The Cellular EJC Interactome Reveals Higher Order mRNP Structure and an EJC-SR Protein Nexus

Guramrit Singh, Alper Kucukural, Can Cenik, John D. Leszyk, Scott A. Shaffer, Zhiping Weng and Melissa J. Moore

SUPPLEMENTAL INFORMATION

Supplemental Figure and table legends:

Figure S1. Endogenous EJCs and their RNA footprints (related to Figure 1).

A. Tetracycline (Tet)-mediated induction of stably expressed FLAG-tagged proteins. Western blots showing expression levels of FLAG-tagged proteins and their endogenous counterparts in stably transfected HEK293 Flp-In cells. Cells were induced for ~16 hr with Tet concentrations indicated above each lane. Tet concentration used for induction of each FLAG-tagged protein during the study is indicated in bold.

B. RNase- and salt-sensitivity of EJC factor self-association. Western blots of proteins (listed on the right) in IPs (lanes 2-5, 7-12) or 5% total extract (TE, lanes 1 and 6) from cells expressing the FLAG-tagged protein indicated on the top. The amount of NaCl and RNase I in each IP is indicated above each lane. hnRNP A1 served as a negative control.

C. EJC factor self-association under increased strigency. Western blots of the proteins on the right in 5% total extract (TE) or α-FLAG IP as indicated from cells expressing either FLAG peptide as control (lanes 1 and 3) or FLAG-eIF4AIII (lanes 2, 4-11). All IPs were treated with 1U/ul RNase I while the amount of NaCl

(lanes 3-11), Empigen BB (EBB; lanes 6 and 7), Heparin (Hep; lanes 8 and 9) and Urea (lanes 10 and 11) in each IP is given above each lane.

D. Effect of cycloheximide on EJC footprints. An autoradiogram showing length distribution of 5′-end labeled RNA fragments from FLAG-eIF4AIII:Y14 RIPiT in absence (lane 2) or presence (lane 3) of cycloheximide (CHX).

E. Nuclease and high-salt sensitivity of EJC footprints. Length distribution of 5′ end labeled RNA fragments obtained from FLAG-Magoh:eIF4AIII RIPiTs with increasing amounts of RNase I (lanes 2-5) or micrococcol nuclease (MNase) treatment (lanes 6-9). Samples in lanes 5 and 9 were treated with 750mM NaCl for 10 minutes on ice prior to nuclease treatment. Pixel intensity profiles of each lane are on the right.

F. Effect of heat denaturation on EJC footprints. RNA fragment length distributions determined as in E. RNase I-treatment was done during first or second IP washes, or not done, as indicated on top. Samples in lanes 5 and 9 were heated at 70 °C for 10 minutes before RNase I treatment. * on pixel intensity profile indicates signal from a gel artifact.

Figure S2. Specificity of the EJC proteome and size estimation of HMW EJCs (related to Figure 2).

A. Specificity and stoichiometry of the EJC proteome. The bar graph on the left shows the relative abundance of EJC and EJC-interacting proteins (Figure 2) in a mammalian cell line [HeLa (Nagaraj et al., 2011)]. On the right, relative stoichiometric amounts of each protein with respect to Pinin is indicated in FLAG-

eIF4AIII (blue bars) and FLAG-Magoh (red bars) IPs. The vertical blue and red lines indicate the levels of EJC core in each IP.

B. Western blots showing levels in total extracts (lanes 1-3) or RIPiTs (lanes 4-6) of peripheral EJC proteins either detected or not in mass spec analysis. * indicates IgG heavy chain.

E. RNase-resistence of EJC-SRSF1 interaction. Western blots showing levels of proteins (right) in total extracts (lane 1) or FLAG-eIF4AIII IPs (lanes 2-5). RNA-protein complexes were treated with indicated amounts of RNase I.

F. Phosphorylation status of SR proteins in RIPiTs. Western blots of proteins listed on the right in the total extracts (lanes 1-3) or RIPiTs (lanes 4-6) as in Figures 1C and 2B. For comparison of phosphorylation status of the SR proteins, HEK293 nuclear extracts treated with Calf intestine phosphatase (CIP; lane 9) or untreated nuclear extracts (lanes 7 and 8) were analyzed on the same gel.

G. Sedimentation of FLAG-eIF4AIII complexes on sucrose gradients. Top: Absorbance profile at 260 nm during fractionation of 10-50% sucrose gradient after sedimentation of RNase I-treated FLAG-eIF4AIII immunoprecipitates mixed with total yeast extract. Bottom: Western blots of proteins indicated on the left in total extract (2%TE), FLAG IP (IP) and gradient fractions (lanes labeled 1-12).

Figure S3. Mapping and evaluation of deep sequencing libraries (related to Figure 3).

A. An overview of steps converting RNA fragments into deep sequencing libraries using the SOLiD (Applied Biosystems) small RNA expression kit.

B. Number and percent of reads in individual deep sequencing libraries at steps in alignment pipeline indicated on the top.

C. Length distributions of deep sequencing reads in the FLAG-Magoh:eIF4AIII library before (red) and after (blue) uniquely mapping to the human genomic sequences.

D. Genomic distribution of deep sequencing reads and EJC peaks. Fraction of reads in libraries indicated on the bottom that uniquely map to exonic (red), intronic (dark gray) or intergenic (light gray) regions in the human genome. The stacked bar on the extreme-right shows the fraction of peaks in these genomic regions reproducibly detected in FLAG-eIF4AIII:Y14 and FLAG-Magoh:eIF4AIII libraries. On the left is the fraction of nucleotide-space occupied by exons, introns and intergenic regions in the human genome.

E. Reproducibility of short EJC footprints. Left: Comparison of short EJC footprint reads per kilobase per million for each RefSeq transcript in two biological replicates of a RIPiT (top) or two different RIPiT configurations (bottom). Right: Same as on left except the comparisons are between the number of uniquely mapping reads in the window -15 to -31 nt from exon junctions in the indicated libraries. Pearson’s correlation coefficients are shown in the top right.

F. Comparison of exon mapping reads in spliced and unspliced mRNAs. Total number of exonic reads in FLAG-Magoh:eIF4AIII library for 12,455 mRNAs was plotted versus their transcript RPKM. Transcripts from 11827 intron-containing (blue) and 628 intronless (orange) genes are shown.

G. Z-score distributions of exonic read counts in transcripts from intron-containing and intronless genes. Z-score for each mRNA was calculated after a local loess regression fit (See Supplemental Experimental Procedures). Z-score distributions of transcripts from intron-containing (blue) and intronless (orange) genes are shown.

H. Comparison between short and long EJC footprint libraries. Exonic RPKMs (total exon mapping reads per kilobase per million reads) for each gene were compared between any two of the short (FLAG-Magoh:eIF4AIII) and long EJC footprint (FLAG-Magoh:eIF4AIII-long; FLAG-eIF4AIII-HMW) libraries. Pearson's correlation coefficients are shown.

Figure S4. Canonical EJC (cEJC): occupancy profiles and influencing factors (related to Figure 4).

A. EJC peak distributions. Tables summarizing the number of EJC peaks among different genomic regions (top) and the distribution of exonic peaks among indicated classes of exons (bottom) in the two short EJC footprint peak sets. Numbers in parentheses show the percent of total peaks in a given class.

B. Read distribution in deep sequencing libraries (indicated above each track on the right) along BAT2 gene.

C. Read distribution in indicated libraries along hnRNP A2B1 gene (top) or its spliced mRNA (bottom).

D-G. Interquartile range (IQR) of 5′-splice site scores (D), 3′-splice site scores (E), exon lengths (F) or intron lengths (G) of the cEJC-occupied (gray) or cEJC-

free sites (white) as described in Figure 4E from transcripts with >1 intron. Medians (horizontal black lines) and their confidence interval (notches) are indicated within box-plots. Dashed lines emphasize differences in medians in F and G.

Figure S5. mRNAs enriched in cEJC signal (related to Figure 5).

A. Higher average cEJC occupancy does not correlate with mRNA stability. Distribution of average cEJC occupancy *z*-scores for mRNAs within the indicated Gene Ontology categories. Median mRNA half-lives for orthologous mouse mRNAs in each GO-term category are indicated at bottom [mouse data from (Schwanhausser et al., 2011)]. Box plots show the half-life IQR with whiskers (1.5 times the IQR), outliers (open circles), median (horizontal black lines within box-plots) and its confidence interval (notches) also indicated.

B. Expression levels of AS-NMD (PTC+) and AS only (PTC-) transcripts used for analysis in Figure 5C.

Figure S6. Non-canonical EJC (ncEJC) peak sequence motifs and composition (related to Figure 6).

A. EJC peak distribution on canonical and non-canonical sites. Number and percentage of canonical and ncEJC peaks from the indicated peak set on all exons (top) or after dividing into first and internal exons (bottom).

B. A dendrogram based on the degree of similarity between internal ncEJC C-rich motifs. All detected motifs are shown while the motifs that satisfy cut-offs (at

least 2.5-fold enrichment in ncEJC versus the scrambled sequences; <50% codon bias at any position; shown in Figure 6D) are connected by thicker lines.

C-E. ncEJC motif validation. Codon bias (left) and distance between motif and read centers (right) for ncEJC motifs in Figure 6. C. First exon ncEJC motif (Fig. 6C). D. Internal ncEJC motif (Fig. 6D, top). E. Internal ncEJC motif (Fig. 6D, bottom).

F. Occurrence of ESEs and ESSs in deep seqencing reads. The boxplots show the distribution of the number of uniquely mapping reads from the indicated libraries that contain ESE (green) and ESS (red) hexamers. The numbers below indicate ratio of median ESE to ESS occurrences in a library. Box-plots are as in Figure S5A except that whiskers (at 1.5 times IQR) and outliers (open circles) are also shown.

G. Correlation between average cEJC occupancy and ncEJC read densities. cEJC occupancy is from Figure 5A. ncEJC read densities were calculated only from internal exons for all transcripts with RPKM >1 in our RefSeq.

Figure S7.

A. Correlation between neighboring cEJCs (related to Figure 7). Distribution of correlations from comparison of cEJC occupancy of each gene (RPKM >10) with itself (and hence auto-correlation) after one- (lag 1), two- (lag 2) or three- (lag 3) shifts in aligned exon positions (solid lines). Auto-correlation distributions for each lag in control sets where cEJC occupancy within an mRNA was randomized before introducing lags are also shown (dashed lines). All distributions are

significantly different (Wilcoxon rank sum test) with *p-values* increasing with each lag (lag 1 $<1x10^{-15}$, lag 2 $=7 \times 10^{-11}$, lag 3 $=3 \times 10^{-4}$).

B. Correlation between neighboring ncEJCs. Same as A except presence or absence of ncEJC peaks was used to derive auto-correlations (Wilcoxon rank sum test *p-values*: lag 1 $<1x10^{-15}$, lag 2 $=1 \times 10^{-5}$, lag 3 =0.67).

C. mRNA binding efficiency of proteins upon eIF4AIII knockdown. Western blots from two independent experiments showing protein levels in total extracts (lanes 1, 2, 5 and 6) and in oligo-dT selected RNAs (lanes 3, 4, 7 and 8) from UV-crosslinked HEK293 cells.

Table S1. Mass spectrometric analysis of cellular EJCs (related to Figure 2)

A list of proteins detected in IP samples from FLAG, FLAG-eIF4AIII and FLAG-Magoh cells. In addition to the quantifications from the current study, number of peptides observed for some proteins in spliced mRNPs (Merz et al., 2007; Tange et al., 2005) and *in vitro* assembled EJC (Tange et al., 2005) are also provided. - signifies lack of quantification/detection.

Supplemental Experimental Procedures

Oligonucleotides

PCR oligos
eIF4AIII_BamHI_forward
GGAGAAGGATCCATCATGGCGACCACGGCCACGATGGC
eIF4AIII_ NotI _reverse
GGAGAAGCGGCCGCGATAAGATCAGCAACGTTCATCGGC
Magoh_HindIII_forward
GGAGAAAAGCTTGCCATGGAGAGTGACTTTTATCTGC
Magoh_ NotI _reverse
GGAGAAGCGGCCGCCTAGATTGGTTTAATCTTGAAGTG
PHGDH_BamHI_forward
GGAGAAGGATCCGCCACCATGGCTTTTGCAAATCTGCGG
PHGDH_NotI_reverse
GGAGAAGCGGCCGCTTAGAAGTGGAACTGGAAGGCTTC
SOLiD/Illumina-PCR-5′ oligo
ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCTATGGGCAGTCGGTGAT
SOLiD/Illumina-PCR-3′ oligo
CTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCTGCTGTACGGCCAAGGC
G
Illumina custom oligo
TCTTCTATGGGCAGTCGGTGA

siRNA oligonucleotide
eIF4AIII_268
AACGAGCAAUCAAGCAGAUCA

Plasmids and cell lines

pcDNA5-TetO-FLAG was created by inserting an annealed oligonucleotide encoding the FLAG sequence into the HindIII site of pcDNA5-TetO such that the upstream HindIII site was inactivated. cDNAs encoding eIF4AIII (BamHI-NotI), Magoh (HindIII-NotI), and PHGDH (BamHI-NotI) were inserted into the polylinker of pcDNA5-TetO-FLAG. Plasmids for transient expression of FLAG- and Myc-tagged EJC proteins (eIF4AIII and Magoh) were constructed by cloning the cDNAs in a derivative of pcDNA3.1 plasmid.

To generate stable cell lines, ~5 x 10^6 HEK293 TRex cells (Invitrogen) were seeded on 10-cm plates for 16 hr in Dulbecco's modified eagle medium (DMEM) supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin. A plasmid mix (1μg of a pcDNA5-TetO-FLAG construct along with 9 μg of pOG44) was transfected using 30 μl of HEK293TransIT reagent (Mirus) following manufacturer's instructions. After 24hr, cells were split at 1:10 dilution into new 10-cm dishes. After overnight incubation Blasticidin (15 μg/ml) and Hygromycin (100 μg/ml) containing media was added to cells to select for stably transfected cells. Once individual transfected cells had grown into colonies visible to the naked eye (~5mm diameter), the clonal pool of stably-transfected cells was harvested. In these cells, the expression level of the stably integrated FLAG-tagged protein was optimized by titrating tetracycline (Tet; 0-500 ng/ml) to determine a concentration where exogenous protein expression levels were comparable to its endogenous counterpart.

EJC RIPiT

For each EJC purification, TRex-HEK293 cells containing a stable copy of FLAG-tagged EJC proteins (eIF4AIII and Magoh) or control cells (expressing FLAG-tag only) were grown in three 15-cm plates. Expression of the FLAG-tagged protein was induced with tetracycline for ~16 hr. One hour prior to cell harvesting, cycloheximide (CHX) was added to 100 μg/ml. The monolayer was rinsed and harvested in phosphate-buffered saline (PBS) containing 100 μg/ml CHX. The cells were lysed in 3 ml hypotonic lysis buffer [20 mM Tris-HCl pH7.5, 15 mM

NaCl, 10 mM EDTA, 0.5% NP-40, 0.1% Triton X-100, 10 μg/ml Aprotinin, 1 μg/ml Leupeptin, 1 μM Pepstatin, 1 mM PMSF, 100 μg/ml CHX] for 10 min on ice. The suspension was sonicated (Branson Digital Sonifier-250) at 40% amplitude using a Microtip for a total of 16 seconds (in 2 second bursts with 10 second intervals). NaCl was adjusted to 300 mM and the lysate was cleared by centrifugation at 15,000 xg for 10 min at 4°C. The cleared lysate was diluted to 10 ml in the above lysis buffer with final NaCl concentration of 300 mM. The diluted lysate was incubated for 2 hr at 4°C with 750 µl of anti-FLAG agarose beads (50% slurry, Sigma) pre-washed twice with 10 ml IsoWB [Isotonic wash buffer; 20 mM Tris-HCl pH7.5, 150 mM NaCl, 0.1% NP-40]. The RNA-protein (RNP) complexes captured on beads were sequentially washed twice each (2 x 10 ml) with ice-cold WB300 [20 mM Tris-HCl pH7.5, 300 mM NaCl, 0.1% NP-40] and IsoWB. After the fourth wash, bound RNP complexes were incubated with one bed volume of IsoWB containing 1U/µl of RNase I for 10 min at 37°C with intermittent shaking. RNP complexes were again washed four times with 10 ml IsoWB. FLAG-epitope containing complexes were affinity eluted from the beads in one bed volume of IsoWB containing 250 μg/ml FLAG peptide and 0.2 U/µl of SUPERase with gentle shaking at 4°C for 2 hr. To prepare the recovered elution for input into second IP, its volume was adjusted to 400µl and its composition to that of the lysis buffer above with NaCl at 150 mM. Also, bovine serum albumin (BSA, NEB) was added as a carrier at 1 mg/ml. The suspension was incubated with antibodies against EJC proteins (Y14 (4C4, Sigma) or eIF4AIII (2256C1, Santa Cruz)) that were pre-coupled to 35 µl of ProteinG-Dyna-beads (Invitrogen)

according to manufacturer's instructions. Immunoprecipitation was carried out at 4°C for 2 hr. Captured RNP complexes were washed six times with 1 ml of ice-cold IsoWB and eluted with 40 µl of clear sample buffer [100 mM Tris-HCl pH6.8, 4% SDS, 10 mM EDTA, 100 mM DTT] at 25°C for 5 min, and subsequently at 95°C for 2 min.

For IPs under protein crosslinking conditions, cells were collected and rinsed once in, and then resuspended in PBS+CHX. Formaldehyde was added to 0.1% and the suspension was gently mixed at RT for 10 minutes. One-tenth volume of quenching buffer (2.5 M Glycine, 25 mM Tris-base) was added, cells were pelleted, and lysed in hypotonic lysis buffer supplemented with 0.1% SDS and 0.1% sodium deoxycholate. Following FLAG IP as described above, IP samples were washed twice with IsoWB+0.1% SDS and 0.1% sodium deoxycholate, and then with IsoWB. All subsequent steps were as above.

RIPiT RNA Extraction and size estimation

The volume of RIPiT elution was adjusted to 400 µl with water. It was extracted twice with equal volume of Phenol (pH 4.5):Chloroform:Iso-amyl alcohol (25:24:1) and once with Chloroform:Iso-amyl alcohol (24:1). The recovered aqueous phase was supplemented with 10 µg of glycogen, 300 mM sodium acetate pH 5.2, and 10 mM $MgCl_2$. RNA was precipitated with 3 volumes of 100% ethanol at -20°C overnight. After a wash with 70% ethanol, RNA was re-suspended in 5 µl of water and stored at -80°C for subsequent steps (analysis of RNA footprint profiles and SOLiD cDNA library construction). For visualizing RNA footprints,

one-tenth of the isolated RNA was 5′-end labeled with γ^{32}P-ATP and T4 poly-nucleotide kinase, precipitated with 3 volumes of 100% ethanol, and resolved on a 26% UREA-PAGE gel. A 30-nucleotide poly-U RNA was base hydrolyzed in 0.1 M Sodium bicarbonate (pH 9.2) for 20 min at 90°C, neutralized with 100 mM Tris-HCl pH 7.0, 5′-end labeled as above and co-migrated as a size-marker.

Mass spectrometric identification of EJC associated proteins

FLAG immunoprecipitation: Five 15-cm plates with ~90% confluent HEK293 TRex cells expressing FLAG-tagged eIF4AIII and Magoh proteins, or FLAG-peptide as a control, were cultured and induced as above. Cell lysis, FLAG-immunoprecipitation, RNase I digestion and FLAG-elution steps were also carried out as described above except that lysis buffer (15 ml per IP) was supplemented with 0.5% Empigen BB and 1 μg/ml FLAG-peptide to improve IP specificity as described earlier (Fenger-Gron et al., 2005). 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 15 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 to denature proteins and reduce di-sulfides, and cooled to RT. 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 15.8 µl of 2X 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 ~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, stained with Imperial protein stain (Pierce) for 1 hr and destained in water overnight. The gel piece containing protein was excised and processed for in gel digestion of proteins.

In gel digestion: The gel slices were cut into 1x1 mm pieces and placed in 1.5 ml eppendorf tubes with 1 ml of water for 1 hr. Following removal of the water by SpeedVac, 1 ml of 50 mM ammonium bicarbonate:acetonitrile (1:1) solution was added to each tube and the samples were incubated at room temperature for 1 hr. Following removal of the solution, 200 ml of acetonitrile was added to each tube turning the gel slices opaque white. Following removal of the acetonitrile, gel slices were rehydrated in 50 ml of 2 ng/ml trypsin (Sigma) in a solution of 0.01% ProteaseMAX Surfactant (Promega) in 50 mM ammonium bicarbonate and incubated at 37°C for 21 hr. The supernatant of each sample was then removed to a separate tube and the gel slices were further dehydrated with 60-100 ml of an acetonitrile: 1% formic acid (v/v) (4:1) solution. The supernatants were combined with the previous supernatants for each sample and dried by Speed Vac.

LC/MS/MS: Tryptic peptides were dissolved in 0.1% trifluoroacetic acid (v/v) and directly loaded at 4 ml/min for 7 min onto a custom-made trap column (100 mm I.D. fused silica with Kasil frit) containing 2 cm of 200Å, 5 mm Magic C18AQ particles (Michrom Bioresources). Peptides were then eluted using a custom-

made analytical column (75 mm I.D. fused silica) with gravity-pulled tip and packed with 25 cm 100Å, 5 mm Magic C18AQ particles (Michrom). Peptides were eluted with a linear gradient from 100% solvent A (0.1% formic acid:acetonitrile (95:05)) to 35% solvent B (acetonitrile containing 0.1% formic acid) in 90 minutes at 300 nanoliters per minute using a Proxeon Easy nanoLC system directly coupled to a LTQ Orbitrap Velos mass spectrometer (Thermo Scientific). Data were acquired using a data-dependent acquisition routine of acquiring one mass spectrum from *m/z* 350 -2000 in the Orbitrap (resolution 60,000) followed by 10 tandem mass spectrometry scans in the LTQ linear ion trap.

Data Analysis: The raw data files were processed with Mascot Distiller (version 2.4.2; Matrix Science) by conversion into peak lists and then searched against the human index of the SwissProt database (version 09/21/11) with Mascot (version 2.3.2; Matrix Science) using parent mass tolerances of 10 ppm and fragment mass tolerances of 0.5 Da. Full tryptic specificity with 1 missed cleavage was used and variable modifications of acetylation (protein N-term), pyro-glutamination (N-term glutamine), popionamide alkylation (cysteine), carbamidomethylation (cysteine) and oxidation (methionine) were considered. Label free quantitation using extracted ion chromatograms (XIC) were done using both the replicate (Radulovic et al., 2004) and average methods (Silva et al., 2006) contained in the Mascot Distiller quantitation software. Mascot search results were also loaded into Scaffold (Version 3.3.1; Proteome Software) for

comparative analyses using spectral counting of tandem mass spectra (Searle, 2010).

In vitro protein dephosphorylation

HEK293 cells from one confluent 10-cm plate were rinsed with PBS and lysed in 1.0 ml of RSB-100 (10 mM Tris-HCl [pH 7.5], 100 mM NaCl, 2.5 mM $MgCl_2$, 40 μg/ml digitonin (Calbiochem)) for 5 min on ice. The nuclei were pelleted from the soluble cytosol by centrifugation at 2,000 Xg for 8 min at 4°C. The resulting pellet was re-suspended in 1 ml RSB-100 containing 0.5% (v/v) Triton X-100. Following incubation on ice for 5 min, the Triton-extracted material was sonicated in 2-second bursts for 10 seconds as described above. The lysate was cleared over a 300 µl 30% sucrose cushion (in RSB-100) by a centrifugation at 4,000 xg for 15 min at 4°C. The top layer - nuclear fraction - was subjected to protein dephosphorylation with calf-intestine phosphatase (CIP) in a 50 µl reaction (10X NEBuffer 3 - 5 µl, Nuclear extract - 40 µl, CIP (NEB, 10U/µl) - 5 µl). Reactions were incubated at 30°C for 0-5 minutes. 15 µl of the lysate was immediately mixed with 2X Lammelli SDS load buffer and separated on 15% SDS-PAGE gels followed by western blot analysis.

Gel-filtration and sucrose density gradient fractionation of FLAG-EJC protein complexes

FLAG-tagged eIF4AIII and its associated RNA:protein complexes were immunoprecipitated from total lysates from three 15-cm plates as described

above. The complexes were eluted from FLAG-affinity resin in 500 µl (Gel-filtration) or 250 µl (Sucrose density gradient fractionation) of $IsoWB_{0.01}$ [NP-40 reduced to 0.01%] supplemented with 250 µg/ml FLAG peptide.

For gel filtration, 450 µl of the eluate was loaded on a 30 ml Sephacryl-S400 HR (GE) column (packed in-house in a 16 mm diameter Kontes chromatography column with a flow adapter) with $IsoWB_{0.01}$ at 1 ml/min flow-rate on ÄKTA FPLC system. Twenty-four 1 ml fractions were collected and subjected to protein precipitation with 25% trichloroacetic acid and 0.1% sodium deoxycholate at 4°C. The precipitates were pelleted at 10,000 xg and washed three times with ice-cold 100% acetone. The air-dried pellet was re-suspended in 1X Lammelli SDS load buffer and proteins were analyzed on 15% SDS-PAGE/Western blotting. The levels of proteins were compared to those in 5% total extract or 2% of input IP sample. Gel filtration of 0.5 ml total extract prepared from FLAG-eIF4AIII expressing cells from one confluent 15-cm plate was also performed as above.

For sucrose density gradient fractionation, FLAG-eIF4AIII complexes IPed above were mixed with 300 µl of total yeast extract prepared from 100 ml of log-phase (OD_{600}=0.5) *Saccharomyces cerevisiae* cells treated with CHX at 100 µg/µl for 2 minutes prior to lysis in Glass bead lysis buffer [Polysome gradient buffer (below) supplemented with 1 mg/ml Heparin, 1 mM PMSF, 10 µg/ml Aprotinin, 1 µg/ml Leupeptin, 1 µM Pepstatin]. FLAG IP-yeast extract mix was sedimented on 10-50% sucrose gradient (made in polysome gradient buffer [20 mM Tris-HCl pH 8.0, 140 mM KCl, 5 mM MgCl2, 100 µg/µl CHX, 0.5 mM DTT])

at 35,000 rpm for 160 min at 4°C. The gradients were fractionated using ISCO fraction collector at 0.75 ml/min flow rate with a new fraction collected every 70 seconds. Proteins from each fraction were precipitated with TCA and analyzed by SDS-PAGE/Western blotting as described above.

Nuclear mRNP footprints

From TRex-HEK293 cells (expressing only FLAG peptide), we first obtained the soluble nuclear mRNP-containing fraction as described (Mili et al., 2001). ~90% confluent cells from a 15-cm plate were rinsed three times with ice-cold PBS and scraped into 3.0 ml of RSB-100 (10 mM Tris-HCl [pH 7.5], 100 mM NaCl, 2.5 mM $MgCl_2$, supplemented with 40 µg/ml digitonin (Calbiochem)). Following incubation on ice for 5 min, the soluble cytosolic material was removed from the nuclear and digitonin-insoluble fractions by centrifugation at 2,000 Xg for 8 min at 4°C. The resulting pellet containing mainly nuclei was re-suspended in 3 ml RSB-100 containing 0.5% (v/v) Triton X-100. Following incubation on ice for 5 min, the Triton-extracted material was separated by centrifugation at 2,000 xg for 8 min at 4°C. The supernatant thus obtained was the soluble nuclear fraction. Final NaCl concentration of this nuclear extract was adjusted to 250 mM and it was diluted to 10ml with RSBT-250 (10 mM Tris-HCl [pH 7.5], 250 mM NaCl, 2.5 mM $MgCl_2$, 0.5% Triton X-100). The soluble nuclear fractions were incubated with ~30 mg oligo-dT cellulose resin reconstituted and washed in RSBT-250. After allowing binding for 2 hr at 4°C, the resin was washed five times with 3 ml RSBT-250. RNase-resistant mRNP complexes were eluted by RNase I digestion (0.04 U/µl)

in 300 μl of elution buffer (EB: 10 mM Tris-HCl [pH 7.5], 10 mM NaCl, 1 mM EDTA, 0.1% Triton X-100) at 37°C for 10 min with intermittent shaking. The protected RNAs were extracted from the elution, precipitated and re-suspended in 3 μl of water for SOLiD cDNA library construction as described above.

RNA fragmentation for the RNA-Seq library

Total RNA was extracted using 10 ml Tri-reagent (Molecular Research Center) from a ~90% confluent 10-cm plate of FLAG-eIF4AIII expressing cells where the protein was induced for 16 hr with 10 ng/ml of Tet. Poly(A)+ RNA was purified from total RNA using oligo-dT beads (PolyA purist-MAG kit, Ambion) following the manufacturer's instructions. RNA was fragmented in 25 μl reaction set up in a PCR tube with 4 μg poly(A)+ RNA and 0.1M $NaHCO_3$ pH 9.2. The reaction was incubated at 90°C for 20 min on a PCR machine and subsequently neutralized by adding 3 μl of 1M Tris pH7.0. The fragmented RNAs were mixed with 2X formamide load buffer, briefly heat denatured at 95°C (~3 min) and separated along with size markers on a 15% 6M Urea-PAGE gel (0.5X TBE, 8x11 inch gel of 1.5mm thickness). Following staining of the size-separated RNA for 5 min with 1X SYBR Gold in 0.5X TBE, 27-36 nt size RNA fragments were excised. The gel piece was fragmented by extruding through a 3 ml syringe and RNA was eluted overnight in 500 μl of 0.3M Sodium acetate pH 5.2, 0.1mM EDTA by gentle mixing. The gel pieces were removed by passing the slurry through Spin-x column (Corning #8161) at 10,000 xg for 2-5 min until most of the liquid comes out into the collection tube. The RNA was precipitated and re-suspended in 10 μl of water as above.

The dephosphorylation of RNA 3′ ends was performed in a 20 µl reaction at 37°C for 30 min (10X T4 PNK - 2.0 µl, 20mM DTT - 1.0 µl, RNA - 10.0 µl, T4 PNK - 1.0 µl) To phosphorylate the RNA 5′ ends in the same reaction, the following mix was prepared and added to the above reaction. Incubation at 37°C was continued for another 45 min: 10X T4 PNK - 1.0 µl, 20mM DTT - 0.5 µl, 2mM ATP - 6.0 µl, T4 PNK - 1.0 µl, Water to 10 µl. The reaction was diluted to 400 µl, RNA extracted and precipitated as above, and re-suspended in 3 µl of water for SOLiD cDNA library construction.

SOLiD cDNA library construction

The RNA fragments obtained above were converted to cDNA libraries for sequencing on the SOLiD platform by using the SOLiD Small RNA Expression Kit (Applied Biosystems) following the accompanying instructions as described below. All incubation steps were carried out on a PCR machine. For adaptor ligation, 1.5 µl of RNA was mixed with 1 µl Adaptor Mix A, 1.5 µl hybridization solution in an 8 µl reaction in a PCR tube. The reaction was heated at 65°C for 10 min and cooled to 16°C for 5 min. 5 µl of 2X ligation buffer and 1 µl of ligation enzyme mix were added to the reaction on ice, mixed by pipetting and incubated at 16°C for ~16 hr. The adaptor-ligated RNA was reverse transcribed in the same tube by adding 10 µl of the following mix to the ligation mix above: 10X RT buffer (2 µl), 2.5 mM dNTP mix (1 µl), ArrayScript Reverse Transcriptase (0.5 µl) and nuclease-free water (6.5 µl). Samples were mixed and incubated at 42°C for 30 min. To 10 µl of the above RT mix, 1 µl of RNase H was added and the reaction

was incubated at 37°C for 30 min. RT reactions with or without RNase H treatment were stored at -80°C. 1 µl of the RNase H treated RT reaction was used per 100 µl of the following PCR reaction: 10X PCR buffer I (10 µl), SOLiD PCR Primers (one of the 1-10 set; 2 µl), 2.5 mM dNTP mix (8 µl), AmpliTaq DNA polymerase (1.2 µl) and nuclease-free water (77.8 µl). The PCR was carried out for 8-12 cycles depending upon the sample using the following cycling conditions: Denaturation - 95 °C (5 min); PCR cycle - 95 °C (30 sec), 62 °C (30 sec), 72 °C (30 sec); Final extension - 72 °C (7 min). Each sample was carefully optimized for PCR conditions to avoid over-amplification. For each sample, the minimum number of cycles required to yield highest amount of the specific PCR product while undergoing less than 10% reduction in the amount of free primers as compared to their starting amounts was chosen as the optimal number of PCR cycles. 200-300 µl of the PCR reactions were performed for each sample depending upon the yield. The volume of pooled PCR reactions was adjusted to 500 µl with water and DNA was extracted with equal volume of phenol/chloroform/Isoamyl alcohol (25:24:1, pH ~8.0). After a 13,000 xg spin for 5 min at RT, 450 µl of aqueous phase was transferred to a new tube. Equal volume of 5M Ammonium acetate and 4 µl of 20mg/ml glycogen was added following which 0.7 volumes of isopropanol was added to precipitate DNA. The tubes were incubated at RT for 5 min and centrifuged at 13,000 xg for 20 min at RT. The supernatant was removed and the pellet was washed in 1 ml 70% ethanol at least three times (4-5 min spin at 13,000 xg for each wash). After removing supernatant from the last wash, the tubes were spun briefly (10-30 sec)

to remove all liquid. DNA was re-suspended in 15-20 µl of water without any air-drying. A 6X DNA gel-loading buffer was added and the PCR products were separated at 80V on 6% native PAGE gel (10 x 10 cm). The gel slice with appropriate PCR product was excised and shredded by centrifuging the slice through a narrow bore in a PCR tube placed in an eppendorf tube at 10,000 xg for 1 minute at RT. DNA was eluted overnight in 600 µl of 1:1 mix of 5M Ammonium acetate and TE buffer (pH 8.0). The gel pieces were removed by passing the sample through a Spin-X column. The DNA was precipitated from the flow-through by adding 1/100 volume of 20 µg/ml glycogen and 0.7 volumes of isopropanol at RT for 5 min followed by centrifugation at 13,000 xg for 20 min at RT. All the supernatant was removed and pellet was re-suspended in water and stored at -80°C. 2-22 µl of 500 pM PCR amplified DNA libraries were sequenced on ABI SOLiD sequencers (systems 3 and 4) for all samples except the one from FLAG-Magoh:eIF4AIII-long EJC footprints. This library was constructed as above except that two-step PCR amplification (8 cycles with SOLiD/Illumina-PCR-5′ and SOLiD/Illumina-PCR-3′ PCR oligos followed by 12 cycles with Illumina PE1.0 and PE2.0 oligos) was performed with primers that enabled sequencing on Illumina platform (single-end 100 nt read) using the Illumina custom primer.

eIF4AIII knockdown and oligo-dT pull-down of polyA+ RNA-crosslinked proteins

For knockdown of eIF4AIII, or GAPDH as a control, ~60% confluent culture of

HEK293 Flp-In cells growing in 35mm wells (6-well plates) were transfected with 50 nM siRNA oligo (ON-TARGET eIF4AIII_268 [custom siRNA] or ON-TARGETplus GAPD control siRNA [Dharmacon]) was transfected using Lipofectamine RNAiMAX reagent. ~50 hrs post-transfection, cells from 6 wells/knockdown were UV-crosslinked at 800 mJ/cm^2 on an ice-bath and subsequently lysed in 4 ml of 1X oligo-dT binding buffer [10 mM Tris-HCl pH 7.5, 0.5%SDS, 0.5 M NaCl]. Lysates were homogenized by passing 6 times through a 25-guage needle and protein concentration was equalized in the two knockdown samples. PolyA+ RNA was isolated by incubating lysates with 30 mg of oligo-dT cellulose resin (reconstituted and pre-rinsed in 1X oligo-dT binding buffer). Following 2 hr incubation at 25°C, the resin was washed four times with 1 ml of 1X oligo-dT binding buffer and once with non-denaturing buffer [10 mM Tris-HCl pH 7.5, 0.1%NP40, 0.1% Triton-X 100, 0.5 M NaCl]. Proteins pulled-down through RNA were eluted by digestion with RNase A (5 U/µl) + RNase T1 (200 U/µl) cocktail in 400 µl elution buffer [10 mM Tris-HCl pH 7.5, 1 mM EDTA] at 37°C for 1 hr with intermittent shaking. Eluted proteins were TCA-precipitated and analyzed on 15% SDS-PAGE/Western blots as described above. Proteins on western blots were detected either on Odyssey scanner (fluorescence) or by autoradiography (chemiluminiscence). The X-ray autoradiograms were quantified using ImageJ.

Mapping of deep sequencing reads

I. Reference sequence sets

The human genomic reference sequence (version hg18) was downloaded from UCSC. To map reads that originated from mRNA exon-exon junctions, all unique and validated exon junctions in the human RefSeq transcriptome (release 47 downloaded March 2011 from UCSC genome browser) (Pruitt et al., 2007) were assembled as described before (Wang et al., 2008). Specifically, sequences 25 nt upstream and downstream of each unique exon junction in mature RefSeq transcripts were joined into a 50 nt long reference sequence in fasta format. In cases where the downstream or upstream exon lengths were shorter than 25 nt, sequences of next consecutive exons were used to create 50 nt reference sequence.

II. Preprocessing of the SOLiD reads and alignment to the human genome

Sequences of small RNA fragments from EJC footprints (12-24 nt), RNA-Seq (25-35 nt), nuclear mRNP protection fragments (15-35 nt) and high molecular weight EJC footprints (24-36 nt) were obtained in color space from SOLiD platform. To trim the adapter sequence and to convert them from color space to sequence space we employed the SOLiD small RNA pipeline (version 0.5.0 custom-fitted with human genomic and exon junction reference sequences) and Ma2Gff conversion tool (version 0.2.06). This pipeline yielded reads in fasta format with error correction by genome referencing. Bowtie (0.12.7) (Langmead et al., 2009) was used to subsequently map the reads with up to one error in the following steps. First, the reads that map to structural RNAs (ribosomal RNAs, tRNAs, snRNA) and repetitive elements were filtered out. Second, we obtained reads that uniquely map to the reference bowtie index containing human genome

and exon junction sequences described above. For the exon junction mapping reads, only those reads where at least four nucleotides cross either side of the junction were considered. Since reads from 25 nt regions abutting the exon junctions will not map uniquely to the reference index in the previous step, we finally mapped the rest of the reads to human genome again. Thus, we obtain all the reads mapping uniquely to human transcriptome. Figure S4B shows a flowchart of this pipeline with the number of reads processed in each step for all the libraries shown in Fig. S4C.

The FLAG-Magoh:eIF4AIII-long footprint reads obtained from Illumina platform were mapped to the human genome and splice-junction reference sequences with Bowtie as described above. For direct comparison to short footprints, only the first 18 nucleotides of reads from the long footprint libraries (tandem-IP long footprints and gel-filtration HMW long footprints) were mapped to the reference sequences.

III. Assigning reads to genomic regions

To assign reads or called peaks (below) into exonic, intronic and intergenic regions, we first merged all known unique exonic regions annotated in RefSeq, preserving the strand information using the mergeBed function of BedTools (Quinlan and Hall, 2010). The regions between the transcript start and end positions that remained un-represented were put into a separate bed file for intronic regions. Within each gene, all annotated coding regions were similarly merged and the remaining sequences were defined as 5′ and 3′UTRs. Similarly, regions between transcripts comprised the intergenic set. The read counts for

exonic, intronic, intergenic, 5′ UTR, 3′UTR and coding regions were detected using the intersectBED function from BedTools.

Selection of a representative RefSeq transcript from HEK293 RNA-Seq

To achieve the best estimation of EJC signal within each gene, we assembled a set of mRNAs that contained one representative transcript for each gene detectably expressed in HEK293 cells. Importantly, the representative transcript chosen for each gene was the most abundant isoform based on the calculated transcript RPKM (reads per kilobase per million mapped reads) (Mortazavi et al., 2008) for each isoform using our RNA-Seq data from HEK293 cells.

Analysis of variation in cEJC occupancy

To assess the variable occupancy at the cEJC site, we first selected those mRNAs that have three or more cEJC sites (-15 to -31 nt window from exon junction) with ≥1 RNA-Seq read per nucleotide. First we calculated the mean read number and its standard deviation within all cEJC sites that satisfy the above criteria. Using these, we then calculated the coefficient of variation (C_v=Standard deviation/mean) for each gene. Kernel density of coefficient of variation in EJC, RNA-Seq and nuclear mRNP protection libraries was plotted with R (package 2.11.0) using default bandwidth.

Peak calling and classification

To quantitatively analyze our short EJC footprint libraries, we developed a novel peak-calling algorithm for RNA-Seq data. For this, we converted all mapping outputs to wiggle format. We observed that the level of background noise in exonic regions as measured by total number of reads in ncEJC positions was higher in more abundant transcripts (data not shown). To account for this effect, this algorithm computes expression-sensitive backgrounds for exons, but uses expression-insensitive backgrounds for introns and intergenic regions. We modeled the distribution of read counts in each exon as a Poisson distribution, parametrized by an exon specific λ_b. To estimate λ_b for each exon, we used a maximum likelihood approach as follows. First, abundance of each RefSeq exon (exon RPKM) is calculated from the RNA-Seq library. All exons were then ranked by their RPKM values. For each exon, 1000 exons with the closest RPKM level were ranked based on the number of reads in given EJC library. The exons contained in the top and bottom 2.5% of this list were removed to minimize the outlier effect. We finally used the maximum likelihood estimator of (λ_b) using the number of reads at each position of the remaining 950 exons such that

$$\lambda_b = \sum_i \frac{k_i}{l_i}, i \in [-475, 475]$$

Our preliminary results indicated a strong accumulation of reads in the EJC library at the position 24nt upstream of the exon junctions. To sensitize the algorithm to detect peaks outside this region, we excluded reads centered around positions -15 to -31 upstream of exon junctions from the estimation of (λ_b). Finally, peaks on each exon were detected using its specific λ_b. Given that

the highest λ_b in our datasets were less than 10, we approximated the p-value of the Poisson test as follows:

$$f(k_i, \lambda_b) = Pr(X = k_i) = \frac{{\lambda_b}^{k_i} e^{-\lambda_b}}{k_i!}$$

This probability is equal to the probability of observing *k* reads at the given position assuming it is coming from the background distribution. If this probability is $<10^{-2}$, the peak footprint was extended to the next nucleotide. This iterative process was used until a nucleotide position was reached where the probability dropped below the significance threshold. A final probability for observing the mean number of reads in the peak was then calculated for the entire region of all significant contiguous positions using the background λ_b as above. If a peak was comprised of five or less reads, it was eliminated. Each of the remaining peaks was assigned three values: a *probability-value*, footprint size and peak height equal to the number of unique reads that cross the most highly represented nucleotide in the peak.

For detecting peaks in the intronic and intergenic regions, a general λ_b value was calculated using the total number of reads in each region across the genome and their total length in the genome. The peaks were then called using the method as described above.

From the list of called peaks, those with three or fewer sequence species and one of those species outnumbering the others by tenfold or more were annotated as “block” peaks. These peaks likely represented PCR jackpots and were eliminated from further analysis. Peaks were assigned to exonic, intronic or

intergenic sites using the approach described above. Further, all exonic peaks were classified into three categories: canonical (peaks with centers between -15 and -31 window), wide-canonical (peak with centers in -15 to -31 window with the signal extending beyond -40 position) and non-canonical peaks (peaks with centers outside -15 to -31 window).

Largest EJC peak set: The peaks detected in individual replicates of each short EJC footprint libraries were highly correlated. We combined uniquely mapped reads from different biological replicates of EJC IP configuration (three replicates of FLAG-Magoh:eIF4AIII and two of FLAG-eIF4AIII:Y14) and repeated the peak calling and classification procedures as above. A common approach for combining biological replicates is to consider each as an independent sample and use the mean number of reads at each position. The mean is considered the maximum likelihood estimator of the true expression level at each position/peak. However, in our peak calling algorithm we assumed a Poisson distribution of reads based on count statistics. Poisson distribution is known to underestimate the observed variance in the distribution of the number of deep sequencing (Oshlack et al., 2010). We therefore opted to use the sum of the number of reads at each position in different replicates of the same IP configuration prior to peak calling. Despite being a biased estimator, the sum of reads increases the variance of the estimated background distribution and reduces the false positive peak calls.

Given that FLAG-Magoh:eIF4AIII IP replicates had the highest number of uniquely mapping reads, peak calling on this set yielded the largest set of EJC

peaks (Fig. S5A).

Reproducible EJC peaks: To find the most reproducible EJC peaks, we obtained a simple intersection between the two sets of peaks obtained above from combined replicates of FLAG-Magoh:eIF4AIII and FLAG-eIF4AIII:Y14 libraries. Overlapping peaks from the two IPs were compared and if the shorter of the two overlapped at least 50% with the other peak, the peak in the FLAG-Magoh:eIF4AIII set was designated as a reproducible peak.

Factors influencing cEJC occupancy

We used representative transcripts selected as described above to test for correlation between the presence of a cEJC peak and various parameters. For each parameter, the distribution of values for cEJC-free and cEJC-occupied sites was plotted using R software package.

I. Mappability

Mappability was determined for each nucleotide from positions -24 to -35 nt relative to exon junctions. For each 18 nt window in this range, if the sequence starting at the given position was uniquely mappable to the human genome, it received a score of 1 and a score of 0 otherwise. The scores for all 12 sliding windows were summed to get the total mappability score of the region -24 to -35 from exon junctions (maximum and minimum possible scores: 12 and 0). The distribution of mappability scores of the cEJC-free sites was compared to those of the cEJC-occupied sites from transcripts with RPKM>10 and that contained an intron using the Wilcoxon rank sum test.

Following this analysis, cEJC-occupied and cEJC-free sites with mappability score ≥8 from all intron-containing transcripts with RPKM >10 were included in the analyses described in *II*-*VI* below.

II. Sequence composition

The sequences from the cEJC-free and cEJC-occupied sites were visualized using WebLogo version 2.8. (Crooks et al., 2004).

III. Secondary structure prediction

We selected sequences from either -10 to -50 or -50 to -90 nt upstream of the exon junctions from our reference transcripts. Free energy of folding for each sequence was calculated using the hybrid-ss-min function from the UNAFold suite version 3.8 (Markham and Zuker, 2008). For the calculations, all default options were used in addition to options –l (prohibit 1 bp helices) and –E (report energy scores only).

IV. 5′SS and 3′SS strength calculation

To calculate the splice site strengths, 5′ and 3′ splice site (SS) strengths of the introns were calculated using the MaxEntScan algorithm and the default parameters (Yeo and Burge, 2004). The MaxEntScan was downloaded from: http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html

For 5′SS calculation, sequences 3 nt upstream and 6nt downstream from each annotated exon-intron junction were input into MaxEntScan. Sequences 20 nt upstream and 3 nt downstream of each annotated intron-exon junction were used for 3′SS score calculations.

V. Exon and intron lengths

Lengths of exons with cEJC-free or cEJC-occupied sites, or lengths of downstream introns were based on the representative RefSeq transcript. Exons from transcripts with one or more introns and RPKM>1 were used for the analysis.

VI. Conservation

Conservation scores at each position within the -10 to -50 window was obtained from the phyloP algorithm using the 44-way vertebrate multiple sequence alignment (Pollard et al., 2009) downloaded from the UCSC genome browser. The mean conservation for each window was used for the comparison. For wobble positions, the reading frame was based on the coding region annotation from the RefSeq database and mean conservation at all wobble positions in the window was computed.

Over and under-represented mRNAs in EJC footprint libraries

I. Average cEJC occupancy signal. From the representative transcripts (above), we selected those with RPKM >0 and with at least one intron. For each gene, we calculated mean EJC peak height by dividing total heights of canonical and wide-canonical EJC peaks with the number of splicing events yielding these peaks. This represented the average EJC signal per gene and was plotted as a function of the expression of the representative transcript as measured by RPKM. These data were fit using a local regression approach (loess, R software package). For each data point, standard error (SE) was used to calculate the *z*-score of average EJC occupancy for each gene as follows:

$$z_{score} = \frac{EJC_o - EJC_f}{SE_f}$$

Here EJC_o and EJC_f are the average cEJC peak heights for a given transcript and the expected occupancy given its expression based on the loess fit, respectively. SE_f is the standard error calculated by the loess function for each point on the loess curve fitted to the data. All mRNAs were sorted by the *z*-scores. Top and bottom 5% mRNAs were those with high and low cEJC occupancy, respectively.

II. ncEJC signal. To measure the amount of signal in the FLAG-Magoh:eIF4AIII short footprint EJC library from non-canonical sites, read density in internal exons was calculated after subtracting reads mapping to cEJC sites. These non-canonical read densities for each gene were plotted against transcript RPKM followed by curve-fitting and z-score calculations as described above to identify mRNAs enriched (top 5%) or depleted (bottom 5%) in ncEJC signal.

Functional Analysis

mRNAs displaying high or low EJC occupancy (in human refseq_RNA namespace) were input into the web-based portal for FuncAssociate 2.0 (Berriz et al., 2009): http://llama.mshri.on.ca/funcassociate/

The analysis was carried out in the unordered mode with a significance cut-off of 0.05 (adjusted p-value for multiple hypotheses testing using 1000 different permutations to generate the null distribution). A custom associations file with known AS-NMD targets (Saltzman et al., 2008) as a functional category was generated as described (Berriz et al., 2009).

Signal in the intronless genes

From the FLAG-Magoh:eIF4AIII short footprint library, all exon mapping reads for transcripts from intron-containing (12455) and intronless genes (628) was plotted versus their RPKM. The loess regression fit and z-scores calculations were performed as above. Z-score distributions were plotted for transcripts from intron-containing and intronless genes, and were compared using a Kolmogrov-Smirnov test.

Average cEJC occupancy up- and downstream of alternative splicing events

A list of mRNAs with annotated cassette exons that introduce a PTC upon inclusion or exclusion (PTC+), or those where alternative splicing of a cassette exon does not introduce a PTC (PTC-), were obtained from (Saltzman et al., 2008). After aligning all transcripts in each set by the cassette exon, fraction of cEJC occupied sites at every successive exon in either direction in the set was plotted (number of EJC occupied exons/total number of exons in the set at the position).

Motif finding

For motif discovery at the ncEJC sites, sequences from wide-canonical and non-canonical peaks were used. These peaks were further classified based on their occurrence in the first exons, internal exons, last exons or unspliced mRNAs.

To estimate the relevant background model, we calculated the GC-content for each of the regions above (first exons, internal exons, last exons or unspliced mRNAs) using all transcripts in RefSeq. Given that our sequences of interest are based on RNA, we modified the motif finding program AlignAce (Roth et al., 1998) to search for motifs on the given strand only (Hon Nian Chua, unpublished). The footprint sequences of reproducible peaks were divided into two equal parts. Fasta format sequences from one part along with the calculated GC-content for the background model was input into AlignACE for motif discovery using default parameters. A position-specific scoring matrix (PSSM) was built using the AlignACE output for each motif. Specifically, we counted the occurrences of each nucleotide at each position and added pseudocounts using the relative background frequency of each nucleotide weighted by 0.1. Then we divided the corrected occurrences by the number of observed motifs plus 0.1. Single nucleotide sliding windows in each of the sequences that contain the motif were scored using these PSSM. The maximum score obtained for each sequence was considered its representative score. The minimum PSSM score obtained from the entire set was picked as the threshold for determining motif presence. We then scored sequences that were left out from the discovery phase. The fraction of peaks with motifs was determined using this independent validation set based on the PSSM threshold defined above.

Motif significance analysis: For testing statistical significance of discovered motifs, each parent mRNA sequence was shuffled one thousand times while preserving tri-nucleotide composition. For the internal exonic motifs, which mostly

occur in the coding region, the tri-nucleotides from the annotated translation frame were preserved during shuffling. For the motifs in the first exons, tri-nucleotides in the entire transcript were shuffled. Using the same PSSM score threshold as above, the fraction of randomized sequences with motif occurrences was determined. The motifs with scores at least three-times greater than the median score of the randomized independent set were considered significant and chosen for further analysis.

For each motif containing sequence, the distance from the center of the motif to the center of its parent peak sequence was calculated (if sequences contained even number of nucleotides, last position in the 5′ half was chosen). Also, the motif-start position was compared to the translation frame of the mRNA to calculate the fraction of motif-containing sequences in each frame.

Motif clustering: A Fisher's exact test was performed to estimate the similarity between each motif pair considered for clustering. The parameters used were: total number of unique peak sequences used for finding the two motifs, number of unique peak sequences contributing to each motif and number of common peak sequences between each motif pair. These Fisher's exact test values were used as distance for hierarchical clustering using the R package hclust complete method.

ESE and ESS occurrence frequencies

Hexamers predicted and validated to function as ESEs or ESSs using two different approaches were obtained (Fairbrother et al., 2002; Ke et al., 2011).

The number of raw reads containing these hexamers was counted for each deep sequencing library. These numbers were divided by the total uniquely mapped reads in the library to display the distributions on the same plot.

Correlation between adjacent exon peaks and auto-correlation function

From the observed mean cEJC occupancy within each RPKM bin (1-3, 3-10, 10-20, 20-30, >30), expected frequencies on two adjacent cEJC sites for the following events were calculated: occupied-occupied, occupied-free and free-free. The observed frequencies of the three events were measured from our largest peak set library (FLAG-Magoh:eIF4AIII). The distributions of the observed and expected frequencies among the three classes were compared for statistical significance using a Chi-square test (degrees of freedom=2).

To test for correlation between cEJC peaks in adjacent exons within each mRNA (RPKM>10), a vector containing EJC peak status (present/absent) for each exon of the mRNA was compared to itself (using auto correlation function in R) after one- (lag 1), two- (lag 2) or three- (lag 3) shifts in exon positions. As a control, EJC peak status for all exons of an mRNA was randomized and the auto-correlations for the three lags were calculated as above. Wilcoxon rank sum test was used to compare the distributions of auto-correlation values observed in the actual and randomized sets.

Supplemental References

Berriz, G.F., Beaver, J.E., Cenik, C., Tasan, M., and Roth, F.P. (2009). Next generation software for functional trend analysis. Bioinformatics (Oxford, England) *25*, 3043-3044.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188-1190.

Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science (New York, N.Y *297*, 1007-1013.

Fenger-Gron, M., Fillman, C., Norrild, B., and Lykke-Andersen, J. (2005). Multiple processing body factors and the ARE binding protein TTP activate mRNA decapping. Mol. Cell *20*, 905-915.

Ke, S., Shang, S., Kalachikov, S.M., Morozova, I., Yu, L., Russo, J.J., Ju, J., and Chasin, L.A. (2011). Quantitative evaluation of all hexamers as exonic splicing elements. Genome Res. *21*, 1360-1374.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Markham, N.R., and Zuker, M. (2008). UNAFold: software for nucleic acid folding and hybridization. Methods Mol. Biol. *453*, 3-31.

Merz, C., Urlaub, H., Will, C.L., and Luhrmann, R. (2007). Protein composition of human mRNPs spliced in vitro and differential requirements for mRNP protein recruitment. RNA *13*, 116-128.

Mili, S., Shu, H.J., Zhao, Y., and Pinol-Roma, S. (2001). Distinct RNP complexes of shuttling hnRNP proteins with pre-mRNA and mRNA: candidate intermediates in formation and export of mRNA. Mol. Cell. Biol. *21*, 7307-7319.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621-628.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. *7*, 548.

Oshlack, A., Robinson, M.D., and Young, M.D. (2010). From RNA-seq reads to differential expression results. Genome biology *11*, 220.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2009). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. *20*, 110-121.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *35*, D61-65.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England) *26*, 841-842.

Radulovic, D., Jelveh, S., Ryu, S., Hamilton, T.G., Foss, E., Mao, Y., and Emili, A. (2004). Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. Mol. Cell. Proteomics *3*, 984-997.

Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat. Biotechnol. *16*, 939-945.

Saltzman, A.L., Kim, Y.K., Pan, Q., Fagnani, M.M., Maquat, L.E., and Blencowe, B.J. (2008). Regulation of multiple core spliceosomal proteins by alternative splicing-coupled nonsense-mediated mRNA decay. Mol. Cell. Biol. *28*, 4320-4330.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337-342.

Searle, B.C. (2010). Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. Proteomics *10*, 1265-1269.

Silva, J.C., Gorenstein, M.V., Li, G.Z., Vissers, J.P., and Geromanos, S.J. (2006). Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Mol. Cell. Proteomics *5*, 144-156.

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.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. Nature *456*, 470-476.

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.

Supplemental Figure 1

Figure S1, related to Figure 1

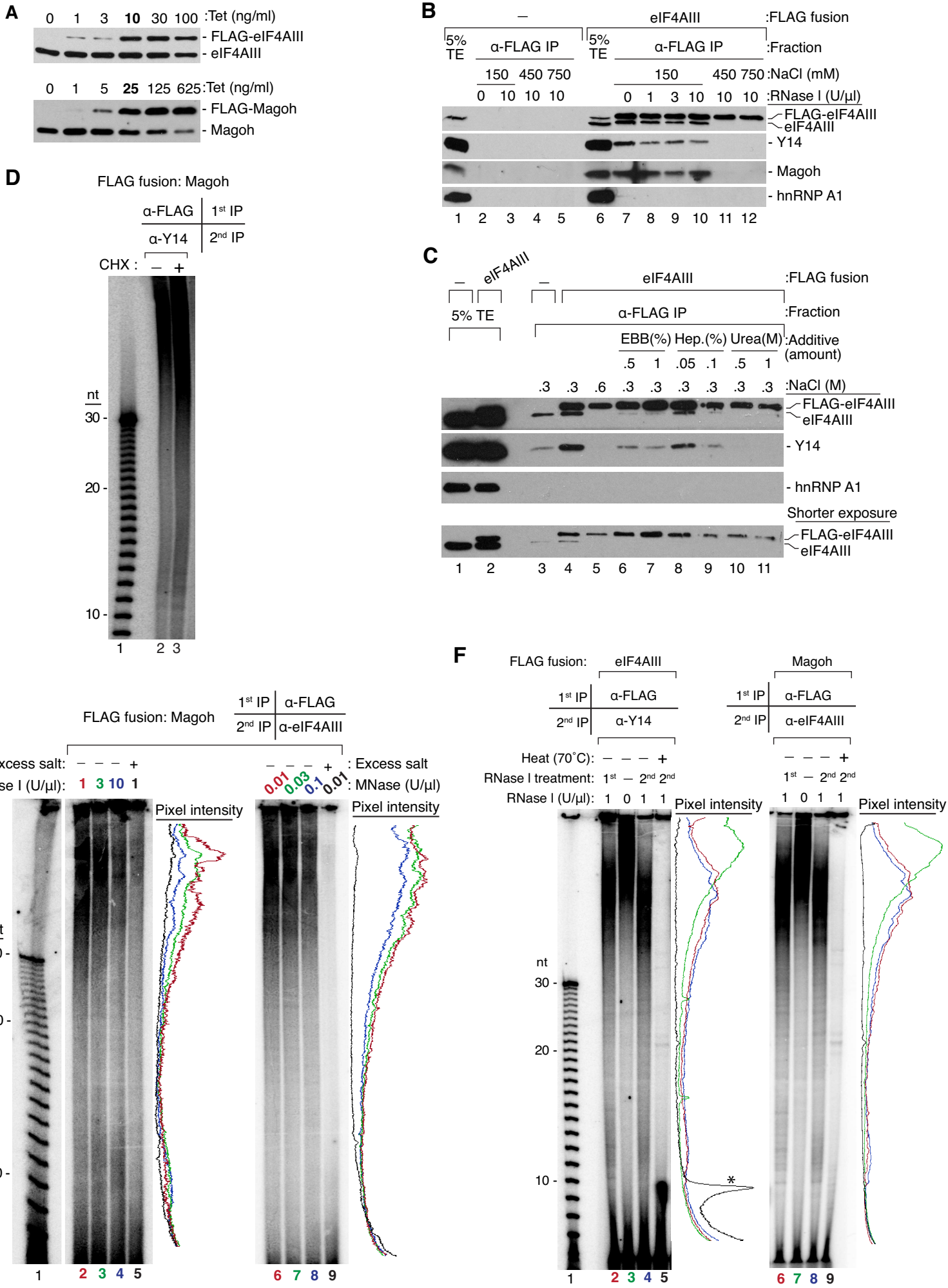
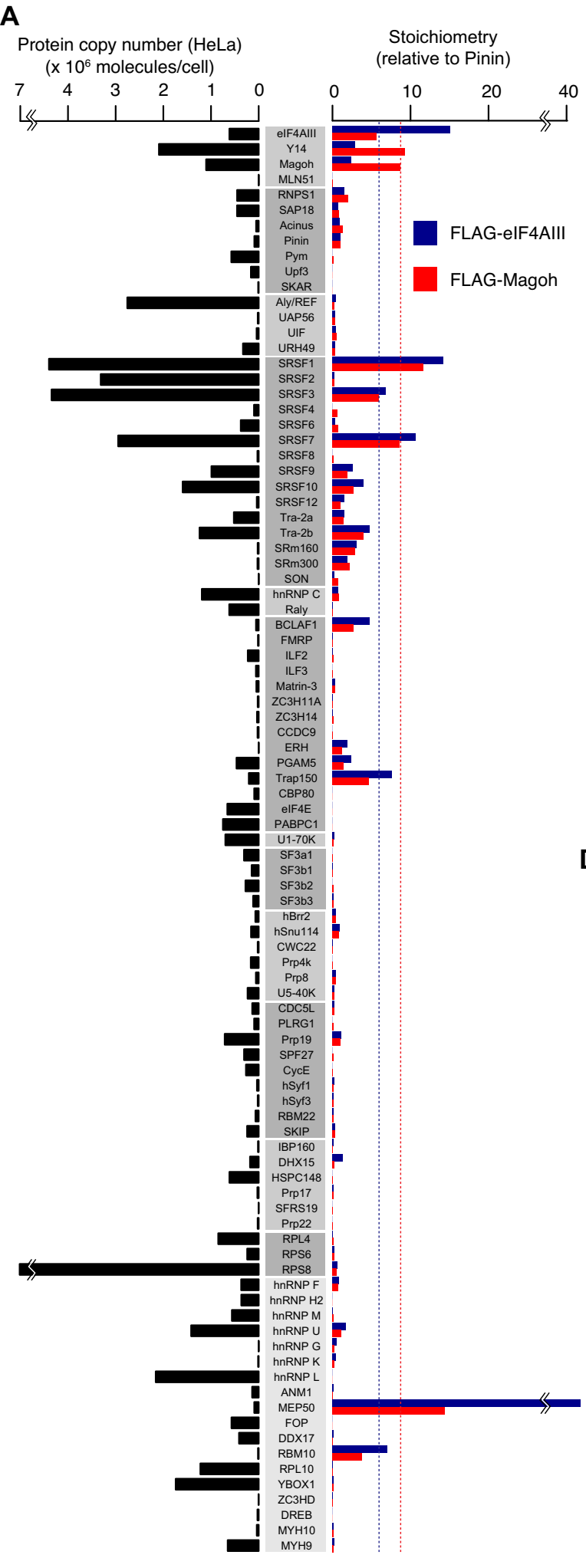
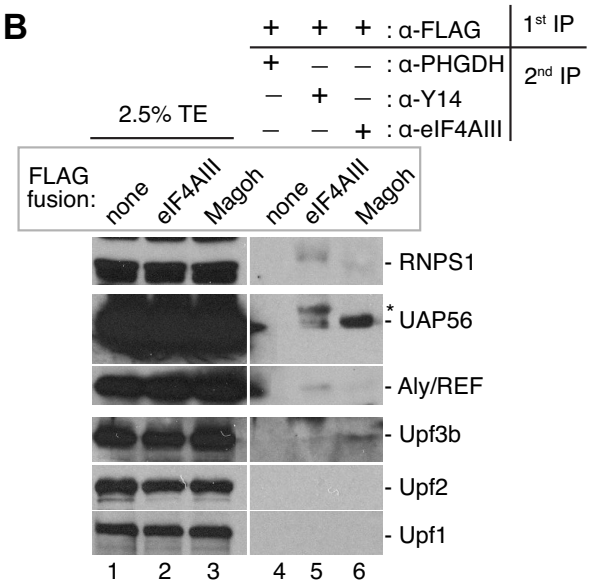
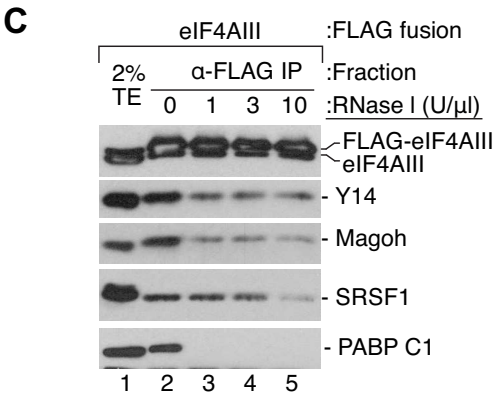
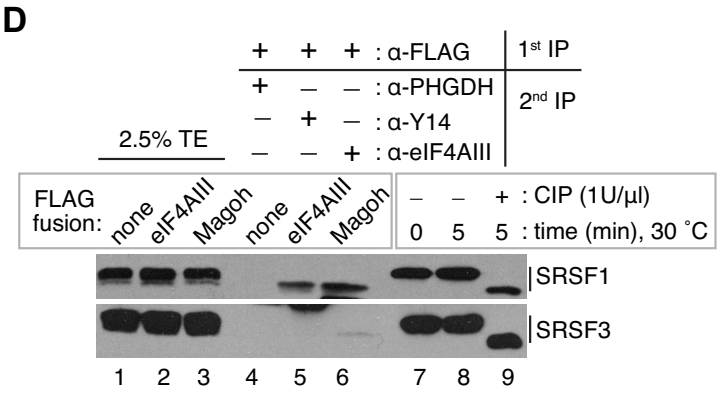


Figure S2 related to Figure 2









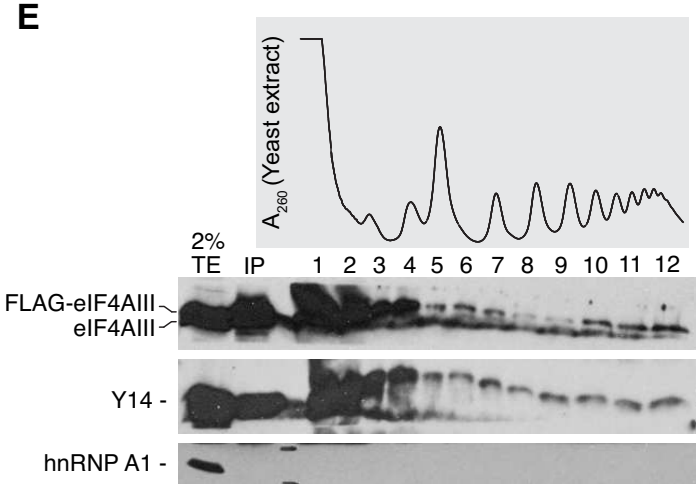
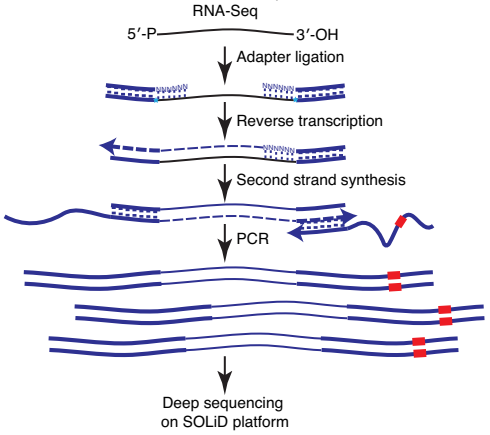


Figure S3, related to Figure S3

Supplemental Figure S3

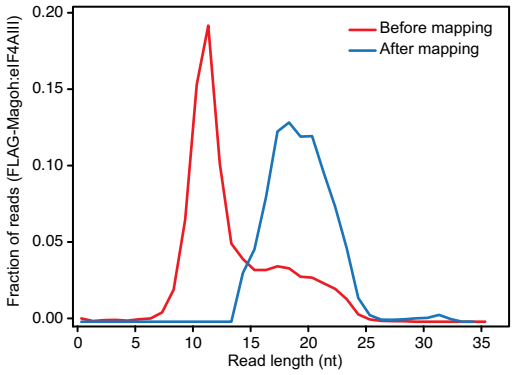
A



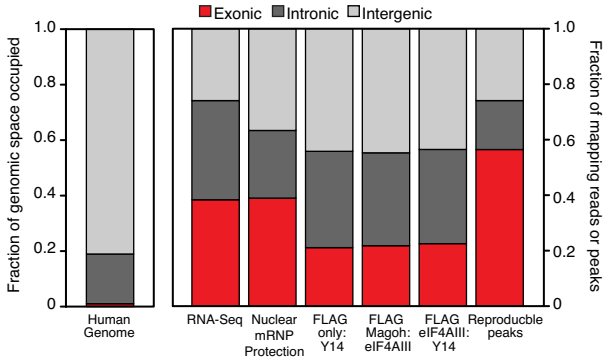
B

	All reads	Processed		Filtered Out		Unique Map	
Deep Sequencing library	**millions**	**millions**	**%**	**millions**	**%**	**millions**	**%**
FLAG only:Y14	22.6	11.0	48.7	1.4	6.2	4.9	21.7
FLAG only:PHGDH	9.6	0.9	9.3	0.0	0.5	0.4	3.9
FLAG-eIF4AIII:Y14 (1)	86.6	19.5	22.5	1.5	1.7	8.9	10.2
FLAG-eIF4AIII:Y14 (2)	26.4	7.1	26.8	0.7	2.6	3.5	13.1
FLAG-eIF4AIII:Y14	113.1	26.6	23.5	2.1	1.9	12.3	10.9
FLAG-Magoh:eIF4AIII (1)	27.7	14.1	50.8	1.2	4.3	6.3	22.9
FLAG-Magoh:eIF4AIII (2)	88.4	23.0	26.1	1.3	1.4	13.0	14.7
FLAG-Magoh:eIF4AIII (3)	83.7	34.6	41.4	3.0	3.6	18.3	21.8
FLAG-Magoh:eIF4AIII	199.8	71.7	35.9	5.5	2.7	37.6	18.8
FLAG-eIF4AIII:Y14 (+ formald	9.4	2.9	30.9	0.3	3.2	1.2	12.8
RNA-Seq	186.4	73.7	39.5	10.2	5.5	41.6	22.3
Nuclear mRNP protection	143.3	41.8	29.1	15.3	10.7	16.2	11.3
FLAG-eIF4AIII-HMW	148.4	66.9	45.1	31.2	21.0	10.1	6.8
FLAG-Magoh:eIF4AIII-long	24.3	24.3	100	7.6	31.1	13.1	53.9

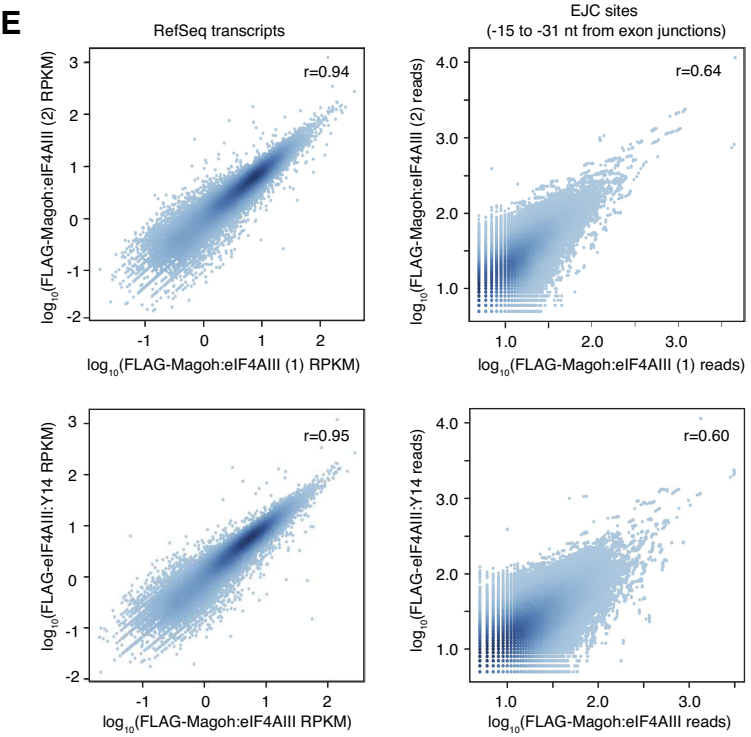
C



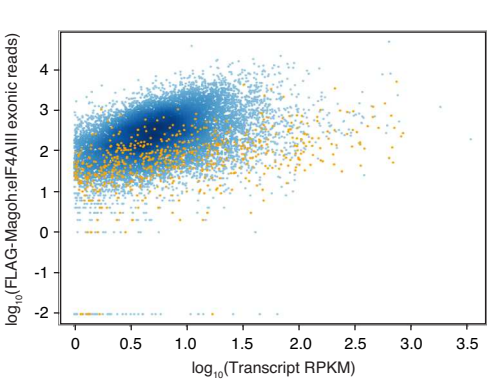
D



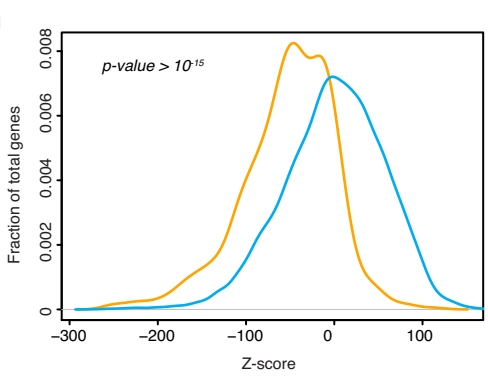
E



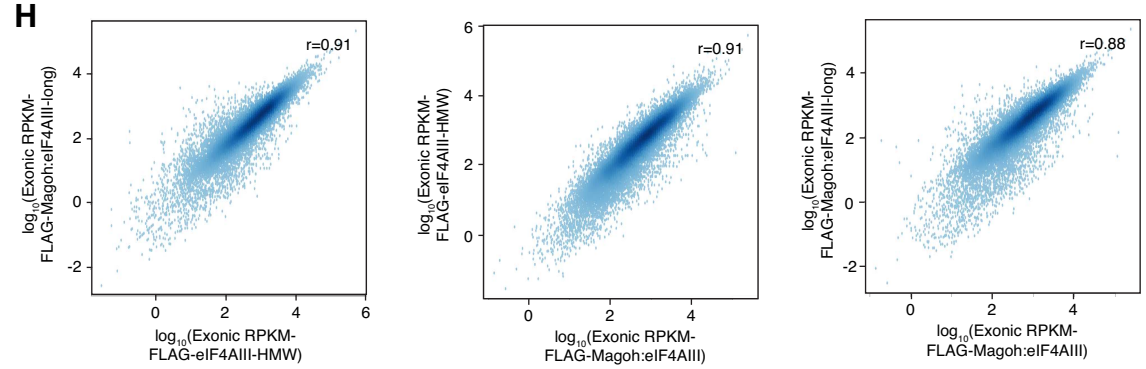
F



G



H



Supplemental Figure 4

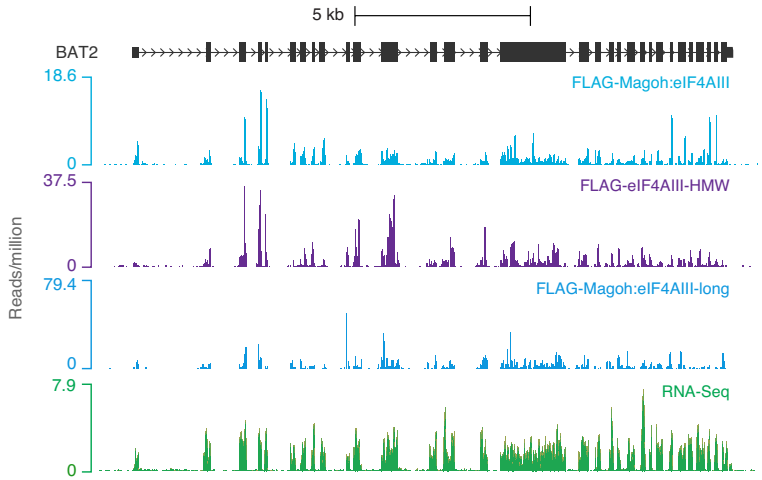
Figure S4, related to Figure 4

A

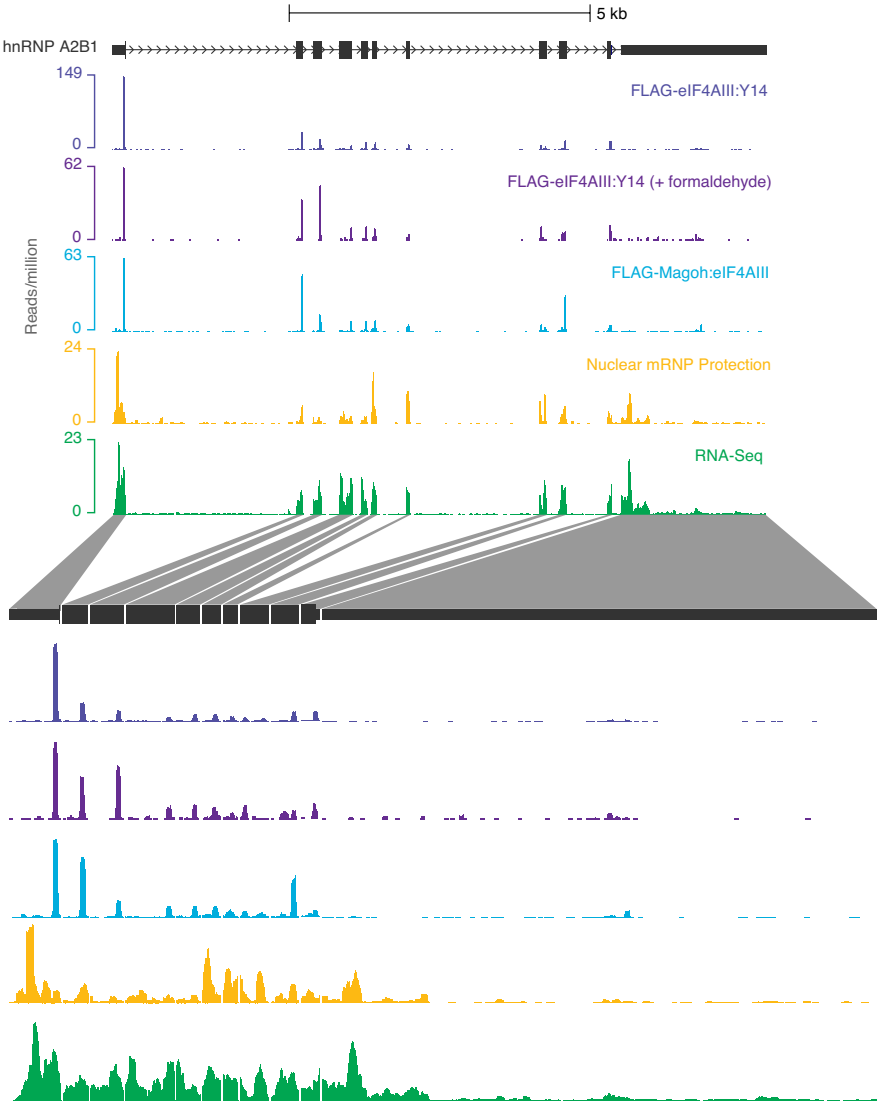
Tandem IP	All peaks			
	Exonic	Intronic	Intergenic	Total
FLAG-Magoh:eIF4AIII	146578 (47.8)	67466 (22.0)	92308 (30.2)	306352
FLAG-eIF4AIII:Y14	65903 (49.8)	27327 (20.7)	39062 (29.5)	132292
Reproducible [shared]	58175 (56.4)	18176 (17.6)	26713 (26.0)	103064

Tandem IP	Exonic peaks				
	First exon	Internal exons	Last exon	Intronless genes	Total
FLAG-Magoh:eIF4AIII	18465 (12.6)	114643 (78.2)	12242 (8.4)	1228 (0.8)	146578
FLAG-eIF4AIII:Y14	9338 (14.1)	53321 (80.9)	2745 (4.2)	499 (0.8)	65903
Reproducible [shared]	7639 (13.1)	48037 (82.6)	2114 (3.6)	385 (0.7)	58175

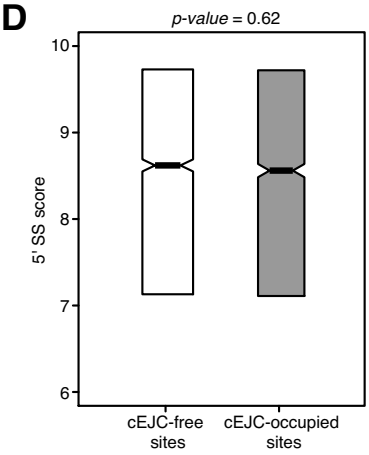
B



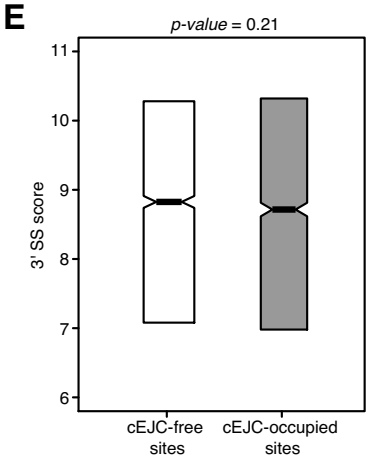
C



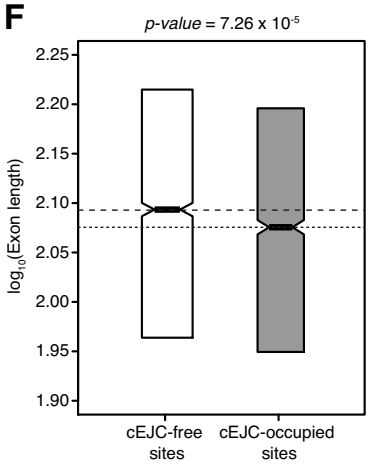
D



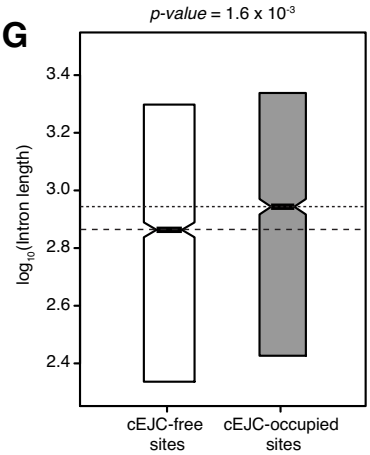
E



F

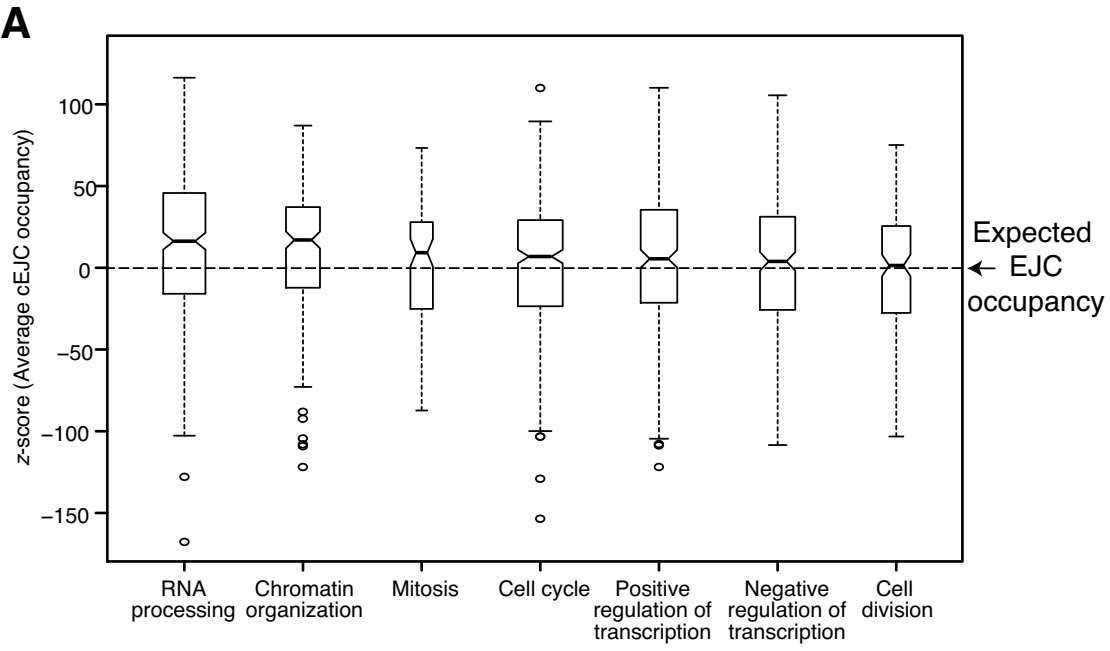


G



Supplemental Figure 5

Figure S5, related to Figure 5



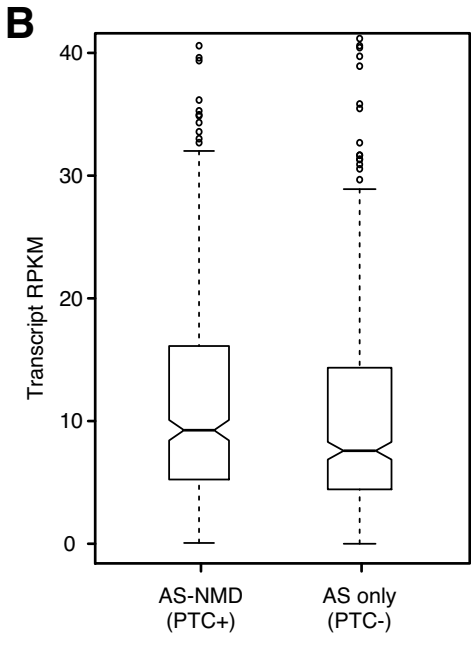


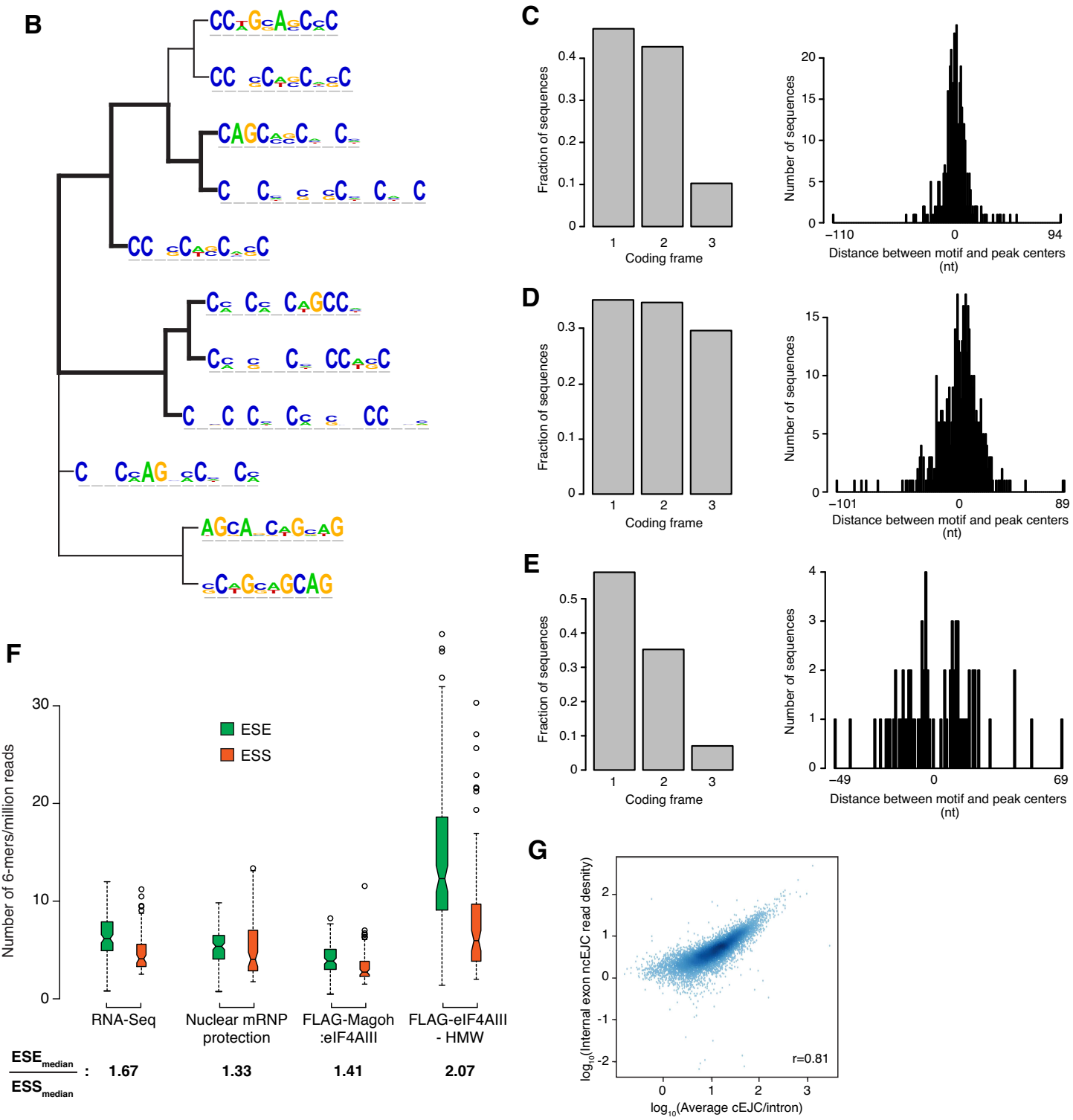
Figure S6, related to Figure 6

Supplemental Figure 6

A

Tandem IP	Exonic peaks (Intron-containing genes)		
	cEJC	ncEJC	Total
FLAG-Magoh:eIF4AIII	63620 (43.8)	81730 (56.2)	145350
FLAG-eIF4AIII:Y14	37409 (57.2)	27995 (42.8)	65404
Reproducible [shared]	35058 (60.7)	22732 (39.3)	57790

Tandem IP	Exonic peaks [downstream intron +]					
	First exon			Internal exon		
	cEJC	ncEJC	Total	cEJC	ncEJC	Total
FLAG-Magoh:eIF4AIII	5275 (28.6)	13190 (71.4)	18465	58345 (50.9)	56298 (49.1)	114643
FLAG-eIF4AIII:Y14	3740 (40.1)	5598 (59.9)	9338	33669 (63.1)	19652 (36.9)	53321
Reproducible [shared]	3370 (44.1)	4269 (55.9)	7639	31688 (66.0)	16349 (34.0)	48037



Supplemental Figure 7

Figure S7, related to Figure 7