Supplemental Information

Supplemental Figure S1: Molecules per cell estimation of a high-coverage intron by qRT-PCR; bioinformatic pipeline for detecting statistically significant intron detention levels.

(A) Left, qPCR of plasmid DNA dilutions containing the cloned region spanning exons 32-34 of the mouse Fn1 gene. Inset shows slope and amplification efficiency of each primer set. Right, estimation of molecules per cell. Total RNA was extracted from counted mESCs and the per cell RNA mass calculated to ~13 pg per cell. Using the known input amount cross-referenced to the amplification of known quantities from the data in the left panel, the qRT-PCR signal (graphed as mean N=3, +/- SEM) was converted to estimated molecules per cell. (B) Schematic of computational pipeline for determining statistically significant intron detention. The gray box indicates the process by which raw RNA-seq data are mapped and processed. The brown box shows the process for extracting a non-overlapping set of introns for quantification. Below, in the yellow box, are the steps using reads mapped to introns for identifying statistically-significant detained introns.

Supplemental Data S1&2: Genomic coordinates of human and mouse detained introns

For each detained intron, chromosome, start, end, name, and strand are indicated in the .bed format. Human intron names come from the Aceview annotation corresponding to the intron, while mouse intron names were determined by overlap of sequenced junctions with known Ensembl transcriptional start and end coordinates. Field 5 in the human .bed file is a place holder, while in the mouse .bed, the alternative splice type and whether the intron shows increased or decreased splicing or no change (More, Less, or NC respectively) in response to 2 hours of CB19 treatment is indicated, separated from splice class by and underscore "_", (e.g. Constitutive_More indicates a constitutive intron that is spliced more in the CB19 treatment compared to DMSO). U11/U12 (minor spliceosome) dependent introns are indicated by the notation (U11/U12) after splicing class. Note these files can

be viewed in the Broad IGV or UCSC Genome browsers. All human coordinates are from the hg19 human genome assembly and all mouse coordinates from the mm9 assembly.

Supplemental Table S1: Splicing classification of detained introns in 4 human cell lines

Relative fractions of all introns (pie charts at left) and detained introns (pie charts at right) in each category of constitutive/alternative splicing class. Splicing classes were determined independently for each cell line based on RNA-seq mapped splice junctions and the proportions of each class determined. Numbers of introns in each category are indicated in the tables to the right. Categories in which the number of detained introns in that class are higher (Enriched, green) or lower (Depleted, red) than expected by chance, and the p-value for over- or under-representation was determined by a two-sided hypergeometric distribution and Fisher's Exact Test.

Supplemental Figure S2: Structural properties of detained introns (A) Gene ontology analysis of mESC genes containing DIs (left panel). Fold enrichment (observed/expected) for each GO term gene set is plotted on the linear scale (blue), along with the $-\log_{10}$(FDR), (red). Right panel, GO analysis of the genes with introns detained in all four human cell lines (N= 1603) compared against the set of all genes detectably expressed in all four cell lines (N=8066). (B) Distribution of gene-level expression for genes (FPKM), determined by Cufflinks (Trapnell et al., 2010), with no DIs (gray) compared to genes containing DIs (blue). (C) Cumulative distribution function (CDF) plot of the distances between the transcription start site (TSS) and the start (5' splice site) of each constitutive intron. DIs are plotted in red, non-DIs in blue. The average position of all introns is slightly skewed toward the 3' end of the gene, likely due to longer mean intron length at the 5' compared to the 3' end of mammalian genes. A Kolmogorov-Smirnov (KS) test confirmed a positional bias for DIs toward the 3' end of genes (mean relative position 0.69 compared to 0.58 for non-DIs).

Supplemental Table S2: Human genes containing NMD-switch exons flanked by DIs

NMD-switch exons flanked by one or more DIs in human genes. Official gene name and gene aliases are shown, coordinates of DIs, type of splicing event leading to NMD, and coordinates of NMD-switch exon are shown. Genes with an asterisk in the GTF column are included in Supplemental Data 3 and were used for the nuclear-cytoplasmic analysis in Figure 4A. References for experimentally determined NMD targets are indicated in the references column by first author, full references are provided below. Genes are grouped by general functional groups indicated in bold section headings.

Supplemental Data S3: Annotation (.gtf) of 33 human genes with CDS, NMD, and DI isoforms

From the list of 64 genes with NMD-switch exons flanked by DIs, 33 were hand annotated using splice junction data from the 4 human encode cell lines. Catagory of each transcript type (CDS, coding isoform, NMD, NMD-substrate isoform, DI- DI containing isoform) is indicated in the transcript type and tag fields. Positions of Start and stop codons, reading frame of each exon and coordinates of coding sequences are also contained in the .gtf. Note this file can be viewed in the Broad IGV or UCSC Genome browsers. All coordinates are from the hg19 human genome assembly.

Supplemental Figure S3: Detained introns are localized in the nucleus after 30 min. of transcriptional inhibition (A) Nuclear poly(A) transcript levels for the transcripts in the set of annotated human genes containing NMD-switch exons flanked by DIs in three human cell lines (HeLa, HepG2, HUVEC, from ENCODE). CDS, coding transcripts (dark blue), NMD, NMD isoform (light blue), DI, DI-containing transcript (red). The mean FPKM is shown below each category; p-values determined by 2-sided T-test. Whiskers = 1.5x interquantile range. (B) mESCs were treated with either DMSO (black bars) or Flavopiridol (white bars). 30 minutes after treatment began, cells were fractionated into nuclear and cytoplasmic fractions and total RNA was

prepared. Relative RNA levels for each intron or exon-exon junction were measured by qRT-PCR and adjusted to cell-equivalents; intron quantities in both conditions and both exon-exon and intron quantities in flavopiridol treated cells were normalized to the DMSO control exon-exon level. Data plotted is combined mean of two independent fractionations, +/- SEM.

Supplemental Figure S4: CB19 treatment rapidly and specifically induces splicing of a subset of DIs (A) mESCs were treated with CB19 for various times and qRT-PCR used to measure exon-intron junctions of the indicated introns in the Clk1 transcript. Intron levels were normalized to the stable transcript Mylpf and to DMSO-treated controls and quantified relative to their levels at time zero. Introns 3 and 4 (red and yellow) are DIs, 2 and 5 (blue and green) are normally spliced. (B) poly(A) selected RNA-seq read density after 2 hours treatment with DMSO (blue) or CB19 (orange) is plotted for the region of the gene Phc1. Exon/intron structure including the location of the DI is indicated schematically below. (C) mESCs were treated with flavopiridol for 30 minutes followed by DMSO or CB19 for 60 minutes. 2 DIs (red/yellow) and flanking normal introns (blue/light blue) for each transcript were assayed by qRT-PCR and normalized first to *Mylpf* and then to the upstream DI DMSO control signal (mean N=3 +/- SEM, p-values from two-sided T-test indicated). The intron number is indicated below the label. (D) Genes containing CB19-responsive DIs that encode splicing or RNA-processing factors are listed, divided into increased intron detention (blue) or increased splicing (red). Those containing a RS-domain are indicated in bold in the left panels (Calarco et al., 2009; Jones and Caceres,). Genes containing multiple DIs that change in the same direction are indicated by the number of affected introns in parentheses.

Supplemental Figure S5: Transcriptional changes estimated by non-DI abundance do not explain differential DI abundance in response to CB19 treatment (A) For each gene, total read counts falling within all non-detained introns were summed and any genes showing a statistically significant (padj < 0.05) change in this intron read count abundance between DMSO and CB19 at 2 hours post treatment were identified. Genes

were then divided into the following categories: genes that contain a DI(s) that showed an increase in abundance in response to CB19 (DI less spliced), genes that contain a DI(s) that showed a decrease in abundance in response to CB19 (top right, DI more spliced), genes that contain a DI(s) that was unchanged in abundance in response to CB19 (DI no change). and genes that contain no DI. Scatterplots show the $\log_2$ of total genic non-DI introns reads in DMSO (x-axis) and CB19 (y-axis) at 2 hours post treatment. Each point is a gene, and red colored points show those genes in which the total non-DI intron count was statistically significantly different between DMSO and CB19 as determined by DESEq. Genes that are transcriptionally upregulated in response to CB19 are thus found above the diagonal and genes that are transcriptionally downregulated are found below the diagonal. (B) Table summarizing the number of genes falling into each of the above described categories. Introns Up indicates genes with increased total genic non-DI read counts (transcriptionally upregulated); Introns Down, genes in which it decreased (transcriptionally downregulated), and Introns NC genes in which the introns showed no change (no transcriptional difference).

Supplemental Figure S6: Doxorubicin treatment induces splicing changes in DIs from DNA-damage response genes. (A) mESCs were treated with DMSO, CB19 doxorubicin, or CB19 plus doxorubicin for 2,4, or 6 hours. RT-PCR was performed on the indicated transcripts and PCR products visualized on an EtBR-agarose gel. The gel was scanned on a Typhoon Phosphorimager and the ratio of coding isoforms to total transcript was quantified and plotted in the graph at right. Bars represent the average and error bars are standard error of the mean, N=3. (B) mESCs were treated with 1 μM doxorubicin or DMSO as a control for 2,4, or 6 hours. RT-PCR was performed on the indicated transcripts and PCR products visualized on an EtBR-agarose gel and quantified on a Typhoon imager. Numbers below indicate average normalized to time 0, +/- S.E.M., N=3. (C) mESCs were treated with 1 mM doxorubicin or DMSO as a control for 2,4, or 6 hours. Real time qRT-PCR was performed on RNA from doxorubicin or DMSO treated cells. All transcript specific signals were first normalized to Mylpf transcript levels, then normalized to the signal at time zero. DI and total transcript signals were normalized to DMSO treated cells at time zero.

Supplemental Figure S7: Liver and ESCs express common and tissue-specific DIs; a model for detained

introns in the control of gene expression

(A) Genes expressed in both ESCs and liver were divided into three groups and subjected to DAVID GO analysis using all genes expressed in both tissues as the background. Left panel, GO for genes with DIs in both tissues; center, GO for genes expressed in both but with DIs only in ESCs, and right, GO for genes expressed in both but only containing a DI in liver. (B) Models for intron detention in gene expression regulation. The left panel shows the case of NMD-switch associated DIs, whereas the right panel depicts the model for a constitutive DI. During the transcription of a gene, introns can be removed cotranscriptionally, or detained. By increasing the proportion of the transcript pool that contains DIs, pre-mRNA encoding particular proteins can be degraded before maturation and translation, decreasing gene expression. These DI containing transcripts in turn form a "buffer" pool of pre-mRNAs that can be redirected to undergo more rapid splicing and release of mature mRNA to the cytoplasm, thus transiently increasing gene expression. For some DIs, Clk kinase inhibits the completion of splicing by maintaining hyperphosphorylation of SR proteins in early spliceosomes. Inhibition of Clk kinase activity or activation of PP1/2A phosphatases catalyzes dephosphorylation of SR proteins or other spliceosome components to allow completion of splicing, cytoplasmic export and translation of the mRNA, and recycling of SR proteins to bind nascent transcripts. Autoregulation or stress can feed back on the rate of intron detention, in the case of the Clk responsive DIs, through modulation of Clk or SR protein activity.

Extended Materials & Methods

Read Filtering for repeat RNAs

A list of regions for filtering, including repeat RNAs (ribosomal, snoRNA, tRNA, snRNA), pseudogenes, single exon ribosomal proteins, and microRNAs, was compiled from Ensembl BioMart and Repeat Masker (Smit et al., 1996). Following Bowtie mapping as described in the Experimental Procedures, intersectBed from BEDTools

(Quinlan and Hall, 2010) was used to remove any reads present within the filter list.

Splicing Classification

Alternative and constitutive classifications were performed using custom Python scripts, and are agnostic with regard to existing annotations other than known gene boundaries. The workflow takes a set of intron coordinates, assigns them to a gene, and divides them into subgroups based on overlapping coordinates. If no overlapping introns exist for a given intron, it is assigned to the *constitutive* class. The subgroups containing overlapping introns are assigned a splicing classification if the start and end coordinates of all of the constituent introns fall into a pattern representing a known splice type (*cassette, mutually exclusive, alternative 5' splice site, alternative 3' splice site*). The introns within a classified subgroup are designated according to their position within the splicing event :

USMXEa: Intron upstream of the first of two mutually-exclusive exons (exon a)

DSMXEa: Intron downstream of exon a (filtered due to overlap with exon b)

MXEMI: Intron between first and second mutually-exclusive exons

USMXEb: Intron upstream of the second of two mutually-exclusive exons (exon b, filtered due to overlap with exon a)

DSMXEb: Intron downstream of exon b

MXEBS: Intron skipping both mutually exclusive exons (filtered due to overlap with exons)

CAS_US: Intron upstream of cassette exon

CAS_DS: Intron downstream of cassette exon

CAS_Skipped: Intron skipping cassette exon (filtered due to overlap with exon)

DUAL: Intron with alternative 5' and 3' splice sites (grouped with complex)

A3P: Alternative 3' splice site, proximal

A3D: Alternative 3' splice site, distal (filtered due to overlap with exon)

A5P: Alternative 5' splice site, proximal (filtered due to overlap with exon)

A5D: Alternative 5' splice site, distal

T3P: Tandem (NAGNAG) alternative 3' splice site, proximal (grouped with Alt 3' ss)

T3D: Tandem (NAGNAG) alternative 3' splice site, distal (filtered due to overlap with exon)

T5P: Tandem (NAGNAG) alternative 5' splice site, proximal (filtered due to overlap with exon)

T5D: Tandem (NAGNAG) alternative 5' splice site, distal (grouped with Alt 5' ss)

Complex: Intron part of alternative event that cannot be categorized into only one of the primary classes (mutually exclusive, cassette, alt 5'ss, alt 3'ss)

Constitutive: No alternative splice site usage

Identification of polyadenylation sites

To identify polyadenylation sites in introns that might result in false-positive DIs, we used coordinates and read counts from 3'-end sequencing experiments performed in V6.5 mESCs (Almada et al., 2013). Putative polyadenylation sites were clustered using BEDTools cluster (Quinlan et al., 2010) to merge putative sites within a 40 nucleotide window and the read counts within these clusters were summed. To remove low-end noise, we required polyadenylation sites to be present in both of two biological replicates at a threshold read count > 10% of the mean read count for all clusters. This process identified 25,591 polyadenylation sites in mESCs. Introns containing these clusters were removed from the set of introns to be considered for identification of DIs.

Annotation of introns: Inferring maximal set of non-overlapping introns

Our Detained Intron identification pipeline consists of two components. The first is an annotation of the transcriptome within the sample(s) to be analyzed that indicates the coordinates of all introns, and the second is the independent statistical determination of introns showing higher than expected abundance. The annotation is independent in the sense that any annotation scheme can be input, for example, we used the Aceview annotation to identify human DIs. However, the presence of exonic sequence within the annotation, e.g. introns that

partially or completely overlap exons, or small-repeat RNAs within the intronic interval, will increase the rate of false-positives within the list of called DIs. We have therefore used a combination of pre- and post- processing, and in some cases supervised annotation, to derive the sets of intronic coordinates to be input into the statistical analysis as described in the following steps. Stranded, paired-end reads from each lane of sequencing were mapped together using Tophat version 2.0.9 (Trapnell et al., 2009) with a custom junctions file as reference and default parameters. The resulting junctions .bed files from each of two lanes (12 samples) were then combined and collapsed to generate a non-redundant set of splice junctions. These were then filtered to eliminate mapping artifacts by requiring a minimum of 100 reads across all 12 samples to support the splice junction. After filtering, we still identified 7943 junctions not present in any of the Aceview, UCSC, Ensembl, or Refseq annotations. This base set of introns was then filtered to eliminate first and last introns and introns that contained polyadenylation sites based on 3' end sequencing in mouse ES cells (Almada et al., 2013) identified as described above, and a minimum intron length > 100 nucleotides was required. Finally, introns that overlap with each other or with potential exonic sequence were eliminated in two steps. First, we dropped introns with splicing classifications that contain exons within them (cassette skipped, mutually exclusive introns in which the upstream or downstream exon is skipped, and proximal 5' and distal 3' splice sites). These steps eliminate spanning introns (those that might subsume other introns or exons) and provide a non-overlapping set for downstream analyses. This problem is a special case of the more general problem of interval scheduling (Kovalyov et al., 2007). Specifically, a greedy algorithm using a last-fit strategy has been proven to provide an optimal solution to determining the maximal set of non-overlapping intervals in polynomial time (Carlisle and Lloyd, 1991; Faigle and Nawijn, 1994). We employed this approach to derive an optimal set of gene-specific non-overlapping introns. Briefly, introns were ordered according to their genomic end coordinates within each gene. Starting with the first ordered intron in each gene, subsequent introns were selected (or skipped) from the ordered list if their start coordinates exceeded (or did not exceed) the end coordinate of the last selected intron. This resulted in a final high-confidence set of internal introns used to identify detained introns using the *in silico* null model enrichment analysis described below. Lastly, for the mESC intronic annotation, we applied a post-processing filter step to eliminate any remaining exon overlaps within the intronic regions. First, to

eliminate overlapping introns within the *complex* splicing class, the shortest intron with a shared 5' or 3' splice site was chosen, and we then implemented a interval scheduling algorithm. Second,we eliminated coding retained introns by filtering out introns that are completely contained within Refseq coding exons, and which preserve reading frame; introns within untranslated exonic regions were filtered out if they were completely contained within non-coding Refseq introns. Finally, detained introns within a *complex* class unit that contains a first or last intron were visually inspected to ensure they were not inappropriately filtered out. For the human intron data set, Hg19 Aceview annotations were used as the starting point. This set of introns was filtered as described above by splice type category and collapsed by interval sorting to produce the final set of non-overlapping introns for each cell line separately.

In silico null model replicates and intron coverage enrichment

Using the list of non-overlapping introns derived using the above-described custom annotations and interval scheduling approach, we devised an *in silico* scheme to identify introns with statistically significant enrichment in read coverage within RNA-seq datasets. Starting with two RNA-seq replicates per condition for each time-point, we generated two *in silico* replicates under a null model to use as a baseline for enrichment analysis, as follows. Read counts for each intron were determined for each of the two RNA-seq replicates using uniquely mapped, filtered reads as described above. To be counted in an intron, the reads that spanned an exon-intron junction were required to have a minimum overlap of 10 nucleotides within the intron. Paired-end fragments where both reads mapped to the same intron, were counted as a single read. Using DESeq (Anders and Huber, 2010), we determined normalizing constants ("size-factors") for these intronic read counts and normalized the counts in order to account for depth of coverage and data dispersion across the two RNA-seq datasets. For each gene, the sum of normalized intronic read counts was then used to allocate reads to individual introns under a null model. First, an "effective length" for each intron was derived using 100mer mapability tracks (Roderic Guigo Lab, CRG, Barcelona; Myers et al., 2011) from the UCSC Genome Browser (Kent et al., 2002). Positions

of non-unique 100mer alignments were summed and then subtracted from the corresponding intron's length. A variance stabilizing transform (Gao et al., 2011) based on the square root of intron effective length adjusted by RNA-seq read length ($\sqrt{(L-d)}$ where L=mapability adjusted intron length, d = RNA-seq read length) was then used to weight individual introns. The sum of normalized intronic reads per gene in each RNA-seq replicate was then partitioned and allocated to each intron proportional to its weight. This results in two *in silico* null model replicates, one corresponding to each RNA-seq replicate. Differential analysis using DESeq was then used to determine introns enriched in read coverage (in the RNA-seq replicates) compared to the *in silico* null model replicates using an FDR adjusted p-value threshold of 0.01 and fold change threshold of 2 for the human cell lines and mESCs. For the mouse liver samples, intronic sequences were processed from RNA-seq junctions. A nominal p-value of 0.01 and fold change threshold of 2 was also used to identify DIs,, but to account for larger variation in tissues compared to cell lines the FDR was adjusted to 0.1. At an FDR of 0.01, 888 DIs were identified in liver of which 233(26.2%) were also present in mESCs. At an FDR of 0.1, 2347 DIs were identified of which 454 (19.3%) were also found in mESCs. The final list of detained introns were derived from a union of enriched introns in either condition at time 0, and to further eliminate low-end noise were required to pass a filter consisting of an average read count between DMSO and CB19 samples at 2 hours (base mean) to intron length ratio of >= 0.01.

Determination of orthologous introns in human and mouse

The set of Aceview derived human introns (hg19 assembly) that was used to determine DIs in the ENCODE human cell lines was mapped to mouse genome (mm9) using the UCSC genome browser liftover tool (http://genome.ucsc.edu/cgi-bin/hgLiftOver) with hg19.mm9.all.chain and -minMatch=0.1 (default for cross-species mapping). After conversion to mm9 coordinates, human introns were intersected with mouse introns using BEDTools intersectBed (Quinlan and Hall, 2010). Statistical significance of the overlapping sets of introns was then determined using Fisher's Exact Test and confirmed by two-sided hypergeometric distributions in R.

Intron phylogenetic conservation

PhastCons data files were downloaded from the UCSC Genome Browser (ftp://hgdownload.cse.ucsc.edu/goldenPath/mm9/phastCons30way/vertebrate). Custom Python scripts were used to process the data. For each intron, the PhastCons score per nucleotide was compiled accounting for the positions and sizes of gaps. The average score was determined across each intron, assigning a score of zero to each nucleotide within a gap.

Splice site strength

The sequences surrounding the 5' and 3' splice site of each intron were compiled and each was assigned a score using the maximum entropy method with the MaxEnt program (Yeo and Burge, 2004). To examine the differences between detained and non-detained introns, all splice site scores above zero were compared between the two groups and the significance in the difference between population means determined using Welch's T-test.

U11/U12 intron classification

Genomic coordinates (mm6) for a high-confidence set of minor spliceosome-dependent introns were downloaded from the U12DB (http://genome.crg.es/cgi-bin/u12db/u12db.cgi). These were mapped back to mm9 coordinates using the UCSC genome browser liftover tool.

Gene ontology analysis

The list of genes derived from the post-filtering set of introns used to identify detained introns was used as the

background for expressed genes in mouse ES cells, or in the case of the human cell lines, the union of all expressed genes in the four cell lines. The set of detained intron containing genes for mouse ES cells, or the list of detained intron genes common to all four cell lines for the human cells, was compared against the appropriate background set using the DAVID Gene Ontology Tool (Huang et al., 2009). Redundant categories were removed and the representative category with the lowest FDR was kept.

Quantification of intron changes, Upf1 KD and CB19 experiments

Reads were mapped uniquely using Bowtie 1.0.1, and repeat RNAs were filtered out as described above. Paired-end fragments where both reads mapped to the same intron were counted as a single read. Read counts for each region in each replicate sample were derived for each intron using the intersectBed and coverageBed options in BEDTools. Differential read density of individual introns was then determined using the DESeq (Anders and Huber, 2010) package in R, with two biological replicates for each sample being compared (shGFP replicates 1 and 2 vs. shUpf1 replicates 1 and 2, and DMSO (2 hour) replicates 1 and 2 vs. CB19 (2 hour) replicates 1 and 2). Changes in intron read density in either direction exhibiting a FDR adjusted p-value < 0.05 were considered significant.

Cassette exon inclusion quantification

Gff3 annotation files were manually produced or obtained from the provided mm9 skipped exon annotation (http://genes.mit.edu/burgelab/miso/docs/annotation.html;Mouse genome (mm9) alternative events v2.0). The annotations were indexed and combined read alignments for the two biological replicates for each treatment/time point were run through exon-centric analysis with the MISO algorithm (Katz et al., 2010). The resulting data files were then used to produce the Sashimi plots (http://genes.mit.edu/burgelab/miso/docs/sashimi.html).

Gene/isoform expression analysis

Following mapping in aggregate (per lane) with Tophat as described above, gene expression analysis was performed using the Cufflinks suite (Trapnell et al., 2010). Total mapped reads from each lane were first run through Cufflinks with an mm9 Ensembl gtf. The Cufflinks output from the two lanes was then combined using Cuffmerge into a single gtf, which was then used to run Cuffdiff on barcode-separated read alignments for each biological replicate. The FPKM values for all isoforms representing manually annotated coding, nonsense-mediated decay, or DI-containing isoforms were combined and the mean FPKM for both biological replicates +/- pooled 95% confidence intervals were calculated. For genome-wide FPKM estimates (Figure S2B), the gene-level estimates from Cufflinks in the time zero samples were used. For nuclear/cytoplasmic poly(A)-selected RNA seq isoform quantification, ENCODE RNA-seq raw data sets from HeLa, HepG2, and HUVEC were mapped using Tophat as described previously. FPKM estimates for each isoform annotated (see Supplemental Data S3 .gtf) were determined for both nuclear and cytoplasm for each cell line using biological replicates through Cufflinks. Data points for all three cell lines for all isoforms were combined to produce the scatterplots shown in Figure 4A. 50% and 95% concentration eclipses were produced using the CAR package in R.

Gene functional classification

Functional classifications for genes containing CB19-responsive DIs (Figure 5E) were assigned manually based on extensive literature review.

CLIP-seq data processing

Fastq files for GFP, Srsf3, and Srsf4 CLIP experiments (Anko et al., 2012) were downloaded from the Array

Express database (accession number E-MTAB-747). Reads were collapsed to remove PCR duplicates and adaptors trimmed using FastX-trimmer (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Reads were mapped using Bowtie2.1.0 with the --sensitive-local option. CLIP clusters were generated using the Clipper algorithm (Lovci et al., 2013) with pre-mRNA lengths used to determine background read distributions. Clusters were further filtered using a binomial test for the SR-specific CLIP vs. GFP only CLIP. Reads in GFP only CLIP were scaled to match the SR-specific CLIP library size by performing linear regression of crosslink counts within genes. Clusters with SR-specific CLIP enrichment above a q-value threshold of 0.05 were then filtered by peak read value and a minimum cluster length. CLIP clusters associated with small repeat RNAs were removed using BEDTools intersectBed.

Motif analysis

Sequences from indicated regions (e.g., 100 nucleotides upstream of the 5' splice site of an intron) in a particular set of samples (e.g., DIs that increase or DIs that decrease in response to CB19) were extracted from the mm9 assembly using Samtools (Li et al., 2009). A scanning window was used to count the number of occurrences of each pentamer or hexamer within the region. The occurrence frequencies were then used to calculate a Z-score for each motif.

References

Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA. 2013. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499:** 360-363.

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

Anko ML, Muller-McNicoll M, Brandl H, Curk T, Gorup C, Henry I, Ule J, Neugebauer KM. 2012. The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biol* **13:** R17.

Calarco JA, Superina S, O'Hanlon D, Gabut M, Raj B, Pan Q, Skalska U, Clarke L, Gelinas D, van der Kooy D,

Zhen M, Ciruna B, Blencowe BJ. 2009. Regulation of vertebrate nervous system alternative splicing and development by an SR-related protein. *Cell* **138:** 898-910.

Carlisle MC, Lloyd EL. 1991. On the k-coloring of intervals. Advances in Computing and Information - ICCI'91, Lecture Notes in Computer Science, F Dehne, F. Fiala and W.W.Koczkodaj, eds., 497 (Springer, Berlin) 90-101.

Crooks GE, Hon G, Chandonia JM, and Brenner SE. 2004. WebLogo: A sequence logo generator. *Genome Research* **14:** 1188-1190.

Faigle U, Nawijn WM. 1995. Note on scheduling intervals on-line. *Discrete Applied Mathematics* **58:** 13-17.

Gao L, Fang Z, Zhang K, Zhi D, Cui X. 2011. Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics* **27:** 662–669.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc* **4:** 44-57.

Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods* **7:** 1009-1015.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12:** 996-1006.

Kovalyov MY, Ng CT, Edwin Cheng TC. 2007. Fixed interval scheduling: Models, applications, computational complexity and algorithms. *European Journal of Operational Research* **178:** 331–342.

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

Long JC, Caceres JF. 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* **417:** 15-27.

Myers RM, Stamatoyannopoulos J, Snyder M, Dunham I, Hardison RC, Bernstein BE, Gingeras TR, Kent WJ, Birney E and ENCODE Project Consortium .2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9:** e1001046.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841-842.

Smit AFA, Hubley R, Green P. 1996-2010. RepeatMasker Open-3.0.(http://www.repeatmasker.org).

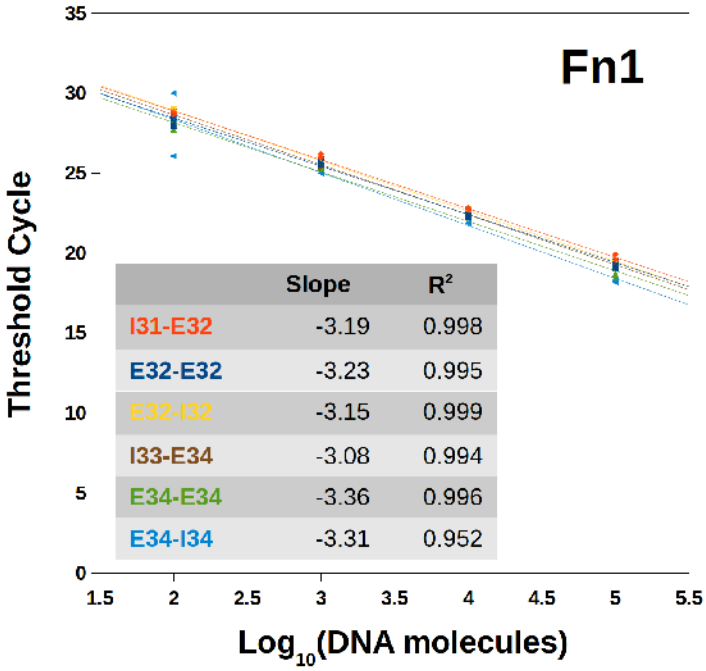
Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25:** 1105-1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511-515.

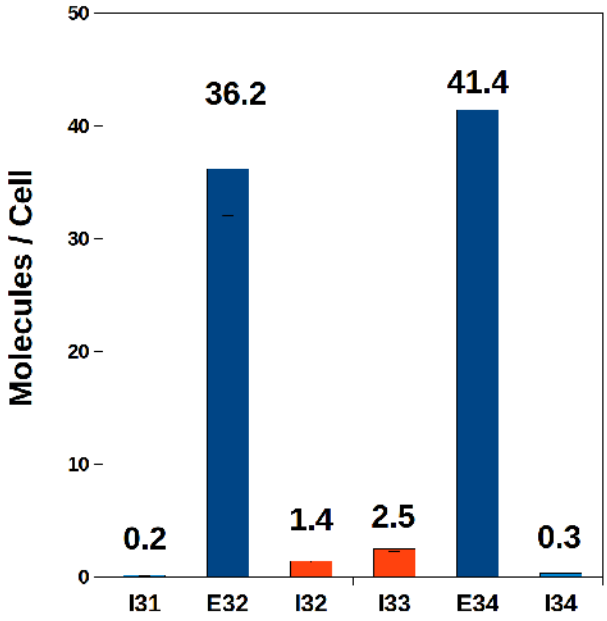
Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11:** 377-394.

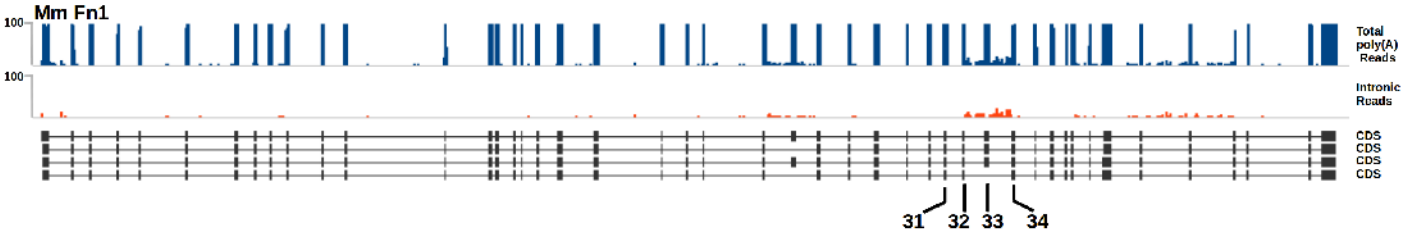
Supplemental Figure S1

A

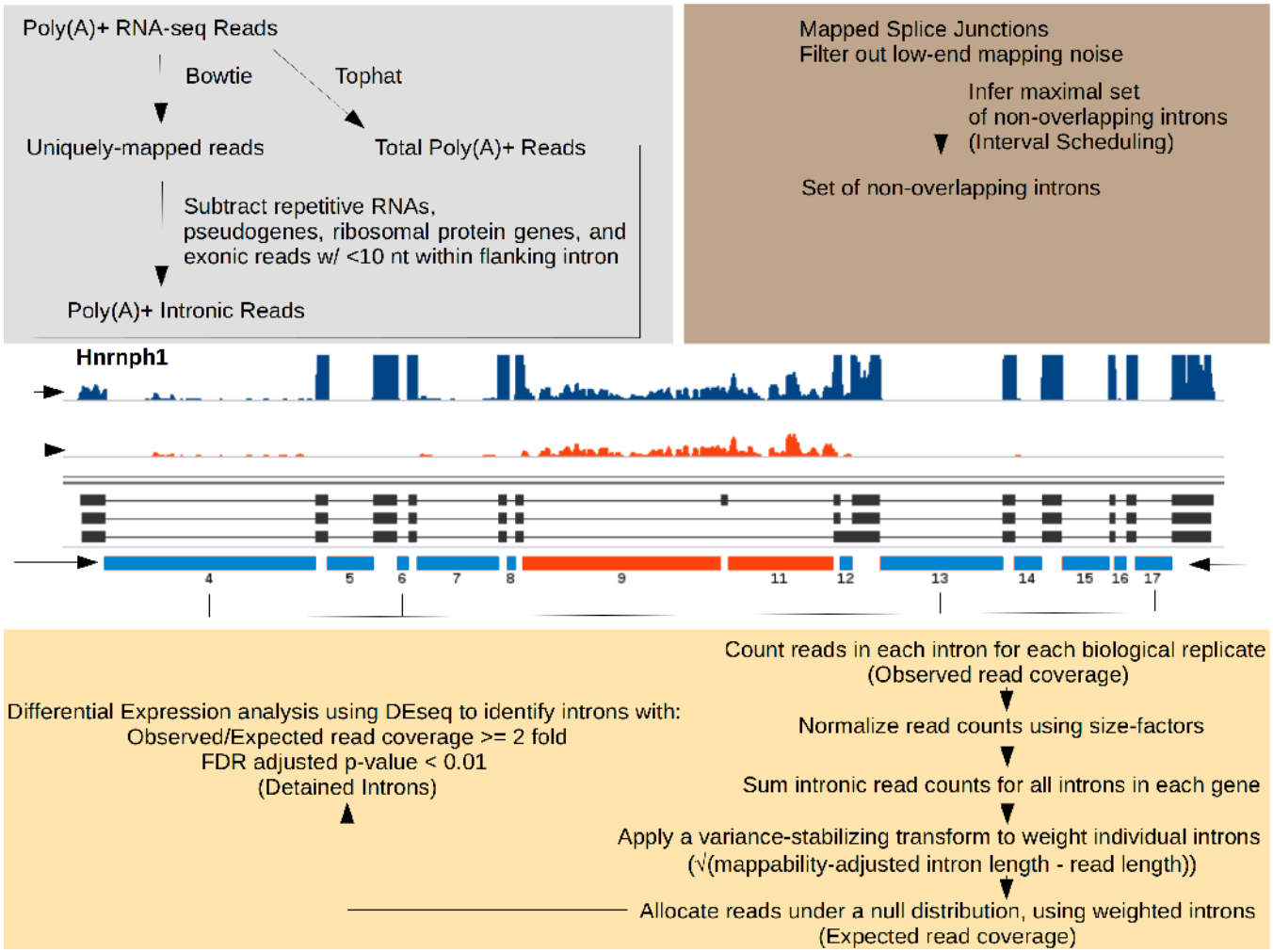


	Slope	R^2
I31-E32	-3.19	0.998
E32-E32	-3.23	0.995
E32-I32	-3.15	0.999
I33-E34	-3.08	0.994
E34-E34	-3.36	0.996
E34-I34	-3.31	0.952

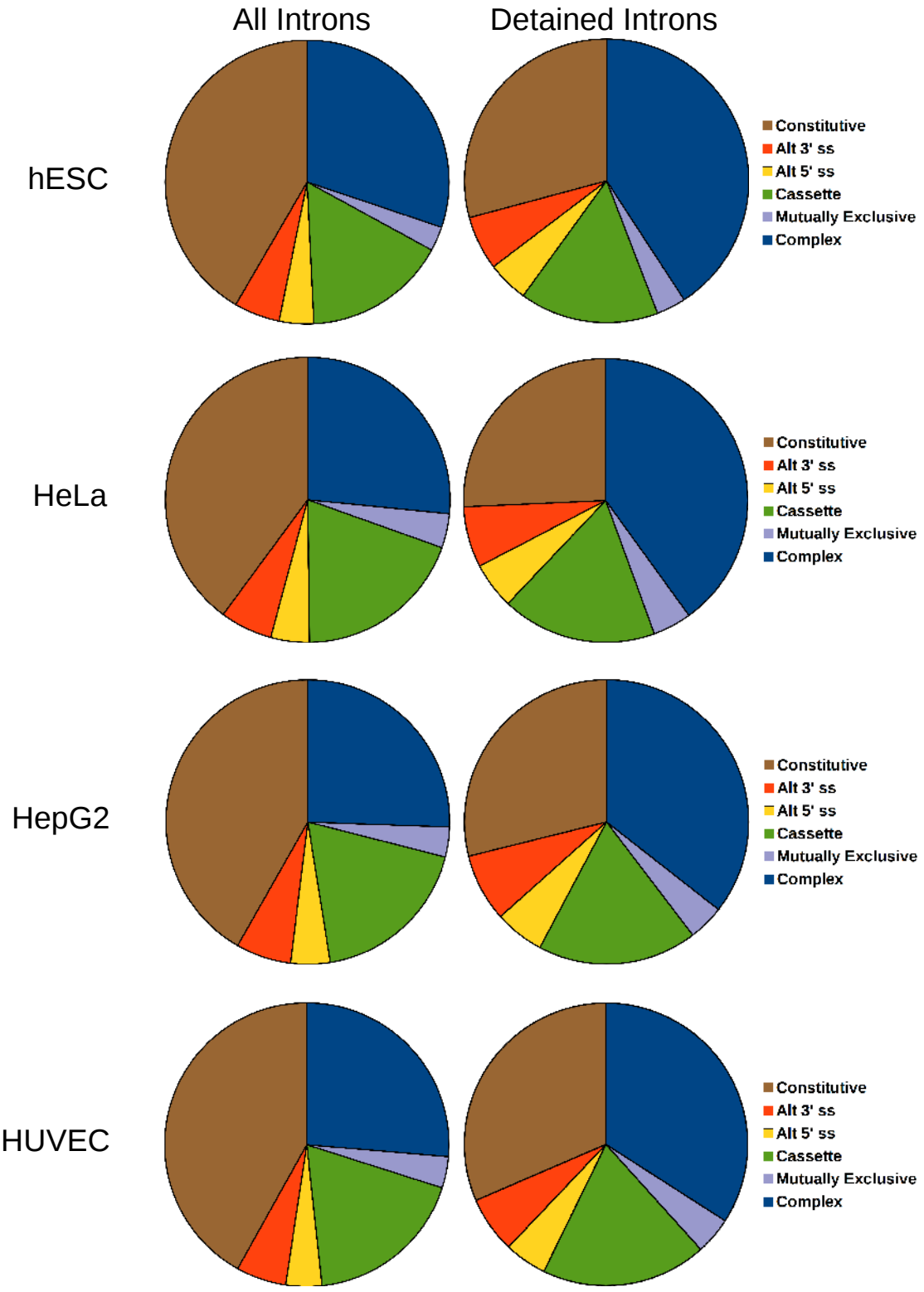




B



Supplemental Table S1



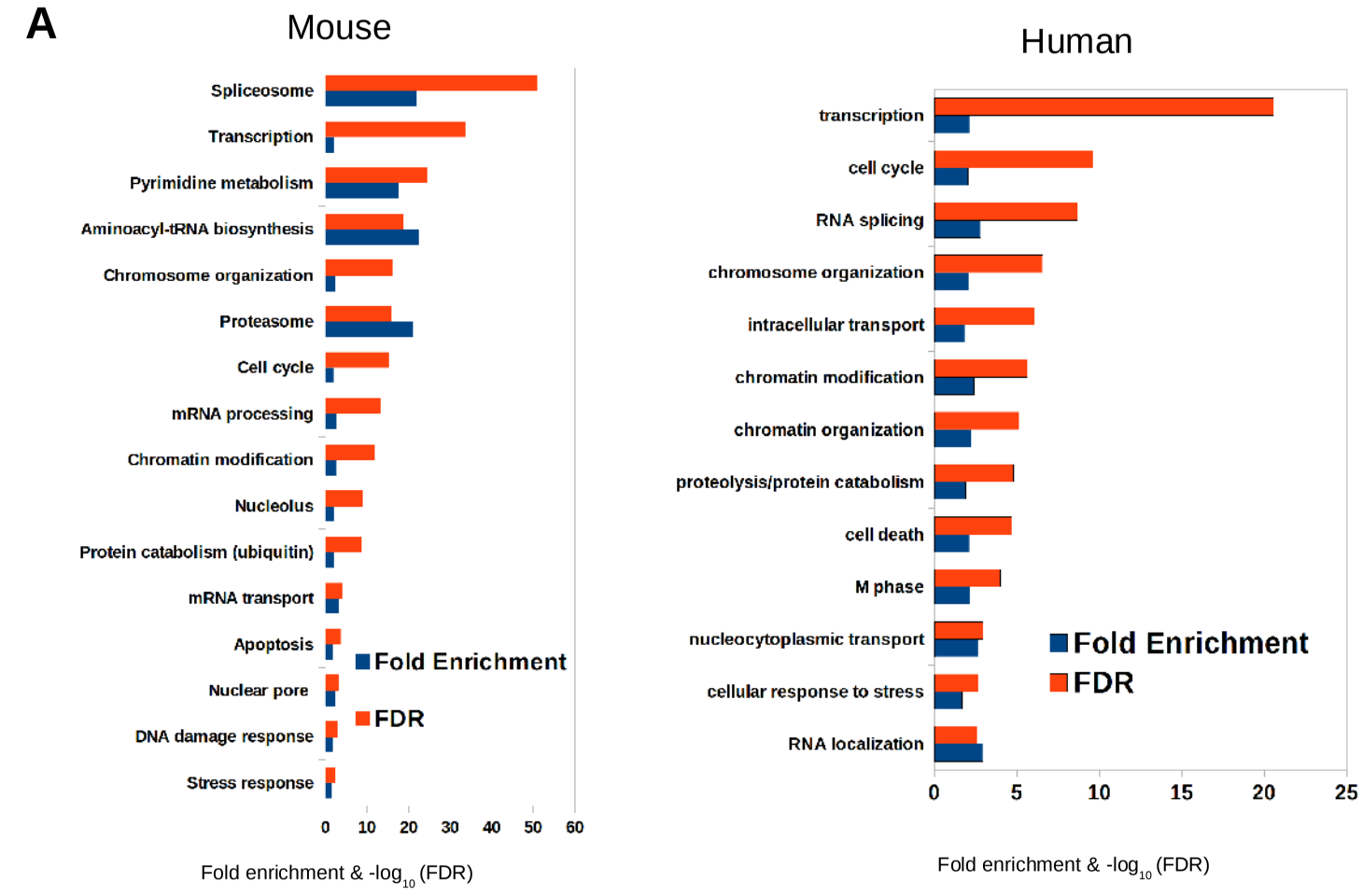
hESC	Not-DI	DI	Total	p-value	
Constitutive	23989	1325	25314	2.72E-072	** Depleted
Alt 3' ss	2897	280	3177	0.003564	** Enriched
Alt 5' ss	2176	212	2388	0.008697	** Enriched
Cassette	9187	718	9905	0.3803	
Mutually Exclusive	1547	152	1699	0.021640	** Enriched
Complex	16496	1853	18349	< 2.2E-016	** Enriched
Total	56292	4540	60832	DI = 7.46% of introns	

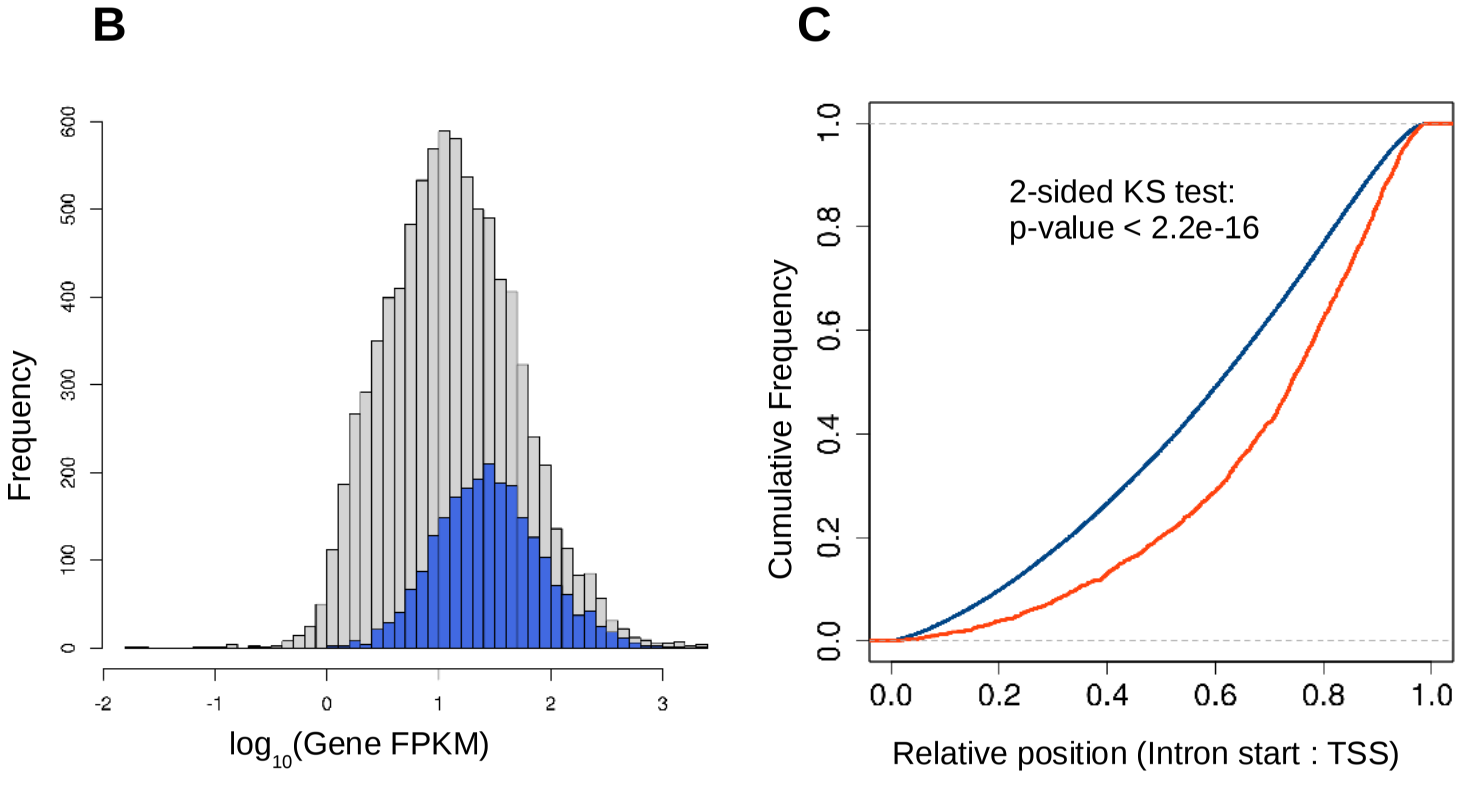
HeLa	Not-DI	DI	Total	p-value	
Constitutive	24632	871	25503	1.57E-070	** Depleted
Alt 3' ss	3558	235	3793	0.01223	** Enriched
Alt 5' ss	2614	179	2793	0.008387	** Enriched
Cassette	11795	597	12392	0.007851	** Depleted
Mutually Exclusive	2330	149	2479	0.1092	
Complex	15635	1356	16991	< 2.2E-016	** Enriched
Total	60564	3387	63951	DI = 5.30% of introns	

HepG2	Not-DI	DI	Total	p-value	
Constitutive	23579	2229	25808	9.65E-138	** Depleted
Alt 3' ss	3292	595	3887	9.13E-008	** Enriched
Alt 5' ss	2335	436	2771	3.02E-007	** Enriched
Cassette	10034	1400	11434	0.3804	
Mutually Exclusive	1783	314	2097	0.0006054	** Enriched
Complex	13056	2745	15801	< 2.2E-016	** Enriched
Total	54079	7719	61798	DI = 12.49% of introns	

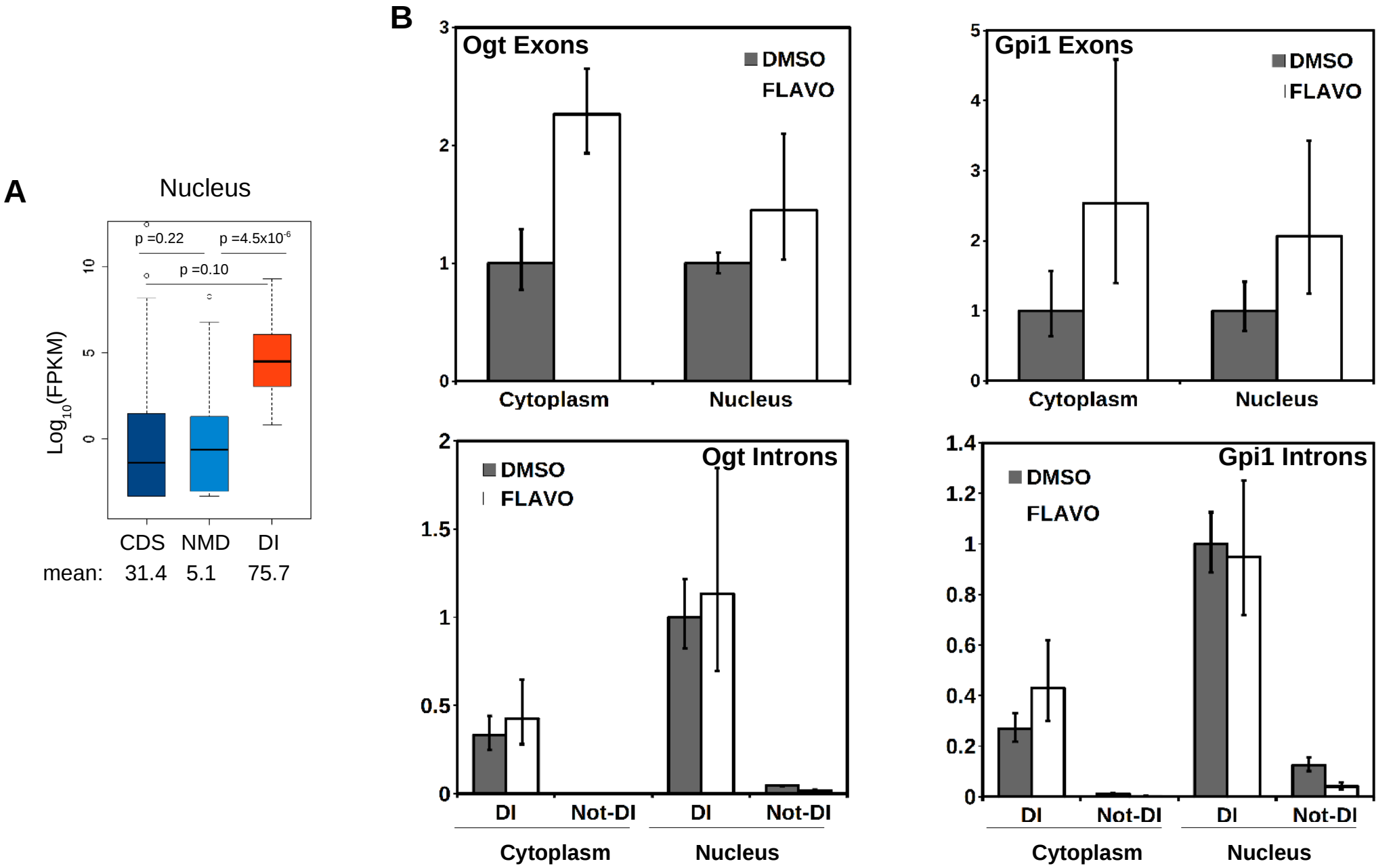
HUVEC	Not-DI	DI	Total	p-value	
Constitutive	21064	2637	23701	2.90E-047	** Depleted
Alt 3' ss	2683	540	3223	1.39E-007	** Enriched
Alt 5' ss	1906	406	2312	3.97E-008	** Enriched
Cassette	8808	1582	10390	1.03E-007	** Enriched
Mutually Exclusive	1663	355	2018	2.38E-007	** Enriched
Complex	12043	2855	14898	< 2.2E-016	** Enriched
Total	50682	7967	58649	DI = 14.81% of introns	

Supplemental Figure S2

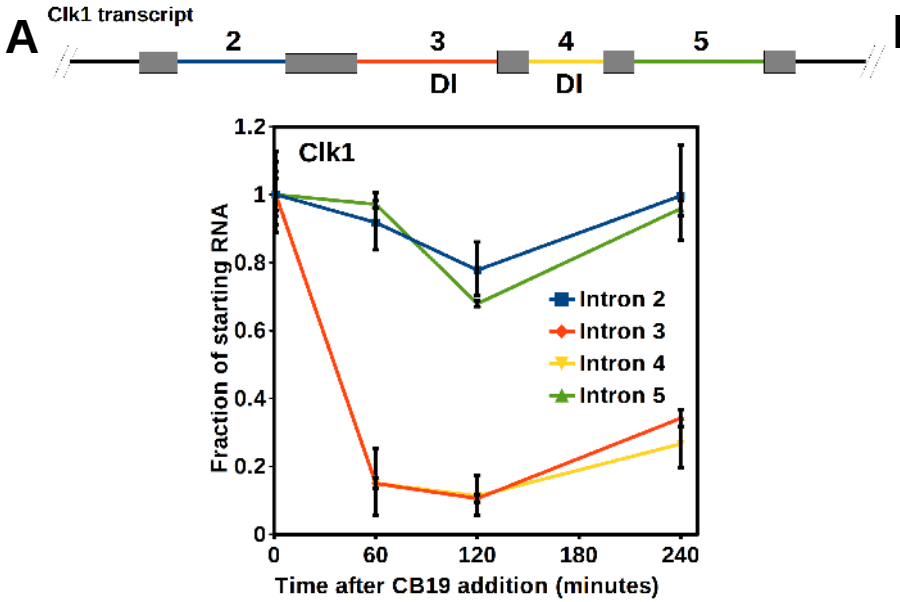


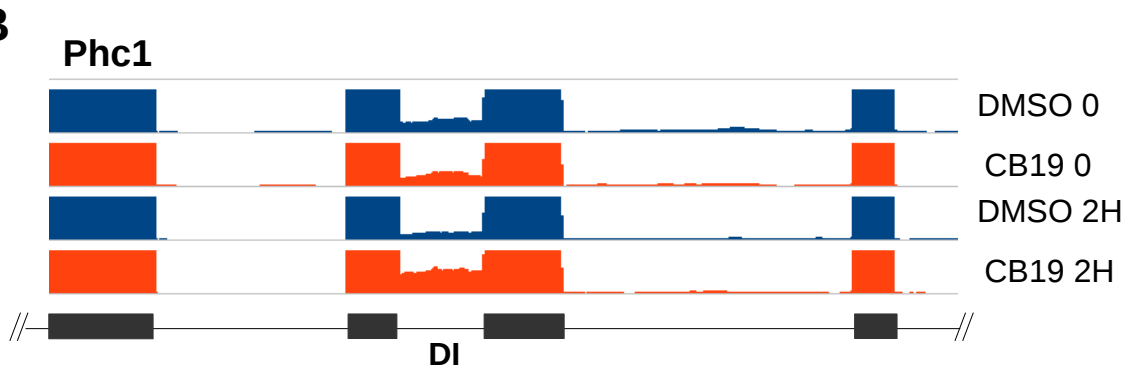


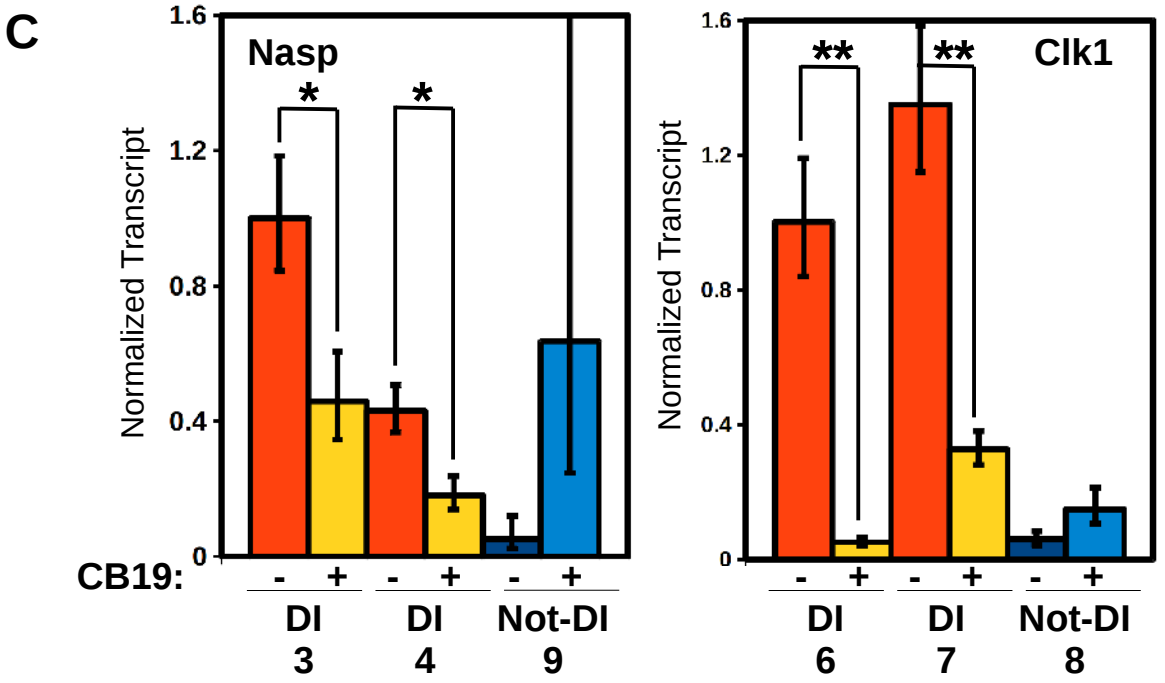
Supplemental Figure S3

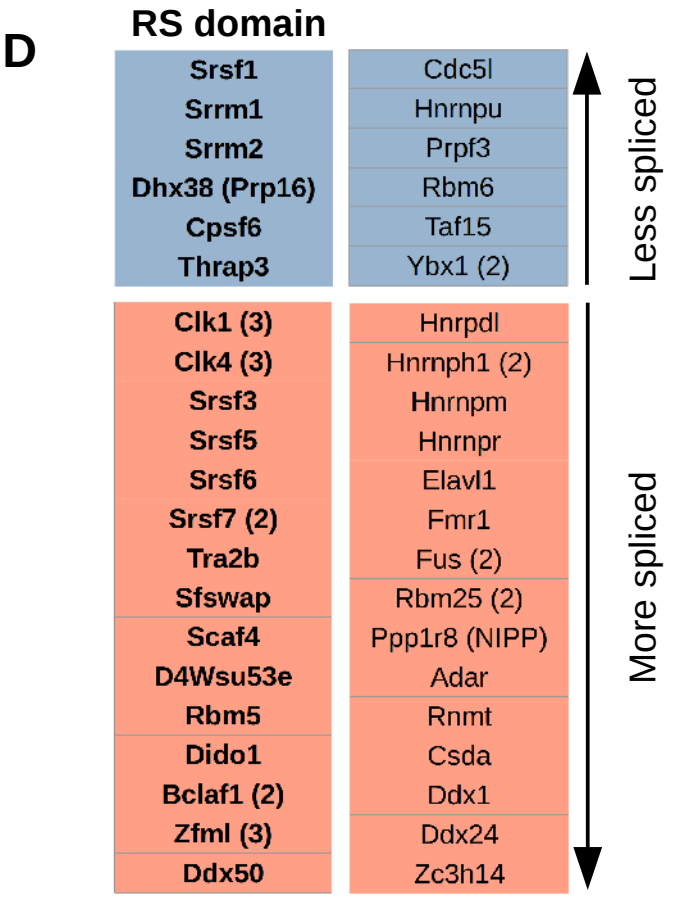


Supplemental Figure S4

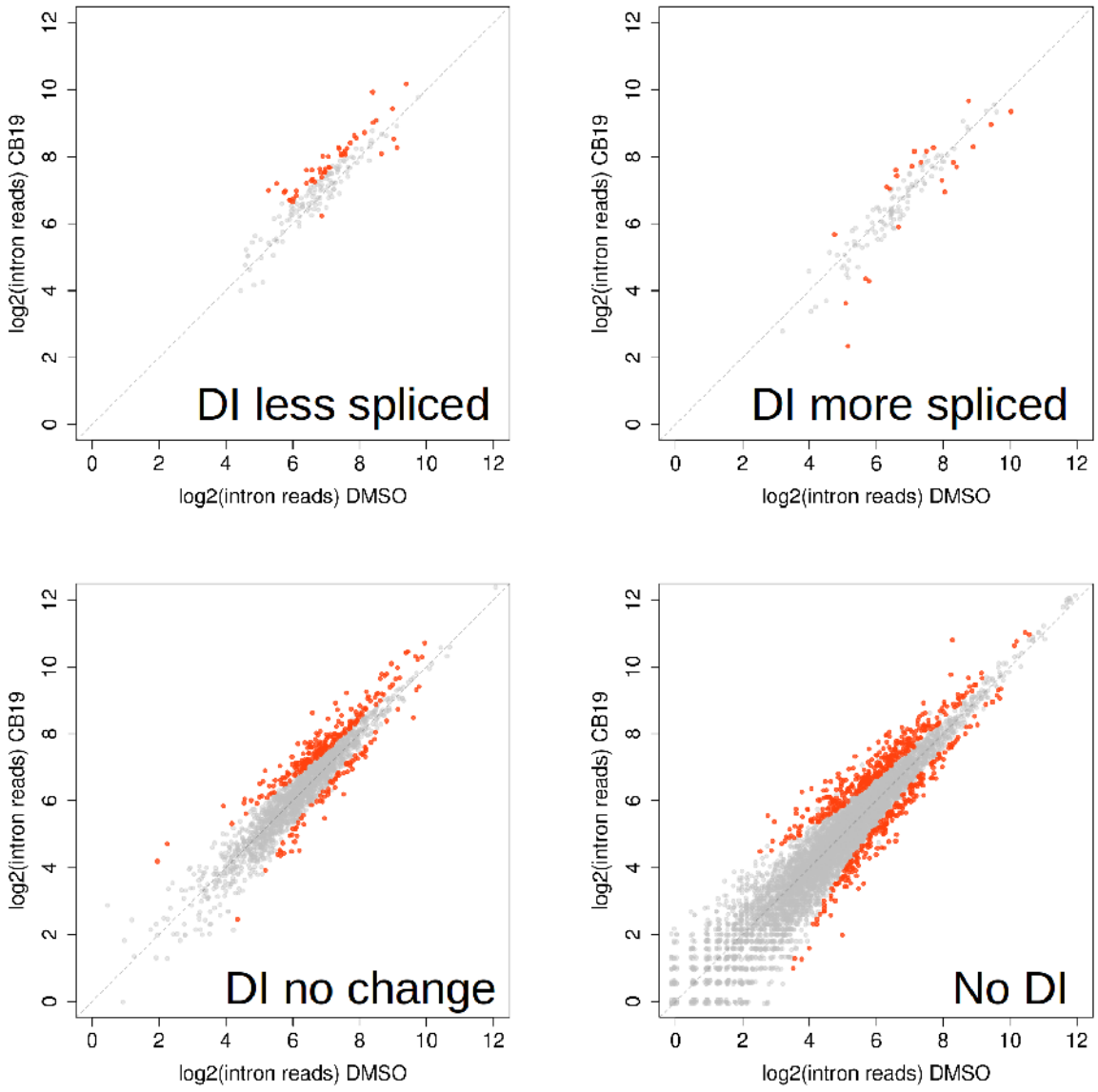








A

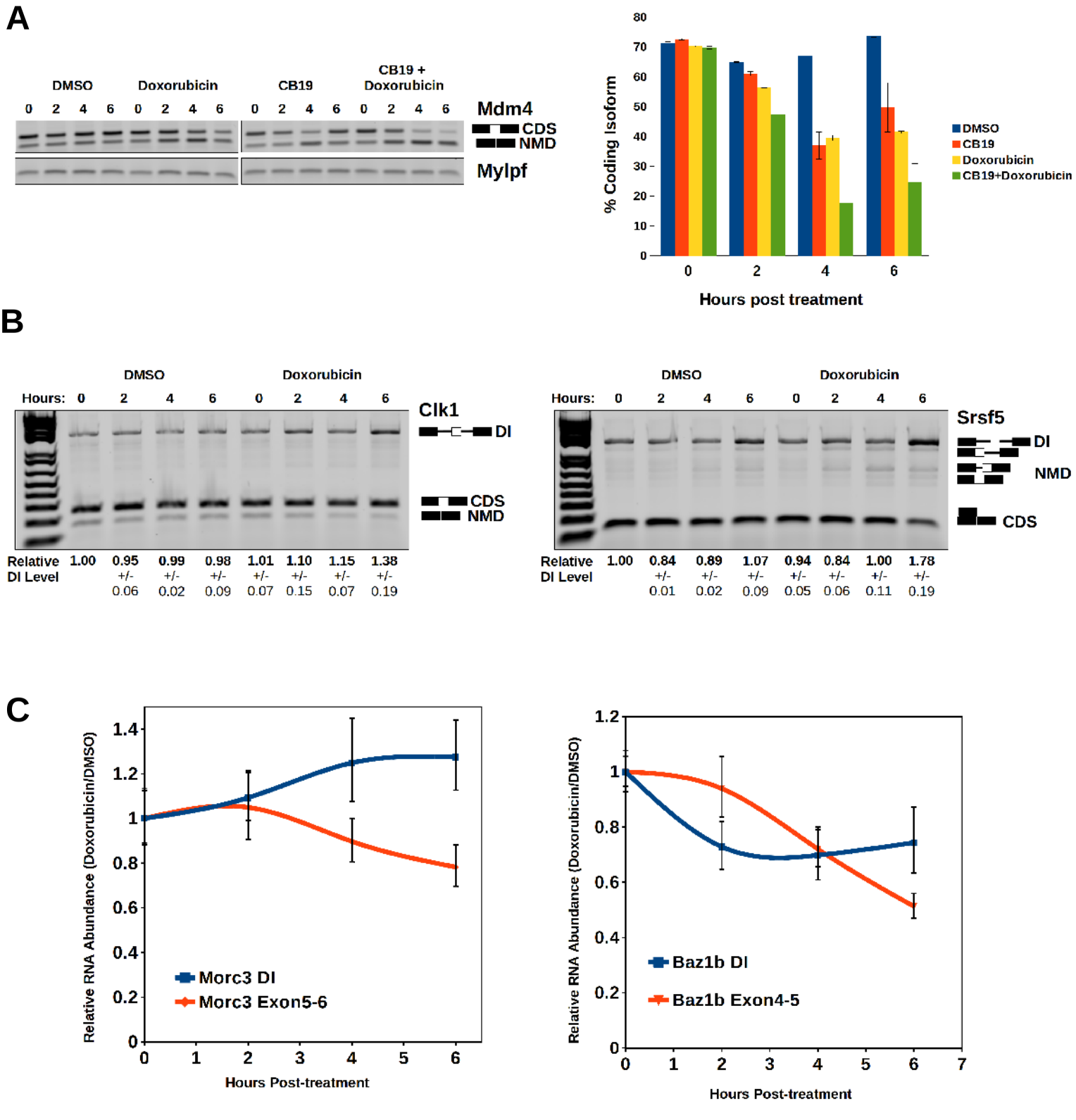


B

All non-detained introns ingene, significant change CB19/DMSO, padj < 0.05

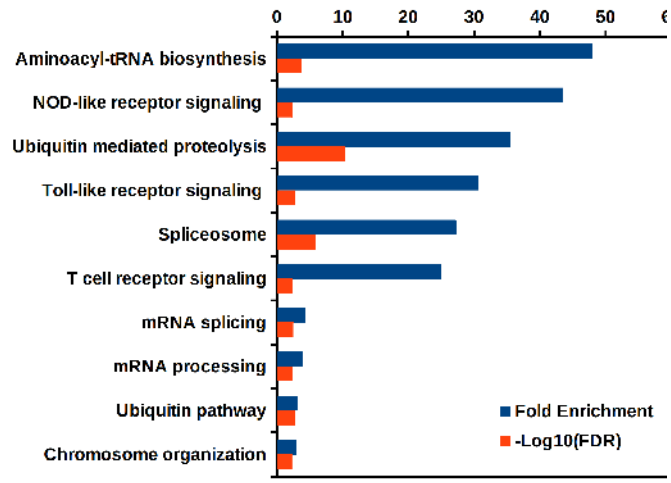
Intron status of gene	Introns up	Introns down	Introns NC
DI less spliced	43	4	122
DI more spliced	11	12	108
DI no change	202	69	1855
No DI	328	215	6128

Supplemental Figure S6

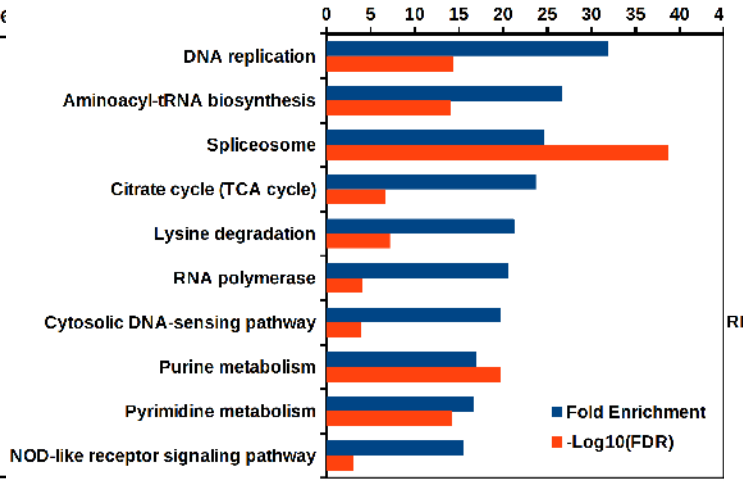


Supplemental Figure S7

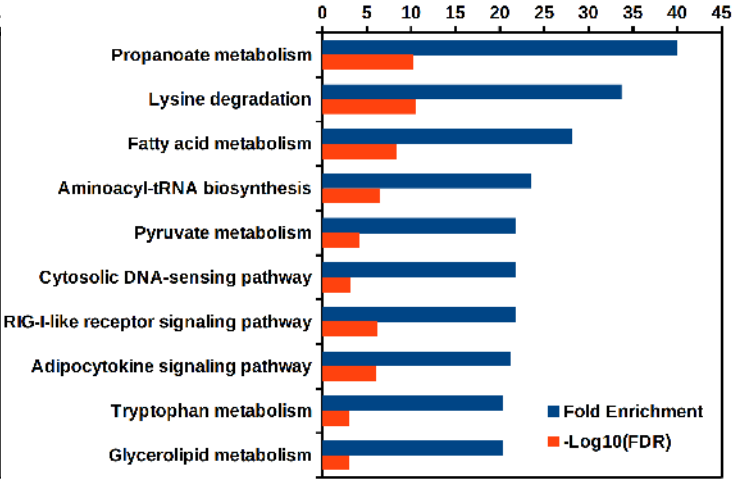
A



Genes with DIs in ESCs and Liver



Genes with DIs in ESCs not Liver



Genes with DIs in Liver not ESCs

B