2019-11-19

Deep sequencing of pre-translational mRNPs reveals hidden flux through evolutionarily conserved AS-NMD pathways [preprint]

Carrie Kovalak
University of Massachusetts Medical School

Et al.

Title Page

Title: *Deep sequencing of pre-translational mRNPs reveals hidden flux through evolutionarily conserved AS-NMD pathways*

Authors and affiliations: Carrie Kovalak[1,2], Mihir Metkar[1,3], Melissa J. Moore[1,2,3*]

[1] RNA Therapeutics Institute, University of Massachusetts Medical School, Worcester, MA 01655, USA

[2] Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01655, USA

[3] Present address: Moderna, 200 Technology Square, Cambridge, MA 02139, USA

* Correspondence: melissa.moore@umassmed.edu

Abstract

Background: The ability to generate multiple mRNA isoforms from a single gene by alternative splicing (AS) is crucial for the regulation of eukaryotic gene expression. Because different mRNA isoforms can have widely differing decay rates, however, the flux through competing AS pathways cannot be determined by traditional RNA-Seq data alone. Further, some mRNA isoforms with extremely short half-lives, such as those subject to translation-dependent nonsense-mediated decay (AS-NMD), may be completely overlooked in even the most extensive RNA-Seq analyses.

Results: RNA immunoprecipitation in tandem (RIPiT) of exon junction complex (EJC) components allows for the purification of post-splicing mRNA-protein particles (mRNPs) not yet subject to translation (pre-translational mRNPs) and translation-dependent mRNA decay. Here we compared EJC RIPiT-Seq to whole cell and cytoplasmic RNA-Seq data from HEK293 cells. Consistent with expectations, we found that the flux through known AS-NMD pathways is substantially higher than what is captured by RNA-Seq. We also identified thousands of previously unannotated splicing events; while many can be attributed to “splicing noise”, others are evolutionarily-conserved events that produce new AS-NMD isoforms likely involved in maintenance of protein homeostasis. Several of these occur in genes whose overexpression has been linked to poor cancer prognosis.

Conclusions: Deep sequencing of RNAs in post-splicing, pre-translational mRNPs provides a means to identify and quantify splicing events without the confounding influence of differential mRNA decay. For many known AS-NMD targets, the NMD-linked AS pathway dominates. EJC RIPiT-Seq also enabled identification of numerous conserved but previously unknown AS-NMD events.

Keywords (3-10)

Exon junctions, mRNA isoforms, AS-NMD, RIPiT-Seq, pre-translational mRNPs, splicing noise

Background

A central mechanism underlying metazoan gene expression is alternative pre-mRNA processing, which regulates the repertoire of mRNA isoforms expressed in various tissues and under different cellular conditions. Extensive deep sequencing of RNA (RNA-Seq) has revealed that ~95% of human protein-coding genes are subject to alternative splicing (AS) [1, 2], with current estimates suggesting ~82,000 different protein-coding mRNA isoforms generated from ~20,000 protein coding genes [3]. Thus, production of alternative mRNA isoforms massively expands the protein repertoire that can be expressed from a much smaller number of genes [4, 5]. But cells also need to control how much of each protein is made. Although transcriptional control is often considered the predominant mechanism for modulating protein abundance, emerging evidence indicates that post-transcriptional regulatory mechanisms are crucial as well.

Not all mRNA variants are protein-coding. Nearly 15,000 human mRNAs in the Ensembl database (release 93) are annotated as nonsense-mediated decay (NMD) targets [3]. NMD is a translation-dependent pathway that both eliminates aberant mRNAs with malformed coding regions (i.e., those containing premature termination codons due to mutation or missplicing) and serves as a key mechanism for maintenance of protein homeostasis [6]. This protein homeostasis function is mediated by AS linked to NMD (AS-NMD), wherein the flux through alternate splicing pathways that result in protein-coding and NMD isoforms is subject to tight control [7]. These NMD isoforms harbor a premature termination codon either due to frameshifting or inclusion of a poison cassette exon. Because NMD isoforms are rapidly eliminated after the first or “pioneer” round of translation, only protein-coding isoforms result in appreciable protein production (**Figure 1A, bottom**). Thus increasing or decreasing flux through the NMD splicing pathway decreases or increases protein production, respectively. Although AS-NMD was originally described as a mechanism by which RNA binding proteins (e.g., SR and hnRNP proteins) could autoregulate their own synthesis, recent work indicates that AS-NMD is much more pervasive, tuning abundance of many other proteins such as those involved in chromatin modification and cellular differentiation [8].

The true extent to which AS-NMD contributes to protein homeostasis can only be appreciated by determining the flux through the protein-coding and NMD splicing pathways. Transcriptome-wide assessment of mRNA isoform abundance generally relies on RNA-Seq of whole cell or cytoplasmic RNA. Such methods provide a static snapshot of the species present in the sample at the time of collection. Because NMD isoforms are so rapidly decayed, they are generally underrepresented in RNA-Seq datasets. Thus a single RNA-Seq snapshot is generally uninformative as to synthetic flux through protein-coding and NMD splicing pathways.

An alternate means to assess protein-coding and NMD pathway flux is to capture newly synthesized mRNAs after splicing completion but prior to translation. Late in the splicing cycle, the exon junction complex (EJC) is deposited upstream of at least 80% of exon-exon junctions (canonical; cEJCs) and multiple other sites throughout the length of spliced exons (noncanonical; ncEJCs) [9, 10]. Upon nucleocytoplasmic export, the pioneer round of translation removes EJCs within the 5′ UTR and CDS regions, with EJCs remaining downstream of stop codons being key mediators of NMD [11]. Pre-translational mRNPs can be selectively isolated by tandem immunoprecipitation of epitope-tagged and untagged EJC components, a technique known as RNA:protein immunoprecipitation in tandem (RIPiT) [12]. Deep sequencing library preparation from RIPiT samples (RIPiT-Seq) has previously enabled us to map the positions of canonical and noncanonical EJCs on spliced transcripts [9] and to investigate the RNA packing principles within pre-translational mRNPs [13].

Here, we compare libraries from pre-translational mRNPs (EJC RIPiT), unfractionated RNA (whole cell RNA-Seq) and RNA post subcellular fractionation (cytoplasmic RNA-Seq) (**Figure 1A**). As expected, EJC RIPiT libraries are enriched for transcript isoforms destined for translation-dependent decay. By providing a window into the repertoire of transcripts generated by splicing but prior to translation-dependent decay, EJC RIPiT libraries provide a more accurate record of the flux through various alternative processing pathways than does standard RNA-Seq. Importantly, EJC RIPiT libraries enabled us to identify numerous new evolutionarily-conserved poison cassette exons that had previously eluded annotation based on even highly extensive RNA-Seq data analyses.

Results

EJC, whole cell and cytoplasmic libraries

In our recent study investigating the organizing principles of spliced RNPs [13], we generated three biological replicates from HEK293 cells of EJC-bound RNAs partially digested with RNase T1 during RNP purification (**Figure 1A**). Paired-end deep sequencing of these EJC RIPiT libraries resulted in 19-25 million mate pairs each (**Supplemental Table 1**). For comparison to RNA-Seq libraries, we chose rRNA-depleted whole cell and cytoplasmic HEK293 RNA-Seq datasets (two biological replicates each) previously published by Sultan et al. [14]. We chose these particular libraries based on their similarity in cell treatment and library preparation to our EJC libraries, their clean cellular fractionation, and sequencing depth (51-57 million mate pairs each).

All libraries were downloaded from their respective repositories (see Declarations) and processed in parallel. Reads were aligned to the Genome Reference Consortium Human Build 38 (GRCh38.p12) [3] using STAR (v2.5.3a) [15] after first filtering out those mapping to repeat RNAs [16]. To minimize the effect of misalignment in ensuing analyses, mismatches were limited to three per read, with gaps caused by deletions or insertions being strongly penalized. These strict mapping parameters resulted in 6-10 million and 30-43 million aligned pairs for the EJC RIPiT and RNA-Seq libraries, respectively (**Supplemental Table 1**). For quantification, we limited all analyses to unique reads with high mapping quality (MAPQ ≥ 5). For all libraries, we used Kallisto (v0.44.0) to derive expression values for the ~200,000 annotated transcripts in GRCh38.p12 [3]. Examination of per-transcript abundance revealed high concordance (≥ 0.93 to 0.99) among all biological replicates (**Supplemental Figure 1A**). Therefore, all subsequent quantitative analyses utilized merged biological replicate data.

EJC libraries are enriched for spliced transcripts and translation-dependent decay targets

To assess the relative representation of NMD targets in EJC and RNA-Seq libraries, we first examined read coverage on known AS-NMD genes. The SR proteins TRA2B and U2AF2 negatively regulate their own expression by promoting inclusion of a highly-conserved poison cassette exon containing a premature termination codon (**Figure 1B, C**). Although these poison exons were detectable in all library types, they were much more abundant in the EJC libraries. As expected due to NMD, cytoplasmic RNA-Seq libraries exhibited the lowest poison exon inclusion (percent spliced in; PSI) values (16% and 4%, respectively), with the whole cell libraries being somewhat higher (29% and 6%, respectively). Yet, the EJC RIPiT libraries indicate much higher inclusion percentages (averaging 94% and 73%, respectively). Thus, for both TRA2B and U2AF2, the predominant splicing pathway in HEK293 cells under standard growth conditions is poison exon inclusion. Similar trends were observed for other known AS-NMD targets (**Supplemental Figure 1B-D**), including hnRNPA1 where the AS-NMD isoform results from splicing in the 3′UTR as a consequence of alternative polyadenylation (**Figure 1D**). The substantial differences between the EJC RIPiT and RNA-Seq quantitations for these previously documented AS-NMD isoforms clearly illustrate the advantage provided by the EJC RIPiT libraries for more accurately assessing flux through alternative processing pathways that result in mRNA isoforms with widely different decay rates.

In GRCh38.p12, every transcript isoform is given a specific annotation [17]; relevant annotations in protein-coding genes are “protein-coding”, “NMD”, “NSD”, “retained intron”, and “processed transcript”, with the latter being a catch-all for transcripts not clearly attributable to any other category. NSD (non-stop decay) is another translation-dependent mRNA degradation pathway that eliminates transcripts having no in-frame stop codon [18]. When exported to the cytoplasm, transcripts containing one or more retained introns are also usually subject to translation-dependent decay due to the presence of in-frame stop codons in intronic regions. For transcripts detectable in our libraries [TPM >0 in all replicates of a particular library type (EJC, whole cell or cytoplasmic)], the number of exon junctions (i.e., positions at which introns were removed) per protein-coding and NMD isoform ranged from 0 to >100 and 1 to 69, respectively (**Figure 2A**). As expected, protein-coding isoforms having no exon junctions were less abundant in the EJC libraries than in either RNA-Seq library (**Figure 2B, top**). In contrast, spliced protein-coding isoforms containing 5 or more exon junctions were enriched in EJC libraries, with the degree of enrichment increasing with exon junction number. For each exon junction number bin (i.e., 1-4, 5-10 and 10+), NMD isoforms were even more enriched in EJC libraries than were protein-coding isoforms (**Figure 2B, bottom**). EJC library enrichment was also readily discernible in per-transcript scatter plots for NMD, NSD, retained intron, and processed transcript isoforms (**Figure 2C, D and Supplemental Figure 2A, B**). All of these observations are consistent with the notion that EJC-associated RNAs are enriched for spliced transcripts subject to subsequent elimination by translation-dependent decay.

Because they are not translated, long intergenic non-coding RNAs (lincRNAs) are not subject to translation-dependent decay. As expected, lincRNAs lacking exon junctions (e.g., MALAT1, RMRP, NEAT1 and NORAD) were substantially depleted from EJC libraries, whereas those containing exon junctions were of similar or higher abundance in EJC than RNA-Seq libraries (**Figure 2E** and **Supplemental Figure 2C**). Particularly notable was XIST, the most highly represented Pol II transcript in our EJC libraries [13]. XIST is both spliced and exclusively nuclear. Reflecting this, median abundance of the eight major XIST isoforms was five- and forty-fold greater in EJC than in whole cell and cytoplasmic libraries, respectively.

EJC libraries capture new exon junctions

Having established that spliced transcripts known to be eliminated by translation-dependent decay are enriched in EJC libraries, we next wondered whether EJC libraries might contain new transcript isoforms that had previously eluded detection due to their low abundance in RNA-Seq. Such isoforms should contain previously unannotated exon junctions. To identify all previously annotated exon junctions, we integrated the RefSeq (hg38) [19], Ensembl (GRCh38.p12) [3], GENCODE (v29) [20] and Comprehensive Human Expressed SequenceS (CHESS) transcriptome annotations to create a comprehensive reference file containing 575,976 known introns (**Supplemental Table 2**). CHESS is derived from 9,795 RNA-Seq samples from diverse cell types in the GTEx collection, so represents the most complete compendium of human transcripts reported to date [21]. Yet while CHESS found 118,183 new exon junctions not previously annotated in RefSeq, Ensembl or GENCODE, 82,918 other junctions present in RefSeq, Ensembl and/or GENCODE were not returned by the CHESS pipeline (**Figure 3A**). This lack of concordance with respect to annotated junctions shows that even the most comprehensive RNA-Seq data analyses are unlikely to capture all bona fide splicing events.

To identify annotated and unannotated exon junctions in our EJC, whole cell and cytoplasmic libraries, we considered only those reads that cross an exon junction. The position of an exon junction in an individual read can be found by examining the “N operation” in the CIGAR string, which indicates the locations and lengths of gaps inserted during alignment to genomic DNA (**Supplemental Figure 3A**). We further required that any candidate junction: (1) occur within an annotated gene; (2) have reads with ≥15 nt aligning on both sides of the junction (≥90% exact sequence match on each side); (3) be detectable in all replicates of a particular library type (EJC, whole cell or cytoplasmic); and (4) have a mean read count ≥2 per library type (**Supplemental Figure 3A**). Using these criteria, we identified 151,072 junctions contained in the RefSeq/Ensembl/GENCODE/CHESS reference file (annotated junctions) and 5,917 previously unannotated junctions. MEME analysis of the latter revealed the 5′ and 3′ splice site consensus motifs for the major spliceosome, although at somewhat lesser strength (bits) than annotated junctions (**Figure 3B**). To limit our analysis to events most likely representing real splicing events (as opposed to mapping artifacts), we subsequently only considered the 5,412 previously unannotated junctions where the putative intron began and ended with dinucleotides expected for either the minor (AT-AC) or major (GT-AG) spliceosome. Of these, only three had AT-AC termini, indicating that the vast majority (>99.9%) of the unannotated events we detected are due to intron excision by the major spliceosome.

The majority (73%) of previously-annotated exon junctions meeting our detection criteria in protein coding genes (**Supplemental Figure 3A**) were present in all three library types (**Figure 3C, left**). There was less concordance, however, with respect to unannotated junctions, with the EJC libraries having many more unannotated junctions than either whole cell or cytoplasmic RNA-Seq (**Figure 3C, right**). Consistent with the expectation that EJC libraries should be enriched for exon junctions, both annotated and unannotated junctions were supported by more reads per million mapped (RPM) in the EJC libraries (**Figure 3D**). Also as expected, annotated junctions were generally supported by more reads than unannotated junctions in all library types. The major class (49%) of the new junctions were new alternative 5′ or 3′ splice sites (i.e., that combined a known 3′ or 5′ splice site with a previously unannotated 5′ or 3′ splice site, respectively) (**Figure 3E**). Other categories were previously unannotated exon skipping events (34%), new cassette exons (14%) and new introns (4%).

Relationship of new splicing events to reading frame

Previous analyses of low abundance, unannotated splicing events in RNA-Seq data have revealed a strong tendency for such events to maintain reading frame [22, 23]. To investigate whether this is due to some inherent ability of the splicing machinery to detect reading frame in the nucleus [24, 25], or simply due to translation-dependent decay of out-of-frame events in the cytoplasm, we determined the distance from each previously unannotated splice site meeting our selection criteria to the nearest annotated splice site observed in any of our three library types. In all, 126 and 273 unannotated 5′ and 3′ splice sites, respectively, occurred within 15 nts of an annotated 5′ or 3′ splice site. Comparison of unannotated-to-annotated splice site distance aggregation plots between the three library types revealed both similarities and differences (**Figure 4A**). Around annotated 5′ splice sites, all three libraries displayed similar patterns, with the greatest unannotated usage being at intron position +5, consistent with the preference for a G and a T at positions +5 and +6, respectively, in the human 5′ splice site consensus sequence (**Figure 3B**) and the prevalence of GT dinucleotides at this position in this set of 126 5′ splice sites (dotted gray line in **Figure 4A**). More notable was the pattern near 3′ splice sites, where positions +3 and +4 in the downstream exon exhibited the highest unannotated usage. Strikingly, whereas the RNA-Seq libraries were strongly skewed toward position +3, both positions +3 and +4 in the EJC libraries were highly represented, with their usage closely reflecting the number of available AG's at these positions (dotted gray line in **Figure 4A**). Comparison of fractional abundance [unannotated read counts/(unannotated + annotated read counts)] at individual sites confirmed that whereas the EJC and RNA-Seq libraries exhibited similar utilization at position +3, utilization of position +4 was much more prominent in the EJC than either RNA-Seq library (**Figure 4B**). These observations strongly support a model in which out-of-frame splicing events are rapidly eliminated by NMD, resulting in their underrepresentation in both whole cell and cytoplasmic RNA-Seq libraries. Because utilization of AGs at positions +3 and +4 in the EJC libraries so closely paralleled their availability, we conclude that (at least with regard to 3′ splice sites) the splicing machinery has no ability to read frame.

Evolutionary conservation versus splicing noise

Regardless of reading frame, most unannotated splicing events are likely due to "splicing error" [26] or "splicing noise" [23]. Splicing noise results from spurious utilization of cryptic splice sites that are not evolutionarily conserved. To assess both evolutionary conservation and splice site strength, we calculated mean basewise phyloP 30-way vertebrate conservation [27] and MaxENT (a generally accepted measure of how well a particular splice site matches the consensus) [28] scores for both annotated and unannotated splice sites, using the same 5′ and 3′ splice site window sizes (9 and 23 nts, respectively) for both calculations (**Figure 5A**). We also calculated conservation and MaxENT scores for sequences chosen at random from inside annotated genes and containing either GT or AG at the appropriate position within the 5′ or 3′ splice site window, respectively. Plotting MaxENT versus conservation revealed markedly different distributions between annotated splice sites and random GT- and AG-containing sequences (**Figure 5B, Supplemental Figure 4**), with annotated sites being significantly skewed toward higher values for both measures. In contrast, whereas unannotated splice sites were similarly distributed as annotated splice sites with regard to MaxENT, the majority exhibited conservation scores more similar to random than annotated splice sites (**Figure 5C**).

For the random sequences, 95% had 5′ and 3′ splice site conservation scores below 1.03 and 0.63, respectively. Using these values as cutoffs to filter out the majority of events likely due to splicing noise (although this may be unnecessarily conservative for 3′ splice sites due to the high degree of overlap between the annotated and random conservation scores) left us with 252 (12%) and 630 (26%) evolutionarily-conserved unannotated 5′ and 3′ splice sites, respectively (**Supplemental Table 3**). The majority of these occurred within annotated protein-coding exons, so their conservation is likely driven by amino acid conservation and not as a requirement for recognition by the splicing machinery (see **Supplemental Figure 4C** for an example). Almost all of the new evolutionarily conserved introns (i.e., both the 5′ and 3′ splice sites were previously unannotated, but exhibited high conservation) also fell into this category. For the new introns, calculation of percent intron retention (PIR) in the EJC libraries revealed highly inefficient splicing (mean PIR = 0.93), and individual examination of those exhibiting the highest number of exon junction reads in the EJC libraries led to no findings of particular note. Thus the new introns likely constitute splicing noise due to low level spliceosome assembly on sites within exons that by happenstance resemble splice site consensus sequences. In contrast, examination of unannotated 3′ splice sites occurring within introns uncovered a conserved alternative splicing event in the HECTD4 (HECT domain E3 ubiquitin protein ligase 4) gene that adds 9 amino acids into the middle of the protein (**Supplemental Figure 4D**); this spliced isoform is currently annotated in mouse RefSeq and GENCODE, but not in humans. Other alternative 3′ splice sites in the CNOT1 and EEA1 genes generate AS-NMD isoforms (**Figure 5D, E**), the latter due to creation of a new poison cassette exon.

New evolutionarily-conserved poison cassette exons

Having found examples of new AS-NMD isoforms generated by unannotated 3′ splice sites, we were interested to investigate which of the new cassette exons identified here might also function in this capacity. Of the 445 new cassette exons (**Figure 3E**), 412 (93%) occurred in protein-coding genes; the remainder occurred in pseudogenes and ncRNAs. Based on the data in **Figure 1**, poison exons should exhibit higher abundance in EJC than in RNA-Seq libraries. Consistent with this, 315/412 (76%) were solely detectable in the EJC libraries, with the remainder averaging 12- and 13-fold higher abundance in the EJC libraries than in whole cell or cytoplasmic RNA-Seq, respectively (**Figure 6A**). Of the 377 new cassette exons detectable in EJC libraries, 70% were frameshifting (i.e., not a multiple of 3 nts long). Individual inspection of the 25 most abundant non-frameshifting exons revealed that 80% contained an in-frame stop codon. Therefore, as expected, the vast majority of new exons likely function as poison cassette exons.

To assess whether any of the new cassette exons constitute conserved regulatory elements, we calculated mean phyloP 30-way conservation scores across the entire exon. Combining these exon conservation scores (white to dark blue in **Figure 6B**) with the previously calculated 5′ and 3′ splice site conservation scores (**Figure 5B**) revealed a set of 20 previously unannotated cassette exons exhibiting both high internal (phyloP score ≥ 1) and high splice site (≥ 1 for both splice sites) conservation (**Figure 6B right; Supplemental Table 3**). Among these, the most highly represented in our datasets was a new 94 nt exon within intron 8 of the 22-intron protein tyrosine phosphatase, receptor type A (PTPRA) gene (**Figure 6C**). Reminiscent of the conserved poison exons in TRA2B and U2AF2 (**Figure 1B, C**), inclusion of Protein Tyrosine Phosphatase Receptor Type A (PTPRA) exon 8a was readily observable in the EJC libraries, but nearly undetectable in the RNA-Seq libraries (**Figure 6C**). Other high abundance examples were a 103 nt exon in intron 3 of the 29-intron DNA Polymerase Theta (POLQ) gene (**Supplemental Figure 5A**) and a 69 nt exon in intron 37 of the 39-intron pleckstrin homology domain interacting protein (PHIP) gene (**Figure 6D**). Although PHIP exon 37a does not frameshift, it does contain three highly-conserved in-frame stop codons (**Figure 6D, bottom**). Thus all of the new evolutionarily-conserved cassette exons identified here likely function as poison exons to regulate protein expression from their host gene.

Discussion

Here we demonstrate that deep sequencing of transcripts in pre-translational RNPs provides a means to identify/quantify mRNA isoforms underrepresented in or absent from RNA-Seq libraries due to their rapid elimination by translation-dependent mRNA decay. We captured this pre-translational population by tandem immunoprecipitation (RIPiT) [12] of two core EJC proteins. EJCs are stably deposited upstream of exon junctions late in the pre-mRNA splicing process, and EJCs in 5′ UTRs and coding regions (~98% of all) are necessarily removed during the first or “pioneer” round of ribosome transit. Thus the EJC provides an excellent handle by which to enrich for fully-processed, but not-yet-translated mRNAs (**Figure 1A**). Because they are specifically enriched for spliced transcripts, EJC RIPiT-Seq libraries also better capture low abundance splicing events than traditional RNA-Seq libraries. This enabled us to identify thousands of new exon junctions not currently annotated in any of four major reference datasets based on RNA-Seq. Many of these new splicing events generate isoforms subject to NMD, with some being evolutionarily-conserved AS-NMD regulatory events. Thus EJC RIPiT-Seq constitutes a useful method to query the spliced transcriptome without the confounding effects of differential translation-dependent decay of individual mRNA isoforms.

Flux through AS-NMD pathways

Since its initial description [29, 30], AS-NMD has increasingly emerged as a key post-transcriptional regulatory mechanism [31-33] (refs). Due to their widely different decay rates, however, the flux through the alternative processing pathways resulting in protein-coding and NMD isoforms cannot be determined by traditional RNA-Seq methods. As shown in **Figure 1** the vast majority of TRA2B and U2AF2 transcripts present in RNA-Seq libraries are the protein coding isoforms. Further, the lower poison exon PSI numbers in cytoplasmic than whole cell libraries are consistent with cytoplasmic decay of the NMD isoforms. The EJC RIPiT-Seq libraries, however, tell a very different story. For both TRA2B and U2AF2, the predominant pre-translational isoform is the poison-exon-included isoform, with poison exon PSIs averaging 94 and 73, respectively. Thus alternative splicing flux for both genes strongly favors poison exon inclusion. Similar results were observed for other RNA-binding protein genes known to maintain protein homeostasis by AS-NMD (**Supplemental Figure 1**). Indeed, enrichment of transcripts subject to translation-dependent decay (e.g., isoforms annotated as NMD and NSD) is a general feature of our EJC RIPiT libraries (**Figure 2**). We note, however, that to increase the abundance of pre-translational RNPs, we exposed our HEK293 cells to 2 mg/ml harringtonine for 60 minutes prior to cell harvest and lysis [13]. At least in yeast growing under suboptimal conditions, inhibition of translation can induce rapid transcriptional upregulation of genes involved in ribosome biogenesis [34]; the extent to which this is also true in mammalian cells growing under optimal conditions, and whether transcription and pre-mRNA processing of other gene classes are affected, has yet to be thoroughly explored. One recent study in HeLa cells, however, showed that, whereas a 15 minute exposure to 100 mg/ml cycloheximide had almost no effect on mRNA abundance in whole cell RNA-Seq, multiple mRNAs encoding ribosomal proteins decreased in abundance after a 24 hour cycloheximide treatment (i.e., the opposite of yeast) [35]. Because any transcriptional effects would confound the analysis, elimination of translation inhibitors would be advisable for any future EJC RIPiT-Seq study specifically aimed at quantifying flux through alternative RNA processing pathways in non-perturbed cells.

Identification of novel conserved splicing events

A major goal for this study was to assess the utility of EJC RIPiT-Seq libraries for identifying novel sites of exon ligation that are underrepresented in traditional RNA-Seq libraries. These could be splicing events resulting in either stable, low abundance isoforms or highly unstable transcripts such as NMD and NSD substrates. As illustrated in **Figure 3A**, even the deepest analysis of RNA-Seq to date (CHESS) failed to capture all of the exon ligation events annotated in RefSeq, Ensembl or GENCODE. CHESS combined data from 9,795 GTEx RNA-Seq libraries covering dozens of tissues and comprising just under 900 billion reads. Yet EJC RIPiT-Seq libraries from a single cell type grown under a single condition encompassing only ~60 million reads enabled us to identify thousands of new exon junctions not currently annotated in RefSeq, Ensembl, GENCODE or CHESS (**Figure 3C**). Whereas the majority of these events occur at sites lacking splice site conservation (**Figures 5 and 6**) and so likely constitute splicing noise, hundreds exhibit high sequence conservation among mammals. Among this conserved set, the majority display features expected to generate an AS-NMD isoform (i.e., frameshift or in-frame stop codon).

New poison exons regulate genes linked to cancer

It has now been well established that changes to pre-mRNA splicing patterns can drive cancer initiation and progression [36, 37] . Thus it is of particular note that three of the most conserved, high-abundance AS-NMD events discovered here are poison cassette exons in PTPRA, PHIP, and POLQ (**Figure 6**). All three genes have been linked to poor cancer prognosis when overexpressed [38-43]. While protein overexpression in cancer often results from gene duplication or transcriptional dysregulation, decreased flux through a splicing pathway leading to poison exon inclusion would have the same effect. Previous studies examining the links between NMD and cancer have mainly focused on loss of tumor suppressor genes due to increased NMD [44, 45] or the advantageous effects of NMD in eliminating mRNA isoforms encoding neoepitopes that would otherwise be recognized by the immune system [46]. But our findings suggest that decreased poison exon inclusion should also be considered as a contributor to the mechanisms underlying cancer. An obvious means to alter splicing flux is a cis-acting mutation that disrupts splice site recognition and, thereby, poison exon inclusion. Although our examination of The Cancer Genome Atlas (Release 19) [47] database revealed no instances of splice site mutations associated with any of the new conserved poison cassette exons documented here, this possibility should certainly be considered in future hunts for cancer-promoting mutations. Of note, current “exome” sequencing generally captures only DNA covering and surrounding annotated exons [48]. Therefore, the unannotated cassette exons we identify here are likely absent from most DNA sequencing databases.

Conclusions

Sequencing of post-splicing, pre-translational mRNPs provides a powerful new approach to identify and quantify transient species that undergo rapid translation-dependent decay and are therefore under-represented in or completely absent from standard RNA-Seq libraries. The data here constitute just one snapshot of AS flux in HEK293 cells growing under optimal conditions. Future studies examining EJC RIPiT-Seq libraries from more diverse biological samples will undoubtedly lead to discovery of even more previously undocumented AS-NMD pathways. Examination of how flux through such pathways change in response to changing cellular conditions will increase our general understanding of how post-transcriptional mechanisms regulate protein abundance.

Methods

Deep sequencing libraries

All libraries were downloaded from the NCBI GEO GSE115788 (specifically, samples GSM3189985, GSM3189986, and GSM3189987) and the European Nucleotide Archive PRJEB4197 (specifically, runs ERR304485, ERR304486, ERR304487, and ERR304488).

EJC libraries were generated from 200-550 nt fragments by 3′ adaptor ligation and reverse transcription. Paired-end sequencing (150 nt reads) on the Illumina NextSeq platform resulted in 18-24 million mate pairs per replicate [13]. RNA-Seq datasets were obtained by paired-end sequencing (51 nt reads) on the Illumina HiSeq platform of Ribo-Zero-treated libraries generated with a modified Illumina TruSeq protocol [14] containing 100 to 200 nt sized inserts. Each RNA-Seq replicate contained an average of 50 to 60 million mate pairs per library.

Library processing and alignment

Read counts for unprocessed libraries and for the individual processing steps detailed below are provided in **Supplemental Table 1**. Prior to alignment, adaptor sequences and long stretches (≥ 20 nt) of adenosines were trimmed from the 3′ end of sequencing reads. All libraries were filtered using STAR v2.5.3a [15] for reads that aligned to repeat regions, as defined by RepeatMasker [16]. Remaining reads were aligned with STAR on two-pass mode to the human genome, release 93 [3]. This alignment allowed a maximum of 3 mismatches per pair and highly penalized deletions and insertions. Mapped reads were then filtered for low mapping quality (MAPQ < 5) and/or duplicated reads, identified with the MarkDuplicates tool (Picard v2.17.8) [49].

RNA isoform quantification

RNA isoform abundances were determined using Kallisto (v0.44.0) [50], using only reads that passed the filtering and alignment steps described above. Transcript biotypes (i.e., “protein-coding”, “nonsense-mediated decay”, etc.) and intron counts used to categorize transcripts throughout **Figure 2** are based on the transcriptome annotation from Ensembl (GRCh38.p12) [3].

Junction identification pipeline

The custom bioinformatics pipeline designed for our annotated and unannotated junction analysis (**Figures 3 – 6**) is shown in detail in **Supplemental Figure 3A**. Transcriptome annotation files from RefSeq (hg38) [19], Ensembl (GRCh38.p12) [3], GENCODE (v29) [20], and CHESS (v2.1) [21] were combined to create a comprehensive reference file of all annotated introns (**Supplemental Table 2**). Any junction that appears in our libraries but is not annotated in one of the aforementioned transcriptomes is referred to as "unannotated."

To identify unannotated exon junctions, all reads with CIGAR strings containing an "N" operation were isolated and then compared to the annotated intron reference file using Bedtools intersect [51]. Reads that did not match the length or location of a known intron were considered the result of potential unannotated splicing events. These junctions were further filtered based on the following criteria: (i) overlap with a known gene, (ii) reads must have ≥15 nt aligned on both sides of the potential junction, (iii) present in all replicates of any library type, (iv) major spliceosome dinucleotide consensus sequences at the 5′ and 3′ splice sites, and (v) mean read count ≥ 2 per library type.

Nearest annotated splice site analysis

For analysis of new splicing events near annotated exons (**Figure 4**), each unannotated 5′ splice site was paired with its nearest annotated 5′ splice site based on the 3′ splice site used in both splicing events. Similarly, each unannotated 3′ splice site was paired with its nearest annotated 3′ splice site based on the 5′ splice site used in both splicing events. The number of available GT and AG dinucleotides at nucleotide positions -30 to +30 surrounding each annotated splice site in this unannotated/annotated paired dataset.

Splice site strength and conservation

Splice site strength and mean conservation scores for annotated and unannotated splice sites were calculated using MaxEntScan [28] and phyloP 30-way basewise conservation scores [27] (**Figure 5A**). Random sequences of the appropriate length (9 nts for 5′ splice sites and 23 nts for 3′ splice sites) and internal to annotated genes were obtained from the hg38 annotation file [3] using the Bedtools random function [51]. Only those random sequences containing a GT at positions 4 and 5 or an AG at positions 19 and 20 were used to calculate MaxENT and conservation scores for comparison to 5′ and 3′ splice sites, respectively.

Plotting and data visualization

Data visualization was performed in R [52] using ggplot2 [53], ggrepel [54], UpSetR [55], ggseqlogo [56], eulerr [57], and ggridges [58] software packages. The UCSC Genome Browser [59, 60] was used to view sequencing library tracks and for transcript figures throughout the manuscript.

List of abbreviations

AS	alternative splicing
RNA-Seq	deep sequencing of RNA
NMD	nonsense-mediated decay
NSD	non-stop decay
AS-NMD	alternative splicing linked to NMD
EJC	exon junction complex
cEJCs	canonical exon junction complex
nEJCs	noncanonical exon junction complex
RIPiT-Seq	RNA:protein immunoprecipitation in tandem followed by deep sequencing
PSI	percent spliced in
PIR	percent intron retained
lincRNA	long intergenic non-coding RNA
CHESS	Comprehensive Human Expressed SequenceS
RPM	reads per million
TPM	transcripts per million

Declarations

Additional files

Additional file 1: Includes Figures S1-S5 and Table S1. **Figure S1.** Comparison of biological replicates; additional examples of AS-NMD transcript coverage. **Figure S2.** Comparison of EJC and cytoplasmic RNA-Seq coverage across different transcript biotypes. **Figure S3.** Representation of bioinformatics pipeline to classify annotated and unannotated exon junctions. **Figure S4.** Comparison of MaxENT and conservation scores for annotated and randomly located splice sites; additional examples of conserved unannotated splicing events. **Figure S5.** Additional example of an unannotated AS-NMD cassette exon. **Table S1**. Sequencing and alignment information for each replicate of the analyzed libraries.

Additional file 2: Includes Table S2. **Table S2.** Comprehensive file of all intron locations annotated in RefSeq (hg38), Ensembl (GRCh38.p12), GENCODE (v29), and CHESS (v2.1**).**

Additional file 3: Includes Table S3. **Table S3.** Locations and characteristics of highly conserved unannotated exon junctions identified in this study.

Acknowledgments

We thank Alicia Bicknell, Athma Pai, Harleen Saini and Guramrit Singh for critical reading of the manuscript. We thank Weijun Chen for technical advice.

Funding

This work was supported by funding from HHMI and NIH RO1-GM53007. M.J.M. was an HHMI Investigator during the time this work was conducted.

Availability of data and materials

The RIPiT datasets analyzed in this study were downloaded from NCBI GEO under accession number GSE115788 (specifically, samples GSM3189985, GSM3189986, and GSM3189987). RNA-Seq datasets were downloaded from the European Nucleotide Archive under accession number PRJEB4197 (specifically, runs ERR304485, ERR304486, ERR304487, and ERR304488).

Authors' contributions

M.J.M conceived of the project. C.K. and M.J.M were responsible for all data analysis and manuscript writing. M.M. provided technical advice regarding the EJC RIPiT libraries and participated in manuscript editing.

Competing interests

M.M. and M.J.M. are currently employees of and shareholders in Moderna. M.J.M. is also a shareholder and scientific advisory board member for Arrakis Therapeutics. This work was entirely carried out in the M.J.M.'s research group at the University of Massachusetts Medical School.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Author details

M.J.M. is a member of the National Academy of Science (USA) and a fellow of the American Association of Arts and Sciences.

References

1. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008; doi:10.1038/nature07509.
2. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics. 2008; doi:10.1038/ng.259.
3. Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, Cummins C, Davidson C, Dodiya KJ, Gall A, Girón CG, Gil L, Grego T, Haggerty L, Haskell E, Hourlier T, Izuogu OG, Janacek SH, Juettemann T, Kay M, Laird MR, Lavidas I, Liu Z, Loveland JE, Marugán JC, Maurel T, McMahon AC, Moore B, Morales J, Mudge JM, Nuhn M, Ogeh D, Parker A, Parton A, Patricio M, Abdul Salam AI, Schmitt BM, Schuilenburg H, Sheppard D, Sparrow H, Stapleton E, Szuba M, Taylor K, Threadgold G, Thormann A, Vullo A, Walts B, Winterbottom A, Zadissa A, Chakiachvili M, Frankish A, Hunt SE, Kostadima M, Langridge N, Martin FJ, Muffato M, Perry E, Ruffier M, Staines DM, Trevanion SJ, Aken BL, Yates AD, Zerbino DR, Flicek P. Ensembl 2019. Nucleic Acids Res. 2019; doi:10.1093/nar/gky1113.
4. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature. 2010; doi:10.1038/nature08909.
5. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S. Function of alternative splicing. Gene. 2013; doi:10.1016/j.gene.2012.07.083.
6. Kurosaki T, Popp MW, Maquat LE. Quality and quantity control of gene expression by nonsense-mediated mRNA decay. Nature Reviews Molecular Cell Biology. 2019; doi:10.1038/s41580-019-0126-2.
7. Lewis BP, Green RE, Brenner SE. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. PNAS. 2003; doi:10.1073/pnas.0136770100.
8. Nasif S, Contu L, Mühlemann O. Beyond quality control: The role of nonsense-mediated mRNA decay (NMD) in regulating gene expression. Seminars in Cell and Developmental Biology. 2018; doi:10.1016/j.semcdb.2017.08.053.
9. Singh G, Kucukural A, Cenik C, Leszyk JD, Shaffer SA, Weng Z, Moore MJ. The cellular EJC interactome reveals higher-order mRNP structure and an EJC-SR protein nexus. Cell. 2012; doi:10.1016/j.cell.2012.10.007.
10. Saulière J, Murigneux V, Wang Z, Marquenet E, Barbosa I, Le Tonquèze O, Audic Y, Paillard L, Roest Crollius H, Le Hir H. CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. NSMB. 2012; doi:10.1038/nsmb.2420.
11. Maquat LE, Tarn WY, Isken O. The pioneer round of translation: features and functions. Cell. 2010; doi:10.1016/j.cell.2010.07.022.
12. Singh G, Ricci EP, Moore MJ. RIPiT-Seq: a high-throughput approach for footprinting RNA:protein complexes. Methods. 2014; doi:10.1016/j.ymeth.2013.09.013.

13. Metkar M, Ozadam H, Lajoie BR, Imakaev M, Mirny LA, Dekker J, Moore MJ. Higher-order organization principles of pre-translational mRNPs. Molecular Cell. 2018; doi:10.1016/j.molcel.2018.09.012.

14. Sultan M, Amstislavskiy V, Risch T, Schuette M, Dökel S, Ralser M, Balzereit D, Lehrach H, Yaspo ML. Influence of RNA extraction methods and library selection schemes on RNA-seq data. BMC Genomics. 2014; doi:10.1186/1471-2164-15-675.

15. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; doi:10.1093/bioinformatics/bts635.

16. Smit AFA, Hubley R, Green, P. RepeatMasker Open-4.0. 2013-2015. http://www.repeatmasker.org. Accessed 04 Sep 2019.

17. http://vega.archive.ensembl.org/info/about/gene_and_transcript_types.html

18. Klauer AA, van Hoof A. Degradation of mRNAs that lack a stop codon: a decade of nonstop progress. Wiley Interdiscip Rev RNA. 2012; doi:10.1002/wrna.1124.

19. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016; doi:10.1093/nar/gkv1189.

20. Frankish A, Diekhans M, Ferreira A, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, Barnes I, Berry A, Bignell A, Sala SC, Chrast J, Cunningham F, Domenico TD, Donaldson S, Fiddes IT, Girón CG, Gonzalez JM, Grego T, Hardy M, Hourlier T, Hunt T, Izuogu OG, Lagarde J, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Ruffier M, Schmitt BM, Stapleton E, Suner M, Sycheva I, Uszczynska-Ratajczak B, Xu J, Yates A, Zerbino D, Zhang Y, Aken B, Choudhary JS, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Reymond A, Tress ML, Flicek P. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019; doi:10.1093/nar/gky955.

21. Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang YC, Madugundu AK, Pandey A, Salzberg SL. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biology. 2018; doi:10.1186/s13059-018-1590-2.

22. Dou Y, Fox-Walsh KL, Baldi PF, Hertel KJ. Genomic splice-site analysis reveals frequent alternative splicing close to the dominant splice site. RNA. 2006; doi:10.1261/rna.151106.

23. Pickrell JK, Pai AA, Gilad Y, Pritchard JK. Noisy splicing drives mRNA isoform diversity in human cells. PLoS Genetics. 2010; doi:10.1371/journal.pgen.1001236.

24. Miriami E, Motro U, Sperling J, Sperling R. Conservation of an open-reading frame as an element affecting 5' splice site selection. J of Structural Biol. 2002; doi:10.1016/S1047-8477(02)00539-7.

25. Wachtel C, Li B, Sperling J, Sperling R. Stop codon-mediated suppression of splicing is a novel nuclear scanning mechanism not affected by elements of protein synthesis and NMD. RNA. 2004; doi:10.1261/rna.7480804.
26. Fox-Walsh KL, Hertel KJ. Splice-site pairing is an intrinsically high fidelity process. PNAS. 2009; doi:10.1073/pnas.0813128106.
27. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010; doi:10.1101/gr.097857.109.
28. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J of Computational Biol. 2004; doi:10.1089/1066527041410418.
29. Morrison M, Harris KS, Roth MB. smg mutants affect the expression of alternatively spliced SR protein mRNAs in Caenorhabditis elegans. PNAS. 1997; doi:10.1073/pnas.94.18.9782.
30. Mitrovich QM, Anderson P. Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans. Genes and Dev. 2000; doi:10.1101/gad.819900.
31. Zheng S, Gray EE, Chawla G, Porse BT, O'Dell TJ, Black DL. PSD-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2. Nature Neuroscience. 2012; doi:10.1038/nn.3026.
32. Hamid FM, Makeyev EV. Regulation of mRNA abundance by polypyrimidine tract-binding protein-controlled alternate 5' splice site choice. PloS Genetics. 2014; doi:10.1371/journal.pgen.1004771.
33. Yan Q, Weyn-Vanhentenryck SM, Wu J, Sloan SA, Zhang Y, Chen K, Wu JQ, Barres BA, Zhang C. Systematic discovery of regulated and conserved alternative exons in the mammalian brain reveals NMD modulating chromatin regulators. PNAS. 2015; doi:10.1073/pnas.1502849112.
34. Santos DA, Shi L, Tu BP, Weissman JS. Cycloheximide can distort measurements of mRNA levels and translation efficiency. Nucleic Acids Res. 2019; doi:10.1093/nar/gkz205.
35. Kearse MG, Goldman DH, Choi J, Nwaezeapu C, Liang D, Green KM, Goldstrohm AC, Todd PK, Green R, Wilusz JE. Ribosome queuing enables non-AUG translation to be resistant to multiple protein synthesis inhibitors. 2019; doi:10.1101/gad.324715.119.
36. Sveen A, Kilpinen S, Ruusulehto A, Lothe RA, Skotheim RI. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene. 2016; doi:10.1038/onc.2015.318.
37. Climente-González H, Porta-Pardo E, Godzik A, Eyras E. The functional impact of alternative splicing in cancer. Cell Reports. 2017; doi:10.1016/j.celrep.2017.08.012.
38. Tabiti K, Smith DR, Goh HS, Pallen CJ. Increased mRNA expression of the receptor-like protein tyrosine phosphatase alpha in late stage colon carcinomas. Cancer Letters. 1995; doi:10.1016/0304-3835(95)03816-f.
39. Ardini E, Agresti R, Tagliabue E, Greco M, Aiello P, Yang LT, Ménard S, Sap J. Expression of protein tyrosine phosphatase alpha (RPTPα) in human breast cancer correlates with low tumor grade, and inhibits tumor cell growth in vitro and in vivo. Oncogene. 2000; doi:10.1038/sj.onc.1203869.

40. Gu Z, Fang X, Li C, Chen C, Liang G, Zheng X, Fan Q. Increased PTPRA expression leads to poor prognosis through c-Src activation and G1 phase progression in squamous cell lung cancer. Intl J of Oncology. 2017; doi:10.3892/ijo.2017.4055.

41. De Semir D, Nosrati M, Bezrookove V, Dar AA, Federman S, Bienvenu G, Venna S, Rangel J, Climent J, Meyer Tamgüney TM, Thummala S, Tong S, Leong SP, Haqq C, Billings P, Miller JR 3rd, Sagebiel RW, Debs R, Kashani-Sabet M. Pleckstrin homology domain-interacting protein (PHIP) as a marker and mediator of melanoma metastasis. PNAS. 2012; doi:10.1073/pnas.1119949109.

42. Wood RD, Doublié S. DNA polymerase θ (POLQ), double-strand break repair, and cancer. DNA Repair. 2016; doi:10.1016/j.dnarep.2016.05.003.

43. Goullet de Rugy T, Bashkurov M, Datti A, Betous R, Guitton-Sert L, Cazaux C, Durocher D, Hoffmann JS. Excess Polθ functions in response to replicative stress in homologous recombination-proficient cancer cells. Biol Open. 2016; doi:10.1242/bio.018028.

44. Lindeboom RG, Supek F, Lehner B. The rules and impact of nonsense-mediated mRNA decay in human cancers. Nature Genetics. 2016; doi:10.1038/ng.3664.

45. Hu Z, Yau C, Ahmed AA. A pan-cancer genome-wide analysis reveals tumour dependencies by induction of nonsense-mediated decay. Nature Communications. 2017; doi:10.1038/ncomms15943.

46. Pastor F, Kolonias D, Giangrande PH, Gilboa E. Induction of tumour immunity by targeted inhibition of nonsense-mediated mRNA decay. Nature. 2010; doi:10.1038/nature08999.

47. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics. 2013; doi:10.1038/ng.2764.

48. Wang VG, Kim H, Chuang JH. Whole-exome sequencing capture kit biases yield false negative mutation calls in TCGA cohorts. PLoS One. 2018; doi:10.1371/journal.pone.0204912.

49. Picard. http://broadinstitute.github.io/picard. Accessed 04 Sep 2019.

50. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification, Nature Biotechnology, 2016; doi:10.1038/nbt.3519.

51. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; doi:10.1093/bioinformatics/btq033.

52. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2016. http://www.R-project.org/. Accessed on 04 Sep 2019.

53. Wickham H. ggplot2: elegant graphics for data analysis. Springer; 2016.

54. Slowikowski K. ggrepel: Automatically Position Non-Overlapping Text Labels with ‘ggplot2’. R package version 0.8.0. 2018. https://CRAN.R-project.org/package=ggrepel. Accessed on 04 Sep 2019.

55. Conway JR, Lex A, Gehlenborg N. UpSetR: An R Package for the Visualization of Intersecting Sets and their Properties. Bioinformatics. 2017; doi:10.1093/bioinformatics/btx364.

56. Wagih O. ggseqlogo: a versatile R package for drawing sequence logos. Bioinformatics. 2017; doi:10.1093/bioinformatics/btx469.

57. Larsson J. eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses. R package version 5.1.0. 2019. https://cran.r-project.org/package=eulerr. Accessed 04 Sep 2019.

58. Wilke CO. ggridges: Ridgeline Plots in ‘ggplot2’. R package version 0.5.1. 2018. https://CRAN.R-project.org/package=ggridges. Accessed on 04 Sep 2019.

59. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. 2002; doi:10.1101/gr.229102.

60. Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D, Kent WJ. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics. 2014; doi:10.1093/bioinformatics/btt637.

Figure Legends

Figure 1

(A) (Top) mRNA metabolism from transcription to degradation. In this illustration, poison exon skipping and inclusion lead to a Protein-coding isoform (grey) and NMD isoform (blue), respectively, with the NMD isoform containing a premature stop codon (red). EJCs (purple) deposited upstream of exon junctions are cleared by ribosomes during the pioneer round of translation. While Protein-coding isoforms are subject to multiple rounds of translation prior to decay, NMD isoforms are rapidly eliminated. (Bottom) Libraries analyzed in this paper: EJC-bound RIPiT-Seq (purple), whole cell (dark green) and cytoplasmic (light green) RNA-Seq. Colored bars indicate RNA populations captured in each library type. (B-D) Genome browser tracks of library coverage across individual genes (grey: protein-coding isoform(s); blue: NMD isoform) containing poison cassette exons (B, TRA2B and C, U2AF2) or 3′ UTR introns (D, hnRNPA1). Shown are all three EJC RIPiT replicates and replicate 1 for whole cell and cytoplasmic RNA-Seq. Conservation tracks show phyloP basewise scores derived from Multiz alignment of 30 vertebrate species. Numbers below tracks indicate mean reads per million (RPM) spanning each exon junction. Numbers to right in B and C are percent spliced in (PSI) values for poison exon inclusion events; PSI values for RNA-Seq libraries are replicate means.

Supplemental Figure 1

(A) Scatterplots comparing transcripts per million (TPM) between replicates of the same library type. R: Pearson's correlation. (B-D) Genome browser tracks of library coverage across individual genes (grey: protein-coding isoform(s); blue: NMD isoforms) that utilize a poison cassette exon (B, SRSF6), exon skipping (C, SNRPA1) or 3′ UTR introns (D, PRPF4B) to generate NMD isoforms. Shown are all three EJC RIPiT replicates and replicate 1 for whole cell and cytoplasmic RNA-Seq. Conservation tracks show phyloP basewise scores derived from Multiz alignment of 30 vertebrate species. Numbers below tracks indicate mean reads per million (RPM) spanning each exon junction. Numbers to right in B and C are percent spliced in (PSI) values for poison exon inclusion events; PSI values for RNA-Seq libraries are replicate means.

Supplemental Table 1

Sequencing and alignment information for each replicate of the analyzed libraries.

Figure 2

(A) Distribution of the number of exon junctions in all annotated protein-coding (grey) or NMD (blue) transcripts. (B) Distribution of protein-coding (top) and NMD (bottom) transcripts per million (TPM) in each library type (colors as in Figure 1A), binned based on indicated number of exon junctions per transcript. Results of one-way ANOVA and Tukey's *post hoc* significance tests comparing EJC RIPiT-Seq to RNA-Seq libraries are indicated: *P<0.05, **P<0.01, ***P<0.005, ****P<0.0001. (C-D) Scatterplots comparing TPMs between EJC RIPiT-Seq and whole cell RNA-Seq libraries for different isoform types: Protein-coding (C, left), NMD (C, right), non-stop decay (D, left), retained intron (D, middle), processed transcript (D, right), and lincRNA (E). In (C), transcripts from Figure 1 are noted. In (E), O and X indicate spliced and unspliced lincRNA transcripts, respectively; XIST isoforms are indicated as open black circles. N: Number of detected transcripts out (of all annotated transcripts of that type). Dashed black line is the x=y line.

Supplemental Figure 2

(A-C) Scatterplots comparing TPMs between EJC RIPiT-Seq and cytoplasmic RNA-Seq libraries in different isoform types: Protein-coding (A, left), NMD (A, right), non-stop decay (B, left), retained intron (B, middle), processed transcript (B, right), and lincRNA (C). In (A), transcripts from Figure 1 are noted. In (C), O and X indicate spliced and unspliced lincRNA transcripts, respectively; XIST isoforms are indicated as open black circles. N: Number of detected transcripts out (of all annotated transcripts of that type). Dashed black line is the x=y line.

Figure 3

(A) Comparison of annotated exon junctions among the transcriptomes sourced from RefSeq (hg38), Ensembl (GRCh38.p12), GENCODE (v29), and CHESS (v2.1). Horizontal bars: total junctions in each reference set; vertical bars: intersections of indicated reference sets. Bar graphs created with UpSetR [55]. (B) Sequence motifs for 5′ (left) and 3′ (right) splice sites used in annotated junctions observed in at least one analyzed library type (top) and for previously unannotated splice sites in indicated library type (bottom). Sequence logos were generated in R using ggseqlogo [56]; letter height signifies the relative abundance of that nucleotide at each position. N: Number of splice sites contributing to each logo. Note that the number of unannotated junctions (5,917) is greater than the total number of unannotated splice sites because many unannotated junctions combine an annotated and unannotated splice site (i.e., alternative 5′ or 3′ splice sites). (C) Venn diagram of annotated and previously unannotated junctions (numbers indicated) shared between library types. Venn diagrams made with eulerr [57]. (D) Cumulative histogram of exon junction reads (RPM) at annotated (X) and previously unannotated (O) junctions in each library type (colors as in Figure 1A). (E) Schematic of unannotated splicing events separated by event type: Skipped exon (red); alternative 3′ (orange) or 5′ (yellow) splice site; new intron (light blue); new cassette exon (dark blue). N: number of observed events; for new cassette exons, both the number of observed unannotated junctions and number of new exons are shown.

Supplemental Figure 3

(A) Schematic of library processing steps used to identify and analyze reads at annotated and unannotated junctions. Full details of each step are in Results and Methods.

Supplemental Table 2

List of all introns previously annotated by RefSeq (hg38) [19], Ensembl (GRCh38.p12) [3], GENCODE (v29) [20], and/or CHESS (v2.1) [21]. Table includes information on intron location, length, strand, transcript ID (if available, [3]), and annotation origin.

Figure 4

(A) Distribution of of unannotated splice sites relative to the closest annotated splice site observed in analyzed libraries (solid colored lines). Grey dotted line: Frequency of available GT or AG dinucleotides surrounding the annotated 5′ (left) and 3′ (right) splice sites with open circles indicating in-frame positions and solid grey dots indicating out-of-frame positions. (B) Distribution of the ratio of unannotated alternative 3′ splice site use (RPM_{Unanno}) over all events using the same 5′ splice site (RPM_{Unanno}+ RPM_{Anno}) in each library type. (Left) Unannotated alternative 3′ splice sites at the +3 position relative to closest annotated 3′ splice site; (middle) same but at the +4 position. Grey lines show how the top 20% (highest RPM_{Unanno}) of unannotated junctions detected in EJC RIPiT-Seq libraries differ between library types. Results of one-way ANOVA and Tukey's *post hoc* tests comparing EJC RIPiT-Seq to RNA-Seq libraries are indicated; ****P<0.0001. (Right) Median [$\mathrm{RPM}_{\mathrm{Unanno}}/(\mathrm{RPM}_{\mathrm{Unanno}}+\mathrm{RPM}_{\mathrm{Anno}})$] values per library at the +3 and +4 positions.

Figure 5

(A) Regions used to calculate MaxEnt and mean conservation scores surrounding unannotated alternative 3′ and 5′ splice sites and new introns (top) or new cassette exons (bottom). (B and C) Scatterplots comparing MaxEnt scores to mean conservation scores (phyloP, 30-way) at 5′ (left) or 3′ (right) splice sites for (B) annotated and random or (C) observed unannotated events. Smaller points are used to represent splice sites with either score lower than 0 as these may result from splicing noise. Annotated splice sites were downsampled by randomly selection (5′, N = 2,048; 3′, N = 2,456; same as unannotated splice site numbers in C) from the 151,072 observed in our libraries. **Supplemental Figure 4A** shows the same plot for all observed annotated splice sites. (B) also contains 2,048 random GT-containing (left) and 2,456 random AG-containing (right) sites; identical plots for four additional sets of randomized locations are shown in **Supplemental Figure 4B**. The top 5% mean conservation scores of random sites is indicated and marked by a dashed line. Genes for which genome-browser tracks are shown in panels D and E and Supplemental Figure 4C and D are indicated. (D-E) Genome browser tracks of library coverage across CNOT1 (D) and EEA1 (E). Annotated transcripts are shown in grey and unannotated alternative 3′ splice site use in orange. Conservation tracks and annotations are as in **Figure 1B-D**.

Supplemental Figure 4

(A) Scatterplots comparing the MaxEnt score to mean conservation score (phyloP, 30-way) at 5′ (left) or 3′ (right) splice sites for all annotated junctions (N = 151,072). (B) Scatterplots comparing the MaxEnt score to conservation (phyloP, 30-way) at 5′ (left) or 3′ (right) splice sites for multiple sets (N = 5) of randomly selected sequences (5′, N = 2,048; 3′, N = 2,456). (C) Genome browser tracks of library coverage across HSPA8 (C) and HECTD4 (D). Annotated transcripts are shown in grey, unannotated alternative 3′ splicing events in orange, and unannotated introns in light blue. Conservation tracks represent phyloP basewise scores derived from Multiz alignment of 30 vertebrate species, as well as 100 vertebrate species in (C). Numbers below tracks indicate mean reads per million (RPM) spanning each exon junction. Numbers to right in C and D are percent intron retention (PIR) and percent spliced in (PSI) values, respectively; PSI and PIR values for RNA-Seq libraries are replicate means. The translated protein sequences of both the annotated and unannotated transcripts are provided in (D).

Supplemental Table 3

List of highly conserved unannotated splicing events (see **Results** for conservation score cut-offs). Table includes information on exon junction locations (coordinates and transcript features), MaxENT and conservation scores, and calculated PIR values.

Figure 6

(A) Density plot comparing junction-spanning read coverage (RPM) for new cassette exons in EJC and RNA-Seq libraries. Line indicates median expression per library and dots represent individual cassette exons. N: number of observed cassette exons per library. (B) (Left) Scatterplot comparing mean conservation (phyloP, 30-way) at 5′ and 3′ splice sites of new cassette exons. Exons with scores above 0 at both splice sites are colored (white to dark blue) to indicate mean exon conservation and sized by the number of junction-spanning reads supporting that exon in EJC RIPiT-Seq libraries. Diamonds indicate exons that create a frameshift in the resulting mRNA; circles indicate non-frameshifting exons. (Right) Zoomed view of exons with mean 5′ and 3′ splice site conservation scores above 1.03 and 0.63, respectively. (B and C) Genome browser tracks of library coverage across new poison cassette exons in PHIP (B) and PTPRA (C). New cassette exons are shown in blue and numbered according to their placement in the major isoform observed in all libraries. Conservation tracks and annotations are as in **Figure 1B-D**.

Supplemental Figure 5

(A) Genome browser tracks of library coverage across the new cassette exon in POLQ. The new cassette exon is shown in blue and numbered according to its placement in the major isoform observed in all libraries. Conservation tracks and annotations are as in **Figure 1B-D**.

Figure 1

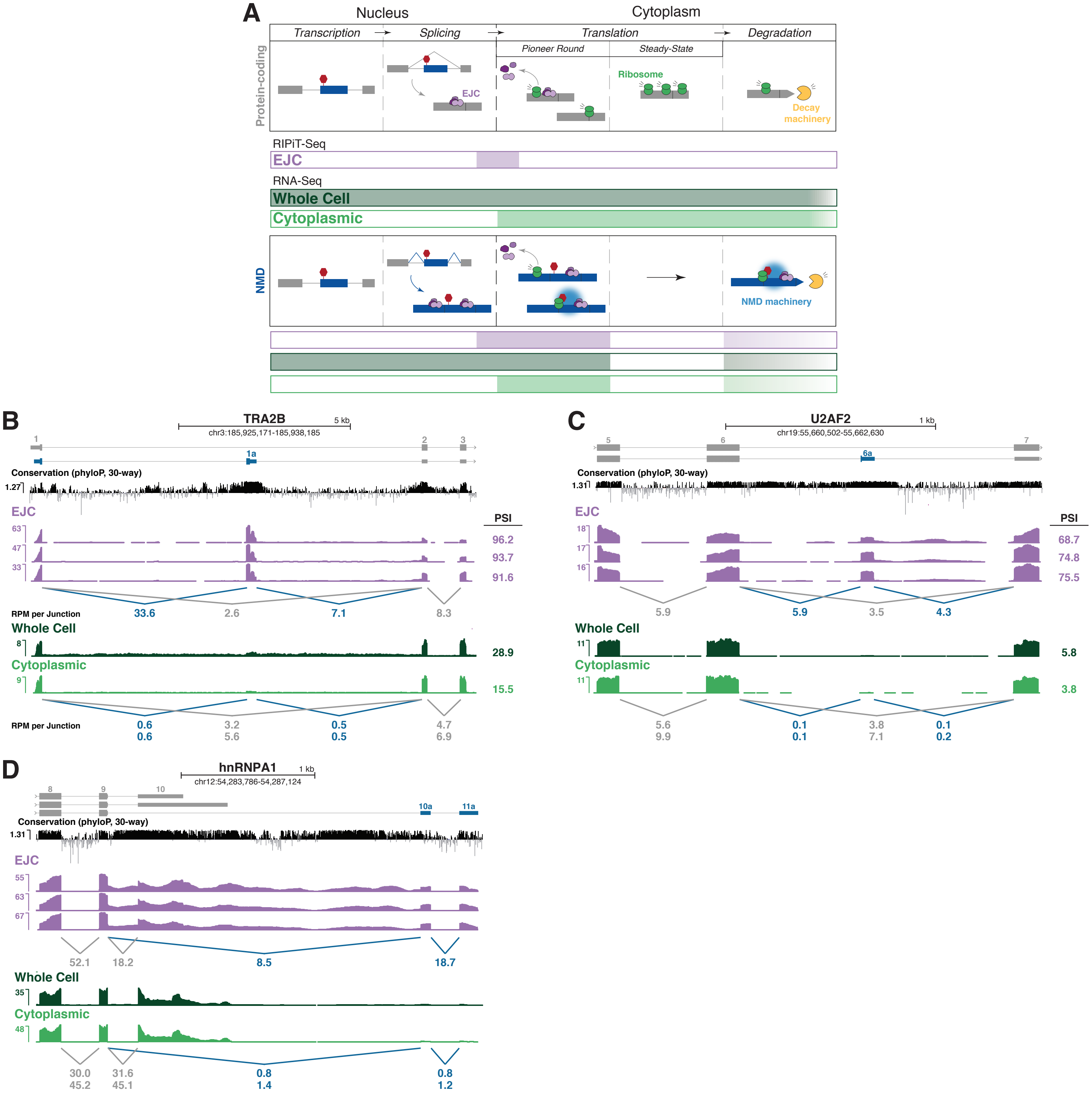
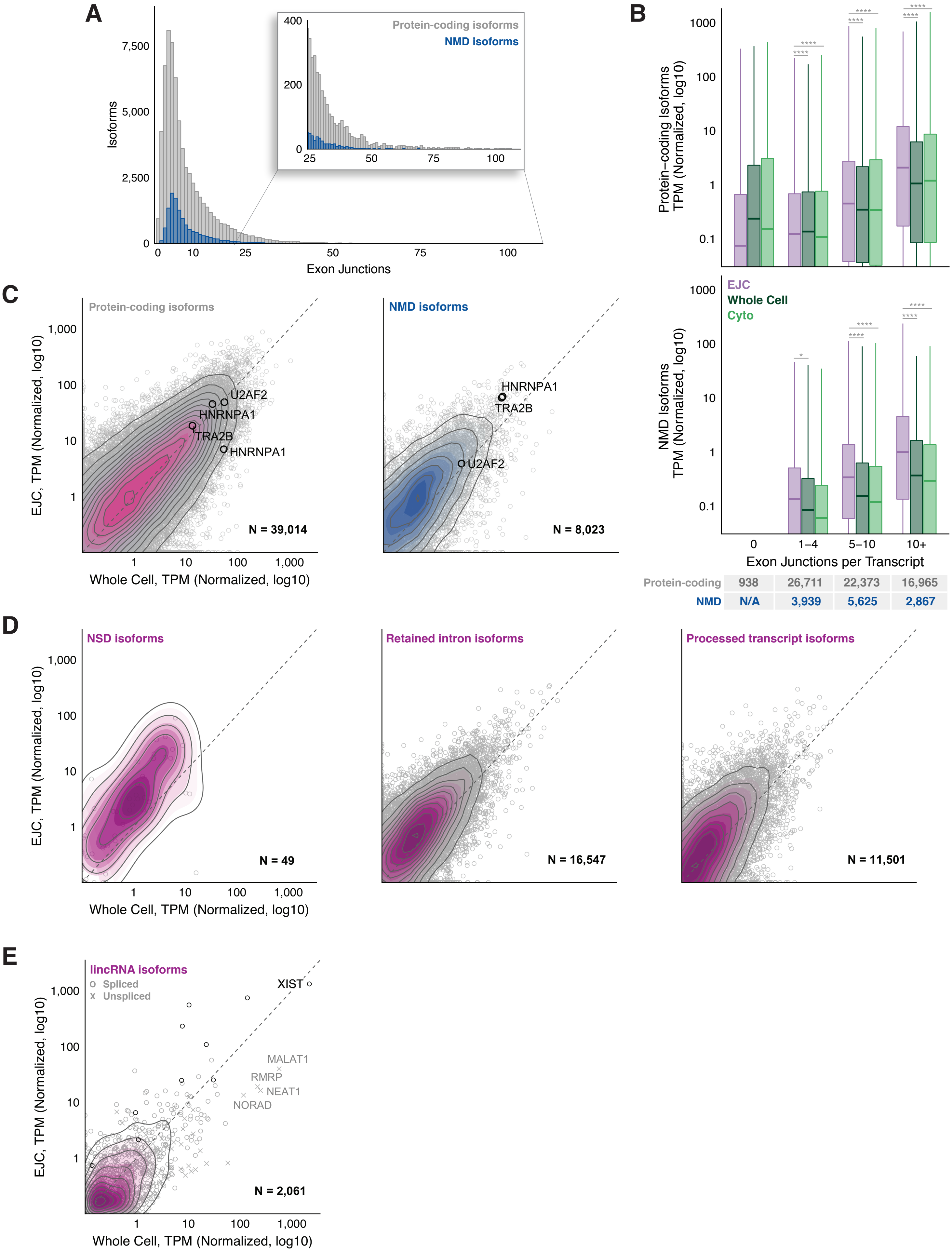

A
Nucleus
Cytoplasm
Transcription
Splicing
Translation
Degradation
Pioneer Round
Steady-State
Protein-coding
EJC
Ribosome
Decay machinery
RIPiT-Seq
EJC
RNA-Seq
Whole Cell
Cytoplasmic
NMD
NMD machinery
B
TRA2B
5 kb
chr3:185,925,171-185,938,185
1
1a
2
3
Conservation (phyloP, 30-way)
1.27
EJC
63
47
33
PSI
96.2
93.7
91.6
RPM per Junction
33.6
2.6
7.1
8.3
Whole Cell
8
28.9
Cytoplasmic
9
15.5
RPM per Junction
0.6
0.6
3.2
5.6
0.5
0.5
4.7
6.9
C
U2AF2
1 kb
chr19:55,660,502-55,662,630
5
6
6a
7
Conservation (phyloP, 30-way)
1.31
EJC
18
17
16
PSI
68.7
74.8
75.5
5.9
5.9
3.5
4.3
Whole Cell
11
5.8
Cytoplasmic
11
3.8
5.6
9.9
0.1
0.1
3.8
7.1
0.1
0.2
D
hnRNPA1
1 kb
chr12:54,283,786-54,287,124
8
9
10
10a
11a
Conservation (phyloP, 30-way)
1.31
EJC
55
63
67
52.1
18.2
8.5
18.7
Whole Cell
35
Cytoplasmic
48
30.0
45.2
31.6
45.1
0.8
1.4
0.8
1.2

Figure 2



	0	1–4	5–10	10+
Protein-coding	938	26,711	22,373	16,965
NMD	N/A	3,939	5,625	2,867

Figure 3

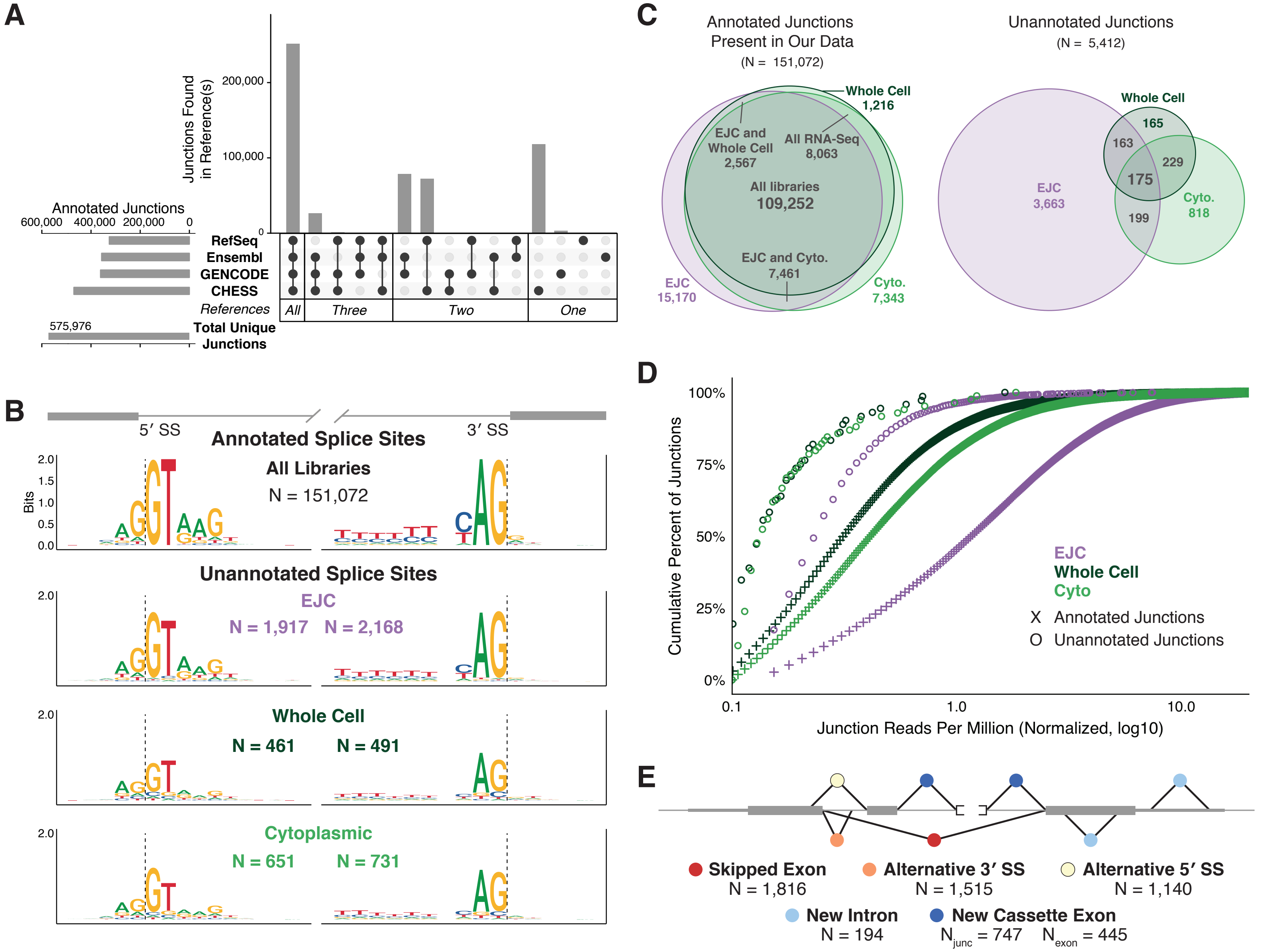
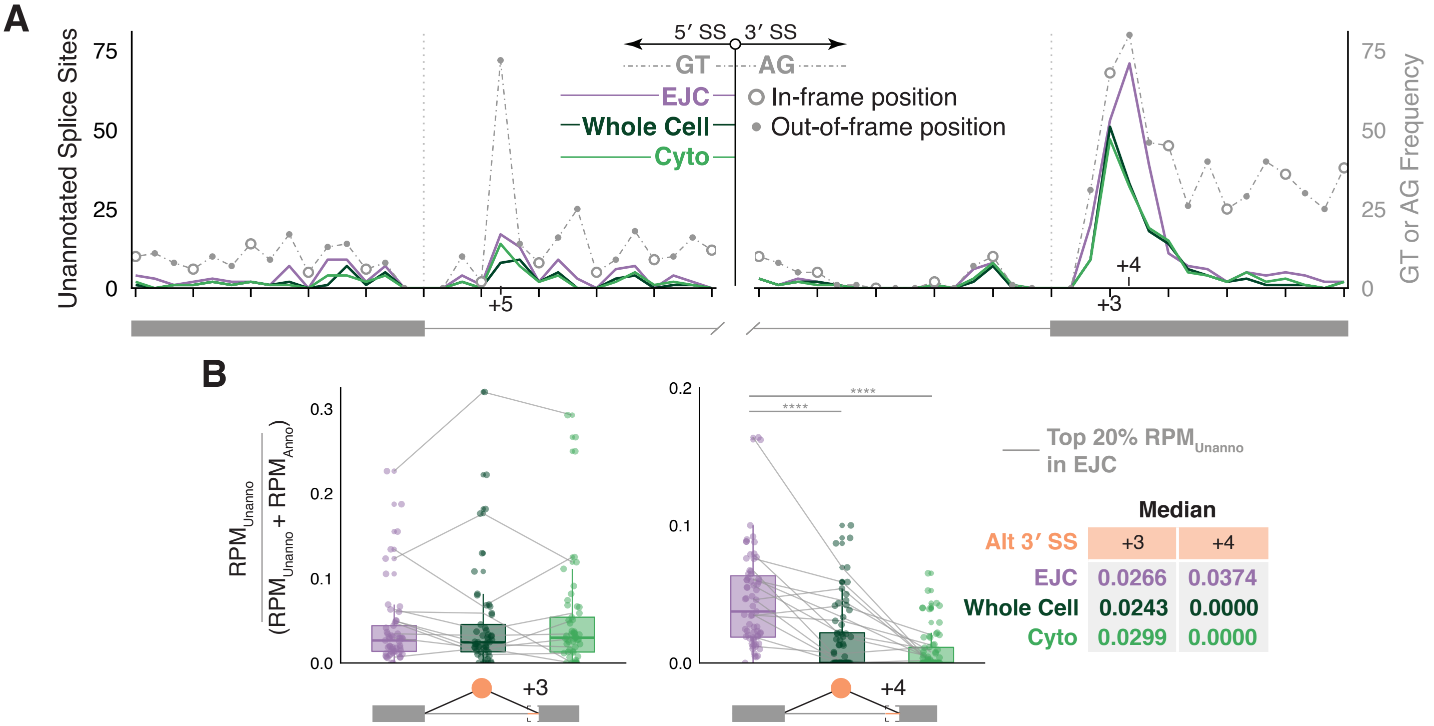

A
Junctions Found in Reference(s)
200,000
100,000
0
Annotated Junctions
600,000 400,000 200,000 0
RefSeq
Ensembl
GENCODE
CHESS
References
575,976
Total Unique Junctions
All
Three
Two
One
C
Annotated Junctions Present in Our Data
(N = 151,072)
Whole Cell 1,216
EJC and Whole Cell 2,567
All RNA-Seq 8,063
All libraries 109,252
EJC and Cyto. 7,461
EJC 15,170
Cyto. 7,343
Unannotated Junctions
(N = 5,412)
Whole Cell
165
163
175
229
EJC 3,663
Cyto. 818
199
B
5′ SS
3′ SS
Annotated Splice Sites
All Libraries
N = 151,072
Bits
2.0
1.5
1.0
0.5
0.0
Unannotated Splice Sites
EJC
N = 1,917
N = 2,168
2.0
Whole Cell
N = 461
N = 491
2.0
Cytoplasmic
N = 651
N = 731
2.0
D
Cumulative Percent of Junctions
100%
75%
50%
25%
0%
0.1
1.0
10.0
Junction Reads Per Million (Normalized, log10)
EJC
Whole Cell
Cyto
X Annotated Junctions
O Unannotated Junctions
E
Skipped Exon
N = 1,816
Alternative 3′ SS
N = 1,515
Alternative 5′ SS
N = 1,140
New Intron
N = 194
New Cassette Exon
N_{junc} = 747 N_{exon} = 445

Figure 4



	Median	
Alt 3′ SS	+3	+4
EJC	0.0266	0.0374
Whole Cell	0.0243	0.0000
Cyto	0.0299	0.0000

Figure 5

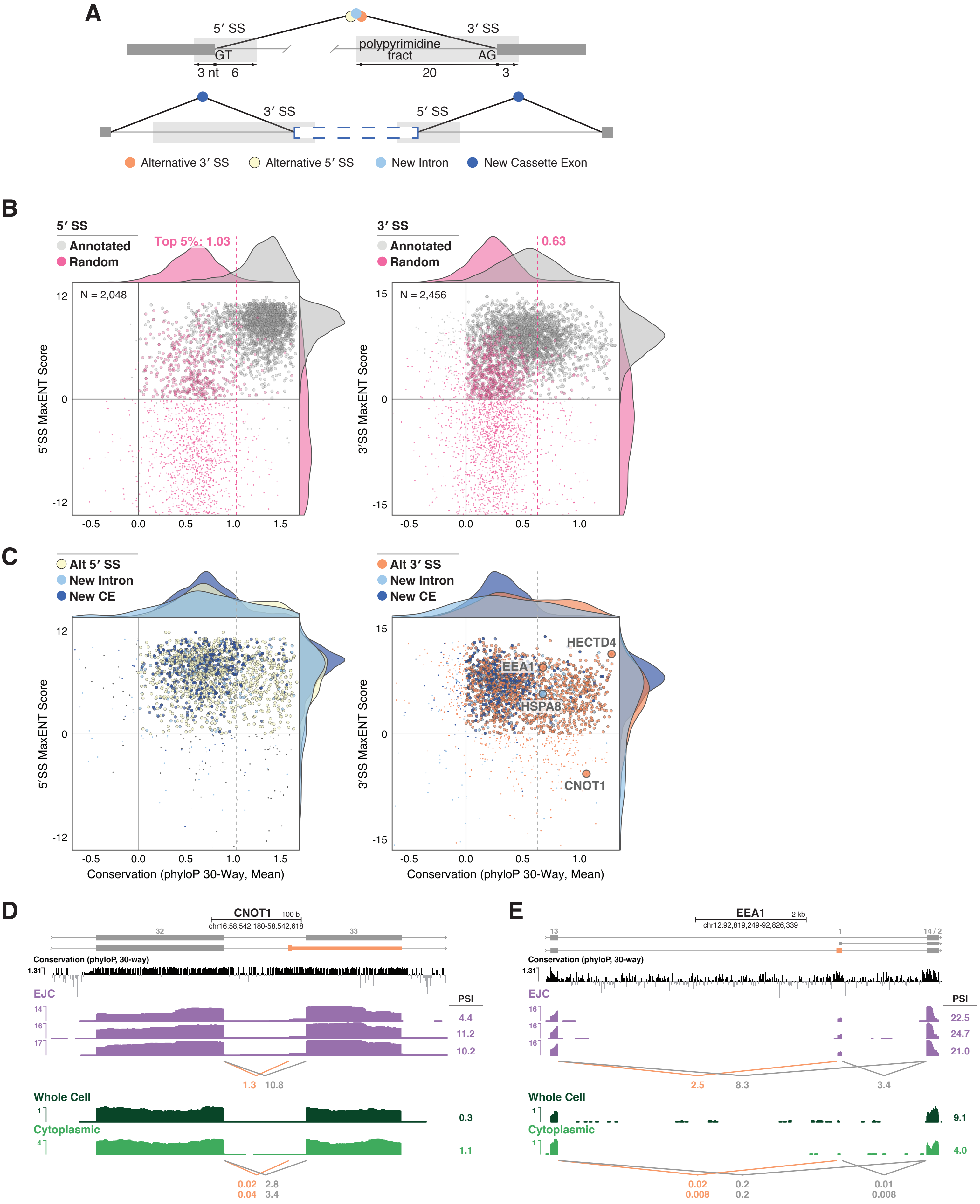
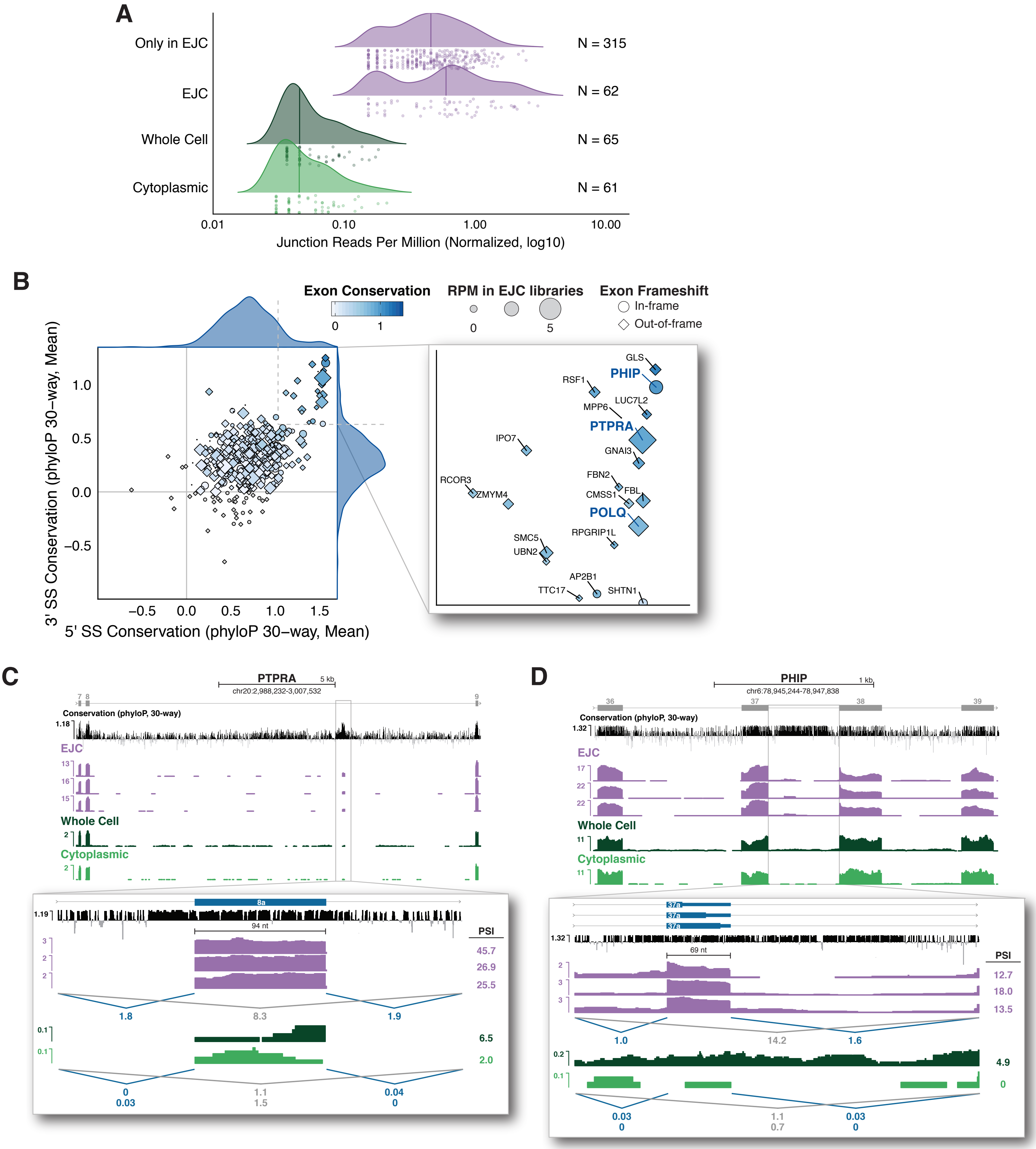
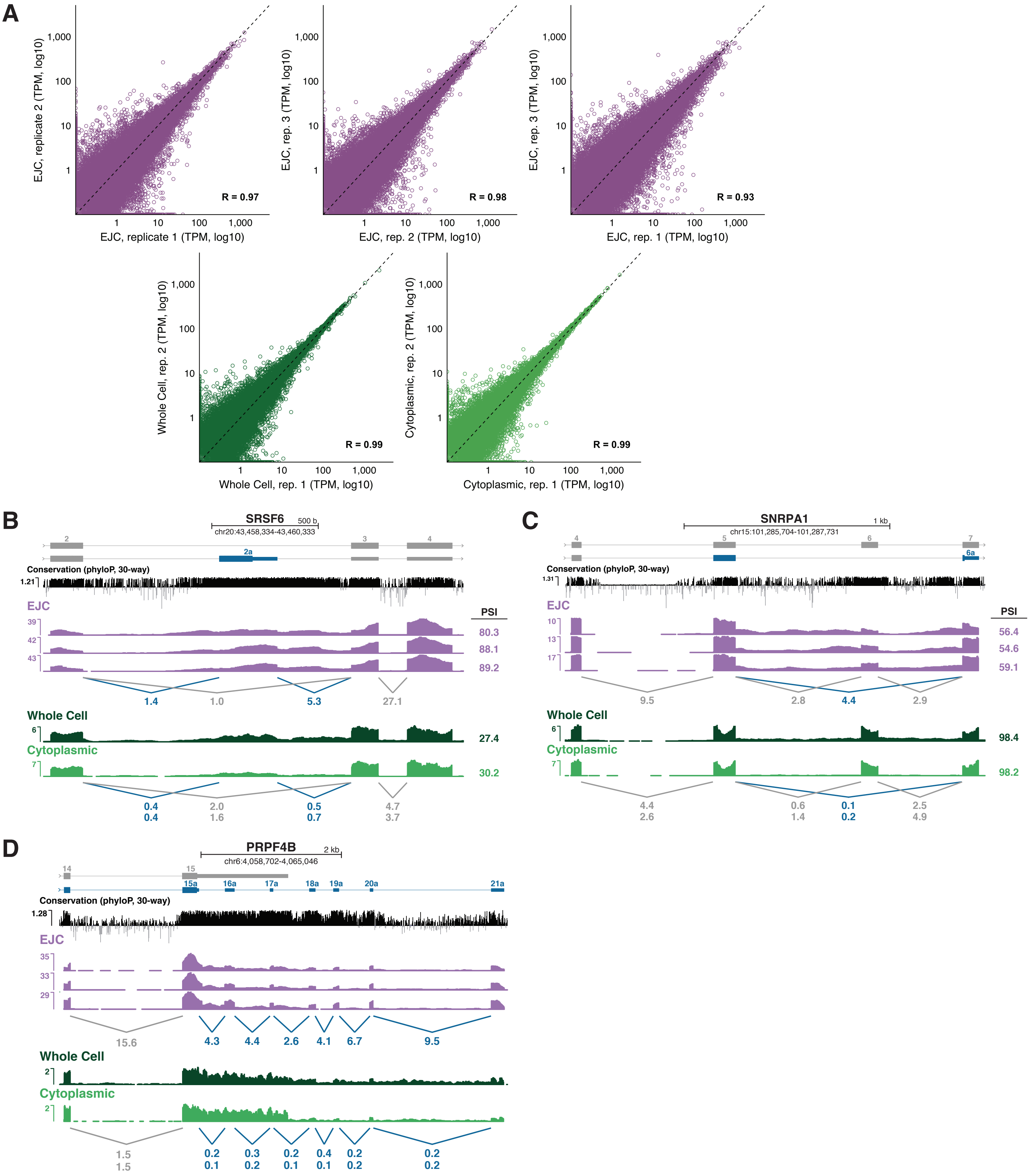

A
5′ SS
3′ SS
GT
polypyrimidine tract
AG
3 nt
6
20
3
3′ SS
5′ SS
Alternative 3′ SS
Alternative 5′ SS
New Intron
New Cassette Exon
B
5′ SS
Annotated
Random
Top 5%: 1.03
N = 2,048
12
0
-12
5′SS MaxENT Score
-0.5
0.0
0.5
1.0
1.5
3′ SS
Annotated
Random
0.63
N = 2,456
15
0
-15
3′SS MaxENT Score
-0.5
0.0
0.5
1.0
C
Alt 5′ SS
New Intron
New CE
12
0
-12
5′SS MaxENT Score
-0.5
0.0
0.5
1.0
1.5
Conservation (phyloP 30-Way, Mean)
Alt 3′ SS
New Intron
New CE
15
0
-15
3′SS MaxENT Score
HECTD4
EEA1
HSPA8
CNOT1
-0.5
0.0
0.5
1.0
Conservation (phyloP 30-Way, Mean)
D
CNOT1
100 b
chr16:58,542,180-58,542,618
32
33
Conservation (phyloP, 30-way)
1.31
EJC
PSI
14
16
17
4.4
11.2
10.2
1.3
10.8
Whole Cell
1
0.3
Cytoplasmic
4
1.1
0.02
0.04
2.8
3.4
E
EEA1
2 kb
chr12:92,819,249-92,826,339
13
1
14 / 2
Conservation (phyloP, 30-way)
1.31
EJC
PSI
16
16
16
22.5
24.7
21.0
2.5
8.3
3.4
Whole Cell
1
9.1
Cytoplasmic
1
4.0
0.02
0.008
0.2
0.2
0.01
0.008

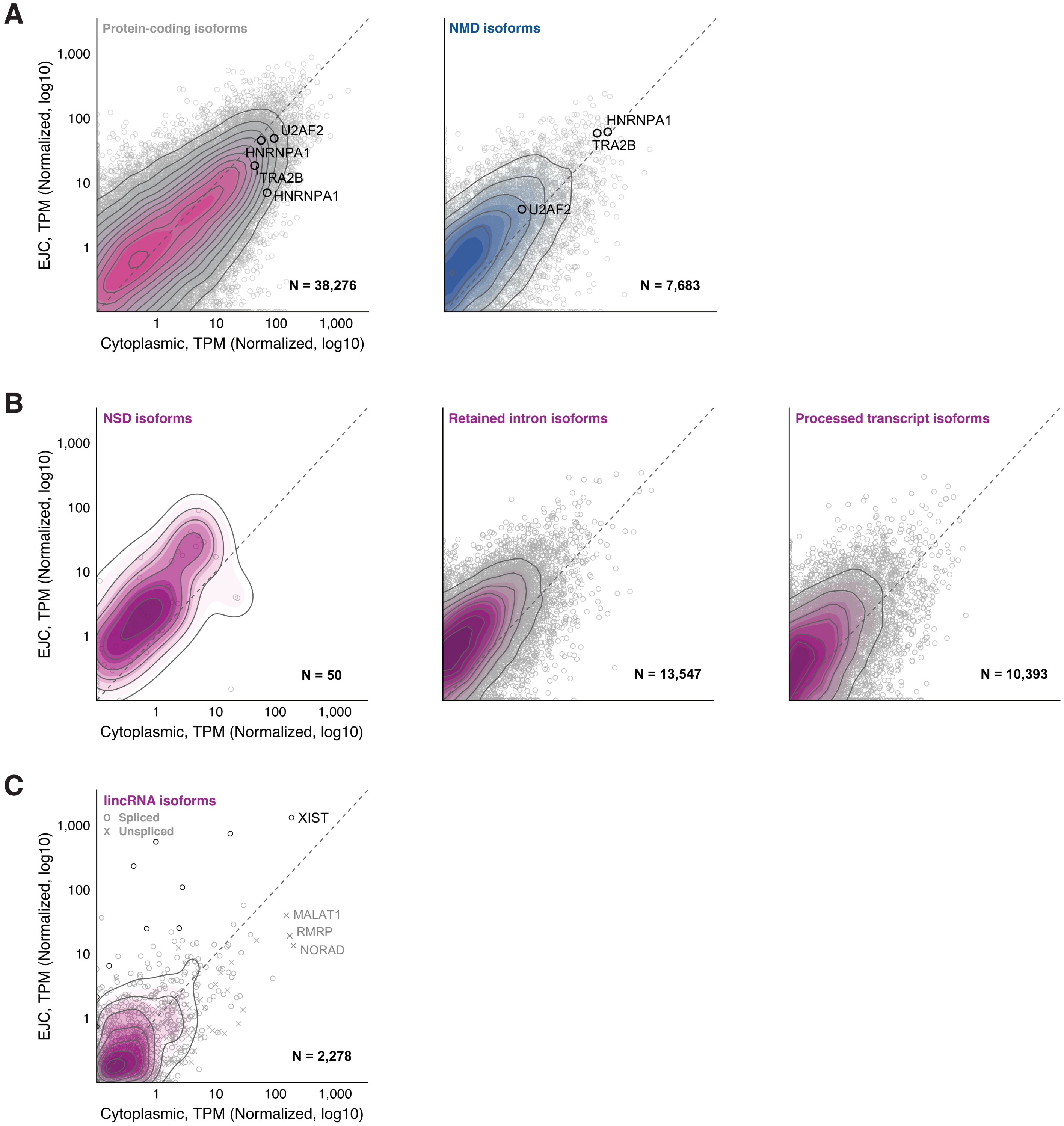
Figure 6


A
Only in EJC
N = 315
EJC
N = 62
Whole Cell
N = 65
Cytoplasmic
N = 61
0.01
0.10
1.00
10.00
Junction Reads Per Million (Normalized, log10)
B
Exon Conservation
0
1
RPM in EJC libraries
0
5
Exon Frameshift
In-frame
Out-of-frame
3' SS Conservation (phyloP 30−way, Mean)
1.0
0.5
0.0
−0.5
5' SS Conservation (phyloP 30−way, Mean)
−0.5
0.0
0.5
1.0
1.5
GLS
PHIP
RSF1
LUC7L2
MPP6
PTPRA
IPO7
GNAI3
FBN2
RCOR3
FBL
CMSS1
ZMYM4
POLQ
RPGRIP1L
SMC5
UBN2
AP2B1
TTC17
SHTN1
C
PTPRA
5 kb
chr20:2,988,232-3,007,532
7 8
9
Conservation (phyloP, 30-way)
1.18
EJC
13
16
15
Whole Cell
2
Cytoplasmic
2
8a
1.19
94 nt
PSI
3
2
2
45.7
26.9
25.5
1.8
8.3
1.9
0.1
0.1
6.5
2.0
0
0.03
1.1
1.5
0.04
0
D
PHIP
1 kb
chr6:78,945,244-78,947,838
36
37
38
39
Conservation (phyloP, 30-way)
1.32
EJC
17
22
22
Whole Cell
11
Cytoplasmic
11
37a
37a
37a
1.32
69 nt
PSI
2
3
3
12.7
18.0
13.5
1.0
14.2
1.6
0.2
0.1
4.9
0
0.03
0
1.1
0.7
0.03
0

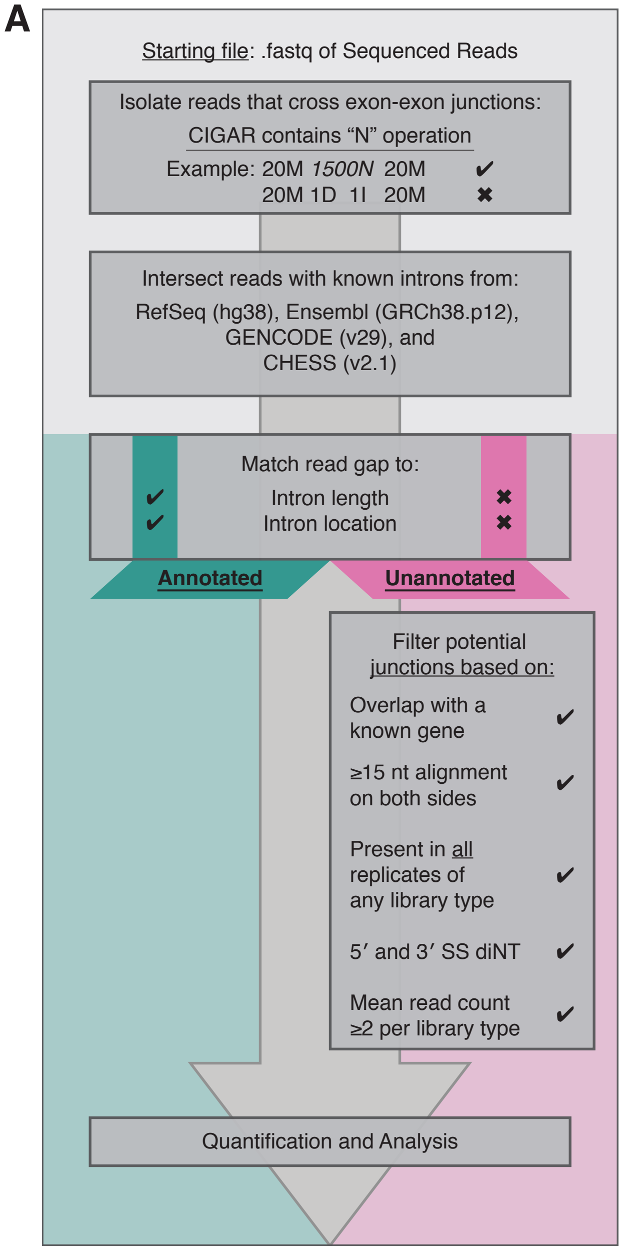
Supplemental Figure 1



Supplemental Figure 2



Supplemental Figure 3



Starting file: .fastq of Sequenced Reads

Isolate reads that cross exon-exon junctions:
CIGAR contains “N” operation
Example: 20M *1500N* 20M ✔
20M 1D 1I 20M ✖

Intersect reads with known introns from:
RefSeq (hg38), Ensembl (GRCh38.p12), GENCODE (v29), and CHESS (v2.1)

Match read gap to:
✔ Intron length ✖
✔ Intron location ✖

Annotated

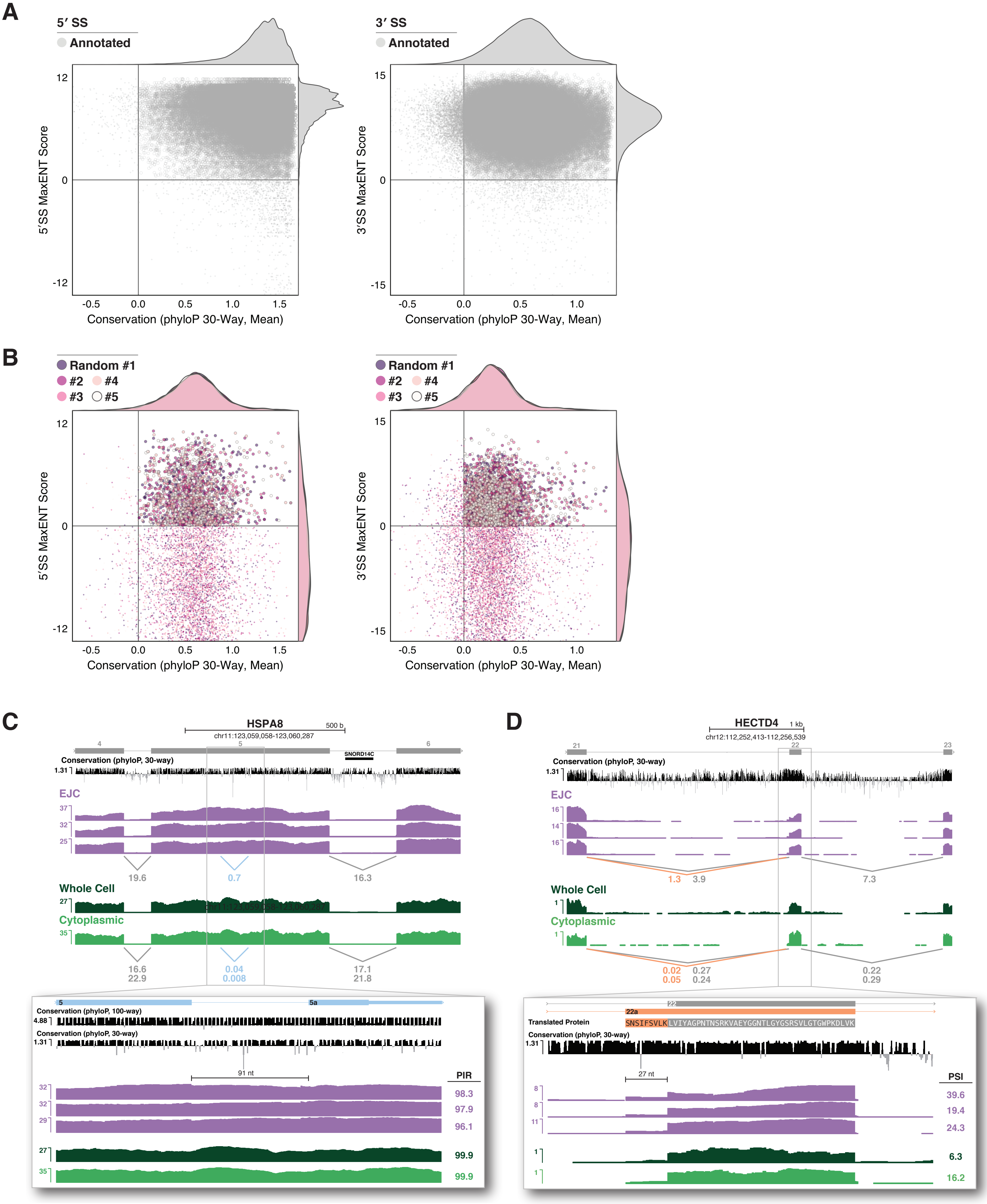
Unannotated

Filter potential junctions based on:

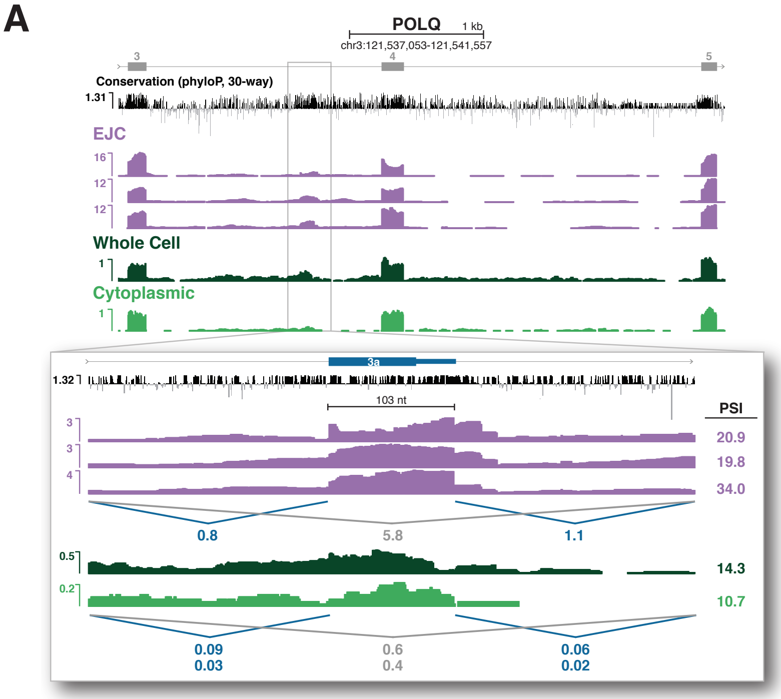
- Overlap with a known gene ✔
- ≥15 nt alignment on both sides ✔
- Present in all replicates of any library type ✔
- 5′ and 3′ SS diNT ✔
- Mean read count ≥2 per library type ✔

Quantification and Analysis

Supplemental Figure 4



Supplemental Figure 5



Supplemental Table 1

Library			Sequencing			Alignment				
Name	*Fraction*	*Replicate*	*Type*	*Insert Size*	*Sequenced Pairs*	*Repeats*	*Aligned Pairs*	*MAPQ ≥ 5*	*Unique Pairs*	*Spliced Reads*
EJC	Total	1	Paired End, 150bp	220 - 500	19 Million	– 3 M	6 M	5.6 M	5.1 M	3.3 M (32%)
		2		220 - 500	25 M	– 4 M	10 M	9.0 M	8.2 M	5.6 M (34%)
		3		220 - 500	23 M	– 4 M	7 M	6.6 M	6.1 M	4.3 M (35%)
RNASeq	**Whole Cell**	1	Paired End, 51bp	100 - 200	57 M	– 15 M	30 M	28.0 M	24.9 M	5.8 M (12%)
		2		100 - 200	55 M	– 15 M	38 M	36.0 M	29.7 M	7.4 M (12%)
RNASeq	**Cytoplasm**	1	Paired End, 51bp	100 - 200	56 M	– 8 M	46 M	33.4 M	33.4 M	11.7 M (18%)
		2		100 - 200	51 M	– 6 M	43 M	32.6 M	32.6 M	11.5 M (18%)

Supplemental Table 2

See file: Supp_Table_2.txt

Supplemental Table 3

See file: Supp_Table_3.txt