EXPERIMENTAL PROCEDURES

Plasmids, Cell lines and cell culture

Plasmids expressing EJC factors and hnRNPA1 were generated by cloning their cDNAs into derivatives of pcDNA3.1 (for transient expression) or pcDNA5-FLAG-FRT/TO (for stable expression) as described previously (Singh et al., 2012). Plasmids expressing β-globin reporters have been described elsewhere (Lykke-Andersen, 2000; Singh, 2007). Human cell lines (HEK293 Flp-In TRex, HeLa CCL2, HeLa Tet-off) were cultured in Dulbecco's modified eagle medium (DMEM) supplemented with 10 % fetal bovine serum (FBS) and 1 % penicillin-streptomycin. Stable cell lines expressing tetracycline-inducible FLAG-tagged proteins were created using HEK293 Flp-In TRex cells as described previously (Singh et al., 2012).

Endogenous Immunoprecipitations

HEK293, HeLa and P19 cells were lysed and sonicated in hypotonic lysis buffer (HLB) [20 mM Tris-HCl pH 7.5, 15 mM NaCl, 10 mM EDTA, 0.5 % NP-40, 0.1 % Triton X-100, 1 x Sigma protease inhibitor cocktail, 1 mM PMSF]. Lysates were sonicated, increased to 150 mM NaCl and treated with RNase A for five minutes. Complexes were then captured on Protein A/G Dynabeads (Life Technologies) conjugated to IgG, α-EIF4A3, α-CASC3 or α-RNPS1 antibodies for 2 hr nutating at 4 °C. Complexes were washed in isotonic wash buffer (IsoWB) [20 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.1 % NP-40] and eluted in clear sample buffer [100 mM Tris-HCl pH 6.8, 4 % SDS, 10 mM EDTA, 100 mM DTT]. Cortical neurons were isolated from FBV wild-type male mice in HBSS buffer. Once isolated, lysate preparation was carried out as stated above.

FLAG Immunoprecipitations

Stable HEK293 cells expressing FLAG-tagged EJC protein were lysed in HLB. Lysates were sonicated, increased to 150 mM NaCl and treated with RNase A. Complexes were then captured on FLAG-beads, washed with IsoWB and FLAG-peptide eluted. IPs were carried out as noted above.

Mass Spectrometry

FLAG Immunoprecipitation

Stably expressed FLAG-MAGOH, FLAG-CASC3, FLAG-RNPS1, or FLAG-peptide were IPed and RNase A digested at 4 °C for 2 hr in lysis buffer supplemented with 1 µg/ml FLAG peptide to improve IP specificity. The FLAG affinity elution was completely dried by vacuum evaporation, re-suspended in 100 µl of water and dialyzed (dialysis buffer: 10 mM Tris-HCl pH7.5, 75 mM NaCl, 0.01 % Triton X-100) for ~6 hr at 4 °C in a MWCO 7,000 Da mini dialysis column (Pierce). The dialyzed sample (90-100 µl) was again completely dried by vacuum evaporation and re-suspended in 20 µl of 0.1 % SDS and 10 mM DTT (prepared from a fresh stock solution). The sample was heated at 95 °C for 5 min and cooled to room temperature. The reduced thiol groups were alkylated by incubating with 0.8 µl of freshly prepared 1M iodoacetamide at room temperature for 45 min in dark. The resulting samples were mixed with 5 µl of 4 x Lammelli SDS load buffer (Bio-Rad) and loaded on 4 %–15 % Mini-PROTEAN TGX gel (Bio-Rad). The samples were migrated until the dye-front had run approximately 1 cm into the gel from the bottom of well. The gel was washed three times with ~200 ml HPLC grade water for 5 min each. The gel piece containing protein was excised and processed for in gel digestion of proteins.

In-Gel Digestion

The gel slices were cut into 1×5×1mm^3 dimension and transferred into a 1.5 mL microcentrifuge tube. Samples were washed with water and dehydrated in acetonitrile :50 mM NH_4HCO_3 (1:1 v/v) for 5 minutes and 100% acetonitrile for 30 seconds. Following drying with a vacuum concentrator (Savant), gel slices were rehydrated in 25 mM dithiothreitol in 50 mM NH_4HCO_3 and incubated for 20 min at 56 °C. With the removal of the supernatant, 55 mM iodoacetamide in 50 mM NH_4HCO_3 was added and the samples were incubated in the dark for 20 min at room temperature. Gel slices were washed and dehydrated as described before. Following removal of the liquid by vacuum concentrator, the sample was rehydrated in 12 ng/µL Trypsin Gold (Promega) in 0.01 % ProteaseMAX™ Surfactant (Promega): 50 mM NH_4HCO_3 and incubated at 50 °C for one hr. Condensate was collected by centrifuging at 14,000 × g for 10 seconds and the solution was transferred to a new tube. Trifluoroacetic acid was added to a final

concentration of 0.5 % to inactivate trypsin and the solution was dried by vacuum concentrator.

LC/MS/MS

The LC-MS/MS tryptic digested peptides were dissolved in 0.1 % formic acid (v/v) and loaded at 15 µL/min for 10 min using a nanoACQUITY C18 Trap column (20×0.18 mm i.d., 5 µm 100Å C18, Waters, 186007238). Peptides were separated using an EASY-Spray LC Column(150×0.075 mm i.d., 3 µm 100Å C18, ThermoFisher, ES800) with a gradient using solvent A (0.1 % formic acid in water) and solvent B (0.1 % formic acid in acetonitrile) at a 0.5 µL/min flow rate. The gradient started with 1 % B for 5 min, followed by 85 min with a linear increase to 35 % B. The gradient was increased to 85 % B in 5 min followed by decreasing back to 1 % B in 5 min as the final wash step. A nanoACQUITY UPLC (Waters) was coupled to a Velos Pro Dual-Pressure Linear Ion Trap mass spectrometer (Thermo Scientific) for data acquisition. A data-dependent acquisition routine was used for a mass spectrum from m/z 300 to 2000 and followed by ten tandem mass spectrometry scans.

Data Analysis

Raw data files were processed using Proteome Discoverer (ThermoFisher). The data were searched against the human Swiss-Port index (09/01/15) using Sequest HT with precursor mass tolerances of 1.5 Da and fragment mass tolerances of 0.8 Da. Maximum missed cleavage sites of full tryptic digestion was two and dynamic modifications of acetylation (N-terminus), carbamidomethylation(cysteine), propionamidation(cysteine) and oxidation(methionine) were considered. The processed mass spectrometry data was analyzed using the Scaffold software (4.4.5). Quantifications from replicate 1 are presented in main figures and from replicate 2 in supplementary figures. For the analysis presented in Figure 1 and S1, spectral quantification was done using the normalized weighted spectral counts with minimum peptide identification threshold at 95 % and protein identification threshold of 95 % with 1 minimum peptide. For analysis presented in Figure 2 and S2, spectral quantification was done using the Normalized Spectral Abundance Factor (NSAF) (Zhang et al.,

2010). A total of 341 (replicate 1) or 259 (replicate 2) proteins were identified at 99 % threshold each with a minimum of 2 peptides. When Scaffold grouped together similar proteins into clusters, only those proteins were retained from a cluster that had non-zero NSAF value in at least one of the four samples. There are several proteins that are detected only in one or the other alternate EJC, and thus had zero values in samples where they were undetected. Also, many EJC specific proteins had zero values in FLAG-only control IPs. A pseudocount of 0.00001 was added to all NSAF values to prevent loss of such proteins when divided by zero. Fold-enrichment over FLAG-only control was calculated for each of the three EJC samples as follows:

Fold-enrichment:

$\log_2$[NSAF(FLAG-EJC+0.00001)] / [NSAF(FLAG-only+0.000001)]

The comparison of fold-enrichment values for all 341 proteins in the EJC core or alternate EJCs was carried out in R package ggplot2 (via scatter plots). Proteins that were >10-fold enriched in one of the three EJC IPs were included in the heatmap (using R package gplots). In this analysis, proteins that most likely were contaminants (e.g. cytoskeleton proteins, histones, metabolic enzymes) were not included in the heatmap but are shown in the >10-fold enrichment table.

Glycerol Gradient Fractionation

Five 15-cm plates with ~90 % confluent HEK293 cells expressing FLAG-tagged MAGOH, CASC3, RNPS1 proteins, or FLAG-peptide as a control, were cultured and induced as above. Cell lysis, FLAG-immunoprecipitation, RNase A digestion and FLAG-elution steps were also carried out as described above. FLAG-IP elution was layered onto pre-cooled continuous 10-30 % glycerol gradients prepared in 11 ml Beckman centrifuge tubes. Gradients were run at 32,000 rpm for 16 hours at 4 °C. Gradients were fractionated by-hand into 500 µl fractions. Proteins were precipitated using TCA and resuspended in 15 µl of 1x SDS loading buffer for analysis on 12 % SDS-PAGEs.

siRNA-mediated Knockdowns

HEK293 siRNA knockdowns

HEK293 cells were seeded into 24-well plates. Cells were transfected 4 hours later following the Mirus TransIT-X2 procedure with 55-60 pmols of control, EIF4A3, CASC3, RNPS1, or UPF1 siRNA. Cells were harvested 48 hr later. Knockdown was checked by Western blot and subsequently purified RNA was used for qRT-PCR analysis. For 96 hr knockdowns, HEK293 cells were transfected as above and were incubated for 24 hr. Cells were transfected again at roughly 70-80 % confluence following the Mirus TransIT-X2 procedure using 50-60 pmol of control, EIF4A3, CASC3, RNPS1, or UPF1 siRNA. Cells were harvested ~48 hr (~96 total hours after first transfection) later. Knockdown was checked by Western blot and subsequently purified RNA was used for qRT-PCR analysis.

HeLa Tet-off siRNA knockdowns and reporter RNA expression

HeLa Tet-off cells in 12-well plates (1.2×10^5 cells/well) were reverse transfected with 15 pmol of control, UPF1, CASC3 and RNPS1 siRNA using RNAiMAX following the manufacturer's protocol. After 24 hr, cells were co-transfected with plasmids (100 ng of pcTET2-βwtβ or pcTET2-β39β, 30 ng of pcβwtGAP3UAC and 70 ng of pezYFP or carrier DNA) and a second dose of 15 pmol of siRNA using JetPrime (PolyPlus). Tetracycline (50 ng/ml) was included at the time of transfection to repress reporter RNA expression, and was removed 20 hr later to induce reporter RNA expression for 6-8 hr. Cells were harvested in clear sample buffer for western blot analysis and TRIzol extraction of RNA, which was analyzed by Northern Blotting.

Overexpression of alternate EJC factors

HEK293 FLAG-EJC overexpression

HEK293 cells expressing FLAG-tagged CASC3, RNPS1 proteins, or FLAG-peptide as a control, were seeded into 6-well plates. Cells were induced with 625 ng/ml of Doxycycline for 48 hr. Overexpression was checked by Western blot and subsequently purified RNA was used for qRT-PCR analysis.

For IPs following alternate EJC protein overexpression, HEK293 cells stably expressing FLAG-tagged CASC3 or RNPS1, or FLAG-peptide as a control, were seeded in 10 cm plates and induced with 625 ng/ml of Doxycycline for 48 hr. Endogenous RBM8A or EIF4A3 IPs were carried out as described above.

HEK293 EIF4A3 or EIF4A3 YRAA Rescue

HEK293 cells were seeded into 24-well plates. Cells were transfected 4 hr later following the Mirus TransIT-X2 procedure with 60 pmol of control, or EIF4A3 siRNA, and 300 ng of pcDNA3 empty, FLAG-EIF4A3 wt or FLAG-EIF4A3 YRAA and 200 ng of pcDNA3 vector. Cells were harvested ~48 hr later. Knockdown and EIF4AIII expression was checked by Western blot and subsequently purified RNA was used for qRT-PCR analysis.

HEK293 cells were seeded into 6-well plates. Cells were transfected four hours later following the Mirus TransIT-X2 protocol with 300 ng of pcDNA3 empty, FLAG-EIF4A3 wt or FLAG-EIF4A3 YRAA and 200 ng of pcDNA3 vector. Cells were harvested in 1x PBS 48 hr later and washed for subsequent FLAG IP as previously described.

β-globin NMD assays

The pulse-chase experiments were performed in HeLa Tet-off cells as described previously (Singh et al., 2007). Briefly, cells growing in 12-well plates were transfected with 10 ng of pcβwtGAP3UAC, 200 ng of pcTET2-β39β and 250 ng of pezFLAG-CASC3 (or pezFLAG empty vector) in the presence of 50 ng/nl tetracycline. 36 hr later, expression of β39 mRNA was induced for 6 hr by removing Tet, and then supressed again via addition of 1 μg/ml Tet. Time points were collected starting ~30 min after addition of Tet for 0 hr time point. Total RNA was extracted using TRIzol and one-half of the RNA sample was analyzed by Northern blots. The autoradiogram signal was scanned using Fuji FLA imager, and quantified using ImageQuant software.

For steady-state assays, HeLa CCL2 cells growing in 12-well plates were transfected with 100 ng of pcβwtGAP3UAC, 100 ng of either pcβwtβ or pcβ39β, and 150 ng of pezFLAG plasmids (expressing RNPS1 or CASC3, or empty vector as a

control). Cells were harvested 48 hr post-transfections and total RNA extracted was analyzed by Northern blots.

qRT-PCR

RNA was isolated from cells using Trizol, DNase treated, purified with Phenol:Chloroform:Isoamyl alcohol (25:24:1, pH 4.5) and resuspended in RNase-free water. RNA was reverse transcribed using oligo-dT (Promega) and Superscript III (Invitrogen). After reverse transcription of RNA the samples were treated with RNase H (Promega) for 30 min at 37 °C. The samples were then diluted to 5 ng/µl before proceeding to qPCR setup. For each qPCR 25 ng of cDNA was mixed with 7.5 µl of 2 x SYBR Green Master Mix (ABS), 0.6 µl of a 10 mM forward and reverse primer mix (defrosted once), and quantity sufficient water for a 15 µl reaction volume. The qPCRs were performed on (Applied Biosystems) in triplicates (technical). Parallel with NMD targets, β-actin or TATA-binding protein (TBP) were used as internal controls. Fold-change calculations were performed by delta-delta Ct method. Fold-change from at least three biological replicates were used to determine the standard error of means. The p-values were calculated using student's t-test.

Cytoplasmic FLAG-IPs of alternate EJCs

One 15-cm plate of HEK293 cells expressing FLAG-tagged MAGOH, CASC3, RNPS1 proteins, or FLAG-peptide as a control, were cultured and induced as above. Cells were lysed in 2 ml RSB-150 buffer [10 mM Tris-HCl pH 7.5, 150 mM NaCl, 5 mM $MgCl_2$, 10 µg/ml Aprotinin, 10 µM Leupeptin, 1 µM Pepstatin, 1 mM PMSF] supplemented with 0.05 % digitonin and 1 mM DTT. Lysates were passed through a 25 gauge needle 5 times before centrifugation at 3,000 x g for 1 min at 4 °C. Supernatant was removed and centrifuged at 21,000 x g for 10 min at 4 °C, and passed through a 0.45 µm filter (Cytoplasmic Fraction). The first pellet was resuspended in RSB-150 and pelleted again at 3,000 x g for 1 min at 4 °C to remove residual cytoplasmic contaminants. The nuclei were then resuspended in 2 ml RSBT (components) and sonicated as previously described, before being centrifuged and purified as done for the cytoplasmic fraction.

Approximately 2 ml of each fraction were collected and used for FLAG-IP as described above.

RIPiT-Seq data analysis

Alternate EJC RIPiTs from HEK293 cells

RIPiTs were carried out with and without formaldehyde crosslinking as described previously (Singh et al., 2014) with the following modifications. For native RIPiTs, total extracts from four 15-cm plates prepared in hypotonic lysis buffer supplemented with 150mM NaCl-containing were used as input into FLAG-IP. For formaldehyde crosslinked RIPiTs, total extracts were prepared from six 15-cm plates in denaturing lysis buffer supplemented with 150 mM NaCl-containing for input into FLAG-IP. Following IP and washes, RNase I (0.006 U/ml in Isotonic wash buffer (IsoWB)) treatments were performed at 4 °C for 10 min. For the second IP, the following antibodies conjugated to protein-A Dynabeads were used: anti-EIF4A3 (Bethyl A302-980A, 10 µg/RIPiT), anti-CASC3 (Bethyl A302-472A, 2 µg/RIPiT), anti-RNPS1 (HPA044014-100UL, 2 µg/RIPiT). RIPiTs were eluted in clear sample buffer and divided into two parts for RNA and protein analysis as described previously. RIPiTs to enrich EJC footprints upon cycloheximide (CHX) treatment were carried out as above except that cells were incubated with 100 µg/ml CHX for 3 hr prior to harvesting. CHX was included at the same concentration in PBS (for washes before lysis) and cell lysis buffers.

High-throughput sequencing library preparation

For RIPiT-Seq, RNA extracted from ~80 % of RIPiT elution was used to generate strand-specific libraries using a custom library preparation method as detailed in (Gangras et al., 2018). For RNA-Seq libraries, 5 µg of total cellular RNA was used as input for ribosomal RNA depletion (Ribozero, Illumina). Purified RNA was then used to generate strand-specific libraries using a custom library preparation method (Gangras et al., 2018). Following PCR amplification, all libraries were quantified using Bioanalyzer

(DNA lengths) and Qubit (DNA amounts). Libraries were sequenced on Illumina HiSeq 2500 in the single-end format (50 and 100 nt read lengths).

Data pre-processing

Adapter trimming and PCR duplicate removal

After demultiplexing, fastq files containing unmapped reads were first trimmed using cutadapt. A 12 nt sequence on read 5' ends consisting of a 5 nt random sequence, 5 nt identifying barcode, and a CC was removed with the random sequence saved for each read for identifying PCR duplicates down the line. Next as much of the 3'-adapter (miR-Cat22) sequence TGGAATTCTCGGGTGCCAAGG was removed from the 3' end as possible. Any reads less than 20 nt in length after trimming were discarded.

Alignment and removal of multimapping reads

Following trimming reads were aligned with tophat v2.1.1 (Trapnell et al., 2009) using 12 threads to NCBI GRCh38 with corresponding Bowtie2 index. After alignment reads with a mapping score less than 50 (uniquely mapped) were removed, i.e. all multimapped reads were discarded. Finally all reads mapping to identical regions were compared for their random barcode sequence; if the random sequences matched, such reads were inferred as PCR duplicates and only one such read was kept.

Removal of stable RNA mapping reads

Next, reads which came from stable RNA types were counted and removed as follows. All reads were checked for overlap against hg38 annotations for miRNA, rRNA, tRNA, scaRNA, snoRNA, and snRNA using bedtools intersect (Quinlan and Hall, 2010), and any reads overlapping by more than 50 % were removed. Reads aligned to chrM (mitochondrial) were also counted and removed.

Human reference transcriptome

The primary reference transcriptome used in all post-alignment analysis was obtained from the UCSC Table Browser. CDS, exon, and intron boundaries were obtained for

canonical genes by selecting Track: Gencode v24, Table: knownGene, Filter: knownCanonical (describes the canonical splice variant of a gene).

Read distribution assignment

Fractions of reads corresponding to exonic, intronic, intergenic, and canonical EJC and non-canonical EJC regions were then computed. Exonic regions were defined by the canonical hg38 genes, with intronic regions defined as the regions between exons in said genes. Bedtools intersect was used to compare reads against these exon and intron annotations, and reads which overlapped the annotation by more than 50 % were counted. Any reads which did not overlap either the exon or intron annotations sufficiently were counted as intergenic. For classification of reads as canonical versus non-canonical EJC footprints, the canonical region for each library was defined using the meta-exon distribution at exon 3'-ends (Figure 3C and S3J). All reads with their 5'ends falling within the window starting at -24 position till the 25 % max height on the 5' side of the canonical EJC peak were counted as canonical reads. Similarly any read whose 5' end was found anywhere between the start of the exon and 10 bp upstream of that 25 % max point was considered non-canonical.

k-mer analysis

Lists of all 6-mers and 3-mers present in reads mapping to exonic regions (as described above) were produced for each RIPiT-Seq sample. The ratio of total 3-mer frequency in RIPiT-Seq samples to RNA-Seq samples was then used to identify 3-mers enriched in alternate EJCs.

Motif enrichment analysis

Motif enrichment analysis was performed by first selecting RNA binding proteins of interest - namely SR proteins - from the position weight matrices (PWMs) available on http://rbpdb.ccbr.utoronto.ca/. For all reads mapped to exonic regions, a score was then generated representing the highest possible binding probability for each protein on that read. For visualization the cumulative distribution of these score frequencies was plotted for both pull down and RNA-Seq replicates, with a relatively higher score frequency at a

positive score implying greater binding affinity. The p-values between RIPiT-Seq and RNA-Seq replicates were also computed for every score using a negative binomial based model, with significant values primarily in positive score regions implying a binding preference for that protein.

Differential enrichment analysis

Differential analysis of exons and transcripts between CASC3 and RNPS1 pull down was conducted with the DESeq2 (Love et al., 2014) package in R. Exons and transcripts with significant differential expression ($p < 0.05$) were selected. All the following analysis was conducted using only the lists of significantly differentially expressed transcripts, unless otherwise noted.

Estimation of nuclear versus cytoplasmic levels

Nuclear and cytoplasmic RNA levels were estimated by first obtaining nuclear and cytoplasmic reads from (Neve et al., 2016). Reads were aligned and mapped to our exonic annotation as described above, and a ratio of nuclear to cytoplasmic reads was then calculated for all transcripts.

Comparison to genes with detained introns

A list of detained and non-detained introns was obtained from (Boutz et al., 2015). To identify detained intron containing and lacking genes in our RNA-Seq data, we carried out analysis using a DESeq2-based pipeline as described by Boutz *et al*. Briefly, using two RNA-Seq replicates we first created artificial data sets containing the same numbers of total reads per replicate, but with counts originating from introns in a given gene spread evenly amongst those introns in the artificial set. By comparing the intron distributions of these artificial replicates to the experimental replicates using DESeq we were able to produce lists of detained introns (introns with significantly ($p < 0.05$) higher expression levels compared to the artificially spread data) and non-detained introns (introns with non-significant ($p > 0.1$) expression levels compared to the artificially spread data). The transcripts containing introns found to be detained/non-detained in both our analysis and the analysis done by Boutz *et al*. make up the stringent lists,

which were used for analysis in Figure 4D. Of the 693 canonical genes containing detained introns reported by Boutz et al. we found 555 in our own analysis (80 %) and of the 5294 canonical genes lacking detained introns we found 812 (15 %).

Features compared between two alternate EJC sets

mRNA half life data was taken from (Tani et al., 2012) with no further processing on our part. Translation efficiency data was similarly obtained from (Kiss et al., 2017).

Gene ontology analysis

DAVID gene ontology tool (Huang et al., 2009) was used to compare the set of genes (canonical Ensembl transcript IDs) predicted by DESeq2 analysis to be significantly enriched in CASC3 or RNPS1 EJCs against a background list containing only those human genes that were reliably detected by DESeq2 (all genes for which DESeq2 calculated adjusted p-values). Only non-redundant categories with lowest p-value (with Benjamini-Hochberg correction) are reported.