWC22-mediated deposition of the DEAD-box helicase eIF4AIII on nascent pre-mRNAs (Alexandrov et al., 2012; Barbosa et al., 2012; Steckelberg et al., 2015) and is followed by recruitment of Mago and Y14, forming a pre-EJC intermediate. The pre-EJC is stably bound to RNA because of the ATPase-inhibiting activity of the (non-RNA-binding) Mago-Y14 heterodimer, which "locks" eIF4AIII helicase in its RNA-bound state (Andersen et al., 2006; Ballut et al., 2005; Bono et al., 2006; Stroupe et al., 2006). Once formed, the pre-EJC is completed by recruitment of Barentsz (Btz), forming mature EJCs (Bono et al., 2006; Bono and Gehring, 2011; Tange et al., 2005). The roles of the EJC in post-transcriptional control of gene expression are manifold. In the nucleus, EJC subunits have a role in splicing (Ashton-Beaucage and Therrien, 2011; Ashton-Beaucage et al., 2010; Roignant and Treisman, 2010), mRNA export (Gatfield et al., 2001), and nuclear retention of intron-containing RNAs (Shiimori et al., 2013). In the cytoplasm, the EJC is reported to play a role in translation (Chazal et al., 2013; Nott et al., 2004), nonsense-mediated decay (NMD) (Buchwald et al., 2010; Gehring et al., 2005; Melero et al., 2012; Okada-Katsuhata et al., 2012; Palacios et al., 2004; Shibuya et al., 2006; Singh et al., 2007), and RNA localization (Ghosh et al., 2012; Hachet and Ephrussi, 2001, 2004; Palacios et al., 2004; van Eeden et al., 2001). Although most EJC functions appear conserved, in *Drosophila* the EJC is not crucial for NMD (Behm-Ansmant et al., 2007), but it is essential for *oskar* mRNA localization within the developing oocyte (Ghosh et al., 2012, 2014; Hachet and Ephrussi, 2001, 2004; Palacios et al., 2004; van Eeden et al., 2001; Zimyanin et al., 2008). To better understand the engagement of the EJC in the fly, we developed a strategy to stabilize mRNA binding proteins (mRBPs) associated with their RNA templates within multi-protein messenger ribonucleoprotein (mRNP) assemblies and set out to define the EJC mRNA interactome in adult *Drosophila melanogaster*. Through the use of the crosslinking agent dithio(bis-) succinimidylpropionate (DSP), our method captures stable and transient protein interactions in close proximity (Lomant and Fairbanks, 1976; Schweizer et al., 1982) and allows definition of the binding sites of specific protein (holo-)complexes associated with their RNA templates (isolation of protein complexes and associated RNA targets [ipaRt]). Our analysis of EJC-protected sites defined by ipaRt reveals that in *Drosophila*, EJC binding occurs at canonical deposition sites (Le Hir et al., 2000), with a median coordinate ~22 nt upstream of exon-exon junctions. Although in mammals EJC-mediated protection outside canonical sides was reported (Saulière et al., 2012; Singh et al., 2012), we find that in *Drosophila* the degree of non-canonical EJC-mediated RNA protection is minimal. We show in *Drosophila* that RNA polymerase II transcripts protected primarily by the EJC derive from genes involved in differentiation or development, while mRNAs protected primarily by mRBPs derive from genes with homeostatic functions. Our analysis suggests that the EJC's bias for transcripts in *Drosophila* is a consequence of several modalities in the genes' architecture, particularly splice site number and intron

length. Moreover, EJC binding is enhanced by adjacent RNA secondary structures and CUG-rich hexamers located 3′ to the EJC binding site. These modalities were not identified in previous studies of mammalian EJC binding (Hauer et al., 2016; Saulière et al., 2012; Singh et al., 2012), reflecting either greater specificity of our method for fully assembled EJCs or differences in EJC binding between flies and human. Our study provides a comprehensive transcriptome-wide view of EJC-RNA interactions in a whole organism and unravels RNA modalities that contribute to the unforeseen biological coherence of the bound transcripts.

RESULTS

Stabilization of the Exon Junction Complex on mRNAs by DSP

The EJC is maintained in its RNA-bound state through direct interaction of the Mago-Y14 heterodimer with the otherwise dynamically binding RNA helicase eIF4AIII (Andersen et al., 2006; Bono et al., 2006; Nielsen et al., 2009; Shibuya et al., 2006; Stroupe et al., 2006; Tange et al., 2005). EJC binding to RNA is labile, as under stringent washing conditions (~1 M salt concentrations), interaction of eIF4AIII and Mago-Y14 is abolished and the RNA is released from the complex (Singh et al., 2012). We therefore hypothesized that introducing covalent bonds between the Mago-Y14 heterodimer and eIF4AIII might stabilize the EJC on its RNA targets and render the protein-RNA complex resistant to the high salt concentrations commonly used in iCLIP (individual-nucleotide-resolution crosslinking and immunoprecipitation) studies. Furthermore, a stabilized EJC complex would enable us to "pull" on EJC subunits other than the RNA-binding eIF4AIII, ensuring isolation of the complex under stringent conditions. To test this we made use of the bivalent crosslinking agent dithio(bis-) succinimidylpropionate (DSP), which reversibly crosslinks primary amino groups of polypeptides in close proximity (Lomant and Fairbanks, 1976; Schweizer et al., 1982). We isolated poly(A)-containing mRNPs on an oligo $d(T)_{25}$ resin (Castello et al., 2012, 2013) from cytoplasm of adult *Drosophila* either untreated or treated with UV, DSP, or UV plus DSP (Figure 1). SDS-PAGE silver staining and western blot analyses of mRNA-RNP precipitates (Figures 1A and 1B) revealed that irradiation of cytoplasmic lysates by UV *ex vivo* only marginally increased co-precipitation of proteins with poly(A)-containing RNAs (Figure 1A, lanes 7 and 8). Upon UV irradiation, only faint signals of known RBPs, such as eIF4AIII and cytoplasmic poly(A) binding protein (PABP), were detected in the poly(A) RNA precipitates. Non-RNA-binding EJC subunits such as Y14 were not detected (Figure 1B, lanes 7 and 8, and Figure 1C), in agreement with previous observations (Castello et al., 2012). In contrast, treatment of cytoplasmic lysates with DSP led to strong protein co-precipitation with mRNAs (Figure 1A, compare lanes 7–10). Western blot analysis of precipitates from DSP- and UV-DSP-treated cytoplasmic lysates revealed strong signals not only for direct mRNA binding proteins such as eIF4AIII and PABP but also mRNP components not directly bound to RNA, such as Y14 (compare Figures 1A and 1B, lanes 7–10, and Figure 1C). In none of the precipitates were cytoplasmic proteins such as kinesin heavy chain (Khc) or the small ribosomal subunit protein RpS6 observed (Figure 1B, lanes 7–10, and Figure 1C), confirming the stringency of the assay. Furthermore, precipitates from the beads-only control were free of all proteins tested (Figure 1A, lane 6), indicating that artifacts due to DSP or UV treatment are unlikely. These observations suggest that for EJC stabilization on RNA, DSP-mediated covalent bond formation between individual EJC subunits is superior to UV crosslinking and support the use of DSP when studying other mRNP assemblies (Figure 1D).

ipaRt: An Approach for High-Quality Isolation of EJC Complexes Associated with RNA Templates

Of the tagged EJC subunits we tested, GFP-Mago showed the highest degree of incorporation into endogenous EJCs (Figure S1B). The eIF4AIII subunit was additionally found to co-sediment with polysome fractions in sucrose density gradients, independently of Mago-Y14 (Figure S1C), suggesting that the DEAD-box helicase might have yet unknown EJC-independent functions in the fly. Therefore, we carried out EJC-specific RNA immunoprecipitation (RIP) from DSP-treated cytoplasmic extracts prepared from GFP-Mago- and GFP tag-expressing flies. By titrating salt and detergent concentrations, we identified stringent washing conditions (see STAR Methods) that yielded high-quality RNA profiles from GFP-Mago RIPs and only scant RNA profiles from GFP control RIPs, compared with standard IP washing conditions (Figure 2A, compare lanes 2–5). To test if the presence of RNA in GFP-Mago precipitates was a consequence of its incorporation into the EJC rather than by virtue of transient interactions of Mago with other RBPs, we subjected immunoprecipitates from DSP-treated or untreated lysates to RNaseI digestion. Western analysis revealed the presence of all tested EJC subunits in the GFP-Mago precipitates (Figure 2B, lanes 7 and 8) but no protein other than GFP itself in the GFP only controls (Figure 2B, lanes 3 and 4). In contrast, we detected no signal for the proteins probed in the beads-only control precipitates (Figure 2B, lanes 2 and 6) and observed signals for the RNA non-binding Khc only in lysate inputs (Figure 2B, compare lanes 1 and 5, 2–4, and 6–8), indicating high stringency of the assay.

The stabilizing effect of DSP on the EJC is evident from the enhanced eIF4AIII signals in GFP-Mago precipitates when cytoplasm was treated with DSP prior to immunoprecipitation (Figure 2B, lanes 5, 7, and 8). Conversely, the PABP signal in the GFP-Mago precipitates disappeared upon incubation with RNase of both DSP-treated and the untreated samples (Figure 2B, lanes 1, 5 2–4, and 6–8). This shows that even when exposed to DSP, proteins whose associations with the EJC are bridged by RNA can be removed by RNA fragmentation.

To confirm the "cleansing effect" of RNaseI, we performed IPs from DSP-treated cytoplasm with or without an RNA fragmentation step and analyzed the precipitates using tandem mass spectrometry (MS). Expression set analysis of MS signals obtained in GFP-Mago and GFP control precipitates identified 45 versus 35 significantly enriched proteins in the untreated and RNase-treated samples, respectively (Figure 2C; Figure S2; Table S4). Although all EJC subunits were enriched independently of RNA integrity (Figure 2C), only upon RNA fragmentation was Btz enriched to a similar degree as Mago, Y14, and eIF4AIII. Except for the poly(A) binding protein Nab2 (Bienkowski et al.,

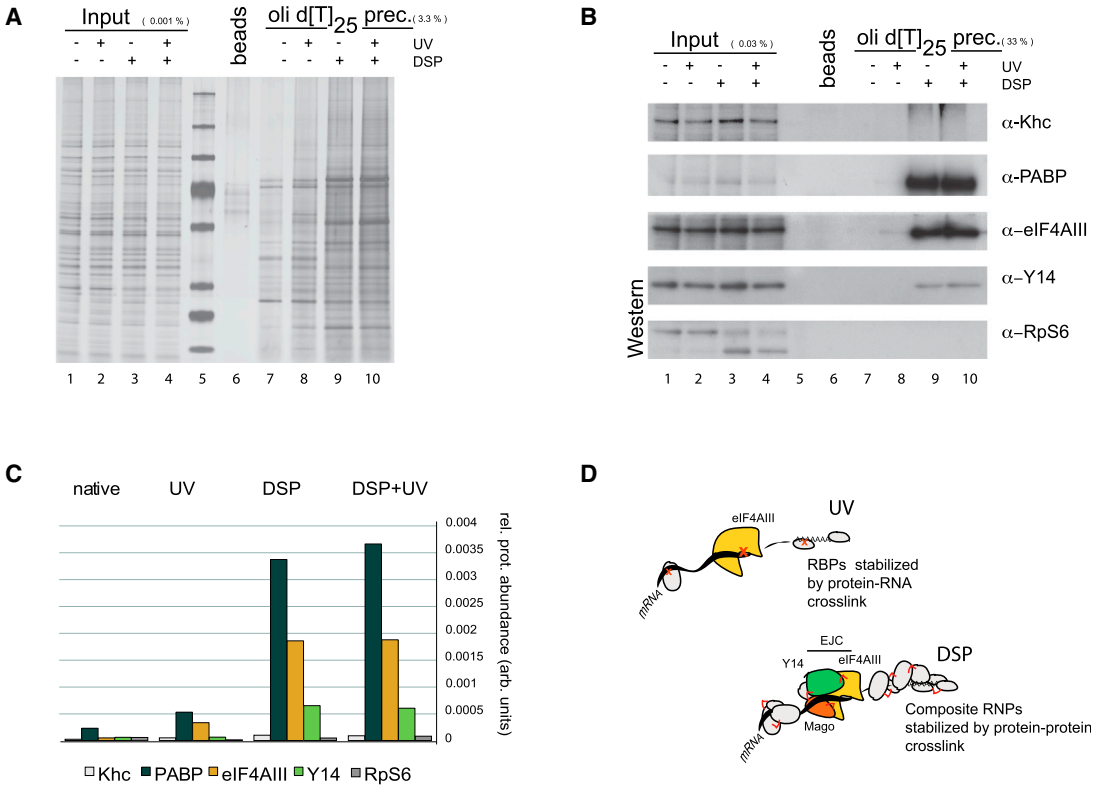


Figure 1. Dithio(bis-)Succinimidyl-Propionate (DSP) Stabilizes EJC in Its mRNA-Bound State

(A) SDS-PAGE and silver staining of proteins co-precipitated with oligo-d(T)$_{25}$ bound mRNAs from cytoplasm. Left to right: input cytoplasmic samples (0.01% of total input; lanes 1–4); protein MW standards (lane 5); beads only control precipitate from all-condition mixture (3.3%; lane 6); and individual oligo-d(T)$_{25}$ precipitates (3.3%; lanes 7–10). Treatment conditions untreated, UV irradiated, DSP supplemented, and UV irradiated-DSP supplemented are indicated (+ or −) above the image. Note that for samples treated with DSP *ex vivo*, UV exposure does not increase the amount of recovered proteins.

(B) Western blot analysis of oligo-d(T)$_{25}$ precipitates. Gel loading order same as in (A); 0.03% of input cytoplasm and 33% of each precipitate were resolved. Blot was probed with antibodies against the proteins indicated at the right of the panel. Khc, kinesin heavy chain; PABP, cytoplasmic poly A binding protein; RpS6, small ribosomal subunit protein S6.

(C) Quantification of proteins detected on the western blot. Crosslinking conditions are indicated in the upper panel. Signals were quantified by densitometry measurement using the Fiji image analysis package. Relative protein abundance is the fraction of the signal in the precipitate relative to the cytoplasm. Note that in contrast to the RNA-binding eIF4AIII, the non-RNA-binding EJC subunit Y14 was detected only when the cytoplasm was treated with DSP.

(D) Schematic of the net effect of crosslinking with UV versus DSP. Exposure of the cytoplasm to UV leads primarily to stabilization of direct protein-RNA interactions. DSP treatment results in efficient retention of proteins associated with RNA by stabilization of polypeptide interactions either within an RBP or between an RBP and other moieties within a complex.

2017), no EJC-unrelated RBPs showed significant enrichment upon RNA fragmentation (Figure 2C), showing that RNA fragmentation by RNaseI increases both sensitivity and specificity of EJC IPs.

We conclude that DSP is a suitable tool for stabilization and isolation of EJC RNA complexes from animal tissues and that our protocol provides a reliable approach for isolation of proteins (or protein complexes) associated with their RNA targets. We termed our experimental strategy "ipaRt."

EJC Binding in *Drosophila* Cytoplasm Maps to Canonical Deposition Sites

To determine which sites in the *Drosophila* transcriptome are protected by the EJC as opposed to other mRNA binding proteins (mRBPs), we performed EJC ipaRt- and oligo(dT)-mediated mRNP capture in parallel, both followed by an RNase digestion step (see STAR Methods), and constructed cDNA libraries of the protein-protected RNA fragments as described for iCLIP (Konig et al., 2011). Analysis of the sequencing results revealed that more than 92% of all reads aligned uniquely to the *Drosophila* genome (Figure S2B). 85% of EJC ipaRt reads mapped to exons, as opposed to 34% in the mRBP footprinting mapped reads, indicating specificity of the ipaRt library (Figure 3A).

To define the median binding coordinates of the protected sites, we determined the sequence coverage ±50 nt of exon-exon junctions in EJC ipaRt and mRNP footprinting (Figure 3B) and averaged the coverage profile over all junctions. The mean coverage profile in mRBP footprinting appeared evenly distributed, indicating an absence of protection bias (Figure 3B). In contrast, the protected sites in EJC ipaRt were located almost exclusively in upstream exons, with an EJC coverage median −21.7 nt 5′ to the exon-exon junction (Figure 3B), consistent

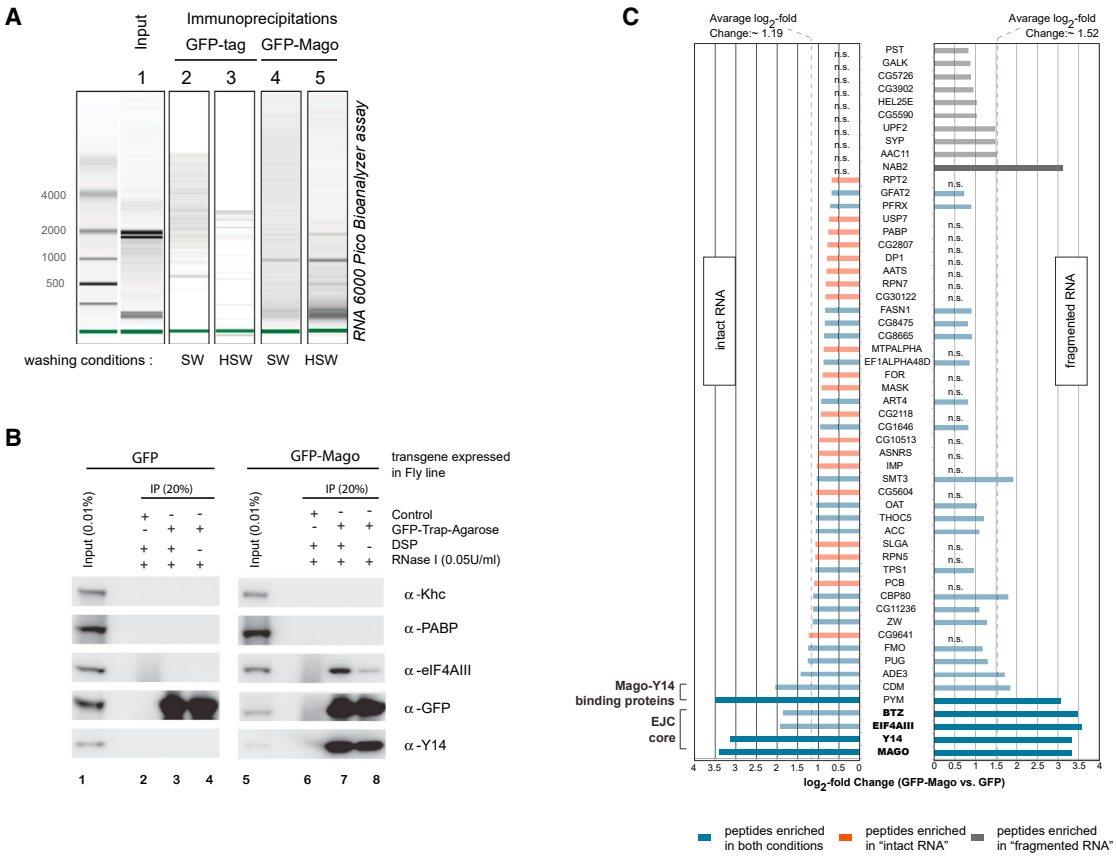


Figure 2. DSP Crosslinking Stabilizes EJC for Stringent Immunoprecipitation and RNA Fragmentation

(A) Comparison of IP washing conditions for RNA isolation. Anti-GFP RNA IP from DSP-treated cytoplasm of GFP-Mago- and GFP tag-expressing flies. 0.02% of RNA isolated from input lysate (lane 1) and 20% from GFP tag (lanes 2 and 3) and GFP-Mago (lanes 4 and 5) RIP precipitates were resolved by capillary gel electrophoresis on an RNA 6000 Pico Chip Bioanalyzer. Washing conditions are indicated at bottom. SW, standard-stringency washing conditions; HSW, high-stringency washing conditions (see STAR Methods).

(B) DSP stabilized EJC core is resistant to RNaseI treatment. Effects of RNaseI treatment on GFP-Mago co-immunoprecipitations in DSP crosslinked and untreated cytoplasm. Western blots of anti-GFP IPs from DSP-treated and from untreated cytoplasm of GFP-Mago (lanes 5–8) and GFP tag (lanes 1–4) expressing flies processed under HSW conditions. Primary antibodies used are indicated on the right. Inputs (0.01%) are shown in lanes 1 and 5. Bead-only control precipitates are shown in lanes 2 and 6. IP precipitates (20%) from DSP-treated (lanes 3 and 7) or untreated cytoplasm (lanes 4 and 8). Details of samples and experimental conditions indicated at top of panel.

(C) RNA fragmentation depletes EJC-unrelated proteins from GFP-Mago co-precipitates. Anti-GFP IPs (HSW condition) from DSP-treated cytoplasm of GFP-Mago- and GFP tag-expressing flies. Precipitates were subjected to isobar labeling, and peptide content was defined by tandem mass spectrometry. Bar plots of protein enrichment are defined by Limma (Ritchie et al., 2015). GFP-Mago-specific protein enrichments in intact RNA and fragmented RNA conditions are highlighted in left and right plots, respectively. The y axis shows individual proteins detected. The x axis shows scale of enrichment (log_2 fold change). Dashed line indicates average enrichment of all significant proteins in each condition. Protein enrichment > 2 times or <2 times average enrichment is indicated by solid or transparent bars, respectively. Enrichment bar color legend is highlighted at the bottom. ns, non-significant (adjusted p value [p.adj.] > 0.05).

with previous studies (Gatfield et al., 2001; Kataoka et al., 2001; Le Hir et al., 2000). Similarly, we estimated a 13-nt-long region of saturated RNA protection in the EJC RNA footprints, from −27 to −15 nt 5′ of the exon-exon junction (Figure S2D) (Ballut et al., 2005; Le Hir et al., 2000; Singh et al., 2012).

Protection Sites Remote of Exon-Exon Junctions in EJC ipaRt Are Not of EJC Origin

Studies of mammalian EJCs have reported a high frequency of EJC-mediated protection outside of canonical binding sites (non-canonical EJC deposition sites) (Saulière et al., 2012; Singh et al., 2012). To test whether this non-canonical distribution is representative of EJC protection across the *Drosophila* transcriptome, we determined coverage maxima for every exon-exon junction protected by EJC. The majority (~95.5%) of protection peaks in EJC ipaRt libraries mapped to sites of canonical EJC binding, proximal to exon-exon junctions (Figures 3C and 3D). The remaining ipaRt protection coverage peaks were located remotely (>50 nt) of canonical EJC binding regions (Figures 3D, 3G, and S4A).

Our analysis shows that proximal peaks map mainly to internal exons (79%), to first exons (20.6%), and only minimally to terminal exons (0.4%), as expected given the splicing-dependent deposition of EJCs upstream of splice junctions. Remote

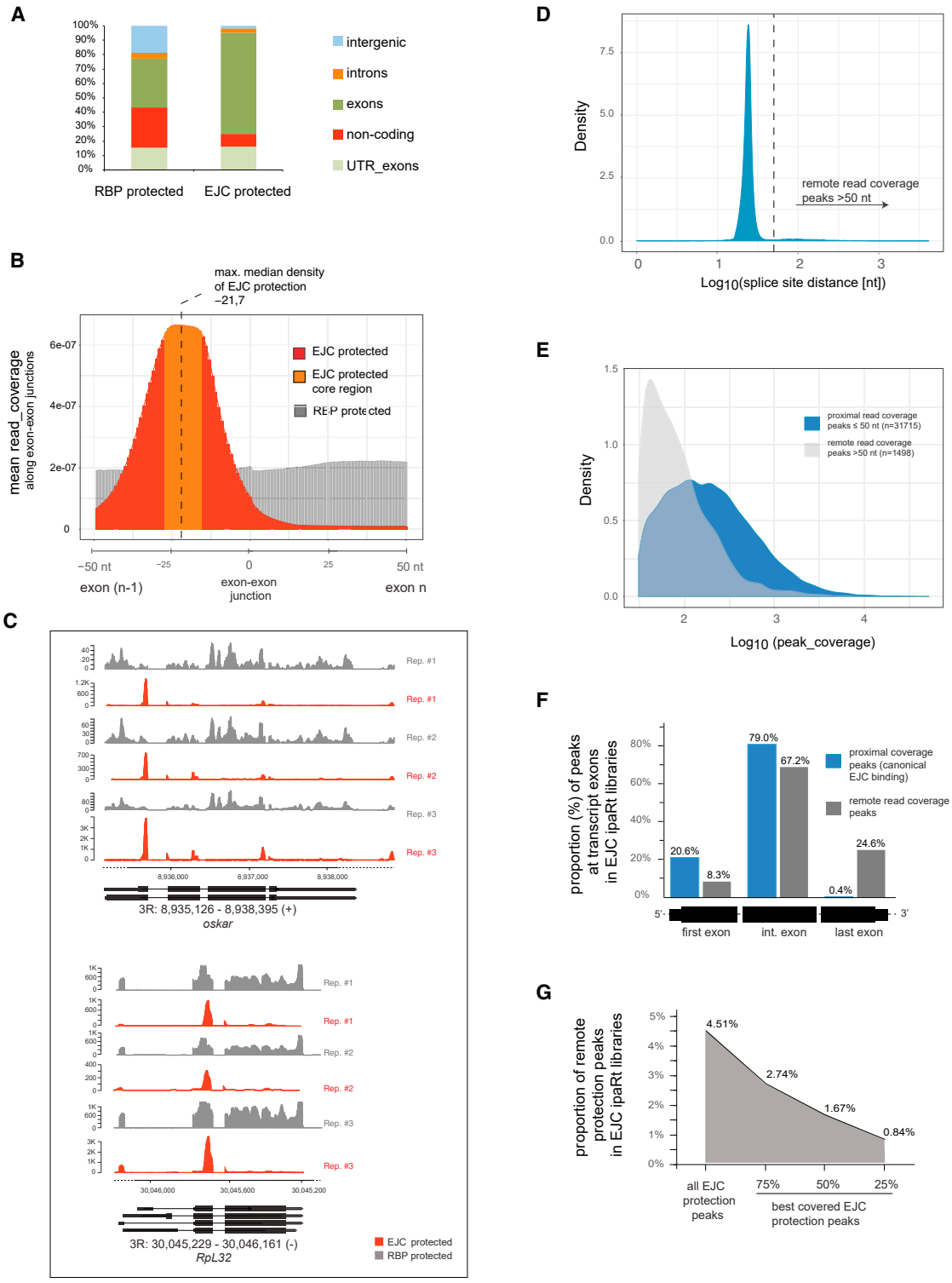


Figure 3. EJC Binding to mRNA in *Drosophila* Cytoplasm Occurs within Exons at Canonical Deposition Sites

(A) EJC ipaRt library reads map to exonic sites in the *Drosophila* genome. Summary of genomic features detected in mRBP footprinting and EJC ipaRt sequencing results. The y axis indicates the proportion (percentage) of uniquely aligning read counts. Color code of genomic features highlighted in the legend on right side of the plot. Note that EJC-protected sites map in majority to exons and UTR exons.

(B) EJC protection median is on upstream exons approximately −22 nt from the 3′ end. Read coverage profiles from EJC ipaRt and mRBP footprinting cDNA libraries. Coordinates of metagene covering +50 nt of exon-exon junctions are indicated on x axis. Note that position 0 defines the last nucleotide of upstream

(legend continued on next page)

protection site peaks mostly mapped to internal exons (67%) and last exons (25%) and only minimally (8%) to first exons of the bound mRNAs (Figure 3F).

Three main features characterize the remote peaks in our EJC ipaRt libraries: low abundance, lower sequencing coverage (Figures 3E and 3G), and relatively low expression compared with proximal peaks (Figures S3C and S3D). Further analysis using Analysis of Motif Enrichment (AME) (Bailey et al., 2009) revealed that the remote peaks in EJC ipaRt are significantly enriched in RNA binding motifs corresponding to known *Drosophila* splicing regulators (Figure S3E), whose binding might be a consequence of DSP crosslinking and co-purification due to direct interaction with the EJC. Our analysis shows that in contrast to EJC in mammals (Saulière et al., 2012; Singh et al., 2012), the proportion of remote EJC binding sites is negligible in *Drosophila*.

EJCs Mark Multi-intron Genes Important for Differentiation and Development

We determined which RNAs are bound preferentially by EJC versus other RBPs using DESeq (Love et al., 2014). This revealed a bias in EJC binding toward mRNAs of protein coding genes comprising greater than 1 exon (Figure S4D) and identified 3,332 enriched and 4,436 depleted genes (false discovery rate [FDR] < 0.05 and $\log_2$ fold change > $\log_2$[1.5]). Gene Ontology (GO) term analysis of EJC-enriched genes (Figure 4A) revealed significant association with genes involved in development or specialized cellular functions such as cell polarity, differentiation, and cell signaling. In contrast, genes with homeostatic functions, involved in cytoskeletal and chromatin organization, in transcription or translation, and metabolic processes are biased for protection by other RBPs (Figure 4B, compare left and right plots). The bias of EJC binding to mRNAs from genes with functions in development and cell polarity suggested that transcripts under spatial or temporal control might also show a preference for EJC binding. Consistent with this hypothesis, analysis of EJC and RBP protection sites on *Drosophila* transcripts annotated in the FlyFISH RNA localization database (Lécuyer et al., 2007; Wilk et al., 2016) revealed that localized maternal mRNAs are more likely to be EJC bound than non-localizing transcripts (Figure 4C).

Gene Architecture Determines EJC Assembly on mRNAs

It was reported that in mammalian cells, EJCs are enriched on mRNAs from alternatively spliced genes (Hauer et al., 2016). In the fly, the EJC has been reported to promote correct splicing of long intron-containing genes (Ashton-Beaucage et al., 2010; Roignant and Treisman, 2010). This relationship between EJCs and gene architecture led us to ask which features could explain the gene-to-gene variation in EJC deposition. We used a multiple regression model (see STAR Methods) to assess how five features (number of introns, maximum intron length, transcript abundance, transcript length, and the degree of alternative splicing) influence gene deposition of EJC. We checked that the effects estimated from the model hold given underlying correlations of the features. For example, after accounting for intron number, which has a strong effect on EJC binding (Figure S5A), we determined that intron length also has a strong positive effect (Figure S5B), while alternative splicing has only a minimal effect on EJC binding (Figure S5C).

It is noteworthy that exon-exon junctions of transcripts from genes comprising at least one large intron (≥10,00 bp) were significantly more enriched than those of genes lacking large introns (Figure S5F). Within transcripts of long intron-containing genes, EJC assembly was biased neither toward junctions formed upon large-intron splicing nor toward neighboring junctions (Figure S5G). Instead, exon-exon junctions within these transcripts showed a general elevation in EJC binding

exons. Scale of mean sequencing read coverage for all normalized biological replicates is indicated on the y axis. Profiles from RBP-protected and EJC-protected junctions are presented in gray and red, respectively. Note that mRBP footprinting sequencing reads show a homogeneous RBP protection profile over the entire metagene body. On the contrary, sequencing reads from EJC ipaRt indicate a modal distribution of EJC protections with a median at −21.5 nt upstream of the exon-exon ligation point. Region of maximal EJC protection (protection core) spans from −27 to −15 nt and is highlighted in orange.

(C) Genomic coverage profile visualized by IGB viewer. Coverage profiles of three individual sequencing replicates from EJC ipaRt and mRNP footprinting along the *oskar* and *RpL32* genes. Size and position of gene regions are indicated. Exons and introns are indicated below the coverage profiles as boxes or lines, respectively. Coverage depth (normalized reads) indicated on the y axis. Note that cDNA reads from EJC-protected fragments map upstream of and in direct proximity to splice sites, while reads from RBP-protected fragments distribute over whole exon bodies. Sequencing coverage depth of EJC-protected sites at exon-exon junctions is variable.

(D) Protection site peaks (modes) in EJC ipaRt libraries cluster primarily within 50 nt of splice junctions. Density plot highlighting distribution of coverage peaks in EJC ipaRt with respect to the splice site distance. The y axis defines estimated density of peaks at defined splice site distances. The x axis indicates $\log_{10}$-transformed distance to splice site. Vertical dashed line highlights border between proximal and remote protection peak estimates. Note all peaks in ipaRt were defined within exons at a 20 nt frame resolution. Peaks chosen for the analysis were 2 times more covered in ipaRt than in mRBP footprinting and had coverage >30 reads.

(E) Proximal (canonical) protection peaks in EJC ipaRt libraries are stronger covered than remote protection peaks. Density plot showing the distribution of estimated coverage of splice junctions proximal (blue) and remote (gray) protection peaks. The y axis defines estimated densities and x axis the peak coverage. Note that peak coverage stands for the sum of all reads within a ±10 nt window surrounding a protection site peak.

(F) Peaks within canonical EJC binding sites map to first and internal exons. Bar plot showing proportion of peaks mapping to first, internal, and last exons. Peaks mapping within or in direct proximity to canonical EJC binding regions (proximal peaks) highlighted in dark blue. Peaks remote from EJC binding regions highlighted in dark gray. Estimated peak proportions within exon classes are indicated in the plot. The y axis indicates proportion as a percentage; the x axis indicates the tested exon classes of a metagene. Note that proximal peaks are nearly exclusively found in first and internal exons, while remote peaks are detected in all three classes of exons.

(G) Proportion of remote protection peaks in EJC ipaRt libraries decreases with sequencing coverage. Plot highlighting the proportion of remote peaks among all EJC ipaRt sequencing coverage peaks and among sequencing coverage subsets of the best 75%, best 50%, and best 25% covered protection peaks. The y axis defines proportion (as a percentage) of remote peaks. The x axis highlights peak coverage subsets. Note that the proportion of EJC protection peaks reduces significantly when increased coverage cutoffs are chosen.

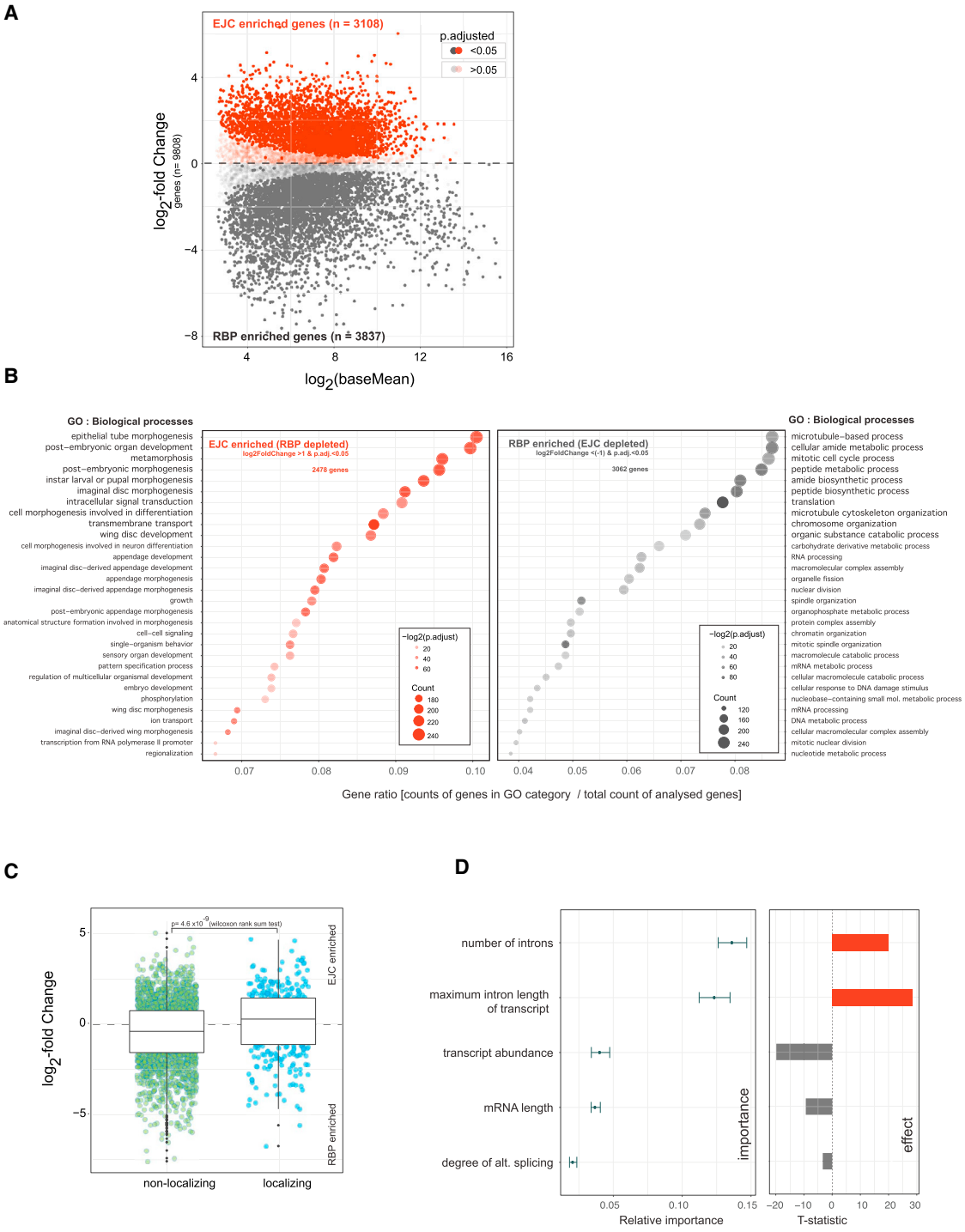


Figure 4. Preferential Recruitment of EJC to mRNAs of Genes with Specialized Cellular Functions Is Determined by Gene Architecture

(A) MA plot of DESeq results from EJC ipaRt and mRBP footprinting. Genes that are either enriched or depleted for EJC (RBP enriched) are indicated in red or gray, respectively. Genes not significantly different (p.adj. > 0.05) between the EJC ipaRt and mRBP libraries are transparent. Relative enrichment (log_2-fold change) is indicated on the y axis. Base mean of signal is highlighted on the x axis. Dashed line defines log_2 fold change = 0.

(B) GO term analysis of EJC-enriched and EJC-depleted transcripts $|log_2$ fold change$| > 1$. GO terms for biological processes of EJC enriched transcripts (left plot) and of EJC depleted transcripts (right plot) are presented to the left and the right of the plots, respectively. The y axis list GO categories of the biological processes most highly represented. Gene counts in individual categories versus overall count of analyzed genes (gene ratios) are shown on the x axis. Legend indicating

(legend continued on next page)

probability, pointing to a "global" large intron-mediated effect on EJC assembly (Figure S5G).

Finally, we assessed the relative importance of each of the five features from the multivariate regression analysis (Figure 4D). Two features dominated our model of EJC deposition, the number of introns per gene and the maximum intron length of the gene (Figure 4D), are positively associated with EJC deposition. This indicates that a gene's architecture is a main determinant of EJC binding.

Splice Site Strength and Hexamer Composition Influence EJC Deposition

We estimated EJC enrichment at the junction level using reads within ±50 nt of the splice site and observed a strong dependency on the gene EJC estimate (Pearson correlation = 0.66, $p < 2e\text{-}16$). This suggests that the junction's EJC profile is primarily determined by its parent gene architecture. We next tested whether any exon-exon junction deviates significantly from its parent gene EJC binding. About 31% of detected junctions have an enrichment that deviates significantly from the gene level (Figure 5A). Furthermore, we observed that EJC reads cover only a subset of junctions within a gene and show higher read coverage coefficient of variation than reads protected by other RBPs (Figures S5D and S6E).

We asked if a known sequence context related to splicing might be responsible for the variability in EJC binding to exon-exon junctions within a gene. We found that strong 5′ and 3′ splice site signals and the presence of 5′ intronic splicing enhancers (5′ISEs) correlate with increased EJC deposition, while the presence of an 5′ exonic splicing silencers (5′ESSs) correlates with reduced EJC deposition (Table S1; Figure 7A), consistent with the fact that EJC assembly is dependent on splicing. Surprisingly, 5′ and 3′ exonic splicing enhancers (ESEs) and intronic ISEs at 3′ splice sites have no effect. Next, we tested the effects of unannotated hexamers, while accounting for ESS, ISE, and splice strength. We detected 63 hexamers associated with a change in EJC deposition. By clustering the hexamers on the basis of sequence similarity, two major groups with either a negative effect or a positive effect on EJC assembly emerged (Figure 5B). Strikingly, 5 of the 16 hexamers associated with increased EJC deposition contain the trinucleotide CUG, and 28 of the 47 hexamers associated with decreased EJC deposition contain the trinucleotide UUU (Figure 5B). The strongest effect of these CUG or UUU trinucleotides is observed when they are present around the region downstream of EJC binding (approximately −16 to −18 nt). This indicates that the sequence composition of this region is a strong determinant of EJC binding (Figure 5C).

RNA Structure Modulates the Degree and Position of EJC Binding in *Drosophila*

Deposition of an EJC at the first exon-exon junction and presence of a structured element next to the deposition site are required for localization of *oskar* mRNA at the posterior pole of the *Drosophila* oocyte (Ghosh et al., 2012; Simon et al., 2015). Interestingly, we observed 2.38-fold stronger enrichment of the first *oskar* exon-exon junction than anticipated from *oskar* mRNA enrichment in our data, indicating that structures near EJC binding sites might affect EJC assembly.

To test if RNA structures might affect EJC binding, we estimated for every junction the probability of base-pairing for each nucleotide −37 to +28 bp of the splice site. We observed three distinct average base-pairing probability (bpp) profiles for exon-exon junctions with unaffected, positively correlated, or negatively correlated EJC binding (Figure 6A). Two regions showed significantly different bpps for junctions with a positive versus a negative effect on EJC binding (Figure 6B). The first region, located in the canonical EJC binding site, showed a decreased bpp for junctions with a positive EJC binding effect (Figures 6A and 6B) and increased bpp for junctions with negative EJC binding effect (Figures 6A and 6B). Surprisingly the second region, located directly downstream of the canonical deposition site (Figures 6A and 6B), showed an elevated bpp near junctions with positive EJC binding but decreased bpp at junctions with negative EJC binding (Figure 6A). This result indicates that although EJC binding in *Drosophila* occurs on single-stranded RNA (ssRNA), in agreement with previous reports (Andersen et al., 2006; Bono et al., 2006), EJC binding to RNA may be enhanced by RNA secondary structures proximal to the EJC binding site.

Given the redundant information between bpp of each nucleotide pair, we performed dimension reduction on bpp profiles within the −24 to −11 region (STAR Methods; Figure 6B) using a Gaussian mixture model, to facilitate subsequent analysis of EJC binding. We obtained four folding categories (Figure 6C) and observed an association between significant positive EJC binding ($\log_2$ fold change > $\log_2[1.5]$) and junctions harboring folding categories 2 and 3, which contain a bpp elevation downstream of, or surrounding, EJC binding sites (Figures 6C and 6D). Junctions with an unstructured profile (folding category 1) show no such bias, and junctions with bpp elevation in the EJC binding site (folding category 4) are associated with a negative EJC

number of genes falling into a particular GO term (size of circles) and its significance of association (color of circles) are (circles) shown on the plots is highlighted in the center.

(C) Scatter-box plot showing DESeq enrichment estimates of *Drosophila* mRNAs with known localization patterns in early embryos. Gene products in DESeq analysis were subset to maternally expressed mRNAs, which localize or do not localize to specific foci in early embryos (Lécuyer et al., 2007). Relative enrichment ($\log_2$ fold change) is indicated on y axis. Localization categories indicated on x axis. Non-localizing and localizing gene products are highlighted in green and blue, respectively. Horizontal solid line in corresponding box plots indicates median $\log_2$ fold change for each category. Dashed line highlights $\log_2$ fold change = 0. Highlighted p value has been estimated by double-sided Wilcoxon rank sum test.

(D) Result of multiple regression model for EJC enrichment highlighting parameters that contribute most to preferential binding of EJC to mRNA. Plots showing the relative contribution and effect of each factor (listed on the y axis) to EJC enrichment, as estimated from the full regression model. For the plot showing relative importance (left), estimates were obtained by bootstrapping the data; the dot indicates the median and whiskers indicate the 95% confidence interval. For the plot showing the effect of each factor (right), the t statistic of each factor was calculated from the estimate and the SE of each coefficient obtained after fitting the full model. The p value has been calculated using hypergeometric test in R package clusterProfiler (Yu et al., 2012).

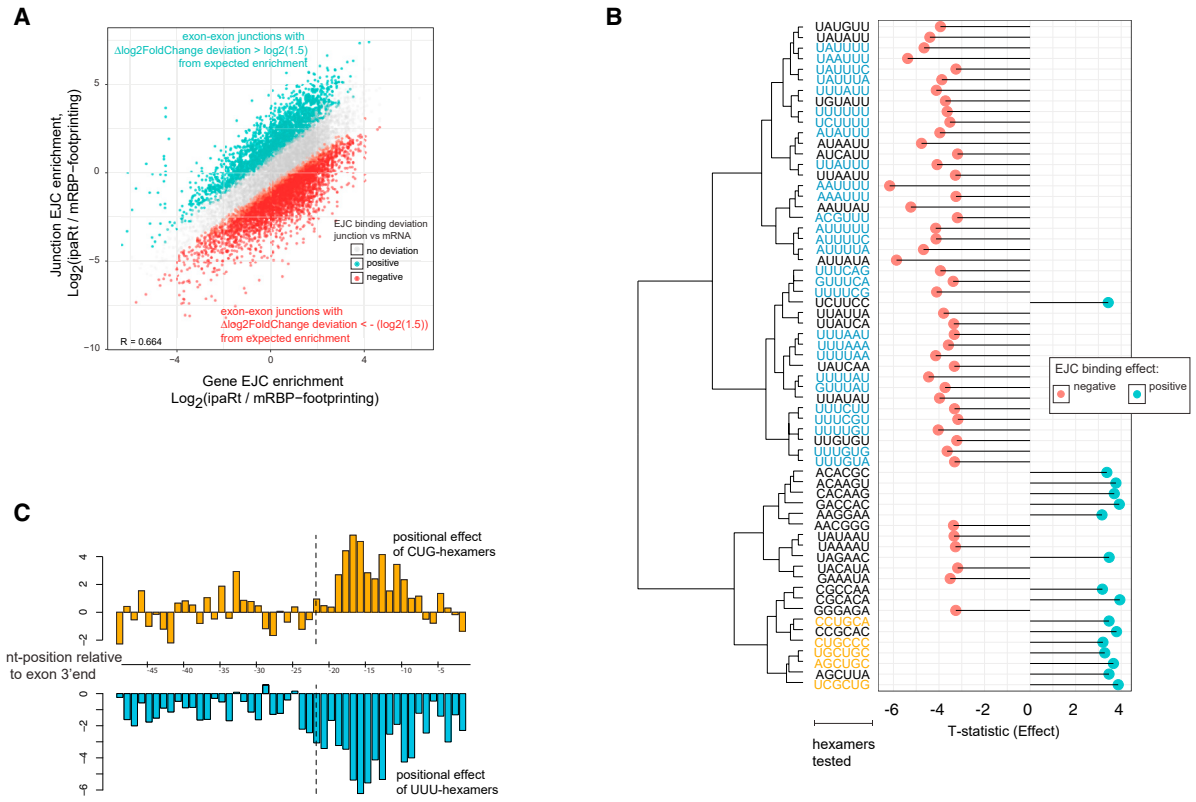


Figure 5. EJC Assembly Reveals a Sequence Bias

(A) EJC enrichment: observed versus expected. Scatterplot showing EJC enrichment defined by DESeq for individual exon-exon junction relative to DESeq estimates of corresponding templates. Enrichment for each exon-exon junction (log_2 fold change exon-exon junction) was calculated from read counts falling within 50 bp upstream and 50 bp downstream of the exon-exon junction. EJC enrichment for the exon-exon junction was compared with its associated gene, and a *Z* score and p value were calculated. Exon-exon junctions whose EJC enrichment was not significantly different from that of its associated gene (p.adj. > 0.05) are shown in gray. Exon-exon junctions with significantly higher EJC enrichment than their associated genes are highlighted in red, and those with significantly lower EJC enrichment are highlighted in blue. The exon-exon junction EJC enrichment is highly correlated with the corresponding gene EJC enrichment (Spearman correlation = 0.664), and twice as many exon-exon junctions show significantly lower EJC enrichment than their corresponding gene product: 4,334 versus 2,275, out of a total of 15,186 junctions tested).

(B) Hexamers containing CUG and UUU contribute most to deviation between observed and expected EJC enrichment. Plot showing the effect of each hexamer on EJC enrichment for each exon-exon junction, after accounting for splice strength and presence of an ESS. For each hexamer, the deviation (of the exon-exon junction EJC enrichment from its gene EJC enrichment) was regressed against (1) the presence of hexamer, (2) splice strength, and (3) ESS counts. The x axis shows the associated t statistic obtained for the hexamer effect after fitting such a model for each individual hexamer. Dendrogram of hexamers was generated using hierarchical clustering with Ward's method, on the basis of pairwise Damerau-Levenshtein distance modeling between all hexamers. CUG-containing hexamers show a concordant positive effect, and UUU-containing hexamers show a concordant negative effect on EJC enrichment.

(C) CUG and UUU effect EJC binding is position biased. Plot showing the positional effect of the tri-nucleotides CUG and UUU. Similar to the analysis performed for hexamers, for each possible position of the tri-nucleotide, the deviation (of the exon-exon junction EJC enrichment from its gene EJC enrichment) was regressed against (1) the presence of CUG and UUU at the position, (2) splice strength, and (3) ESS counts. The y axis shows the associated t statistic obtained for the position effect after fitting such a model for each individual position upstream of the exon-exon junction. The x axis defines the nucleotide coordinates in the upstream exon. Both CUG and UUU show a highly positive or negative t statistic around −18 bp upstream of exon-exon junction, indicating a strong position-specific effect.

binding effect. This suggests that when located near EJC deposition sites, RNA structures may positively affect EJC binding in *Drosophila* (Figures 6C and 6D, lanes 2 and 3) and could explain the enhanced binding of EJC to the first exon-exon junction in *oskar* mRNA.

Previous *in vitro* experiments have shown that RNA secondary structures affect EJC assembly site coordinates (Mishler et al., 2008). We asked whether predicted stable RNA stem structures within or flanking a canonical EJC binding site at exon-exon junctions have any impact on the precise coordinates of EJC assembly in the fly. For this we estimated potential double-stranded RNA (dsRNA) content in region A, spanning from −30 to −21 and for region B spanning from −23 to −14 within exon-exon junctions (see sketch in Figure 6E). We focused on junctions with a positive EJC binding effect to allow robust estimates on the basis of junctions with high coverage. The analysis revealed a strong downstream shift of EJC assembly coordinates when dsRNA content in region A was high and low in region B, and only a minor shift when the dsRNA content was low in region A and high in region B (Figure 6F). Taken together, these

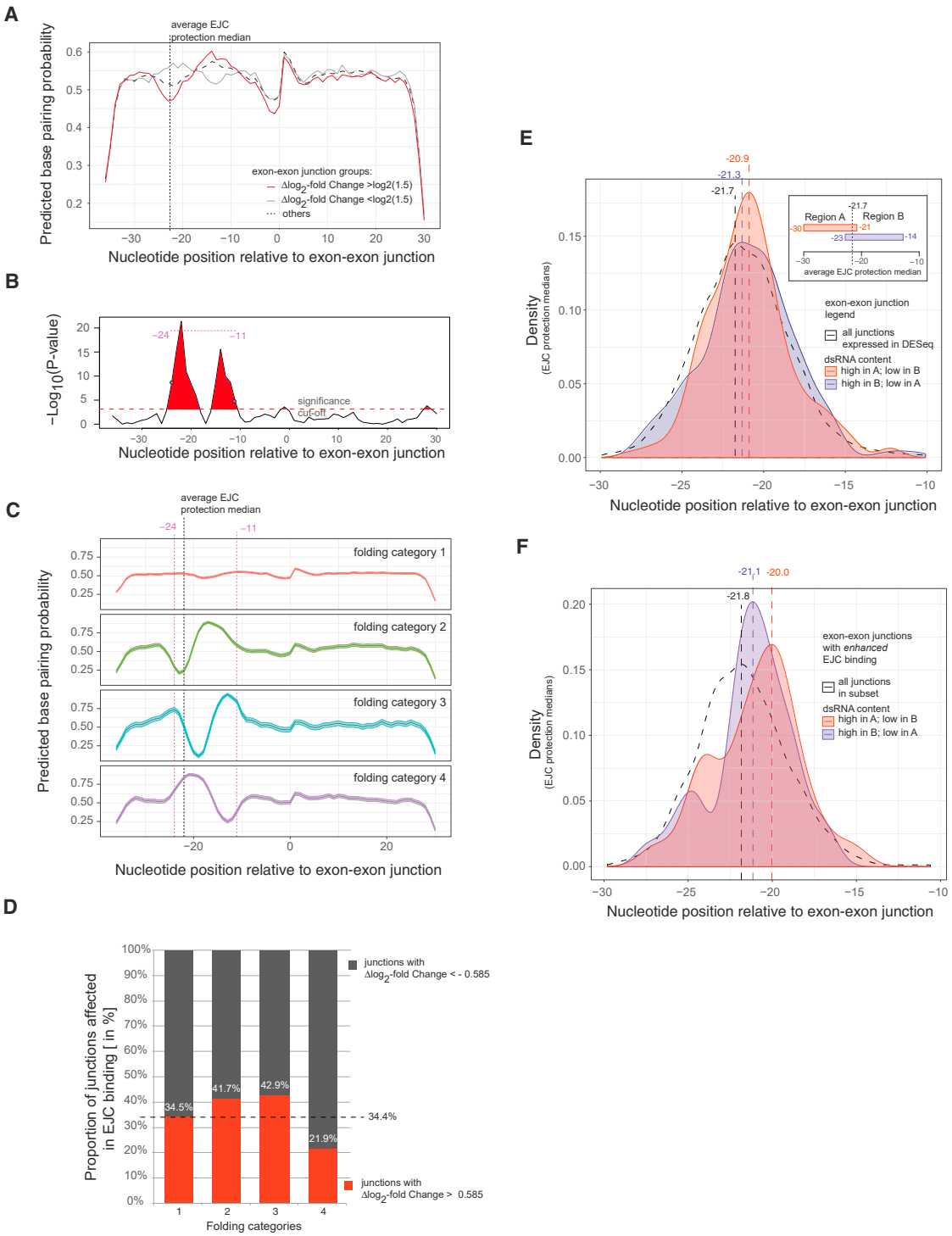


Figure 6. mRNA Secondary Structures Modulate EJC Binding

(A) Pairing probability profiles of EJC bound exon-exon junctions. Predicted base-pairing probability (bpp) profiles, by RNAplfold from Vienna RNA package, of junctions with enhanced, inhibited, and unaffected EJC binding. Black dashed, red, and gray solid lines highlight average bpp profiles of junctions that are unaffected, enhanced, and inhibited for EJC binding, respectively. Note that the region used for RNA-fold analysis covers the last 37 nt of the upstream exon and the first 28 nt of the downstream exon. The y axis shows predicted bpp. The x axis highlights nucleotide positions relative to exon-exon junction. Position 0 represents last RNA nucleotide of upstream exons. Dashed vertical line highlights the coordinate (−21.7) of the average EJC protection median.

(legend continued on next page)

observations confirm that the structural context of exon-exon junctions not only affects EJC binding efficiency but also directs the site of EJC assembly.

Comparison of Human and *Drosophila* Datasets Reveals Common Factors that Influence EJC Enrichment

Previous studies of the EJC in *Drosophila* and human reported differences between these two species in terms of EJC protein components and cellular function. We asked whether any of the factors we identified in *Drosophila* would also influence EJC deposition in human cell lines. Using the tools and models described in previous sections, we analyzed CLIP enrichment of the EJC component BTZ in HeLa cells (Hauer et al., 2016). Results from the multiple regression analysis showed that the number of introns is a major determinant of the gene-to-gene variation in EJC deposition (Figure S7A). This agrees with our finding in *Drosophila* and suggests that this mechanism is conserved between *Drosophila* and humans. In contrast to our observations in the fly, we found that in mammals the extent of alternative splicing in genes has a small but positive effect on EJC deposition (Figure S7A), in agreement with previous findings that alternatively spliced genes are over-represented in EJC-enriched genes (Hauer et al., 2016; Saulière et al., 2012). However, in contrast to *Drosophila*, in humans intron length did not facilitate but rather antagonized EJC mRNA binding (Figure S7A).

Next, we investigated the factors that determine variability in EJC deposition within genes (Table S2; Figure S7D). Similar to our *Drosophila* findings, high 5′ and 3′ splice strength and presence of an ISE at the 5′ junction enhances EJC deposition within a junction. In this analysis, presence of ESEs in the upstream and of ESS in the downstream 50 nt strongly affects EJC deposition, something we did not observe in *Drosophila* (Figure 7A). A possible explanation for this is that ESE, ESS, and ISE sequences are annotated based on experiments in human cell lines and may not exert a similar effect in *Drosophila*. Among the 238 ESEs, we observed 28 hexamers containing AGAA, which is similar to a motif (GAAGA) found in a previous EJC CLIP study (Hauer et al., 2016; Saulière et al., 2012). We separated ESEs according to the presence of AGAA and indeed found that the ESEs containing AGAA are associated with stronger EJC deposition. We asked whether bpp profiles differ between junctions with positive and negative EJC binding. We found that an overall negative bpp (−24 to −18) around the EJC deposition site favors EJC deposition (Figures S7B and S7C). Unlike *Drosophila*, we did not find other regions whose bpps are associated with increased EJC deposition in mammalian cells. Taken together, these results indicate that across *Drosophila* and human, not only do conserved regulatory mechanisms such as intron counts, splice strength, or structural hindrance within EJC deposition sites influence EJC deposition, but also divergent regulatory factors such as ESEs in mammals or RNA folding in proximity of EJC deposition sites in *Drosophila* can affect EJC deposition.

Features that Inform on EJC Binding May Predict mRNA Localization

The bias of EJC binding to mRNAs from genes with functions in development and cell polarity suggested that EJC binding might be indicative of transcripts under spatial or temporal control. Consistent with this hypothesis, analysis of EJC and RBP protection sites on *Drosophila* transcripts annotated in the FlyFISH RNA localization database (Lécuyer et al., 2007; Wilk et al., 2016) revealed that localized maternal mRNAs are more bound by EJC than non-localizing transcripts (Figure 4C). We postulated that modalities underlying EJC enrichment at the gene and junction level can inform us about the localization of a transcript and applied decision tree learning on RNA localization using the R package rpart (Therneau and Atkinson, 2018). Given the imbalance between localized to non-localized dataset, we used Cohen's kappa coefficient to assess the predictive value of the model using different data groups. The use of gene features and features of the most enriched EJC junction is sufficient to achieve predictive accuracy comparable with a model

(B) Definition of bpp regions distinct between exon-exon junctions with enhanced and inhibited EJC binding. To define a cutoff for significantly distinct regions in exon-exon junctions with enhanced and inhibited EJC binding, the bpp significance for each nucleotide in the two junction groups was tested using ANOVA and then subjected to permutation testing. Note that at alpha of 0.05, the regions −11 to −16 and −20 to −24 are significantly distinct in bpp between junctions enhanced and depleted in EJC binding. The y axis defines $-\log_{10}$-transformed ANOVA p values. Horizontal red dashed line indicates cutoff at $\alpha = 0.05$ of the permutation test. The x axis highlights nucleotide position relative to exon-exon junction.

(C) Folding categories of analyzed junctions. To define folding categories, exon-exon junctions were clustered on the basis of bpp estimates within the region −24 to −11 (see B). Clustering was performed with MClust using Gaussian mixture model assuming four clusters, and from the resulting categories the mean bpp profile of all cluster members are shown as individual bpp profile plots. Shaded region around solid lines indicates 95% confidence interval estimated using normal approximation to binomial distribution. Vertical dashed lines indicate borders of selected region for bpp profile clustering. The y axis shows predicted bpp estimates, and the x axis highlights nucleotide positions relative to exon-exon junction.

(D) Impact of RNA structures on EJC binding. Segmented bar plot showing proportion of significantly affected exon-exon junctions enhanced for EJC binding. Horizontal dashed line indicates overall proportion of junctions with enhanced EJC binding. Segments referring to proportion of junctions with enhanced or inhibited EJC binding are highlighted in red or gray, respectively. The y axis defines proportion (as a percentage) of all junctions significantly effected in EJC binding. The x axis indicates plotted folding categories (shown C). Percentage estimates for junctions with enhanced EJC binding are highlighted within the bar segments. Note that folding categories 2 and 3 have a positive, while folding category 4, which has an increased bpp within EJC deposition sites, has a negative impact on EJC binding.

(E and F) RNA stem structures impact EJC binding coordinates. Plots showing distribution densities of EJC protection site medians. Double-stranded (ds) RNA content in regions A and B was identified using the mean of rounded bpp estimates (0 and 1) in the corresponding sequence stretches. High and low dsRNA content are defined as >0.7 and <0.3, respectively. The structural exon-exon junction conditions tested, as well as the corresponding color coding, are indicated in the plot legend. RNA stem structural impact on EJC deposition coordinates in all junctions analyzed by DESeq is presented in (E). RNA stem structural impact on EJC deposition coordinates in junctions with enhanced or inhibited EJC binding is highlighted in (F). Vertical dashed lines represent coordinates of estimated mean EJC protection site medians in each of the conditions. The y axis defines EJC protection median density, and the x axis indicates nucleotide coordinates relative to the exon 3′ end. Note that the strongest structural impact on EJC binding coordinates is present at exon-exon junctions with enhanced EJC binding.

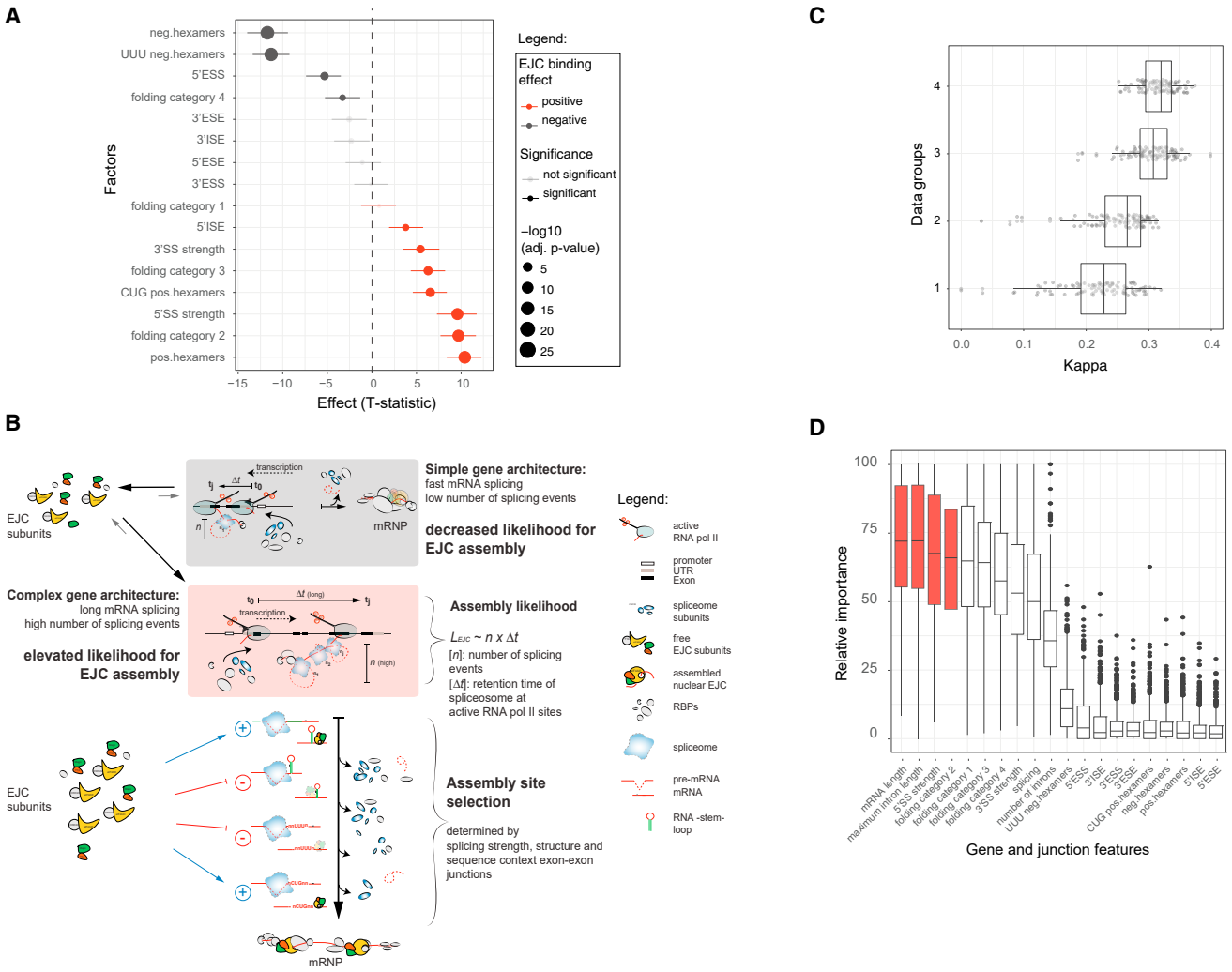


Figure 7. Summary of All Factors Regulating EJC Assembly at Exon-Exon Junctions in *Drosophila*

(A) Factors in *Drosophila* that determine EJC binding variability within a transcript. Plot comparing the effects of the different variables on junction level EJC enrichment in *Drosophila*. We fitted a full linear model: $\Delta\log_2$ fold change ~ splice site (SS) strength + ESE + ESS + ISE + hexamer categories + folding categories. $\Delta\log_2$ fold change is the difference between junction EJC enrichment and gene EJC enrichment. 5′SS and 3′SS would be the splice strength present at 5′ and 3′ splice junctions. Hexamer and folding categories were defined in Figures 5B and 6C, respectively. ESE, ESS, and ISE stand for the number of indicated splicing regulatory elements present in the upstream exon or downstream intron of splice site, respectively. The p value of each term was obtained using a t test (summary.lm function in R), corrected for multiple testing (p.adj. in R with "BH"), and plotted. The effect and SE of the term (sign of the coefficient) is reflected by color of the bar. Red indicates a positive effect, meaning that this variable has a positive effect on EJC deposition at the junction level, relative to the gene level. The y axis indicates factors (terms) tested, and the x axis highlights the result of the t test. Summary legend is indicated to the right of the plot.

(B) Model of EJC assembly and variable binding along the gene's transcript. Co-transcriptional recruitment is affected by speed of mRNA production. mRNAs of simple genes with low complexity and small number of introns (n) are transcribed and processed faster than mRNAs from genes with diverse intron sizes and large number of introns. Longer processing increases retention time (Δt) of assembled spliceosomes at the site of transcription and thereby increases the likelihood of EJC assembly ($L_{EJC} \sim n \times \Delta t$). Variable binding of EJC to mRNAs is a consequence of mRNA structure and sequence-mediated hindrance or facilitation of EJC binding. RNA is highlighted in red, and potential base-pairing regions are highlighted in green. Note that binding of EJC is enhanced in proximity of dsRNA regions but repelled within dsRNA moieties. Legend is indicated at the bottom of the figure.

(C and D) Gene and junction features inform on mRNA localization.

(C) Cohen's kappa coefficient for different models. Box plots showing the distribution of Cohen's kappa coefficient obtained after fitting tree model for classification by recursive partitioning (over 100 bootstraps of four different data groups) using R package "rpart" (http://CRAN.R-project.org/package=rpart) (Therneau and Atkinson, 2018). Group 1 includes gene features of the mRNA that influence EJC deposition from Figure 4E. Group 2 includes junction features (from Figure 7A) of the junction with the highest EJC deposition. Group 3 includes all variables in groups 1 and 2. Group 4 includes all the variables from group 3 and EJC estimates ($\log_2$ fold change for the gene and highest EJC deposited junction and the delta value). Median values are 0.246 (data group 1), 0.267 (group 2), 0.306 (group 3), and 0.315 (group 4).

(D) Relative variable importance for all features. Box plots showing the distribution of variable importance for all variables obtained after fitting tree model for classification, rpart over 1,000 bootstraps of data group 3. The variable importance is an estimate of the usefulness of the variable in splitting the different classes in the regression tree. Highlighted in red are variables that are likely more useful in the classification compared with the other variables.

incorporating all possible features (Figure 7C). Using the gene and highest EJC covered junction information as a working model, we asked which variables are important in this model. To prevent bias in estimate, the data were bootstrapped 1,000 times. We observed that transcript length, maximum intron size in the gene, and 5′ splice strength and folding of the most enriched junction are more useful predictors (Figure 7D) compared with other variables. This suggests that features of gene architecture that orchestrate EJC deposition can distinguish localizing and non-localizing mRNA and might be important for mRNA localization.

DISCUSSION

ipaRt: A Method for High-Confidence Identification of Protein Binding Sites on RNAs *In Vivo*

We have profiled the landscape of EJC binding across the transcriptome of a whole animal, *Drosophila* melanogaster, and determined the parameters that influence the distribution of the complex on RNAs in the organism. Previous knowledge of EJC-RNA interactions was based on UV-crosslinking experiments in specific cell types grown as homogeneous cultures for the individual studies (Hauer et al., 2016; Ince-Dunn et al., 2012; König et al., 2010, 2011; Licatalosi et al., 2008; Modic et al., 2013; Saulière et al., 2012; Tollervey et al., 2011; Ule, 2009; Ule et al., 2003; Wang et al., 2010). Although UV crosslinking remains a method of choice for identification of protein binding sites on nucleic acids, because of the inefficient penetration of UV light into tissues and organisms, the method is most useful when applied to cells in culture. In contrast, our analysis of EJC distribution in the tissues of whole *Drosophila* flies was made possible by ipaRt, which uses the crosslinking agent DSP to freeze protein-protein interactions within otherwise dynamic RNP complexes, such as the EJC.

We have demonstrated that DSP-mediated covalent bond formation between the RNA helicase eIF4AIII and the Mago-Y14 heterodimer preserves EJCs in their "locked" state on mRNAs (Andersen et al., 2006; Ballut et al., 2005; Bono et al., 2006; Stroupe et al., 2006) and that efficient recovery of the bound RNAs does not require their crosslinking to eIF4AIII using UV light. Our "ipaRt" approach, like CLIP and iCLIP (König et al., 2010; Ule et al., 2003), enables highly stringent washing of the samples. In support of the robustness and reliability of our DSP-based assays, we demonstrated high reproducibility not only among technical but also biological replicates of EJC ipaRt, as well as mRBP footprinting sequencing results (Figure S4A).

Furthermore, ipaRt allows enables the use of non-RNA-binding subunits of the EJC, such as Mago, as immunoprecipitation baits. This is highly relevant in the context of the EJC, as we and others have shown that its RNA-binding subunit, the RNA helicase eIF4AIII, may have other, EJC-independent functions in the cell (Figure S1C) (Alexandrov et al., 2011; Choudhury et al., 2016). ipaRt afforded us the option of using Mago as our EJC bait, and indeed this is a main reason for the high-quality definition of the EJC binding landscape in the fly cytoplasm that we have achieved. The protection site reads we obtained from EJC ipaRt map almost exclusively to canonical EJC deposition sites (Le Hir et al., 2000) with a median protection ~22 nt of the upstream exon's 3′ end. In contrast to mammalian EJC CLIP and RIP studies, in which eIF4AIII served as an immunoprecipitation bait (Saulière et al., 2012; Singh et al., 2012), EJC ipaRt reads mapping to regions distant from canonical deposition sites are of low abundance and sequencing coverage. Although this discrepancy could reflect differences in EJC engagement in humans and *Drosophila*, it more likely reflects the choice of bait or the cell compartment in which the analysis was executed. Indeed, a recent study in human cells revealed that when the cytoplasmic EJC component Btz was chosen as the bait rather than eIF4AIII, the proportion of non-canonical EJC deposition sites was negligible (Hauer et al., 2016).

Finally, in ipaRt the DSP crosslinker is applied *ex vivo* during tissue disruption and does not require inhibition of translation *in vivo*. We therefore consider ipaRt a method of choice for functional investigations of protein-RNA complexes in fully developed organisms and tissues.

Regulation of EJC Assembly in *Drosophila*

Through our analysis, we defined factors that contribute to or inhibit EJC assembly on mRNAs and at individual exon-exon junctions in *Drosophila*. From this we deduce that the landscape of EJC binding to RNAs is sculpted through regulation of EJC assembly at two levels in the fly (Figure 7B).

At the upstream regulatory level, the degree to which EJCs are assembled on an mRNA is dictated by the complexity of the gene's architecture: mRNAs produced from genes of simple architecture are marked by fewer EJCs, while mRNAs from genes of complex architecture, comprising multiple splice sites and long introns, are EJC bound to a higher degree (Figures 4D and S6B). Given that EJCs assemble on mRNAs concomitantly with splicing (Alexandrov et al., 2012; Barbosa et al., 2012; Le Hir et al., 2000; Steckelberg et al., 2015), it is not surprising that mRNAs of genes containing a greater number of introns are more likely to be EJC bound. However, our finding that the enhancing effect on EJC binding provoked by large introns is not restricted to flanking junctions but occurs at junctions mRNA-wide is unexpected (Figure S5F). Loss-of-function experiments indicate that the EJC participates in exon definition during splicing of long intron-containing genes in *Drosophila* (Ashton-Beaucage et al., 2010; Roignant and Treisman, 2010), particularly in definition of exons proximal to the long introns (Hayashi et al., 2014; Malone et al., 2014). Our data exclude any significant bias toward EJC assembly in proximity to long-intron splice junctions. Instead they reveal a general enhancement of EJC binding at exon-exon junctions throughout transcripts of long-intron genes (Figure S5G). Therefore, we conclude that stable binding of EJCs within mRNAs of long-intron genes is not the result of EJC engagement in exon definition. Instead, we propose that the high degree of EJC binding to long-intron transcripts derives from the increased number and resting time of co-transcriptionally assembled spliceosomes on the nascent transcripts, which would increase the probability of CWC22-dependent eIF4AIII recruitment to pre-mRNAs during splicing (Figure 7B).

At the downstream regulatory level, after EJC assembly rates at transcripts are defined, deposition of EJCs along mRNA exon-exon junctions is modulated by the structural and sequence context of the splice sites (Figures 7A and 7B). dsRNA stem

structures in exon-exon junctions of *Drosophila* mRNAs either antagonize EJC assembly when present within canonical EJC deposition sites or enhance EJC assembly when located in the vicinity of the EJC deposition site (Figures 6D and 7A). Absence of dsRNA within the EJC binding moiety is in agreement with reported preference of EJCs for ssRNA (Andersen et al., 2006; Bono et al., 2006; Mishler et al., 2008). It remains to be elucidated how and why EJC binding is positively affected when RNA stem structures are found in its direct proximity on the bound template.

Although it is likely that the structural context of exon-exon junctions in *Drosophila* directly influences the degree of EJC assembly, sequence composition-derived effects on EJC binding to mRNA are a consequence of the assigned roles of these sequences during pre-mRNA splicing. We have demonstrated that exon-exon junctions with strong 5′ and strong 3′ splice sites (SSs) are biased toward junctions with enhanced EJC binding (Table S1; Figure 7A). For the regulation of weak 5′ and 3′ SSs, which commonly occur at alternatively spliced junctions, *cis*-acting splicing regulatory elements (SREs) were shown to be of importance (Brooks et al., 2011; Koren et al., 2007; Shepard et al., 2011). In light of the negative impact of alternative splicing at the level of EJC mRNA binding (Figure 4D), it is not surprising that conventional ESEs and ESSs hardly affect EJC binding at the level of individual exon-exon junctions. Whether the position-dependent bias mediated by the UUU-triplet- and CUG-triplet-containing hexamers toward inhibited or enhanced EJC binding that we have discovered in our *Drosophila* dataset (Figures 5B and 5C) is due to a direct or indirect influence of these hexamers on splicing remains to be addressed. UUU-triplet-containing hexamers, which are strongly biased against EJC binding, could potentially function as yet undefined 5′ESS in *Drosophila*. Interestingly, CUG-triplet-containing hexamers, which are strongly biased toward enhanced EJC binding, share sequence similarity with a previously predicted CUG containing 5′ESE of short intron splice sites (Brooks et al., 2011). It appears likely that the CUG-triplet and UUU-triplet hexamers exert their effect on EJC binding as a yet undefined class of SREs.

RNA Modalities Influencing EJC Binding in Mammals and Fly

In agreement with reports in mammals (Hauer et al., 2016; Saulière et al., 2010, 2012; Singh et al., 2012), the extent of EJC occupancy varies between mRNAs and exon-exon junctions also in *Drosophila*. The splice site score next to a junction correlates with increased EJC deposition in the fly, and this relationship between splicing efficiency and EJC deposition has also been proposed in mammalian studies (Custódio et al., 2004; Gudikote et al., 2005). Analysis of published mammalian Btz iCLIP data (Hauer et al., 2016) revealed several modalities that correlate with the increased binding landscape of the EJC on mRNAs in both mammals and *Drosophila*, including the large number of introns, high transcript abundance, and sequence context of individual exon-exon junctions (Figure S7). Interestingly, the presence of long introns has a slightly negative effect and the amount of alternative splicing a slightly positive effect on EJC occupancy in mammals (Figures 4D and S7); the latter agrees with previous observations (Hauer et al., 2016; Saulière et al., 2012; Singh et al., 2012). Studies in cultured mammalian cells have reported that EJC-enriched junctions contain a relatively high proportion of "non-canonical" protection sites, which were enriched for RBP consensus sequences of the SR protein family (Saulière et al., 2012; Singh et al., 2012). Our analysis of mammalian Btz iCLIP data (Hauer et al., 2016) confirms that presence of ESEs in upstream exons and 5′ISEs in introns correlates with enhanced EJC binding (Table S2). Moreover, we have identified a group of junctions in mammals containing AGAA hexamers that are biased for enhanced EJC binding (Figure S6A), but their effects are not especially strong near the canonical EJC deposition site (Figure S6B). These hexamers match the AGAA-encompassing consensus sequence of the mammalian SR protein SRSF10, known to function as splicing enhancers (Cléry et al., 2011; Tsuda et al., 2011), and have been found previously in EJC bound exon-exon junctions (Hauer et al., 2016; Saulière et al., 2012; Singh et al., 2012). Not only do our *in silico* results agree with these reports and support the proposed cooperative binding of EJC with SR proteins (Hauer et al., 2016; Saulière et al., 2012; Singh et al., 2012), they also partially explain the EJC's preference in mammals for alternatively spliced mRNAs.

One observation deriving from our analysis of published mammalian Btz iCLIP datasets is surprising. Although we observed junctions in *Drosophila* to be enhanced or inhibited in EJC binding by specific base-pairing probability (bpp) profiles, thus by specific RNA folding categories (Figures 6A–6D), we could not detect any striking difference between overall bpp profiles of exon-exon junctions with enhanced or inhibited EJC binding in mammals (Figure S7B). Indeed, the only aspect of RNA structure shared by mammals and *Drosophila* is the negative effect of dsRNA when directly overlapping with the canonical EJC deposition site (Figure S7C) (Mishler et al., 2008). In *Drosophila*, however, the presence of dsRNA close to canonical deposition sites enhances EJC binding, an effect that is not observed in mammalian cells.

Insights into Evolution and Divergence of EJC Functions

Our findings regarding the differences in the RNA modalities enriched at highly occupied mammalian and *Drosophila* EJC sites provide insight into the expansion of functions of the EJC during eukaryotic evolution. Spliceosome catalyzed splicing reactions are bidirectional, and efficient formation of exon-exon junctions during RNA maturation is achieved by Prp22-induced release of spliceosomes from mRNAs (Hoskins and Moore, 2012; Smith and Konarska, 2008; Tseng and Cheng, 2008). The EJC is absent in organisms with low rates of RNA splicing, such as *Saccharomyces cerevisiae*, but present in organisms with high splicing rates, such as *Schizosaccharomyces pombe* (Goffeau et al., 1996; Wen and Brogna, 2010; Wood et al., 2002). This suggests that with the increased demand for splicing accuracy in higher eukaryotes, the EJC evolved to function as an exon-exon junction "lock" hindering spliceosome reassembly at spliced exon-exon junctions. Because EJC binding in the fly is enhanced at strong splices sites, but is not affected by splicing enhancer elements, and is not biased toward alternatively spliced mRNAs, we propose that the EJC preserved its primary function as such a lock in *Drosophila*. Two recent studies provide evidence that also in mammals bound EJCs hinder spliceosome assembly

(Blazquez et al., 2018; Boehm et al., 2018), suppressing recursive splicing (RS) of RS exons. The previously reported importance of EJC for splicing fidelity (Hayashi et al., 2014; Malone et al., 2014), and our observations on the mode of EJC binding to transcripts in the fly revealing its independence from splicing regulatory elements indeed supports that the EJC's most conserved function is to ensure splicing irreversibility.

The EJC further evolved to become a central component of the NMD pathway (Buchwald et al., 2010; Gehring et al., 2005; Melero et al., 2012; Okada-Katsuhata et al., 2012; Palacios et al., 2004; Shibuya et al., 2006; Singh et al., 2007) in mammals, in which more than 95% of all genes are alternatively spliced (Gromadzka et al., 2016). This may explain why EJCs in mammals are enriched on alternatively spliced transcripts. In *Drosophila*, in which only 30% of all genes appear to be alternatively spliced (Gibilisco et al., 2016), the EJC is not a component of the main NMD pathway (Behm-Ansmant et al., 2007). We propose that although the EJC-NMD pathway evolved before segregation of the proto- and deuterostome clades, it gained importance by complementing the faux 3′UTR-NMD pathway during the evolution of vertebrates (Eberle et al., 2008), for which RNA surveillance and spatiotemporal control of gene expression are essential.

Similarly, recruitment of the EJC and interacting proteins upon splicing to facilitate mRNA localization so far seems exclusive to *Drosophila*. Two *Drosophila*-specific features that modulate EJC binding, namely, the presence of a large intron within a gene and secondary structure near the junction, are also predictive of mRNA localization. Although the precise strength of association between these features and mRNA localization remains to be verified with larger and more quantitative datasets, previous studies with the SOLE in *oskar* RNA have shown that RNA structure and EJC binding are indeed crucial for *oskar* mRNA localization (Ghosh et al., 2012).

STAR★METHODS

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

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Topo and Gateway cloning
  - Preparation of cytoplasmic lysates
  - Cellular fractionation quality control
  - Precipitation of mRNA-protein complexes
  - Immunoprecipitation and RNase treatment
  - Choice of EJC bait: GFP-Mago
  - Sucrose density gradient centrifugation
  - EJC sedimentation: eIF4AIII is a poor bait
- WESTERN ANALYSIS AND ANTIBODIES
  - Mass spectrometry
  - EJC ipaRt and mRBP footprinting
  - EJC ipaRt and L3-App adaptor ligation
  - mRBP-footprinting and L3-App adaptor ligation
  - cDNA library preparation
  - mRNA Sequencing
  - Computational data processing
- PROCESSING OF MASS SPECTROMETRY DATA
  - Sequencing read mapping
  - Assay quality control: gene class enrichment
  - Sequencing read coverage across exon junctions
  - EJC protection site peak calling
- DIFFERENTIAL EXPRESSION SET ANALYSIS
  - Gene feature annotation
  - Splicing analysis and ISE, ESE, ESS
  - Assessing features that explain EJC binding
  - RNA structure prediction and clustering
  - Tree model for mRNA localization
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND CODE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.celrep.2019.06.088.

ACKNOWLEDGMENTS

We thank Marco Blanchette, Matthias Hentze, Isabel Palacios, and Jean-Yves Roignant for antibodies and fly stocks. We thank Sandra Müller, Alessandra Reversi, and Anna Cyrklaff for technical assistance. We thank Vladimir Benes and the EMBL Gene Core Facility for their advice and service. We thank Mandy Rettel, Frank Stein, and the EMBL Proteomics Core Facility for their service and assistance with mass spectrometry analysis. We are grateful to members of the Ephrussi lab for discussions during the course of the work. A.O. was supported by postdoctoral fellowships from the Swedish Vetenskapsradet (registration number 2010-6728), by Marie Curie Actions (FP7-PEOPLE-IEF number 2763207), and by a Deutsche Forschungsgemeinschaft (DFG) Network Grant (FOR 2333). This study was funded by the DFG (FOR 2333) and the European Molecular Biology Laboratory.

AUTHOR CONTRIBUTIONS

A.O. conceived, designed, and performed wet lab experiments. A.O., G.L., and N.H. contributed reagents, materials, and analysis tools. G.L., A.O., and N.H. conceived and designed the *in silico* analyses. A.O., G.L., N.H., J.U., and A.E. analyzed and interpreted the results. A.O., G.L., N.H., J.U., and A.E. contributed to the writing of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 28, 2018
Revised: April 30, 2019
Accepted: June 24, 2019
Published: July 30, 2019

REFERENCES

Alexandrov, A., Colognori, D., and Steitz, J.A. (2011). Human eIF4AIII interacts with an eIF4G-like partner, NOM1, revealing an evolutionarily conserved function outside the exon junction complex. Genes Dev. *25*, 1078–1090.

Alexandrov, A., Colognori, D., Shu, M.D., and Steitz, J.A. (2012). Human spliceosomal protein CWC22 plays a role in coupling splicing to exon junction complex deposition and nonsense-mediated decay. Proc. Natl. Acad. Sci. U S A *109*, 21313–21318.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Andersen, C.B., Ballut, L., Johansen, J.S., Chamieh, H., Nielsen, K.H., Oliveira, C.L., Pedersen, J.S., Séraphin, B., Le Hir, H., and Andersen, G.R. (2006). Structure of the exon junction core complex with a trapped DEAD-box ATPase bound to RNA. Science *313*, 1968–1972.

Ashton-Beaucage, D., and Therrien, M. (2011). The exon junction complex: a splicing factor for long intron containing transcripts? Fly (Austin) *5*, 224–233.

Ashton-Beaucage, D., Udell, C.M., Lavoie, H., Baril, C., Lefrançois, M., Chagnon, P., Gendron, P., Caron-Lizotte, O., Bonneil, E., Thibault, P., and Therrien, M. (2010). The exon junction complex controls the splicing of MAPK and other long intron-containing transcripts in Drosophila. Cell *143*, 251–262.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. *37*, W202-8.

Ballut, L., Marchadier, B., Baguet, A., Tomasetto, C., Séraphin, B., and Le Hir, H. (2005). The exon junction core complex is locked onto RNA by inhibition of eIF4AIII ATPase activity. Nat. Struct. Mol. Biol. *12*, 861–869.

Barbosa, I., Haque, N., Fiorini, F., Barrandon, C., Tomasetto, C., Blanchette, M., and Le Hir, H. (2012). Human CWC22 escorts the helicase eIF4AIII to spliceosomes and promotes exon junction complex assembly. Nat. Struct. Mol. Biol. *19*, 983–990.

Behm-Ansmant, I., Gatfield, D., Rehwinkel, J., Hilgers, V., and Izaurralde, E. (2007). A conserved role for cytoplasmic poly(A)-binding protein 1 (PABPC1) in nonsense-mediated mRNA decay. EMBO J. *26*, 1591–1601.

Bienkowski, R.S., Banerjee, A., Rounds, J.C., Rha, J., Omotade, O.F., Gross, C., Morris, K.J., Leung, S.W., Pak, C., Jones, S.K., et al. (2017). The conserved, disease-associated RNA binding protein dNab2 interacts with the fragile X protein ortholog in Drosophila neurons. Cell Rep. *20*, 1372–1384.

Blazquez, L., Emmett, W., Faraway, R., Pineda, J.M.B., Bajew, S., Gohr, A., Haberman, N., Sibley, C.R., Bradley, R.K., Irimia, M., and Ule, J. (2018). Exon junction complex shapes the transcriptome by repressing recursive splicing. Mol. Cell *72*, 496–509.e9.

Boehm, V., Britto-Borges, T., Steckelberg, A.-L., Singh, K.K., Gerbracht, J.V., Gueney, E., Blazquez, L., Altmüller, J., Dieterich, C., and Gehring, N.H. (2018). Exon junction complexes suppress spurious splice sites to safeguard transcriptome integrity. Mol. Cell *72*, 482–495.e7.

Bono, F., and Gehring, N.H. (2011). Assembly, disassembly and recycling: the dynamics of exon junction complexes. RNA Biol. *8*, 24–29.

Bono, F., Ebert, J., Lorentzen, E., and Conti, E. (2006). The crystal structure of the exon junction complex reveals how it maintains a stable grip on mRNA. Cell *126*, 713–725.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Breiman, L., Friedman, J., Stone, C.J., and Olshen, R.A. (1984). Classification and Regression Trees, First Edition (CRC Press).

Brooks, A.N., Aspden, J.L., Podgornaia, A.I., Rio, D.C., and Brenner, S.E. (2011). Identification and experimental validation of splicing regulatory elements in Drosophila melanogaster reveals functionally conserved splicing enhancers in metazoans. RNA *17*, 1884–1894.

Buchwald, G., Ebert, J., Basquin, C., Sauliere, J., Jayachandran, U., Bono, F., Le Hir, H., and Conti, E. (2010). Insights into the recruitment of the NMD machinery from the crystal structure of a core EJC-UPF3b complex. Proc. Natl. Acad. Sci. U S A *107*, 10050–10055.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., et al. (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell *149*, 1393–1406.

Castello, A., Horos, R., Strein, C., Fischer, B., Eichelbaum, K., Steinmetz, L.M., Krijgsveld, J., and Hentze, M.W. (2013). System-wide identification of RNA-binding proteins by interactome capture. Nat. Protoc. *8*, 491–500.

Chazal, P.E., Daguenet, E., Wendling, C., Ulryck, N., Tomasetto, C., Sargueil, B., and Le Hir, H. (2013). EJC core component MLN51 interacts with eIF3 and activates translation. Proc. Natl. Acad. Sci. U S A *110*, 5903–5908.

Choudhury, S.R., Singh, A.K., McLeod, T., Blanchette, M., Jang, B., Badenhorst, P., Kanhere, A., and Brogna, S. (2016). Exon junction complex proteins bind nascent transcripts independently of pre-mRNA splicing in *Drosophila melanogaster*. eLife *5*, e19881.

Cléry, A., Jayne, S., Benderska, N., Dominguez, C., Stamm, S., and Allain, F.H. (2011). Molecular basis of purine-rich RNA recognition by the human SR-like protein Tra2-β1. Nat. Struct. Mol. Biol. *18*, 443–450.

Custódio, N., Carvalho, C., Condado, I., Antoniou, M., Blencowe, B.J., and Carmo-Fonseca, M. (2004). In vivo recruitment of exon junction complex proteins to transcription sites in mammalian cell nuclei. RNA *10*, 622–633.

Duffy, J.B. (2002). GAL4 system in Drosophila: a fly geneticist's Swiss army knife. Genesis *34*, 1–15.

Eberle, A.B., Stalder, L., Mathys, H., Orozco, R.Z., and Mühlemann, O. (2008). Posttranscriptional gene regulation by spatial rearrangement of the 3′ untranslated region. PLoS Biol. *6*, e92.

Franken, H., Mathieson, T., Childs, D., Sweetman, G.M., Werner, T., Tögel, I., Doce, C., Gade, S., Bantscheff, M., Drewes, G., et al. (2015). Thermal proteome profiling for unbiased identification of direct and indirect drug targets using multiplexed quantitative mass spectrometry. Nat. Protoc. *10*, 1567–1593.

Gatfield, D., Le Hir, H., Schmitt, C., Braun, I.C., Köcher, T., Wilm, M., and Izaurralde, E. (2001). The DExH/D box protein HEL/UAP56 is essential for mRNA nuclear export in Drosophila. Curr. Biol. *11*, 1716–1721.

Gehring, N.H., Kunz, J.B., Neu-Yilik, G., Breit, S., Viegas, M.H., Hentze, M.W., and Kulozik, A.E. (2005). Exon-junction complex components specify distinct routes of nonsense-mediated mRNA decay with differential cofactor requirements. Mol. Cell *20*, 65–75.

Ghosh, S., Marchand, V., Gáspár, I., and Ephrussi, A. (2012). Control of RNP motility and localization by a splicing-dependent structure in oskar mRNA. Nat. Struct. Mol. Biol. *19*, 441–449.

Ghosh, S., Obrdlik, A., Marchand, V., and Ephrussi, A. (2014). The EJC binding and dissociating activity of PYM is regulated in Drosophila. PLoS Genet. *10*, e1004455.

Gibilisco, L., Zhou, Q., Mahajan, S., and Bachtrog, D. (2016). Alternative splicing within and between Drosophila species, sexes, tissues, and developmental stages. PLoS Genet. *12*, e1006464.

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 genes. Science *274*, 546–, 563–567.

Groemping, U. (2006). Relative importance for linear regression in R: the package relaimpo. J. Stat. Softw. *17*, 27.

Gromadzka, A.M., Steckelberg, A.L., Singh, K.K., Hofmann, K., and Gehring, N.H. (2016). A short conserved motif in ALYREF directs cap- and EJC-dependent assembly of export complexes on spliced mRNAs. Nucleic Acids Res. *44*, 2348–2361.

Gudikote, J.P., Imam, J.S., Garcia, R.F., and Wilkinson, M.F. (2005). RNA splicing promotes translation and RNA surveillance. Nat. Struct. Mol. Biol. *12*, 801–809.

Hachet, O., and Ephrussi, A. (2001). Drosophila Y14 shuttles to the posterior of the oocyte and is required for oskar mRNA transport. Curr. Biol. *11*, 1666–1674.

Hachet, O., and Ephrussi, A. (2004). Splicing of oskar RNA in the nucleus is coupled to its cytoplasmic localization. Nature *428*, 959–963.

Hauer, C., Sieber, J., Schwarzl, T., Hollerer, I., Curk, T., Alleaume, A.M., Hentze, M.W., and Kulozik, A.E. (2016). Exon junction complexes show a distributional bias toward alternatively spliced mRNAs and against mRNAs coding for ribosomal proteins. Cell Rep. *16*, 1588–1603.

Hayashi, R., Handler, D., Ish-Horowicz, D., and Brennecke, J. (2014). The exon junction complex is required for definition and excision of neighboring introns in Drosophila. Genes Dev. *28*, 1772–1785.

Hoskins, A.A., and Moore, M.J. (2012). The spliceosome: a flexible, reversible macromolecular machine. Trends Biochem. Sci. *37*, 179–188.

Hughes, C.S., Foehr, S., Garfield, D.A., Furlong, E.E., Steinmetz, L.M., and Krijgsveld, J. (2014). Ultrasensitive proteome analysis using paramagnetic bead technology. Mol. Syst. Biol. *10*, 757.

Ince-Dunn, G., Okano, H.J., Jensen, K.B., Park, W.Y., Zhong, R., Ule, J., Mele, A., Fak, J.J., Yang, C., Zhang, C., et al. (2012). Neuronal Elav-like (Hu) proteins regulate RNA splicing and abundance to control glutamate levels and neuronal excitability. Neuron *75*, 1067–1080.

Kataoka, N., Diem, M.D., Kim, V.N., Yong, J., and Dreyfuss, G. (2001). Magoh, a human homolog of Drosophila mago nashi protein, is a component of the splicing-dependent exon-exon junction complex. EMBO J. *20*, 6424–6433.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. *17*, 909–915.

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2011). iCLIP—transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. J. Vis. Exp. (*50*), 2638.

Koren, E., Lev-Maor, G., and Ast, G. (2007). The emergence of alternative 3′ and 5′ splice site exons from constitutive exons. PLoS Comput. Biol. *3*, e95.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput. Biol. *9*, e1003118.

Le Hir, H., Izaurralde, E., Maquat, L.E., and Moore, M.J. (2000). The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions. EMBO J. *19*, 6860–6869.

Lécuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T.R., Tomancak, P., and Krause, H.M. (2007). Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. Cell *131*, 174–187.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature *456*, 464–469.

Lomant, A.J., and Fairbanks, G. (1976). Chemical probes of extended biological structures: synthesis and properties of the cleavable protein cross-linking reagent [35S]dithiobis(succinimidyl propionate). J. Mol. Biol. *104*, 243–261.

Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Malone, C.D., Mestdagh, C., Akhtar, J., Kreim, N., Deinhard, P., Sachidanandam, R., Treisman, J., and Roignant, J.Y. (2014). The exon junction complex controls transposable element activity by ensuring faithful splicing of the piwi transcript. Genes Dev. *28*, 1786–1799.

Melero, R., Buchwald, G., Castano, R., Raabe, M., Gil, D., Lazaro, M., Urlaub, H., Conti, E., and Llorca, O. (2012). The cryo-EM structure of the UPF-EJC complex shows UPF1 poised toward the RNA 3′ end. Nat. Struct. Mol. Biol. *19*, 498–505, S491–S492.

Mishler, D.M., Christ, A.B., and Steitz, J.A. (2008). Flexibility in the site of exon junction complex deposition revealed by functional group and RNA secondary structure alterations in the splicing substrate. RNA *14*, 2657–2670.

Modic, M., Ule, J., and Sibley, C.R. (2013). CLIPing the brain: studies of protein-RNA interactions important for neurodegenerative disorders. Mol. Cell. Neurosci. *56*, 429–435.

Newmark, P.A., Mohr, S.E., Gong, L., and Boswell, R.E. (1997). mago nashi mediates the posterior follicle cell-to-oocyte signal to organize axis formation in Drosophila. Development *124*, 3197–3207.

Nielsen, K.H., Chamieh, H., Andersen, C.B., Fredslund, F., Hamborg, K., Le Hir, H., and Andersen, G.R. (2009). Mechanism of ATP turnover inhibition in the EJC. RNA *15*, 67–75.

Nott, A., Le Hir, H., and Moore, M.J. (2004). Splicing enhances translation in mammalian cells: an additional function of the exon junction complex. Genes Dev. *18*, 210–222.

Okada-Katsuhata, Y., Yamashita, A., Kutsuzawa, K., Izumi, N., Hirahara, F., and Ohno, S. (2012). N- and C-terminal Upf1 phosphorylations create binding platforms for SMG-6 and SMG-5:SMG-7 during NMD. Nucleic Acids Res. *40*, 1251–1266.

Palacios, I.M., Gatfield, D., St Johnston, D., and Izaurralde, E. (2004). An eIF4AIII-containing complex required for mRNA localization and nonsense-mediated mRNA decay. Nature *427*, 753–757.

R Core Team (2017). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

Reichel, M., Liao, Y., Rettel, M., Ragan, C., Evers, M., Alleaume, A.-M., Horos, R., Hentze, M.W., Preiss, T., and Millar, A.A. (2016). In planta determination of the mRNA-binding proteome of Arabidopsis etiolated seedlings. Plant Cell *28*, 2435–2452.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

Roignant, J.Y., and Treisman, J.E. (2010). Exon junction complex subunits are required to splice Drosophila MAP kinase, a large heterochromatic gene. Cell *143*, 238–250.

Rørth, P. (1998). Gal4 in the Drosophila female germline. Mech. Dev. *78*, 113–118.

Saulière, J., Haque, N., Harms, S., Barbosa, I., Blanchette, M., and Le Hir, H. (2010). The exon junction complex differentially marks spliced junctions. Nat. Struct. Mol. Biol. *17*, 1269–1271.

Saulière, J., Murigneux, V., Wang, Z., Marquenet, E., Barbosa, I., Le Tonquèze, O., Audic, Y., Paillard, L., Roest Crollius, H., and Le Hir, H. (2012). CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. Nat. Struct. Mol. Biol. *19*, 1124–1131.

Schweizer, E., Angst, W., and Lutz, H.U. (1982). Glycoprotein topology on intact human red blood cells reevaluated by cross-linking following amino group supplementation. Biochemistry *21*, 6807–6818.

Scrucca, L., Fop, M., Murphy, T.B., and Raftery, A.E. (2016). Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J. *8*, 289–317.

Shepard, P.J., Choi, E.-A., Busch, A., and Hertel, K.J. (2011). Efficient internal exon recognition depends on near equal contributions from the 3′ and 5′ splice sites. Nucleic Acids Res. *39*, 8928–8937.

Shibuya, T., Tange, T.O., Stroupe, M.E., and Moore, M.J. (2006). Mutational analysis of human eIF4AIII identifies regions necessary for exon junction complex formation and nonsense-mediated mRNA decay. RNA *12*, 360–374.

Shiimori, M., Inoue, K., and Sakamoto, H. (2013). A specific set of exon junction complex subunits is required for the nuclear retention of unspliced RNAs in Caenorhabditis elegans. Mol. Cell. Biol. *33*, 444–456.

Simon, B., Masiewicz, P., Ephrussi, A., and Carlomagno, T. (2015). The structure of the SOLE element of oskar mRNA. RNA *21*, 1444–1453.

Singh, G., Jakob, S., Kleedehn, M.G., and Lykke-Andersen, J. (2007). Communication with the exon-junction complex and activation of nonsense-mediated decay by human Upf proteins occur in the cytoplasm. Mol. Cell *27*, 780–792.

Singh, G., Kucukural, A., Cenik, C., Leszyk, J.D., Shaffer, S.A., Weng, Z., and Moore, M.J. (2012). The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. Cell *151*, 750–764.

Smith, D.J., and Konarska, M.M. (2008). Mechanistic insights from reversible splicing catalysis. RNA *14*, 1975–1978.

Steckelberg, A.L., Altmueller, J., Dieterich, C., and Gehring, N.H. (2015). CWC22-dependent pre-mRNA splicing and eIF4A3 binding enables global deposition of exon junction complexes. Nucleic Acids Res. *43*, 4687–4700.

Stroupe, M.E., Tange, T.O., Thomas, D.R., Moore, M.J., and Grigorieff, N. (2006). The three-dimensional arcitecture of the EJC core. J. Mol. Biol. *360*, 743–749.

Tange, T.O., Shibuya, T., Jurica, M.S., and Moore, M.J. (2005). Biochemical analysis of the EJC reveals two new factors and a stable tetrameric protein core. RNA *11*, 1869–1883.

Therneau, T., and Atkinson, B. (2018). rpart: Recursive Partitioning and Regression Trees.https://rdrr.io/cran/rpart/.

Tollervey, J.R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., König, J., Hortobágyi, T., Nishimura, A.L., Zupunski, V., et al. (2011). Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. Nat. Neurosci. *14*, 452–458.

Tseng, C.K., and Cheng, S.C. (2008). Both catalytic steps of nuclear pre-mRNA splicing are reversible. Science *320*, 1782–1784.

Tsuda, K., Someya, T., Kuwasako, K., Takahashi, M., He, F., Unzai, S., Inoue, M., Harada, T., Watanabe, S., Terada, T., et al. (2011). Structural basis for the dual RNA-recognition modes of human Tra2-β RRM. Nucleic Acids Res. *39*, 1538–1553.

Ule, J. (2009). High-throughput sequencing methods to study neuronal RNA-protein interactions. Biochem. Soc. Trans. *37*, 1278–1280.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212–1215.

van Eeden, F.J., Palacios, I.M., Petronczki, M., Weston, M.J., and St Johnston, D. (2001). Barentsz is essential for the posterior localization of oskar mRNA and colocalizes with it to the posterior pole. J. Cell Biol. *154*, 511–523.

Wang, Z., Rolish, M.E., Yeo, G., Tung, V., Mawson, M., and Burge, C.B. (2004). Systematic identification and analysis of exonic splicing silencers. Cell *119*, 831–845.

Wang, Z., Kayikci, M., Briese, M., Zarnack, K., Luscombe, N.M., Rot, G., Zupan, B., Curk, T., and Ule, J. (2010). iCLIP predicts the dual splicing effects of TIA-RNA interactions. PLoS Biol. *8*, e1000530.

Wen, J., and Brogna, S. (2010). Splicing-dependent NMD does not require the EJC in Schizosaccharomyces pombe. EMBO J. *29*, 1537–1551.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis (Springer-Verlag).

Wilk, R., Hu, J., Blotsky, D., and Krause, H.M. (2016). Diverse and pervasive subcellular distributions for both coding and long noncoding RNAs. Genes Dev. *30*, 594–609.

Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., et al. (2002). The genome sequence of Schizosaccharomyces pombe. Nature *415*, 871–880.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS *16*, 284–287.

Zimyanin, V.L., Belaya, K., Pecreaux, J., Gilchrist, M.J., Clark, A., Davis, I., and St Johnston, D. (2008). In vivo imaging of oskar mRNA transport reveals the mechanism of posterior localization. Cell *134*, 843–853.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rat polyclonal anti-Y14*(Drosophila)*	Laboratory of M. Blanchette	N/A
Rabbit polyclonal anti-Mago*(Drosophila)*	Laboratory of M. Blanchette	N/A
Rabbit polyclonal anti-eIF4AIII*(Drosophila)*	Laboratory of I. Palacios	N/A
Rabbit polyclonal anti-PABP (poly A binding protein) *(Drosophila)*	Laboratory of M. Hentze	N/A
Rabbit polyclonal anti-RpL32 (ribosome large subunit protein L32) *(Drosophila)*	Laboratory of M. Hentze	N/A
Mouse monoclonal anti-RpS6 [clone: 54D2] (ribosome small subunit protein S6)	Cell Signalling	Cat#2317
Rabbit polyclonal anti-GFP	Laboratory of T. Pines	N/A
Rabbit polyclonal anti-KHC (kinesin heavy chain)	Cytoskeleton	N/A (discontinued)
Mouse monoclonal anti-cMyc [clone: 9E10]	Santa Cruz Biotechnology	Cat#SC-40
Mouse monoclonal anti-Flag M2	Sigma Aldrich	Cat#F3165
Chemicals, Peptides, and Recombinant Proteins		
Dithio(bis-)succinimdylproprionate (DSP)	Thermo Fisher	Cat#22585
Trizol LS	Thermo Fisher	Cat#10296010
Heparin	Sigma-Aldrich	Cat#H4784
Western Blocking Reagent	Roche	Cat#11921681001
Lithium dodecyl sulfate	Sigma-Aldrich	Cat#L9781
IGEPAL CA-630	Sigma-Aldrich	Cat#I3021
CompleteMini Protease Inhibitor Cocktail	Roche	Cat#11836170001
Ribolock	Fermentas	Cat#EO0381
RNaseOUT	Thermo Fisher	Cat#10777-019
RNAseI	New England Biolabs	Cat#M0243S
TurboDNase	Thermo Fisher	Cat# AM2239
T4 RNA Ligase 2 (truncated)	New England Biolabs	Cat#M0242S
CircLigase II	Epicentre	Cat#CL9025K
PEG400	Sigma-Aldrich	Cat#8074850050
Linear Acrylamide	Thermo Fisher	Cat#AM9520
Protein A Agarose	Roche	Cat#11719408001
Protein G Agarose	Roche	Cat#11719416001
GFP trap Agarose	Chromotek	Cat#gta-100
Anti-Flag Agarose	Sigma-Aldrich	Cat#A1205
Anti-cMyc Agarose	Sigma-Aldrich	N/A
oligo-dT_{25} coated magnetic Dynabeads (for mRNA Seq)	Thermo Fisher	Cat#61002
oligo-$d(T)_{25}$ Magnetic Beads (for mRBP pulldown assay)	New England Biolabs	Cat#S1419S
Amicon Ultra 10K	Merck Millipore	Cat#UFC201024
MinElute Gel Extraction kit	Qiagen	Cat#28604
QIAquick PCR purification kit	Qiagen	Cat#28104
Critical Commercial Assays		
Illumina TruSeq RNA Sample Preparation v2 Kit	Illumina	Cat#RS-122-2001
TMT10plex Isobaric Labelling	Thermo Fisher	Cat# A34808
SuperScript III Reverse Transcriptase Kit	Sigma-Aldrich	Cat#18080044
Phusion Flash High-Fidelity PCR Master Mix	Thermo Fisher	Cat#F-548L

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Gateway LR Clonase Enzyme Mix	Thermo Fisher	Cat#11791-019
pENTR/D-TOPO Cloning Kit	Thermo Fisher	Cat#K240020
Deposited Data		
FASTQ files of ipaRt, mRBP footprinting and mRNA-Seq library sequencing are deposited at European Nucleotide Archive (ENA)	European Nucleotide Archive (ENA)	accession number PRJEB26421 https://www.ebi.ac.uk/ena/data/view/PRJEB26421
Drosophila melanogaster: reference proteome (UP000000803)	UniProt	ProteomeID: UP000000803 https://www.uniprot.org/proteomes/UP000000803
Drosophila melanogaster: reference genome (fasta file)	Ensembl	ftp://ftp.ensembl.org/pub/release-81//fasta/drosophila_melanogaster/dna/Drosophila_melanogaster.BDGP6.dna_sm.toplevel.fa.gz
Drosophila melanogaster: reference anntotation (gtf file)	Ensembl	ftp://ftp.ensembl.org/pub/release-81/gtf//drosophila_melanogaster/Drosophila_melanogaster.BDGP6.81.gtf.gz
Experimental Models: Organisms/Strains		
D. melanogaster: wildtype [w1118]	Bloomington Drosophila Stock Center	BDSC:3605 Flybase ID: FBst0003605
D. melanogaster: ubiquitous GFP expression (stable line) Genotype: *y1w*;P{w[+mC]=act5C-GAL4}25FO1/CyO, P{w[+mW.hs]=ubi-GFP.S65T}PAD1*	Bloomington Drosophila Stock Center	BDSC:4888 Flybase ID: FBst0004888
D. melanogaster: GFP-Mago expression (stable strain) Genotype: *bw⁻,mago-GFP::Mago;If/Cyo;Sb/TM3Ser*	Newmark et al., 1997	N/A
D. melanogaster: [3A2] UAS mediated FLAG-HA eIF4AIII and MYC-Y14 expression (unstable line) Genotype: *bw⁻, mago-GFP::Mago; P{UASp-FlagHA::eIF4AIII]}/CyO; P{UASp-MYC::Y14}, P{w[+mC]=tubP-GAL4}LL7 /TM3 Ser¹*	This paper	N/A
D. melanogaster: UAS mediated Flag-Myc-eGFP expression (unstable line) Genotype: *w⁻; P{UASp-FLAG::MYC::eGFP}/CyO; P{w[+mC]=tubP-GAL4}LL7 /TM3 Ser¹*	This paper	N/A
D. melanogaster: tub-Gal4 driver line (stable line) Genotype: *y¹w*; If/CyO; P(w[+mC]=tubP-GAL4)LL7/TM3, Sb¹ Ser¹*	Bloomington Drosophila Stock Center	BDSC:5138 Flybase ID: FBst0005138
Oligonucleotides		
eGFP CDS F: caccATGGTGAGCAAGGGC	This paper	N/A
eGFP CDS R: CTTGTACAGCTCGTCCATGC	This paper	N/A
Y14 CDS F: caccATGGCCGATGTGTTGGACATTG	This paper	N/A
Y14 CDS R: TCTGCGACGCTTTTCGGACTT	This paper	N/A
5' pre-Adenylated - L3 App adapter (HPLC purified) 5rApp/AGATCGGAAGAGCGGTTCAG/3ddC/	Konig et al., 2011	N/A
CUT oligo (HPLC purified): GTTCAGGATCCACGACG CTCTTCaaaa	Konig et al., 2011	N/A
P3 primer (HPLC purified): CAAGCAGAAGACGGCATAC GAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTC CGATCT	Konig et al., 2011	N/A
P5 primer (HPLC purified): AATGATACGGCGACCACC GAGATCTACACTCTTTCCCTACACGACGCTCTTCC GATCT	Konig et al., 2011	N/A
Recombinant DNA		
pPFMW (Gateway destination Vector for Drosophila expression)	The Drosophila Gateway Vector collection	https://emb.carnegiescience.edu/drosophila-gateway-vector-collection
pPFHW (Gateway destination Vector for Drosophila expression)	The Drosophila Gateway Vector collection	https://emb.carnegiescience.edu/drosophila-gateway-vector-collection
pPMW (Gateway destination Vector for Drosophila expression)	The Drosophila Gateway Vector collection	https://emb.carnegiescience.edu/drosophila-gateway-vector-collection

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
Samtools 0.1.18	Li et al., 2009	https://anaconda.org/bioconda/samtools
Tophat 2.1.1	Kim et al., 2013	https://anaconda.org/bioinfo/tophat2
RNAfold 2.3.5	Lorenz et al., 2011	https://anaconda.org/bioconda/viennarna
AME 4.10.2	Bailey et al., 2009	http://meme-suite.org
Kallisto 0.42.4	Bray et al., 2016	https://pachterlab.github.io/kallisto/download
R version 3.3.1	R Core Team, 2017	https://www.r-project.org
DESeq2 1.12.4	Love et al., 2014	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
Mclust 5.2.3	Scrucca et al., 2016	https://cran.r-project.org/web/packages/mclust/index.html
Rpart 4.1-13	Therneau and Atkinson, 2018	https://cran.r-project.org/web/packages/rpart/
GenomicAlignments 1.8.4	Lawrence et al., 2013	https://bioconductor.org/packages/release/bioc/html/GenomicAlignments.html
clusterProfiler 3.9	Yu et al., 2012	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
Fqtrim 0.9.4	John Hopkins center for computational biology	https://ccb.jhu.edu/ssoftwar/fqtrim/
Other		
iCLIP primer sequences for ipaRt and mRBP footprinting cDNA-library preparation are attached as a list to supplementary information	Konig et al., 2011	N/A

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Anne Ephrussi (ephrussi@embl.de).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All *Drosophila melanogaster* stocks were maintained at 25°C, and throughout the study, samples were prepared from a mix of male and female flies. For the generation of transgenic flies from Gateway destination vectors, pUASp-based destination plasmids containing tagged EJC subunits were injected into fly embryos together with helper plasmid as described previously (Hachet and Ephrussi, 2004). To establish a control line for ubiquitous expression of epitope-tagged fusion proteins, the transgenic fly line *w⁻; P[UASp-FLAG::MYC::eGFP]/CyO;Sb/TM3 Ser¹* was crossed with the driver line *y¹w*; If/CyO; P(w[+mC] = tubP-GAL4)LL7/TM3, Sb¹ Ser¹* (Bloomington Stock 5138). To compare the degree incorporation of transgenic GFP-Mago, Myc-Y14 and FLAG-HA-eIF4AIII into endogenous EJCs, a fly line carrying all three transgenic EJC subunits and a transgene for ubiquitously expressed GAL4 was established by standard genetic crosses.

Briefly, the driver line *y¹w*; If/CyO; P(w[+mC] = tubP-GAL4)LL7/TM3, Sb¹ Ser¹* (Bloomington Stock 5138) was crossed with the established transgenic line Myc-Y14 *w⁻;If/CyO;P(UASp-MYC::Y14)/TM3 Ser¹*. Upon recombination and balancing of Myc-Y14 and the GAL4 driver transgene on the 3rd chromosome, the ubiquitously expressing Myc-Y14 fly line *w⁻;If/CyO;UASp-MYC::Y14,P(w[+mC] = tubP-GAL4)LL7 /TM3 Ser¹* was further crossed with *w⁻;P(UASp-FLAG::HA::eIF4AIII)/CyO;Sb/TM3 Ser¹* and *bw, mago-GFP::Mago;If/Cyo;Sb/TM3Ser* (Newmark et al., 1997) to obtain the stable line *bw⁻,mago-GFP::Mago; UASp-FlagHA::eIF4AIII/CyO; UASp-MYC::Y14, P(w[+mC] = tubP-GAL4)LL7 /TM3 Ser¹*, in which Myc-Y14, eIF4AIII were expressed under control of the UAS-GAL4 expression system and GFP-Mago under control of its own promoter. For sucrose density gradients, mRBP footprinting, and EJC ipaRt experiments, male and female flies of the following genotypes were used: *w¹¹¹⁸* (wild-type), *bw⁻,mago-GFP::Mago;If/Cyo;Sb/TM3Ser* expressing GFP-Mago under control of the *mago* promoter (Newmark et al., 1997), and the GFP tag-only expressing fly line *y¹w*;P(w[+mC] = act5C-GAL4)25FO1/CyO,P(w[+mW.hs] = ubi-GFP.S65T)PAD1*.

METHOD DETAILS

Topo and Gateway cloning

For entry clone preparation, the full-length Y14 coding region was PCR amplified from *Drosophila melanogaster* cDNAs using *5′caccATGGCCGATG-TGTTGGACATTG3′* as the forward primer and *5′TCTGCGACGCTTTT-CGGACTT3′* as the reverse primer. Similarly, the eGFP full-length coding region was PCR amplified from plasmid pPWG (see below, destination vector recombination) using *5′caccATGGTGAGCAAG-GGC′3* as the forward primer and *5′CTTGTACAGCTCGTCCATGC3′* as the reverse primer. Directional Topo cloning into pENTR/SD/D-TOPO plasmid (Thermo Fisher) was performed according the Manufacturer's protocol. Preparation of an entry clone for the full-length eIF4AIII coding region was described elsewhere (Ghosh et al., 2014).

Recombination of entry clones with destination vectors from the *Drosophila* Gateway vector collection (gift of Terence Murphy, Carnegie Institution for Science; see link https://emb.carnegiescience.edu/) was performed according the Gateway cloning protocol (Thermo Fisher). Y14, eIF4AIII and eGFP entry clones were used for recombination with the destination vectors pPFMW and pPFHW for N-terminal FLAG tagging, and pPMW for N-terminal MYC tagging of Y14 and eIF4AIII, respectively. For C-terminal FLAG and GFP tagging of eIF4AIII, the destination vectors pTWF, pTWG and pPWF, pPWG were utilized in the recombination reaction.

Preparation of cytoplasmic lysates

Fly stocks were amplified and two day-old flies were sedated by CO_2, quick-frozen and pulverized with a mortar under liquid nitrogen. Fly powder was taken up in lysis buffer [20mM HEPES pH8.0, 125mM KCl, 4mM MgCl2, 0.1% IGEPAL (Sigma-Aldrich), 1 Unit/ml Ribolock (Fermentas) and 1x CompleteMini Protease Inhibitor Cocktail (Roche)] in a 1:8 weight:volume ratio and subsequently transferred to glass tissue grinder (Kontes Glass, USA). All subsequent steps were carried out on ice. During a first round of grinding, a loose bulb pestle (Kontes Glass, USA) was utilized until a homogenate of fly tissues and lysis buffer was obtained. For the second round of grinding, a tight pestle (Kontes Glass, USA) was utilized to ensure sufficient rupturing of the cells. Nuclei and crude cytoplasm were separated by centrifugation for 10 min at 900 g at 4°C. Nuclei were washed once in lysis buffer and quick-frozen for later use. For initial poly-A tail containing mRNA-RBP precipitation assay crude cytoplasmic supernatant was either left untreated or supplemented with DSP (Thermo Scientific) to 1mM final concentration and incubated for 1h at 4°C with constant shaking. Upon completion of the DSP cross-linking reaction, both the treated and untreated crude cytoplasmic fractions were blocked with 25mM Tris-HCl pH7.5 and centrifuged at 25000 g for 30 min at 4°C. Clarified cytoplasmic lysates were quick-frozen and stored at −80°C for later use.

Cellular fractionation quality control

Cellular lysates obtained from "wild-type" *w1118* flies were subjected to western blot analysis (see western blot analysis and antibody section). Western results of the separated cellular fractions confirmed the reliability of the assay (Figure S1A). The cytoskeletal motor protein Kinesin heavy chain (Khc) was present in cytoplasmic and absent from nuclear fractions; chromatin components such as histone 3 (H3) was detected primarily in nuclear and was nearly absent from cytoplasmic fractions (Figure S1B). Finally, known shuttling proteins such as the EJC subunits eIF4AIII and Y14 were present in both cellular fractions (Figure S1B).

Precipitation of mRNA-protein complexes

All procedures were carried out at 4°C or on ice, except for the final elution step. To precipitate proteins in complex with poly-A containing mRNAs (mRNPs), 15ml of DSP treated and untreated (native) cytoplasm were evenly distributed in Petri dishes and exposed to UV radiation (254 nm) using a UV Crosslinker (Stratalinker) at an energy setting of 2x 150mJ/cm^2. UV irradiated and unexposed cytoplasmic aliquots (DSP treated and native) were complemented with LiCl, for a final salt concentration of 150mM KCl and 500mM LiCl. For mRNP precipitation, 100µl equilibrated bead slurry of oligo-d(T)$_{25}$ Magnetic Beads (NEB) was mixed with 15ml salt-adjusted cytoplasmic lysate. Hybridization of poly-A containing mRNPs with oligo d(T)$_{25}$ beads was allowed to proceed for 2h, at 4°C. Nucleotide mediated precipitation of mRNPs was stopped by pelleting oligo d(T)$_{25}$ beads 2 min on a magnet. Pelleted beads were mixed with 15ml high salt wash buffer [20mM HEPES pH 7.8, 750mM LiCl, 0.2% IGEPAL CA-630, 0.1% LiDS, 1mM DTT, 5 mM EDTA, 1x CompleteMini Protease Inhibitor Cocktail (Roche)] and incubated for 30 min at 4°C. Washed beads were resuspended in 1.5ml high salt wash buffer and washing was repeated two additional rounds for 10 min at 4°C. After high salt washing was completed, beads were resuspended in medium salt washing buffer [20mM HEPES pH 7.8, 0.5M LiCl, 0.2% IGEPAL CA-630, 0.1% LiDS, 1mM DTT, 5 mM EDTA, 1x CompleteMini Protease Inhibitor Cocktail (Roche)], followed by two subsequent washing steps in medium salt washing buffer. To minimize contamination by unrelated rRNPs or other RNA-protein complexes, two additional washing steps were performed in 1.5ml low salt LiDS buffer [20mM HEPES pH7.8, 250 mM LiCl, 0.2% IGEPAL CA-630, 0.05%LiDS 1mM DTT, 5 mM EDTA, 1x CompleteMini Protease Inhibitor Cocktail (Roche)] and one washing step in 1.5ml low salt buffer without LiDS [20mM HEPES pH7.8, 250mM LiCl, 0.2% IGEPAL CA-630, 1mM DTT, 5mM EDTA, 1x CompleteMini Protease Inhibitor Cocktail (Roche)]. For elution of mRNPs, the hybridized to oligo d(T)$_{25}$ magnetic beads were mixed with 20µl HE elution buffer [20mM HEPES pH8.0, 1mM EDTA] and incubated for 5 min at 30°C, with agitation at 1200rpm. Elution was repeated two additional rounds. Supernatants were pooled and stored on ice for later use.

Immunoprecipitation and RNase treatment

All procedures were carried out at 4°C or on ice. Whenever indicated, cytoplasmic extracts utilized were either untreated (native) or treated with DSP. Herein, DSP was directly added to the cell-homogenate prior to nuclear-cytoplasmic fractionation. DSP crosslinking was allowed in total for 1h, during cell fractionation and cytoplasmic clarification step, and finalized by addition of 25mM Tris-HCl pH7.5. For pre-clearing of individual immunoprecipitation (IP) reactions, 10-15ml cytoplasmic lysate (4-5mg/ml protein-conc.) was incubated 1h at 4°C with 150µl slurry of Protein A/G agarose-beads (Roche) to reduce resin mediated stickiness in all subsequent steps. After pre-clearing was completed, beads were depleted by centrifugation for 10min at 900 g and 4°C. The reaction supernatant was transferred to fresh 15ml flacon tubes and stored on ice for immediate use. Prior to use, 20µl IP beads specific for the Flag or Myc epitope (Sigma) or 10µl GFP-trap beads (Chromotek) were incubated for 1h at 4°C in lysis buffer containing 1% western blocking reagent (Roche) and 0.1mg/ml heparin (Sigma-Aldrich). Blocked IP beads were collected by centrifugation 5 min at 900 g, 4°C prior their use in subsequent IP reactions. IP reactions on pre-cleared lysates was allowed to proceed for 3-4h at 4°C with continuous agitation. Reactions were stopped by 5 min centrifugation at 900 g, 4°C.

For washing of IP beads with standard stringency, collected beads from IP reaction were resuspended in 1.5ml ice cold high salt wash buffer [20mM HEPES pH8.0, 400mM KCl, 1mM MgCl2, 0.2% IGEPAL CA-630, 1mM DTT, 0.1 Unit/ml Ribolock (Fermentas) and 1x CompleteMini Protease Inhibitor (Roche)] Beads were washed for 10min at 4°C with continuous agitation. Washed beads were collected by centrifugation at 4°C, 900 g for 5min, resuspended high salt wash buffer and subjected to two additional rounds of high salt washing steps. The washed beads were transferred to fresh tubes and mixed with 1.5ml low salt wash buffer [20mM HEPES pH8.0, 250mM KCl, 1mM MgCl2, 0.2% IGEPAL CA-630, 1mM DTT, 0.1 Unit/ml Ribolock (Fermentas) and 1x CompleteMini Protease Inhibitor Cocktail (Roche)]. Subsequent washing steps were executed as for high salt washing. To further minimize risk of protein contaminants, washed beads were transferred to new reactions tubes, rinsed with ice-cold 1ml PBS and placed on ice for later use.

For washing of IP beads with high stringency, each individual wash was performed for 15 min. Three series of IP bead washes were applied using high salt wash buffer [20mM HEPES pH 7.8, 750mM NaCl, 0.1% IGEPAL CA-630, 0.1% Na-deoxycholate, 0.1% SDS, 1mM DTT, 0.02mg/ml Heparin, 1x CompleteMini Protease Inhibitor Cocktail (Roche)], medium salt wash buffer [20mM HEPES pH 7.8, 250mM NaCl, 0.1% IGEPAL CA-630, 0.1% Na-deoxycholate, 0.02% SDS, 1mM DTT, 0.02mg/ml Heparin, 1x CompleteMini Protease Inhibitor Cocktail (Roche)] and low salt wash buffer [20mM HEPES pH 7.8, 150mM NaCl, 0.02% IGEPAL CA-630, 0.01% Na-deoxycholate, 1mM DTT, 0.02mg/ml heparin, 1x CompleteMini Protease Inhibitor Cocktail (Roche)]. Herein for each washing series four consecutive washing steps were executed. Upon completion of each washing series, the IP beads were transferred into new clean reaction tubes. Finally, as with standard stringency washings, the washed beads were transferred to new reactions tubes, rinsed with ice cold 1ml PBS and placed on ice for later use. To distinguish direct protein to protein interactions from interactions mediated by RNA templates, beads from the IP reaction were resuspended in 15ml PBS before the washing was performed and supplemented with 100 units of RNaseI (Thermo Fisher). Initial round of RNA fragmentation was allowed for 1h at 4°C under permanent agitation. Subsequent washing was performed under highly stringent washing conditions as described above. After finalization of the last washing step IP beads were resuspended in 1ml PBS and completed with 0.2 units of RNaseI. Final RNA fragmentation was incubated for 5 min at 37°C, with agitation at 1300rpm. The reaction was stopped on ice and IP beads were subjected to an additional round of high stringency washing. After the final wash, the IP beads were rinsed with ice cold 1ml PBS, transferred into fresh tubes and stored on ice for later use.

Choice of EJC bait: GFP-Mago

To test which epitope-tagged EJC subunit, would serve as the optimal IP bait, we established a fly line in which transgenes encoding GFP-Mago, MYC-Y14 and FLAG-HA-eIF4AIII were co-expressed (see Key Resources Table). Expression of GFP-Mago was under control of the mago promoter, while the expression of the other transgenes was driven globally by UAS-GAL4 system (Duffy, 2002; Rørth, 1998). To monitor any treatment-specific effects, experiments were executed simultaneously on dithio(bis-)succinmidyl propionate (DSP) cross-linked (Lomant and Fairbanks, 1976) and native cytoplasmic lysates.

Results obtained from western blotting (Figure S1B) suggested that DSP treatment did not lead to unspecific protein stickiness, as judged from the lack of signal for any of the proteins probed in the control IPs (Figure S1B; lanes 3 and 8). Further confirming stringency of the assay, the non-RNA binding Khc was exclusively detected in lysate "inputs" but not in any of the SDS-PAGE resolved precipitates (Figure S1B; lanes 1, 3-6, 8-11). Independent of the conditions tested, in anti-MYC precipitates we observed strong signals of MYC-Y14 itself and of transgenic GFP-Mago, whereas endogenous eIF4AIII and its transgenic FLAG-HA tagged counterpart were absent or only marginally detected (Figure S1B; lanes 4 and 9). Anti-FLAG precipitates specific for FLAG-HA-eIF4AIII showed strong signals for the transgenic bait itself, but signals for its transgenic and endogenous interaction partners Y14 and Mago were weak when compared with the signals in the input lysate (Figure S1B; lanes 5 and 10). In our hands, the only transgenic bait that displayed sufficient incorporation into the EJC, was GFP-Mago: anti-GFP precipitates showed strong signals for all probed EJC subunits, including GFP-Mago, endogenous eIF4AIII, Y14, as well as the transgenic MYC-Y14. anti-GFP precipitates also showed reproducibly weak signals for FLAG-HA-eIF4AIII upon DSP cross-linking, but only to an extent similar to what was observed within anti-MYC IPs (Figure S1B; lanes 6 and 11). Interestingly in all IPs except the control we detected signals for the genuine mRNP component poly-A binding protein PABP. (Figure S1B; lanes 1, 3-6, 8-11). For GFP-specific IPs this was not surprising, due to the efficient incorporation of GFP-Mago into cytoplasmic EJCs. However for IPs in which FLAG-HA-eIF4AIII served as bait and no other EJC subunits were detected, this result was not anticipated (Figure S1B; lanes 5 and 10), and questioned the restrictedness

of eIF4AIII to EJC (see EJC Sedimentation). Conversely, the efficient co-IP of EJC subunits with GFP-Mago qualified to us GFP-Mago as the bait of choice for subsequent EJC specific IP assays.

Sucrose density gradient centrifugation

Pulverized flies were resuspended in lysis buffer supplemented with 0.4mM cycloheximide (Sigma-Aldrich). All subsequent steps for cellular fractionation were executed as described above. To resolve cytoplasm into fractions containing mRNPs, ribosomes or polysomes, typically 150μl of lysate were loaded on a 3.4ml 10%–50% linear sucrose gradient. Sucrose gradients were made using 20mM HEPES pH8.0, 150mM KCl, 4mM MgCl2, 0.4mM Cycloheximide, 0.0.5% IGEPAL CA-630, 0.2% Na-deoxycholate, 1 Unit/ml Ribolock (Fermentas), and 1x CompleteMini Protease Inhibitor (Roche) using a gradient mixer (Biocomp) and fraction collector (Biocomp), following the manufacturer's protocol. Protein-RNA complexes were separated by ultracentrifugation in a SW60Ti rotor (Beckman) at 50000rpm, at 4°C for 30 min. Fractions were collected batch-wise in 100μl aliquots. Nucleic acid content in the fractions was measured manually at 254nm. For protein analysis, fractions were supplemented with trichloracetic acid (30%) and kept on ice for 15 min to ensure protein precipitation. Protein pellets were subsequently washed with ice-cold acetone and denatured for 10 min at 95°C in 2x LDS sample buffer (Thermo Fisher), 10mM DTT.

EJC sedimentation: eIF4AIII is a poor bait

Since not only FLAG-HA (see Pilot IP) but also other tagged versions of eIF4AIII failed to incorporate efficiently into endogenous EJCs (data not shown), we wondered whether the RNA-binding DEAD-box helicase might also be associated with RNAs in an EJC-unrelated manner. We therefore compared sedimentation profiles of endogenous cytoplasmic Y14 and eIF4AIII obtained by ultracentrifugation of cytoplasmic lysates in sucrose density gradients, and monitored the content of nucleic acids by UV absorption and accompanying content of proteins by western blotting (Figure S1C). Supporting the notion that the EJC is a component of mRNPs, Y14 co-sedimented with eIF4AIII in light mRNP fractions (Figure S1C; lanes 2-5). Contradicting, however, the exclusive function eIF4AIII in EJC related processes, we also observed eIF4AIII (but not Y14) co-sedimenting with ribosomal subunit proteins such as RpS6 and RpL32 in the 40/48 s and 60 s fractions, and with high-density polysomes (Figure S1C, lanes 2-17, 28-30, 32). This unexpected result indicated to us a yet undefined function for *Drosophila* eIF4AIII outside of an EJC context. This for us disqualified the DEAD-box RNA helicase eIF4AIII as bait for EJC specific immunoprecipitations.

WESTERN ANALYSIS AND ANTIBODIES

Washed IP beads were mixed with 2x LDS sample buffer (Thermo Fisher) containing 50mM DTT (Sigma-Aldrich). Eluted mRNPs from oligo-d$(T)_{25}$ precipitates samples as well as input samples were mixed in 1:1 ratio with 4X LDS sample buffer, 100mM DTT. To allow reduction of covalent DSP mediated bonds, the samples were incubated for 20min at 42°C prior to denaturation for 10min at 95°C. To test protein content and presence of candidate proteins by western blotting in inputs and precipitates, samples were resolved by standard SDS-PAGE and visualized by silver staining (GE Healthcare) or blotted onto PVDF membrane (Millipore) and blocked, following instructions in the manufacturer's protocols.

All antibodies utilized for western detection were diluted in PBS, 5% dry milk, 0.1% TWEEN (Thermo Fisher). The primary antibodies (and dilutions) were: rat anti-Y14 (1:2500), rabbit anti-Mago (1:2000, gift of M. Blanchette), rabbit anti-eIF4AIII (1:4,000; gift of I. Palacios), rabbit anti-PABP (1:4000; gift of M. Hentze), rabbit anti-RpS6(1:3000, Cell Signaling) rabbit anti-RpL32 (1:2000; gift of M. Hentze), rabbit anti-Kinesin heavy chain (KHC, 1:25000; Cytoskeleton), rabbit anti-GFP (1:2000, Torrey Pines). Goat anti-rabbit (1:2500) and anti-rat (1:2500) conjugated with HRP (GE Healthcare) were used as secondary antibodies.

Mass spectrometry

For tandem mass spectrometry, immunoprecipitates were submitted for further preparation and analysis to the EMBL Proteomics Core facility. All reagents were prepared in 50 mM HEPES (pH 8.5). For reduction of cysteines, dithiothreitol was added (56°C, 30 minutes, 10 mM); further alkylation was performed using iodacetamide (10 mM, for 30 minutes in the dark, at room temperature). Samples were prepared for LC-MS/MS using the SP3 protocol (Hughes et al., 2014), digestion was performed using trypsin (1:50 enzyme:protein ratio) at 37°C overnight. TMT10plex Isobaric Labeling (ThermoFisher) was performed according the manufacturer's instructions. The OASIS® HLB μElution Plate (Waters) was used for sample clean-up. Offline high pH reverse phase fractionation was performed as described previously (Reichel et al., 2016). In brief, the samples were run on an Agilent 1200 Infinity high-performance liquid chromatography (HPLC) system equipped with a Gemini C18 column (3 μm, 110 Å, 100 × 1.0 mm, Phenomenex). The solvent system consisted of 20 mM ammonium formate (pH 10.0) as mobile phase-A, and 100% acetonitrile as mobile phase-B. After offline fractionation, peptides were separated using the UltiMate 3000 RSLC nano LC system (Dionex) fitted with a trapping cartridge (μ-Precolumn C18 PepMap 100, 5μm, 300 μm i.d. x 5 mm, 100 Å) and an analytical column (Acclaim PepMap 100 75 μm x 50 cm C18, 3 μm, 100 Å). The outlet of the analytical column was coupled directly to a QExactive plus (Thermo) using the proxeon nanoflow source in positive ion mode. Solvent A was water, 0.1% formic acid and solvent B was acetonitrile, 0.1% formic acid. Trapping time was 6 minutes at a constant flow of solvent A at 30 μL/ min onto the trapping column. Peptides were eluted via the analytical column at a constant flow of 0.3 μL/ min. During elution, the percentage of solvent B increased in a linear fashion from 2% to 4% B in 4 minutes, from 4% to 8% in 2 minutes, then 8% to 28% for a further 96 minutes, and finally from 28% to 40% in another

10 minutes. Column cleaning at 80% B followed, lasting 3 minutes, before returning to initial conditions for the re-equilibration, lasting 10 minutes. The peptides were introduced into the mass spectrometer (QExactive plus, ThermoFisher) via a Pico-Tip Emitter 360 μm OD x 20 μm ID; 10 μm tip (New Objective) and a spray voltage of 2.3 kV was applied. The capillary temperature was set at 320°C. Full scan MS spectra with mass range 350-1400 m/z were acquired in profile mode in the FT with resolution of 70,000. The filling time was set at maximum of 100 ms with a limitation of $3x10^6$ ions. DDA was performed with the resolution of the Orbitrap set to 35000, with a fill time of 120 ms and a limitation of $2x10^5$ ions. Normalized collision energy of 32 was used. A loop count of 10 with count 1 was used and a minimum AGC trigger of $2e^2$ was set. Dynamic exclusion time of 30 s was applied. The peptide match algorithm was set to 'preferred' and charge exclusion 'unassigned', charge states 1, 5 - 8 were excluded. Isolation window was set to 1.0 m/z and 100 m/z set as the fixed first mass. MS/MS data was acquired in profile mode.

EJC ipaRt and mRBP footprinting

For EJC ipaRt and mRBP footprining assays, DSP was added directly to Lysisbuffer prior to preparation of the fly cell homogenate (compare with Cytoplasmic Lysate preparation). Crosslinking was allowed for 1h at 4°C and stopped by addition of 25mM Tris-HCl pH7.5 prior clarification of cytoplasmic lysates. All steps for cDNA library preparation were adapted and modified from the previously published iCLIP protocol described by König and co-workers (Konig et al., 2011). Barcoded primer sequences for reverse transcription, cut-oligo sequence and cDNA amplification primers for Illumina sequencing are summarized in Table S3 in the section supplementary information.

EJC ipaRt and L3-App adaptor ligation

For isolation of protein (holo-)complexes associated with RNA templates (ipaRt), DSP treated cytoplasmic extracts were supplemented with 0.02% heparin. Subsequently, cytoplasmic fractions from GFP-Mago and GFP expressing flies were mixed with 100μl equilibrated bead slurry of protein A/G agarose (1:1 mixture) for pre-clearing. Pre-clearing proceeded for 1h at 4°C with continuous agitation. GFP-trap beads (Chromotek) and Protein A/G (Roche) the ''beads-only'' control were blocked for 1h at 4°C in 20mM HEPES pH7.8, 125mM KCl, 0.1% IGEPAL CA-630, 0.1mg/ml Heparin, 4mM $MgCl_2$, 1x CompleteMini Protease Inhibitor Cocktail (Roche), 20 Units/ml Ribolock. Next, 15ml of heparin complemented and pre-cleared cytoplasmic lysate were mixed with 20μl immunoprecipitation resin. IP reactions were incubated for 4h at 4°C with constant agitation. Washing buffer compositions in ipaRt were the same as those used for washing IPs under highly stringent conditions (see above). However, each complete wash consisted of a series of 5 repeated washes. To reduce the likelihood of transfer of nucleic acid contaminants, at the end of each a wash step the IP bead slurry was transferred into a clean new reaction tube. After the final washing step bead aliquots were transferred into clean new reaction tubes and adjusted to 1ml isotonic reaction buffer [20mM HEPES pH 7.8, 150mM KCl, 4mM $MgCl_2$, 0.02% IGEPAL CA-630, 1x CompleteMini Protease Inhibitor Cocktail (Roche)] containing 1 Unit RNaseI and 4 Units TURBO DNaseI (Thermo Fisher). The RNase-DNase reaction was allowed to proceed for 5min at 37°C with continuous agitation (1200rpm). To stop the RNase reaction, ipaRt bead aliquots were placed on ice, immediately rinsed with high salt wash buffer and transferred to new clean reaction tubes. Subsequently, ipaRt beads were subjected to two rounds of alternating high salt and low salt washing steps. After every individual washing step, the ipaRt beads were transferred into new clean reaction tubes. When washing was completed, beads were rinsed twice with 1ml isotonic reaction buffer [20mM HEPES pH 7.8, 150mM KCl, 4mM $MgCl_2$, 0.02% IGEPAL CA-630, 1x CompleteMini Protease Inhibitor Cocktail (Roche)] and placed on ice for later use. For 3′ end de-phosphorylation of protected RNA fragments, ipaRt beads were directly taken up in 50 μl T4-PNK reaction mix [70mM Tris HCl pH6.5, 10mM $MgCl_2$, 5mM DTT, 10 Units T4-PNK (Fermentas), 20 Units Ribolock (Fermentas)]. De-phosphorylation proceeded for 45min at 37°C and continuous shaking at 1100rpm. The T4 PNK reaction was stopped by placing the samples on ice. Immediately thereafter, the beads were exposed to two rounds of repetitive high salt and low salt washing steps (as above) and finally rinsed with ice-cold PBS. For the ligation of L3-App DNA adaptor (Konig et al., 2011) to co-immunoprecipitated 3′ end dephosphorylated RNA fragments, ipaRt beads were resuspended in 40μl T4 RNA ligation mix composed of 50mM Tris-HCl pH 7.5, 10mM $MgCl_2$, 10mM DTT, 5% PEG400 (Sigma-Aldrich), 20 Units RNaseOUT (Sigma-Aldrich), 1.5μM L3-App adaptor and 10 Units T4 RNA Ligase 2 (truncated) (NEB). Ligation was allowed to proceed for 15h at 16°C, with constant agitation at 1100rpm. To remove non-ligated L3-App DNA adaptor after ligation, ipaRt beads were subjected to 2x high salt, 2x medium salt and 2x low salt washing steps, with a transfer of the beads into new clean reaction tubes every second wash. To extract the L3-App-RNA ligation products, ipaRt beads were mixed with 200μl proteinase K reaction buffer [100mM Tris-Cl pH 7.4; 50mM NaCl; 10mM EDTA] complemented with 200μg proteinase K (Fermentas). The proteolysis reaction was incubated for 20min at 37°C, 1100rpm, then supplemented with 200μl proteinase K urea reaction buffer [100mM Tris-HCl pH 7.4, 50mM NaCl, 10mM EDTA, 6M urea] and incubated for an additional 45min at 37°C, with constant shaking at1100rpm. For extraction of RNA-DNA ligation products, the reaction samples were vortexed 5 min with 400μl phenol/chloroform/isoamylalcohol (25:24:1) following standard DNA extraction protocols. RNA-DNA ligation products were buffered with sodium acetate pH5.2 and finally subjected to ethanol precipitation at −20°C over-night using 5μg linear acrylamide (Thermo Fisher) as carrier.

mRBP-footprinting and L3-App adaptor ligation

mRBP footprinting on poly-A tail precipitated mRNPs (see above) was performed using RNaseI. To do so, mRNP elution aliquots were mixed and adjusted to 1ml isotonic buffer [20mM HEPES pH 7.8, 150mM KCl, 4mM $MgCl_2$, 0.02% IGEPAL CA-630, 1x CompleteMini Protease Inhibitor Cocktail (Roche)] containing 1 Unit RNaseI (NEB) and 4 Units TurboDNaseI (Thermo Fisher).

RNase-DNase digestion was allowed to proceed for 5min at 37°C with constant agitation (1200rpm). RNA digestion was stopped by placing samples on ice and adding 0.5% LiDS. To separate short RNA digestion products from mRNA fragments in complex with RBPs, samples were placed into an Amicon Ultra 10K concentrator columns (cutoff 10kDa, Millipore) and centrifuged at 4°C for a final volume of 50μl according to the manufacturer's instructions. The resulting concentrates were dissolved in 10 volumes of HE buffer containing 0.5% LiDS and subjected to a second round of concentration. Finally, 50μl mRBP-RNA concentrates were supplemented with 5μl 1M DTT, incubated 20 min at 42°C to reverse the DSP mediated protein-protein cross-linking. Upon reduction of covalent bonds, samples were supplemented with 200μl proteinase K reaction buffer [100mM Tris-HCl pH 7.4, 50mM NaCl, 10mM EDTA, 0.2% SDS], 10μl proteinase K (20mg/ml, Fermentas) and incubated for 20min at 37°C. Next, 200μl proteinase K urea reaction buffer [100mM Tris-HCl pH 7.4, 50mM NaCl, 10mM EDTA, 6M urea, 0.2% SDS] was added and protein digestion was allowed to proceed for additional 40 min at 37°C. To stop protein digestion, samples were diluted with 3 volumes of Trizol LS (Thermo Fisher). With the exception of the overnight RNA precipitation at −20°C, all steps to isolate RBP protected mRNA fragments were performed according instructions in the Trizol LS manual (Thermo Fisher). For removal of 3′ phosphorylated ends, after RNaseI digestion RNA the pellets were resuspended directly in 50 μl ddH_2O, and adjusted to a final volume of 100μl with 70mM Tris HCl pH6.5, 10mM $MgCl_2$, 5mM DTT, 10 Units T4-PNK (Fermentas), 20 Units Ribolock (Fermentas). The PNK reaction was allowed to proceed for 45 min at 37°C, with agitation at 1100 rpm and stopped by addition of 3 volumes of Trizol LS (Thermo Fisher). As described above, all steps for recovery of RNA except for overnight RNA precipitation at −20°C were performed following the instructions in the Trizol LS manual (Thermo Fisher). Pellets of de-phosphorylated and cleaned RNA fragments were resuspended in 20μl ddH_2O. For ligation of the L3-App adaptor (Konig et al., 2011, see oligonucleotide list in Supplementary Information), RNA samples were adjusted to a 40μl ligation reaction composed of 50mM Tris-HCl pH 7.5, 10mM $MgCl_2$, 10mM DTT, 5% PEG400 (Sigma-Aldrich), 20 Units RNaseOUT (Sigma-Aldrich), 1.5μM L3-App adaptor and 10 units of T4 RNA Ligase 2 (truncated) (NEB). Ligation proceeded for 15h at 16°C with constant agitation (1100rpm). To reduce the content of non-ligated L3-App adaptor in the samples after completion of the ligation reaction, aliquots were mixed with 1ml HE elution buffer [20mM HEPES pH8.0, 1mM EDTA] and concentrated in Amicon Ultra 10K concentrator columns until a volume of 100μl was reached. RNA-DNA ligation products were extracted with a mix of phenol, chloroform and isoamyl alcohol (25:24:1) mixed with 0.1 volumes 3M Na-Acetate pH 5.2, 5μg linear acrylamide (Thermo Fisher) and finally precipitated in 2 volumes ethanol over night at −20°C.

cDNA library preparation

Barcoded primer sequences for reverse transcription, cut-oligo sequence and cDNA amplification primers for Illumina sequencing are listed in Supplementary Information and in the published iCLIP protocol by König and co-workers (Konig et al., 2011). First, pellets of RNA-L3-App products were washed in 80% ice-cold ethanol and subsequently resuspended in 10μl ddH_2O. For reverse-transcription, the SuperScript III First-Strand Synthesis SuperMix kit (Invitrogen) was utilized. Herein 8μl of sample solution were mixed with 1μl solution of RTClip-Primer mixture (each 0.5 μM) and 1μl Annealing cocktail (Invitrogen). To anneal the RTClip primer with complementary sites in the L3-App adaptor sequence, sample-primer mixture aliquots were denatured for 3 min at 70°C and allowed to anneal at 25°C for 10 min. Subsequently, samples were supplemented with 10μl 2x Reaction Mix (Invitrogen) and 2μl Superscript III/RNaseOUT™ enzyme mix (Invitrogen). Reverse transcription was performed in three steps using a C1000 Touch Thermal Cycler (Biorad): 1st for 20 min at 25°C, 2nd for 30 at 42°C and 3rd for 60min at 50°C. The RT reaction was stopped with a 3 min pulse of denaturation at 90°C. To avoid re-hybridization of RNA and cDNA, reverse transcribed samples were immediately placed on ice, topped up to a volume of 100μl with ice-cold ddH_2O, and supplemented with 500μl guanidine-HCl-containing PB buffer from the PCR purification kit (QIAGEN). Samples were thereafter transferred into QIAquick columns (QIAGEN) and washed according to the manufacturer's instructions. Elution was performed in two consecutive steps using at each step 10μl warm (60°C) elution buffer (QIAGEN). For circularization of the cDNA products, purified samples were adjusted to a 40μl final reaction volume composed of 1X CircLigase Buffer II (Epicenter), 500μM $MnCl_2$, 60 Units CircLigase II (Epicenter). cDNA circularization proceeded for 1h at 60°C. After circularization, the samples were mixed with 11μl ddH2O, 4.5μl 10x restriction Buffer 4 (NEB), 1.5μl Cut-Oligo [10μM, see table] and incubated in a C1000 Touch Thermal Cycler (Biorad) applying a program comprising an initial 4min, 85°C denaturation followed by sequential cooling to 25°C in 1°C steps every 10 s. After hybridization, the samples were completed with 3 μL BamHI HF (NEB) for digestion of Cut-Oligo RTClip primer hybridization sites. Linearization was carried out for 30min at 37°C. Linearized barcoded cDNAs were subjected for extraction to the QIAquick PCR purification protocol (QIAGEN). cDNA elution was performed in two consecutive steps using at each step 14μl warm (60°C) elution buffer (QIAGEN). For cDNA amplification 14μl eluted sample were mixed with an Illumina sequencing compatible P3/P5 primer mixture [1μM final concentration each] and 1x Phusion Flash High-Fidelity PCR Master Mix (Thermo Fisher) to a final volume of 20μl. Amplification was performed in a C1000 Touch Thermal Cycler (Biorad). Initial denaturation was performed for 20 s at 98°C, followed by 23-26 amplification cycles. Each cycle comprised 1 s denaturation at 98°C, 10 s annealing at 65°C and 15 s elongation at 72°C. The final elongation step was extended to 1 min. PCR products were desalted and cleaned using the QIAquick PCR purification protocol (QIAGEN). Purified PCR products were resolved by standard gel-electrophoresis in 1x TBE, 4% MetaPhor Agarose (Lonza) gels at 4°C. To obtain an optimal insertion size and exclude primer duplicates, cDNA amplicons migrating between 200 and 300 bp were excised and extracted by MinElute Gel Extraction kit (QIAGEN). Depending on the yield, 5-12ng of purified cDNA library were obtained and submitted to the EMBL GeneCore Facility. Strand specific libraries were sequenced single-end with 55 bp on an Illumina HiSeq2000.

mRNA Sequencing

Total mRNA was extracted from whole flies using Trizol LS (Thermo Fisher) according to the manufacturer's instructions. 10μg of purified total RNA were depleted of rRNA and small RNAs through two consecutive poly-A mRNA capture steps, using oligo-dT_{25} coated magnetic Dynabeads (Sigma-Aldrich) according to the manufacturer's protocol. 500ng purified mRNA were submitted for library preparation. Barcoded stranded mRNA-seq libraries were prepared using the Illumina TruSeq RNA Sample Preparation v2 Kit (Illumina, San Diego, CA, USA) implemented on the liquid handling robot Beckman FXP2. Obtained libraries were pooled in equimolar amounts; 1.8 pM solution of this pool was loaded on an Illumina NextSeq 500 sequencer and sequenced bi-directionally, generating ~500 million of paired reads, each 85 bases long.

Computational data processing

Computational analysis was executed in R (R Core Team, 2017). Representative plots were generated with Microsoft Excel or with ggplot2 (Wickham, 2009) package in R. For individual R packages see Key Resources Table and Results section.

PROCESSING OF MASS SPECTROMETRY DATA

Acquired data were processed by IsobarQuant (Franken et al., 2015) and Mascot (v2.2.07), searched against a Uniprot *Drosophila melanogaster* proteome database (UP000000803) containing common contaminants and reversed sequences. The data were searched with the following modifications: Carbamidomethyl (C) and TMT10 (K) (fixed modification), Acetyl (N-term), Oxidation (M) and TMT10 (N-term) (variable modifications). The mass error tolerance for the full scan MS spectra was set to 10 ppm and for the MS/MS spectra to 0.02 Da. A maximum of 2 missed cleavages was allowed. For protein identification a minimum of 2 unique peptides with a peptide length of at least seven amino acids and a false discovery rate below 0.01 were required on the peptide and protein level. To define relative enrichment of candidates in GFP-Mago precipitates versus GFP-control, detected proteins were subjected to expression set analysis by Limma (Ritchie et al., 2015). Limma results, including enrichment, significance and adjusted p value, are summarized in the Table S4: Limma Results from TMT labeled mass spectrometry.

Sequencing read mapping

mRBP footpring and ipaRt reads were demultiplexed using the sample barcode and the unique molecular barcode was appended to the read name in the fastq file. These reads are then trimmed using fqtrim (https://ccb.jhu.edu/software/fqtrim/) with the adaptor sequence AGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG. Trimmed reads shorter than 20bp were discarded. The ipaRt, mRBP-footprinting and RNA-seq reads from Illumina deep sequencing were mapped to the *Drosophila melanogaster* reference genome (dm6, gtf from BGDP version 81) using the transcriptome alignment option in TopHat2 (Kim et al., 2013). Only uniquely mapped reads were used, and in the ipaRt and mRBP-footprinting samples duplicates were removed using the random barcode at the start of the forward read. For analysis of human iCLIP and mRNA-seq data from Hauer et al. (2016), reads from GEO accession GEO: E-MTAB-4215 were processed the same way, using hg38 and gtf annotation 82 from ensembl.

Assay quality control: gene class enrichment

Stabilization of protein-protein interactions by DSP cross-linking in ipaRt and mRBP footprinting (Figure S1D) might result in the trapping of unwanted proteins and RNPs such as ribosomes (and RNAs within) within larger mRNP composite complexes. Although we observed a general reduction of EJC unrelated proteins in GFP-Mago precipitates upon RNaseI fragmentation (see main result section and Figure S2A), we found that approximately 8% of all sequencing reads from EJC ipaRt and mRBP footprinting were mapping non-uniquely in the *Drosophila* genome (Figure S2B) to intergenic, intronic and non-coding RNA sites (Figure S2C). This suggests that, to a certain extent, mRNP-unrelated RNP complexes were co-precipitated in both EJC ipaRt and mRBP footprinting assays.

To test whether the minor abundance of reads mapping at multiple site in the genome was due to DSP cross-linking, we analyzed sequencing results from EJC ipaRt, mRBP footprinting and mRNA-Seq (Figure S4A) for gene-class enrichment by expression set analysis using DESeq (Anders and Huber, 2010; Love et al., 2014). DESeq analysis of EJC ipaRt and mRNA-Seq expression analysis revealed an EJC enrichment for non-coding transcripts such rRNAs, tRNAs, snRNAs, while the protein-coding and non-coding gene classes transcribed by RNA Pol II appeared either unaffected or depleted, as was the case of single exon-genes (Figure S4B). Just as with EJC ipaRt, the DESeq analysis of sequencing results from mRBP-footprinting and mRNA-Seq showed an mRBP enrichment bias for rRNA and small non-coding RNAs (Figure S4C). Nevertheless, when we defined gene class enrichment by direct comparison of EJC ipaRt with mRBP footprinting, both of which were treated with DSP, primarily mRNAs of multi-exon protein coding genes showed a bias for EJC enrichment, while mRNAs of single-exon genes and genes of non-RNA Pol II transcripts were either unaffected or depleted (Figure S4D).

These findings suggest that for reduction of false positive estimates in RNA immunoprecipitation assays in which protein-protein crosslinking is implemented, a direct comparison of specific IP libraries with mRBP protected fragments is more appropriate than comparing with mRNA-Seq libraries. Given the strong enrichment of protein coding multi-exon genes in the EJC associated versus mRBP protected RNA fragments, we focused all subsequent analyses on this class of genes.

Sequencing read coverage across exon junctions

For exons of at least 100nt, coverage measurement within 50nt upstream and 50nt downstream of exon-exon junction was calculated from each bam file using coverage and Views function of the GenomicAlignments package (Lawrence et al., 2013) in R, while for exons of > 200nt length, coverage measurement within 100nt upstream and downstream of exon-exon junction was calculated. This per nt coverage was utilized for definition of coverage median coordinates within exon-exon junctions in EJC ipaRt and mRBP footprinting assays. When a 5′ splice site was observed to undergo splicing reaction with 2 alternate 3′ splice sites, we assigned coverage within 5′ exons to corresponding exon-exon junction isoforms based on the ratio of reads that span either junction. The reciprocal approach was executed when a 3′ splice site was observed to undergo splicing with alternate 5′ splice sites.

EJC protection site peak calling

Replicates were combined and the genomic coverage was calculated for the ipaRt samples. The genomic coverage was translated onto the transcriptome coordinates using GenomicFeatures (Lawrence et al., 2013) in R. We considered only transcripts that have a TPKM > 1. For each transcript, regions with more than 2x the mean coverage of the transcript were identified, and the maxima position called out for each of these regions, and defined as peak position. To estimate the $\log_2$-fold change of a peak coordinate in ipaRt over mRBP footprinting, we took the surrounding 20nt (+- 10nt) of each peak position, and calculate the $\log_2$fold change in read counts across all replicates. For final consideration of ipaRt peaks, we retained only peaks with a coverage of at least 30 and $\log_2$-fold change over mRBP > 1.

DIFFERENTIAL EXPRESSION SET ANALYSIS

Genes expression was estimated by counting reads that overlapped with all annotated genes in BGDP version 81 gtf, using the countOverlaps function in GenomicAlignments. DESeq2 (Love et al., 2014) was used for estimation of $\log_2$-fold change between ipaRt and mRBP footprinting or total mRNA. Samples were normalized using total library size for analysis across all genes (Figures S4B–S4D). For assessing EJC enrichment in protein coding genes, the default median normalization in DESeq2 was used. To estimate EJC enrichment at exon-exon junctions, reads that overlapped within 50bp upstream and 50bp downstream of exon-exon junction were counted similarly using the countOverlaps function, and DESeq2 was used on this count table.

Gene feature annotation

For each gene, we estimated the most abundant transcript using Kallisto (Bray et al., 2016) and used that as a representative of the gene. From this representative transcription, we calculated the number of introns, transcript length, maximum intron length of the gene. The transcript abundance of the gene was tpkm of the most abundant transcript. The Shannon entropy of each gene was used as an estimate of the degree of splicing for each gene. This was calculated using the transcript abundance estimate obtained from Kallisto.

Splicing analysis and ISE, ESE, ESS

Prediction of 5′ splice strength was performed using MaxEntScan. (Yeo and Burge, 2004). Annotation of junctions with respect to ISE, ESE and ESS was done using hexamers provided from Brooks et al. (2011) and Wang et al. (2004).

Assessing features that explain EJC binding

For assessing how different features explain EJC binding at the RNA level, a linear model was fitted, with the $\log_2$-fold change (estimated from DESeq) between ipaRt and mRBP footprinting regressed against number of introns, maximum intron length of transcript, transcript abundance, mRNA length and degree of alternative splicing. These features were estimated from RNaseq data (see Gene feature annotation section) and relative importance of each feature was estimated using the R relaimpo package (Groemping, 2006). For assessing different features that explain EJC binding at the junction level, a linear model was fitted, with the delta $\log_2$-fold change (estimated from DESeq between junction and gene EJC enrichment) regressed against splicing related features (see Splicing analysis and ISE, ESE, ESS), folding related features (see Prediction of RNA structures and transformation using Gaussian Mixture Model) and counts of the hexamers we identified. All features were scaled and centered before linear regression using lm() function in R.

RNA structure prediction and clustering

RNA structure prediction was performed using Vienna RNA Package (Lorenz et al., 2011). For each exon-exon junction in *Drosophila*, the 37nt upstream of the junction and the 28nt downstream of the junction were merged. Secondary structure of this 67nt long sequence was then predicted by using the following command line: "RNAplfold -T 21 -W 66 -u 1," which assumes folding at 21°C. We obtained the pairing probability of each base pair and performed logit transformation on these probabilities. To reduce dimension on this bpp data, we fitted a Gaussian Mixture Model assuming equal variance of the clusters ("EII" model). The probabilities of the junctions falling into each category would be used in subsequent analysis as a variable for predicting junction EJC deposition. We tried 2 to 6 clusters, at each try regressing the $\Delta\log_2$-fold change against these probabilities. At 4 clusters, we observed no substantial increase in R^2 and used this number of clusters for subsequent analysis.

Tree model for mRNA localization

Applied decision tree learning (Breiman et al., 1984) was applied on various data groups using the R package rpart (Therneau and Atkinson, 2018) with RNA localization as a binary response. The ratio of localized to non-localized dataset is ~1:9, so Cohen's Kappa coefficient was used to assess the agreement between the model's prediction and observed data. Different weights were assigned to localizing and non-localizing mRNAs to ensure the model emphasize on getting localizing mRNAs correct. To obtain an optimal weight, we performed 1000 bootstraps (with replacement) across all data groups and in each bootstrap, we ensure that the non-localizing and localizing ratios are preserved. This bootstrapping was used for subsequent analysis as well. Based on getting the average Kappa across all data groups, a final weight of 1 for non-localizing mRNA and 3.4 for localizing mRNA was used.

QUANTIFICATION AND STATISTICAL ANALYSIS

Software and statistical analysis details can be found in the Key Resources Table, Results, and Computational data processing section (see above). All statistical analyses were performed in R using the stats package (version 3.3.3 and 3.5.0), with the numbers tested indicated in the main or supplementary figures. Changes in EJC enrichment were analyzed using the DESeq2 package (Love et al., 2014). Two-sided t.test for group comparisons were performed using the t.test() function, correlation was estimated using the cor.test() function and the ANOVA analysis used the lm() and anova() functions. Linear regression was performed using lm() function and t-statistics for each coefficient was obtained using the summary.lm() function. Boxplot elements show the median (black line) and quantile values (box denotes 25th and 75th quantile), with outliers shown as black dots outside of the box whiskers. Violin plots show median (black dot), 25th and 75th quantile (black line) and distribution of the groups. For Figures 6A and 6A, 2275 significantly enriched junctions (FDR < 0.05 and delta $\log_2$-fold change > log2(1.5)) were tested against 4334 depleted junctions (FDR < 0.05 and delta $\log_2$-fold change > log2(1.5)). For Figure 6C, the number of junctions in each folding category is as follows: 1. 10220, 2. 2220, 3. 1009, 4.1737.

DATA AND CODE AVAILABILITY

The accession number for the FASQ files of ipaRt, mRBP footprinting and mRNA-Seq reported in this paper is European Nucleotide Archive PRJEB26421: https://www/ebi.ac.uk/ena/data/PRJEB26421.