



## Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring

Kyu-Hyeon Yeom, Zhicheng Pan, Chia-Ho Lin, et al.

*Genome Res.* published online April 8, 2021

Access the most recent version at doi:[10.1101/gr.273904.120](https://doi.org/10.1101/gr.273904.120)

---

<b>P&lt;P</b>	Published online April 8, 2021 in advance of the print journal.
<b>Accepted Manuscript</b>	Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.
<b>Creative Commons License</b>	This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see <a href="https://genome.cshlp.org/site/misc/terms.xhtml">https://genome.cshlp.org/site/misc/terms.xhtml</a> ). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <a href="http://creativecommons.org/licenses/by-nc/4.0/">http://creativecommons.org/licenses/by-nc/4.0/</a> .
<b>Email Alerting Service</b>	Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or <a href="#">click here</a> .

---



---

To subscribe to *Genome Research* go to:  
<https://genome.cshlp.org/subscriptions>

---

Published by Cold Spring Harbor Laboratory Press

**Tracking pre-mRNA maturation across subcellular compartments identifies developmental gene regulation through intron retention and nuclear anchoring.**

Kyu-Hyeon Yeom<sup>1,†</sup>, Zhicheng Pan<sup>2,3,†</sup>, Chia-Ho Lin<sup>1</sup>, Han Young Lim<sup>1,4</sup>, Wen Xiao<sup>1</sup>, Yi Xing<sup>3,5,\*</sup>, Douglas L. Black<sup>1,\*</sup>.

Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, Los Angeles, CA 90095

Bioinformatics Interdepartmental Graduate Program, University of California, Los Angeles, CA 90095

Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA 19104

Molecular Biology Interdepartmental Doctoral Program, University of California, Los Angeles, Los Angeles, CA 90095

Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104

† Equal Contributions

\* Correspondence: [xingyi@email.chop.edu](mailto:xingyi@email.chop.edu), [dougb@microbio.ucla.edu](mailto:dougb@microbio.ucla.edu)

RUNNING TITLE: Tracking mRNA maturation during differentiation

KEYWORDS

Subcellular compartment, neuronal differentiation, intron retention, chromatin-associated RNA, *Gabbr1*, splicing, posttranscriptional gene regulation, RNA processing

## ABSTRACT

Steps of mRNA maturation are important gene regulatory events that occur in distinct cellular locations. However, transcriptomic analyses often lose information on the subcellular distribution of processed and unprocessed transcripts. We generated extensive RNA-seq datasets to track mRNA maturation across subcellular locations in mouse embryonic stem cells, neuronal progenitor cells, and postmitotic neurons. We find disparate patterns of RNA enrichment between the cytoplasmic, nucleoplasmic and chromatin fractions, with some genes maintaining more polyadenylated RNA in chromatin than in the cytoplasm. We bioinformatically defined four regulatory groups for intron retention, including complete cotranscriptional splicing, complete intron retention in the cytoplasmic RNA, and two intron groups present in nuclear and chromatin transcripts but fully excised in cytoplasm. We find that introns switch their regulatory group between cell types, including neuronally excised introns repressed by polypyrimidine track binding protein 1 (PTBP1). Notably, transcripts for the neuronal gamma-aminobutyric acid type B receptor subunit 1 (GABBR1) are highly expressed in mESC but are absent from the cytoplasm. Instead, incompletely spliced *Gabbr1* RNA remains sequestered on chromatin, where it is bound by PTBP1 similar to certain long non-coding RNAs. Upon neuronal differentiation, *Gabbr1* RNA becomes fully processed and exported for translation. Thus, splicing repression and chromatin anchoring of RNA combine to allow posttranscriptional regulation of *Gabbr1* over development. For this and other genes, polyadenylated RNA abundance does not indicate functional gene expression. Our datasets provide a rich resource for analyzing many other aspects of mRNA maturation in subcellular locations and across development.

## INTRODUCTION

After transcription initiation, the maturation of pre-messenger RNA (pre-mRNA) requires splicing, polyadenylation, and release of the RNA from the chromatin template, before export to the cytoplasm for translation. For many genes, the bulk of expressed RNA exists in the cytoplasm as mature mRNA, while nascent, intron-containing transcripts are limited to small nuclear puncta at the sites of transcription (Vargas et al. 2011; Coulon et al. 2014). For other genes, unspliced introns may remain after transcript completion but are ultimately excised to allow export (Girard et al. 2012; Popp and Maquat 2013; Stewart 2019). These nuclear transcripts are not necessarily found at their gene loci but some polyadenylated transcripts, including many non-coding RNAs, are tightly associated with chromatin (Quinn and Chang 2016). Although proteins affecting processes such as DNA template release, RNA export, and nuclear RNA decay have been identified (Schmid and Jensen 2018; Stewart 2019), the global distribution of RNA transcripts between subcellular compartments, and the alteration of their maturation and location with development have not been well studied.

In earlier studies, we examined the kinetics of transcription, splicing, and nuclear export for macrophage transcripts induced by inflammatory stimuli (Bhatt et al. 2012; Pandya-Jones et al. 2013). By following inflammatory gene transcripts, we found that partially spliced but polyadenylated transcripts in the chromatin fraction completed splicing over time, and were released to the soluble nucleoplasmic fraction before appearing in the cytoplasm as functional mRNAs (Bhatt et al. 2012; Pandya-Jones et al. 2013). These studies focused on introns whose slow splicing impacted the rate of inflammatory gene expression. However, polyadenylated, partially spliced RNA has been long been observed in nuclei where its interactions and localization are largely unknown.

The above analyses used a fractionation procedure to enrich for nucleoplasmic or chromatin-associated RNA (Wuarin and Schibler 1994; Pawlicki and Steitz 2008; Pandya-Jones and Black 2009; Khodor et al. 2012; Herzog and Neugebauer 2015; Yeom and Damianov 2017). Nucleoplasmic and chromatin compartments are operationally defined as the supernatant and pellet fractions, respectively, after nuclear lysis in a stringent buffer containing NP-40, Urea, and NaCl. This solubilizes many components such as the U1 snRNP, while leaving other molecules associated with the high molecular weight chromatin pellet (Wuarin and Schibler 1994). The cytoplasmic fraction is enriched for mature mRNA, while the nucleoplasmic fraction contains recently matured transcripts released from the chromatin that have not yet reached the cytoplasm (Bhatt et al. 2012; Pandya-Jones et al. 2013), as well as some mature mRNAs associated with ER and mitochondria (Yeom and Damianov 2017). The chromatin pellet is enriched for nascent RNA bound by elongating RNA Pol II, but also contains substantial polyadenylated RNA, including the *Xist* non-coding RNA tightly bound to chromatin (Pandya-Jones et al. 2020), and the *Malat1* non-coding RNA, which is enriched in nuclear speckles that are adjacent to chromatin but only partially in contact with it (Hutchinson et al. 2007; Fei et al. 2017).

The consequences of intron retention are diverse and complex to dissect. Splice sites and binding of spliceosomal components can prevent nuclear RNA export (Hautbergue 2017; Stewart 2019; Garland and Jensen 2020). Nevertheless, some intron containing transcripts are exported to the cytoplasm as alternative mRNA isoforms that either encode an alternative protein or are subject to altered translation and decay (Jacob and Smith 2017; Wegener and Müller-McNicoll 2018). Other introns slow to be excised relative to transcription are ultimately removed and their transcripts exported as fully spliced mRNAs (Ninomiya et al. 2011; Bhatt et al. 2012; Hao and Baltimore 2013; Pandya-Jones et al. 2013; Frankiw et al. 2019a). Such transcripts can create a nuclear

pool of partially spliced RNA, which acts as a reservoir to feed the cytoplasmic mRNA pool upon splicing. A group of these introns found in genes affecting growth control and cell division were named “detained introns” to distinguish them from classical “retained introns” found in cytoplasmic mRNA (Boutz et al. 2015; Braun et al. 2017). A similar pool of incompletely spliced transcripts affecting synaptic function is found in neurons, where cell stimulation induces their processing to allow transcription independent changes in mRNA pools (Mauger et al. 2016). The term “retained intron” thus encompasses a wide range of molecular behaviors.

Retained introns are more difficult to characterize than other patterns of alternative splicing in whole transcriptome RNA-seq data. Overlapping patterns of alternative processing can be mis-called as intron retention by sequence analysis tools (Wang and Rio 2018; Broseus and Ritchie 2020). Many RNA-seq studies have identified conditions leading to higher levels of unspliced introns across the transcriptome (Wong et al. 2013; Braunschweig et al. 2014; Pimentel et al. 2016; Edwards et al. 2016; Schmitz et al. 2017; Jacob and Smith 2017; Naro et al. 2017; Parra et al. 2018). These studies have not always distinguished between nuclear and cytoplasmic RNA or examined the fate of the partially spliced transcripts, information that is essential to understanding the biological role of these regulatory mechanisms.

Here we undertook a broad examination of how RNAs are distributed between subcellular compartments and how this compartmentalization changes with development. Our goals were to distinguish transcripts in the nucleoplasmic and chromatin-associated RNA pools from cytoplasmic mRNAs and assess how their processing and localization to chromatin tracked with expression of mature cytoplasmic mRNA.

## RESULTS

### **Both coding and non-coding RNAs exhibit defined partitioning between cellular compartments.**

To broadly categorize RNAs enriched in different cellular locations, and gain insight into how this compartmentalization might be regulated across cell types, we generated deep RNA-seq data from mouse embryonic stem cells (mESC), a neuronal progenitor cell line derived from embryonic mouse brain (mNPC), and explanted mouse cortical neurons cultured *in vitro* for 5 days (E15DIV5; mCtx) (Fig. 1A). RNA was isolated from three fractions of each cell: cytoplasm, soluble nucleoplasm, and chromatin pellet as previously described (Wuarin and Schibler 1994; Pandya-Jones and Black 2009; Bhatt et al. 2012; Yeom and Damianov 2017). The quality of subcellular fractionation was assessed by immunoblot for GAPDH and Tubulin alpha-1A (TUBA1A) proteins as cytoplasmic markers, SNRNP70 fractionating with the soluble nucleoplasm, and Histone H3.1 as a chromatin marker (Supplemental Fig. S1A; Supplemental Table S1B).

To provide information on the maturation of transcripts in each cell type and location, RNA was isolated as two separate pools. A total RNA pool depleted of ribosomal RNA (Total) will include nascent incomplete transcripts. A polyadenylated pool (Poly(A)+) includes RNAs whose transcription and 3' processing is complete. Each RNA pool from each fraction was isolated from three separate cultures of each cell type to yield biological triplicates of each experimental condition. The RNA pools were converted to cDNA libraries, sequenced on the Illumina platform to yield 100 nt paired end reads, and aligned to the genome (Supplemental Table S2). Gene expression markers for each of the three cell types confirmed the expected patterns of ESC, NPC or immature neurons (E15DIV5; Supplemental Fig. S1B). Clustering of gene expression values across all the datasets showed the expected segregation by cell type, fraction, and replicate, for both the poly(A)+ and total RNA libraries (Supplemental Fig. S1C). The

resulting 54 datasets constitute an extensive resource for examining multiple aspects of RNA maturation and its modulation during development [GSE159919 for poly(A)+ RNA and GSE159944 for total RNA]. In addition to the libraries used in this study, we also generated libraries of small RNAs (< 200 nt) from all samples. As previously described, these can be used to assess miRNA maturation and other processes (Yeom et al. 2018). These 27 datasets are also available from GEO [GSE159971].

Examining read distributions in the different RNA pools and fractions, we found that the housekeeping gene *Gapdh* (Fig. 1B) yields similar patterns of reads from either the poly(A)+ or the total RNA populations, with the RNA being most abundant in the cytoplasm. The total *Gapdh* RNA on chromatin contains intron reads from the nascent transcripts (Fig. 1B bottom). Although more abundant in the soluble nucleoplasm and especially in the cytoplasm, polyadenylated *Gapdh* transcripts are also found in the chromatin fraction, but in contrast to the total RNA lack intron reads. We also examined the long non-coding RNA *Xist*, which condenses on the inactive X Chromosome in female cells (Fig. 1C). The mNPCs were isolated from female mice, and *Xist* is seen to partition almost completely to chromatin in these cells. The poly(A)+ and the total RNA samples yielded very similar patterns of *Xist* reads indicating that this RNA is largely spliced and polyadenylated (Brockdorff et al. 1992). Other non-coding RNAs yielded more complex patterns of subcellular partitioning that changed with cell type. The paraspeckle lncRNA *Neat1* is more highly expressed in mESC than mNPC or neurons (Supplemental Fig. S2A). The short polyadenylated form (*Neat1\_1*) predominates in ESC and is found mostly with chromatin but also in the nucleoplasm. The longer non-polyadenylated *Neat1* RNA (*Neat1\_2*) is seen in the total RNA samples and is also chromatin-enriched. Whether this is a stable long isoform or nascent RNA is not clear. This longer RNA contributes a larger portion of the *Neat1* transcripts in mNPC and neurons, consistent with observations that *Neat1* cleavage and polyadenylation may be

modulated (Naganuma et al. 2012). Overall, we find that gene transcripts can exhibit diverse patterns of enrichment and processing across the different fractions and cell types.

Because the relative transcript numbers and overall library complexity will differ between fractions, RPM (Reads Per Million) values or other read number normalizations of individual genes cannot be directly compared between different subcellular fractions. Using qRT-PCR in mESC to directly quantify individual transcripts in different fractions, we found that for cytoplasmic enriched transcripts in both the poly(A)+ and the total RNA libraries, RPM values undercounted the RNA abundance in the cytoplasmic fraction relative to the chromatin and nucleoplasm (Supplemental Table S3). On the other hand, for RNAs that are primarily chromatin associated, qRT-PCR quantification yielded cytoplasmic to chromatin ratios that were similar to relative RPM numbers (Supplemental Table S3). Although the absolute transcript levels were not quantifiable by RPM, the ratios of these RPM values did reflect their relative enrichment in each fraction across a variety of genes. As an index for how RNAs partition between the chromatin and cytoplasmic pools, we used DESeq2 (Anders and Huber 2010) to measure the fold change in reads for each gene between the chromatin and cytoplasmic poly(A)+ RNA. This returns the ratio of the averaged read counts for each gene between fractions. For genes whose TPM (Transcripts Per Million) value in chromatin was over the median and which had read counts greater than 0 in the cytoplasm (13,036 genes), this chromatin partition index was distributed over a 100 fold range centered on 1 ( $\text{Log}_2=0$ ). Thus, a typical gene showed equal normalized read counts in chromatin and cytoplasm (Fig. 1D). Examining the Ensembl annotations (V.91) for genes in the left, middle, and right side of this distribution (400 genes each), we found that genes with predominately cytoplasmic reads as well as genes with roughly equal read numbers in cytoplasm and chromatin were annotated almost entirely as protein-coding genes. For

example, on the left edge (Fig. 1D), *Gapdh* RNAs partition much more strongly to the cytoplasm than is typical. In the middle of the distribution, *Rbfox2* RNAs exhibit slightly fewer reads on chromatin than in the cytoplasm, whereas *Cdk8* exhibits 2 to 3 fold more chromatin reads (Fig. 1D). Thus, although the transcripts from protein coding genes are usually most abundant in the cytoplasm, a substantial fraction of a gene's RNA product is often nuclear and chromatin-associated. Comparing qRT-PCR quantification for select genes to their chromatin partition indices, we found that RNAs from genes exhibiting a partition index above 3.6 were actually more abundant in chromatin than the cytoplasm. This included about 3 % of protein coding genes. At the right edge of the curve, the 400 most chromatin enriched transcripts included the expected non-coding RNAs, such as pri-miRNAs, snoRNAs, and lncRNAs, but also many protein-coding genes, including *Clcn2*, *Ankrd16*, and *Gpc2* (Supplemental Figs. S2C, S2D, S6B), and *Gabbr1*, which is analyzed further below. For these protein-coding genes, the majority of the polyadenylated product RNA is chromatin associated where it is presumably inactive for protein expression (Supplemental Figs. S2C, S2D, S6B).

Examination of individual genes whose poly(A)<sup>+</sup> transcripts remain sequestered with chromatin showed that their splicing was modulated across cell types. The chromatin-associated *Meg3* non-coding RNA is well expressed in mESC and neurons but not in mNPC (Supplemental Fig. S2B). *Meg3* is the host transcript for the miRNAs MiR-770 and MiR-1906-1. Mature MiR-770, processed from the last *Meg3* intron, is weakly expressed in neurons but absent from mESC (Supplemental Table S4). This intron is absent from the RNA in mESC where it is apparently efficiently spliced. By contrast in neurons, this intron is abundant in the chromatin fraction of polyadenylated RNA, where its reduced excision might allow more efficient processing of *MiR-770* (Supplemental Fig. S2B). This is consistent with observations that perturbations causing a host transcript to be released from chromatin reduce DROSHA cleavage and miRNA

expression (Pawlicki and Steitz 2008; Liu et al. 2016a). The mESC small RNA data was previously used to examine expression of primary *MiR-124a-1* in mESC whose processing is blocked by PTBP1 in the chromatin fraction (Yeom et al. 2018). For *Meg3*, the processing of *MiR-770* may be modulated by the excision rate of its host intron. The upstream portion of *Meg3* that includes *MiR-1906-1* undergoes complex processing and exhibits more splicing in neurons than in mESC. Thus, an additional product from the gene, possibly *MiR-1906-1*, may also be differentially regulated between mESC and neurons. These introns present in the polyadenylated RNA are not more abundant in the total RNA than adjacent exon sequences, indicating an absence of excised intron, which could also give rise to the miRNAs. Overall the data indicate that splicing of the *Meg3* transcript is regulated on chromatin to allow differential expression of its mature products.

### **Chromatin associated transcripts can be spliced either cotranscriptionally or posttranscriptionally.**

It is expected that most introns will be transient species within the chromatin RNA, with many introns excised prior to transcript completion, while some introns with slow kinetics will be removed later. Various studies estimate that 45 to 84 % of introns are cotranscriptionally excised in mammals (Ameur et al. 2011; Bhatt et al. 2012; Girard et al. 2012; Khodor et al. 2012; Tilgner et al. 2012; Windhager et al. 2012). Several approaches compare read numbers for spliced (exon-exon) and unspliced (exon-intron or intron-exon) junctions in nascent RNA to those in total RNA to measure cotranscriptional excision (Tilgner et al. 2012; Windhager et al. 2012; Herzog and Neugebauer 2015). To ensure that measurements are of the nascent RNA, this requires removal of polyadenylated RNA from the chromatin fraction and prevents parallel analysis of posttranscriptional events. Other studies identified sawtooth patterns of RNA read abundance in total cellular RNA, where reads peak in exons and then decline to the

next exon or recursive splice site. Such a pattern is thought to indicate that the time needed to excise an intron is small relative to the time for RNA synthesis through the next intron downstream (Ameur et al. 2011; Duff et al. 2015; Sibley et al. 2015). While sawtooth read densities can be observed on certain introns in the total chromatin RNA pools (Supplemental Fig. S3), these patterns were infrequent and lost on introns shorter than 50kb, many of which are expected to be cotranscriptionally excised (Ameur et al. 2011).

As an alternative for defining cotranscriptional and posttranscriptional intron excision, we compared the total RNA from chromatin to the poly(A)+ RNA from the same fraction. Introns remaining in polyadenylated RNA must be excised after transcription or be dead-end products. For example, in the *Sorbs1* gene (Fig. 2A) reads are observed across all the introns in the total RNA from chromatin indicating the presence of unspliced introns in the nascent transcripts. In the polyadenylated RNA on chromatin, reads are largely absent from introns indicating that by the time of polyadenylation or shortly after, these introns have been spliced out. However, one intron in *Sorbs1* exhibits substantial read numbers in poly(A)+ RNA on chromatin that are reduced in RNA from the nucleoplasm and absent from the cytoplasm (Fig. 2A). This intron is presumably excised after cleavage/polyadenylation. While most introns are absent from the polyadenylated RNA and likely spliced cotranscriptionally, there are many transcripts with one or more introns that are highly retained in the polyadenylated chromatin associated RNA (Figs. 2A, 2B). The comparison of intron levels in total and poly(A)+ RNA on chromatin provides a simple bioinformatic metric for distinguishing co- versus posttranscriptional excision.

To compare intron levels in the total and poly(A)+ RNA pools, we determined fractional inclusion values (FI; Supplemental Fig. S4A) by counting reads across exon-intron, intron-exon, and exon-exon junctions. Assessing intron retention (IR) by FI value

can be confounded by alternative splicing, polyadenylation or transcription initiation events occurring within the intron being measured (Supplemental Fig. S4B) (Wang and Rio 2018; Broseus and Ritchie 2020). To avoid errors in IR measurements arising from other processes, we defined a set of introns exhibiting a unique Ensembl v91 annotation without alternative processing events (Supplemental Fig. S4B). This set of 149,333 “unique” introns (U introns) across 28,733 genes was used for subsequent analysis. Focusing on the mESC RNA, we determined the FI values of all U introns in the total RNA and the poly(A)+ RNA for genes above the median expression level as measured by kallisto (Bray et al. 2016). We included only introns excised by the major spliceosome with GU/AG splice junctions. Reads from poly(A)+ RNA containing long unspliced introns can be biased toward the 3' ends. To avoid undercounting in the poly(A)+ samples, we removed genes where reads per nucleotide length from the second exon were less than half that of the second to last exon. To filter out introns that were not measurable due to anomalies in the generation of particular junction reads, we removed introns yielding a FI value below 0.1 in the total RNA, and introns with a zero value for one or more of the junction read counts. In mESC, these criteria returned 49,629 U introns within 7,672 genes for analysis.

Of the 49,629 U introns being measured, 34,939 introns (within 6,952 genes) exhibited low FI values in the poly(A)+ RNA ( $FI < 0.1$ ) and are presumably spliced before transcript completion. Conversely, 14,753 introns within 5,550 genes exhibited a FI value greater than or equal to 0.1 in the poly(A)+ RNA. These introns (29.7 %) appear to be excised posttranscriptionally, with many highly unspliced in the chromatin poly(A)+ RNA despite being fully spliced in other fractions. By this analysis, at least 70.3 % of introns within our analysis set are excised cotranscriptionally, similar to estimates made by other methods (Fig. 2C; Supplemental Figs. S4C – S4F; Supplemental Table S5). On the other hand, the majority of genes (5,550 out of 7,672) have at least one

posttranscriptionally spliced intron. Restricting the analysis to the top quartile of expressed genes rather than the top half, the fractions of co- and posttranscriptional splicing change only slightly (70.7 % cotranscriptional). The fraction of cotranscriptionally spliced introns is also essentially the same if the analysis is restricted to the first introns in each transcript or to internal introns. For introns that are the last intron transcribed before the polyadenylation site, a slightly higher fraction is classified as posttranscriptional, presumably because they are polyadenylated more rapidly after intron synthesis (Supplemental Fig. S4F). Thus, posttranscriptional splicing does not appear to be associated with higher or lower gene expression, or with the position of an intron along the gene. Examples of introns defined as co or posttranscriptional by these measures are shown in Fig. 2B. Although in the minority, posttranscriptionally spliced introns are found across a wide range of genes, and often exhibit high FI values in the chromatin fraction, even though the cytoplasmic RNA is completely spliced.

In addition to the U introns analyzed above, we also analyzed a set of introns flanking simple cassette exons that could also be unambiguously measured for FI. Using the same parameters to define co- versus posttranscriptional splicing, we found a reversal in the percentages. Of these introns flanking alternative exons, approximately 67 % exhibit high read numbers ( $FI > 0.1$ ) in the poly(A)<sup>+</sup> RNA and thus appear to be excised posttranscriptionally (Fig. 2C; Supplemental Fig. S4E). This was seen for introns both upstream and downstream of the cassette exon. These data indicate that the majority of regulated splicing events occur with slower kinetics than the excision of typical constitutive introns.

**Retained introns can be classified by their enrichment in the chromatin, nucleoplasmic, and cytoplasmic compartments.**

A variety of fates are possible for transcripts that retain introns after polyadenylation. Intron containing transcripts can be sequestered in the nucleus until they are spliced or can undergo nuclear decay. Other intron containing mRNAs are exported unspliced to the cytoplasm where they can be translated or undergo Nonsense-mediated mRNA decay (NMD). To categorize introns based on both their retention levels and location, FI values for the unique intron set in the polyadenylated RNA of all cells and fractions were subjected to X-means cluster analysis (Fig. 3A; Supplemental Table S6) (Pelleg and Moore 2000). Consistently, in all three cell types the clustering algorithm defined four groups of introns. The largest cluster Group A, containing 49,981 introns in mESC, was almost entirely spliced in all three fractions. Introns in Group B (7,529), exhibited measurable retention in the poly(A)+ RNA from chromatin, but showed nearly complete splicing in the nucleoplasm and cytoplasm (Fig. 3A). Group C introns (1,351), including introns in *Zfp598* and *Neil3* (Fig. 3B), showed higher FI values in the chromatin and nucleoplasm than Group B, but were almost completely excised from the cytoplasmic RNA. The smallest cluster of only 247 introns in mESC, Group D, was almost entirely retained in all three fractions. Each of the other two cell types also generated four clusters with similar splicing levels and similar numbers of introns in each group (Fig. 3A).

Group B and C introns that do not leave the nucleus can be seen to have different properties from Group D introns that also have high retention levels in the cytoplasm. A larger percentage of Group D introns are found in 5' and 3' UTR sequences, where they will not disrupt the primary reading frame, but will likely affect translation and decay (Supplemental Table S7B). Group D introns were also found to be depleted of in-frame premature termination codons (PTC) compared to Groups A, B and C (Fig. 3C), presumably due to selection to prevent NMD in the cytoplasm. These

observations indicate that the different intron clusters arise from selection for different functions in the intron containing RNAs.

We found that among transcripts where all introns could be assigned a group (Supplemental Fig. S4G), RNAs containing at least one Group C intron have a higher average chromatin partition index than transcripts with no Group C intron (Supplemental Fig. S4H). Previous work defined nuclear transcripts in mESC containing what are called detained introns (DI), whose splicing is modulated in cancer and growth control pathways (Boutz et al. 2015; Braun et al. 2017). Of 3,150 detained introns, 1,021 were on our U intron list (Supplemental Tables S7A, S7B). Of these, 1,000 introns passed the filters for FI measurement and are seen to fall predominantly into Groups B and C, in agreement with the earlier studies (Fig. 3D). However, the 1,021 detained introns were only a subset of the nearly 9,000 retained introns identified in Groups B and C (Supplemental Table S8B). Similar to the detained introns affecting growth control, as well as inflammatory and neuronal gene introns also identified previously (Bhatt et al. 2012; Hao and Baltimore 2013; Pandya-Jones et al. 2013; Mauger et al. 2016; Frankiw et al. 2019b), these new retained introns could affect cellular function by altering the movement of material through the gene expression pathway.

### **Predicting retained introns**

To examine whether introns in different groups could be identified by their sequence features alone, we developed a deep learning model for predicting intron behavior. We extracted 1,387 sequence features from the first and last 300 nucleotides of each intron and from the two flanking exons. For introns less than 300 nucleotides, the intron interval includes some adjacent exon sequence. Analyzed features included short motif frequencies, predicted RBP binding elements, propensity to form local secondary structure, splice site strength scores, conservation scores, and nucleosome

positioning scores (Supplemental Table S9A). This feature information was used to train a three-layer deep neural network tasked with predicting whether an intron belonged in Group A, B, C, or D (Fig. 4A).

The performance of the model was assessed using Receiver Operating Characteristic (ROC) curves plotting the false and true positive rates (Fig. 4B). The model was highly predictive in distinguishing Group D introns from A, yielding an Area Under the Curve of 0.94 (AUC = probability that any true positive will rank higher than any true negative). Group D introns could also be distinguished from Group B and C (AUC=0.9 and 0.84, respectively), and Group B and C introns from Group A with reduced accuracy (AUC = 0.68 and 0.76, respectively). Thus, the Group D introns are most different from the introns of other groups.

To assess the features of Group C and D introns that distinguish them from each other and from Group A, we isolated the top 15 features predictive of intron retention or its absence and used a t-distributed stochastic neighbor embedding algorithm (t-SNE) to project them onto two dimensions (Fig. 4C; Supplemental Table S9B for top 50 features). As previously observed, high splice site strength scores were predictive of Groups A and C over D, and also Group A over C (Sakabe and de Souza 2007; Braunschweig et al. 2014). Other features redundant with splice site strength scores were also predictive of Groups A or C, including GTAAG count in the 5' portion of the intron and the conservation of the splice site sequences. Translatability of the flanking exons and their spliced product was predictive of Groups A and C over D. This may reflect a greater percentage of Group D introns in 5' and 3' UTR sequences (Supplemental Table S7B). Conversely, the translatability of the exon-intron-exon unit containing the retained intron was predictive of Group D over Group C, in agreement with the Group D introns being depleted of in frame termination codons (Fig. 3C) and adding a coding segment to the

mRNA. Overall, the data indicate that intron retention is controlled by many factors each having relatively small effect.

We examined whether particular sequence elements correlated with the intron group assignments, indicative of regulatory protein binding sites. The model did not clearly identify known elements affecting nuclear localization or intron retention such as constitutive transport elements or decoy exons (Li et al. 2006; Parra et al. 2018). However, the sequence conservation score of the 5' portion of the intron was predictive of Group D over Groups C or A, and conservation of both ends of the intron was predictive of C over A (Supplemental Table S9B). Particular triplet motif frequencies within introns or their flanking exons were also predictive of intron behavior. For example, CGA triplets in the 3' portion of the intron were predictive of Group D over C, whereas TTG and GTT triplets in the 5' intron segment were predictive of Group C over D. The predictive power of intron sequence conservation and of multiple triplets indicate that particular RNA/protein interactions likely determine the retention properties of these groups.

### **Intron retention and chromatin association are regulated with neuronal development.**

Since the X-means analysis yielded four intron clusters in each cell type, these cluster definitions allow bioinformatic analysis of IR regulation between cell types. While many introns maintain their classification between cell types (Fig. 5A, left), some introns switched their group (Fig. 5A, right). One example is *Med22* (Fig. 5B), which contains a highly retained intron 3 (I3) in all three fractions of mESC (Group D). This intron became more spliced in mNPC and was classified as Group C, and then became almost fully spliced as a Group B intron in neurons. The nearby intron 1 (I1) was maintained as a Group A intron in all three cell types. *Med22* encodes a subunit of the transcriptional

mediator complex. The retention or splicing of *Med22* intron 3 creates MED22 proteins with different C-terminal peptides that likely alters mediator function in the two cell types. The group switching introns are presumably part of the extensive alternative splicing programs modulated during neuronal development. Examining their Gene Ontology (GO) functions, we found that the 231 genes containing introns highly spliced in mESC but unspliced in neurons (switching from Group A or B to Group C or D) were enriched in processes such as ribosome biogenesis, organelle assembly, and metabolism. These functional categories may reflect the different proliferation rates and metabolic status of the two cells. In contrast, 413 genes whose introns were unspliced in mESC and became more spliced in neurons (switching from Group C or D to Group A or B) were enriched in GO biological processes of glutamatergic synaptic transmission and organelle localization by membrane tethering, in keeping with gene expression and cell morphology changes in the early neuronal state (Supplemental Fig. S5).

The changes in splicing between mESC, mNPC and neurons are driven by changes in the expression of multiple protein regulators. In previous work, we and others characterized alternative splicing programs controlled by the polypyrimidine tract binding proteins, PTBP1 and PTBP2 (Keppetipola et al. 2012; Vuong et al. 2016). In ESC and other cells, PTBP1 maintains alternative splicing patterns characteristic of non-neuronal cells, and PTBP1 downregulation is a key step in neuronal differentiation. While the cultured NPC's are not true lineage precursors to the immature cortical neurons used here, the depletion of PTBP1 is common to many neuronal lineages. We previously reported neuronal cassette exons regulated by PTBP1 in ESC (Linares et al. 2015), and PTBP1 regulated retained introns, including the *Med22* intron, have been described in a neuronal cell line (Yap et al. 2012). We next examined whether additional PTBP1 targets could be identified in the chromatin compartment of mESC.

To assess PTBP1 regulation, we fractionated cells after *Ptbp1* knockdown and measured the splicing of polyadenylated RNA in the different compartments by RNA-seq. This confirmed the PTBP1 dependence of *Med22* intron 3, which shifted from Group D to Group C with *Ptbp1* depletion (Fig. 5B, right). Examining all the retained introns, we found that many more splicing changes could be observed in the chromatin associated RNA than in the nucleoplasmic and cytoplasmic fractions (Fig. 5C). As shown previously with cassette exons, these PTBP1 dependent introns in ESC also change with neuronal differentiation as PTBP1 levels drop (Fig. 5C). These include introns identified previously (Yap et al. 2012) as well as new introns. Other introns whose splicing changes with neuronal development but are not sensitive to PTBP1 are presumably regulated by other factors.

By examining the chromatin associated RNA, our analysis identified substantially more PTBP1-regulated introns than previously recognized. The transcripts containing these introns may remain in the nucleus, similar to detained introns, or may be exported to the cytoplasm and then lost to NMD. To assess this, we used data from a study of unfractionated polyadenylated RNA after *Upf1* knockdown that globally identified NMD targets in mESC (Hurt et al. 2013). A majority of Group A, B and C introns are predicted to induce NMD, if their parent transcripts were exported to the cytoplasm (Fig. 3C). However, we find that of 871 genes containing PTBP1 dependent retained introns in the chromatin fraction, only 87 exhibited greater than 10 % transcript upregulation after *Upf1* depletion (Supplemental Table S8C). Thus, the majority of the PTBP1 dependent retained intron transcripts likely stay in the nucleus and will be eliminated by nuclear RNA decay pathways.

Looking more broadly at whether NMD might create the apparent nuclear enrichment of some transcripts, we found that protein-coding genes with high chromatin partition indices were actually less likely to show increases after *Upf1* depletion than

other genes across the distribution (Supplemental Table S8D). For the genes in the L, M, and R regions in Fig. 1D, NMD targets constituted 4.2, 7.2, and 1.1 % respectively. Rather than NMD causing the observed nuclear enrichment by depleting the cytoplasmic RNA, the nuclear enrichment may buffer the effect of NMD on the level of total RNA. It would be interesting to assess this by examining the effect of *Upf1* knockdown specifically on the levels of cytoplasmic mRNA.

### **Posttranscriptional repression of *Gabbr1* expression.**

For the most part, transcripts enriched in the chromatin fraction of mESC's (Fig. 1D) were only mildly or unaffected by *Upf1* depletion (Supplemental Table S8D). Rather than cytoplasmic degradation, other processes prevent mRNA expression from these genes. A notable example is *Gabbr1*, which encodes GABBR1, an inhibitory neurotransmitter receptor whose cytoplasmic mRNAs are highly expressed in neurons, moderately expressed in mNPC, but nearly absent in mESC (Fig. 6A). By immunoblot, GABBR1 protein is only observed in neurons (Fig. 6C). In the chromatin fraction of mESC the *Gabbr1* precursor RNA is present at high levels that nearly match those seen in mNPC and neurons (Fig. 6A). This *Gabbr1* RNA is polyadenylated and most introns are excised, but introns 4 and 5 that exhibit a complex pattern of alternative processing in neurons, are largely unprocessed in mESC (Fig. 6A). *Gabbr1* mRNA expression is apparently blocked by a combined process of splicing inhibition and sequestration on chromatin. Upon differentiation into neurons, the chromatin partition index of *Gabbr1* RNA shifts from 4.43 to -0.69, as the RNA becomes fully processed and released from chromatin to appear in the cytoplasm as mature mRNA (Fig. 6A). Other protein coding transcripts, including *Gpc2*, were found to behave similarly to *Gabbr1* with RNA abundant in mESC chromatin but low in cytoplasm. In neurons, this pattern was

reversed with the *Gpc2* partition index shifting from 4.60 in mESC to 1.09 in neurons (Supplemental Fig. S6B).

PTBP1 was previously found to regulate *Gabbr1* exon 15 in a neuronal cell line (Makeyev et al. 2007). To assess introns 4 and 5, we examined iCLIP maps of PTBP1 binding in mESC (Linares et al. 2015), which showed prominent PTBP1 binding peaks in the intron 4-5 region, as well as confirming PTBP1 binding upstream of exon 15 and to the 3' UTR (Figs. 6A, 6B). Examining the fractionated RNA-seq data, we found that *Ptbp1* knockdown led to processing of the *Gabbr1* RNA into the neuronal isoforms, including activation of exon 15 and activation of the exon 5 microexon encoding a 6 amino acid linker of *Gabbr1a* (Figs. 6A, 6B). Some processed *Gabbr1* mRNA was present in the cytoplasm after *Ptbp1* knockdown, but more of this spliced RNA was in the soluble nuclear fraction. Even after *Ptbp1* depletion, a majority of the *Gabbr1* RNA was still in the chromatin fraction and still unprocessed in the intron 4-5 region, despite exon 15 being strongly activated for splicing in this fraction (Supplemental Fig. S6C). GABBR1 protein was also not observed in mESC after *Ptbp1* knockdown (Supplemental Fig. S6D). Thus, although PTBP1 strongly affected the processing of *Gabbr1*, its depletion did not yield the predominantly cytoplasmic RNA seen in neurons. There must be additional factors preventing release of the RNA from chromatin in mESC. *Gabbr1* is highly transcribed in mESC, but its mRNA expression is blocked by a combination of splicing repression, NMD of transcripts that enter the cytoplasm, and sequestration of the unprocessed RNA on chromatin, with the latter mechanism having the largest effect.

## DISCUSSION

### **A resource for the analysis of RNA-level gene regulation.**

We developed extensive datasets to examine RNA maturation events across cellular location and developmental state. Applying these data to analyze intron retention, we compare total and polyadenylated RNA across subcellular fractions and cell types to define classes of introns exhibiting different regulatory behaviors, and we uncover a novel form of gene regulation acting on chromatin associated RNA. We find that a substantial fraction of the polyadenylated RNA product of some genes is incompletely spliced and still associated with chromatin. This points to a limitation for whole transcriptome measurements of gene expression that assess total cellular polyadenylated RNA; The RNA being measured in these studies is not all cytoplasmic mRNA. The presence of nuclear polyadenylated RNA may thus contribute to the observed lack of correlation between RNA and protein levels in global gene expression measurements (Edfors et al. 2016; Liu et al. 2016b).

The isolation of chromatin associated RNA has frequently been used to enrich for nascent pre-mRNAs and other short lived species (Pandya-Jones and Black 2009; Davidson et al. 2012; Herzel et al. 2017). We find that many introns are only observed in the total RNA of this fraction, while others are also present in the polyadenylated RNA. Quantifying this difference, we estimate that 70 % of introns within our analysis set are spliced before the RNA has been completely transcribed. Although this roughly agrees with other studies, we believe it is a lower-bound estimate in our system because the criteria for counting cotranscriptionally excised introns required a measurable presence of the intron in the total RNA. In contrast, we find that introns flanking alternatively spliced cassette exons are mostly spliced posttranscriptionally - exhibiting significant intron retention levels in the polyadenylated RNA. These introns may be spliced more slowly than typical constitutive introns because of the complex regulatory RNP structures

that must assemble onto the sequences flanking alternative exons. By creating a pool of unspliced RNA for these genes, the delayed splicing may allow additional controls over the isoform choice. It will be interesting to examine whether the subset of exons whose inclusion is affected by transcription elongation rates and perturbations of RNA Pol II are among the 30 % that appear to be cotranscriptionally excised (Naftelberg et al. 2015; Saldi et al. 2016; Herzel and Neugebauer 2015).

Our data provide a rich resource for examining other questions of RNA metabolism and its regulation over development. Besides introns, transient species one could observe in chromatin associated RNA include upstream antisense RNAs and extended transcripts downstream from polyadenylation sites (Seila et al. 2008; Flynn et al. 2011; Vilborg and Steitz 2017). These data could also allow more sensitive detection of recursive or back-splicing, and inform studies of regulated RNA export. We have also examined regulated miRNA processing using parallel data from short RNA libraries (GSE159971; Supplemental Table S4) (Yeom et al. 2018).

### **Behaviors of retained introns.**

To characterize incompletely spliced transcripts, we assessed introns based on their retention levels across fractions and cell types. Unsupervised X-means clustering yielded four intron groups in each cell type. The largest cluster (Group A) were completely spliced in the poly(A)<sup>+</sup> RNA, including in the chromatin fraction, and are presumably excised prior to transcription termination. The smallest cluster (Group D) behaved like classical retained introns in being exported to the cytoplasm within the otherwise fully spliced mRNA. Two intermediate clusters of introns (Groups B and C) were fully spliced in the cytoplasm while exhibiting different levels of retention on chromatin and to some extent the nucleoplasm. A deep neural network trained using a well-defined set of introns and a wide range of genomic features was able to distinguish

introns in Group D from those in A or C with high accuracy. Group C introns were also distinguished from Group A with moderate accuracy (Fig. 4B). These data indicate that Groups D and C are functionally distinct and the features which define them should give clues to their regulation. These features include those previously associated with retained introns, such as weak splice sites, conservation, and coding capacity (Sakabe and de Souza 2007; Jaillon et al. 2008; Braunschweig et al. 2014; Dvinge and Bradley 2015; Mauger et al. 2016; Parra et al. 2018). We found that introns of the different groups were defined by enrichment of particular short sequence motifs in their terminal regions and adjacent exons. We have not yet identified proteins whose binding sites might underlie the enrichment of these motifs. This may be because the recognition elements assigned to individual proteins are not sufficiently specific. Introns also may be regulated by so many different proteins that no single binding motif is strongly predictive. Proteins including PTBP1 and others are known to regulate particular retained introns (Yap et al. 2012; Horan et al. 2015; Pendleton et al. 2017; Frankiw et al. 2019b), but there may be many such factors each regulating a subset of introns in a group. The extension of our approach to larger datasets will allow correlation of changes in intron group assignment with the expression of particular RNA-binding proteins.

Groups B and C include several previously described sets of interesting retained introns. Detained introns were defined as partially spliced introns in transcripts affecting growth control, whose excision can be modulated by cellular stimuli (Ninomiya et al. 2011; Boutz et al. 2015; Braun et al. 2017). These detained introns are a subset of the Group B and particularly Group C introns we defined in mESC. Another group of retained introns were shown to be regulated by PTB proteins in a neuronal cell line (Yap et al. 2012). Our analytical strategy identified many new PTBP1 dependent introns that remain as chromatin associated transcripts in mESC. In the total cellular polyadenylated RNA of mature primary neuronal cultures (Mauger et al., 2016), retained introns were

characterized as transient or stable according to their splicing after transcription inhibition. In our data from less mature neurons, we found that the largest portion of transient introns were in Group C (40 %). In contrast, of the stable introns that we could assay in our cultures, about 40 % were in Group D (Supplemental Table S8E), consistent with the stable introns remaining in cytoplasmic mRNA after transcriptional shutoff. Similar to detained introns, Mauger et al. found that synaptic activation could change the splicing level of some retained introns. It will be interesting to examine whether these introns are associated with chromatin, but this will require improved isolation of nuclei from mature neuronal cultures.

#### **Developmental regulation by splicing inhibition and chromatin sequestration.**

In previous studies, we showed how the neuronal specific expression of certain genes is determined by the coupling of a PTBP1 dependent splicing event to NMD. RNAs for the neuronal PTBP2 and PSD-95 proteins are expressed in ESC and other non-neuronal cells, but through the action of PTBP1 are spliced as isoforms that are subject to NMD (Boutz et al. 2007; Makeyev et al. 2007; Spellman et al. 2007; Zheng et al. 2012; Linares et al. 2015). A similar mechanism affects *Gabbr1* through regulation of exon 15 by PTBP1 (Makeyev et al. 2007), but the change in RNA with loss of NMD is small (Hurt et al. 2013). Most protein-coding transcripts exhibiting chromatin enrichment were not seen to be upregulated by *Upf1* depletion, while some were modestly affected similar to *Gabbr1*. The nuclear pools of these RNAs may reduce the observed efficiency of NMD on total RNA levels, where transcripts exhibit only partial depletion by the decay pathway even though near complete loss of protein is observed. Here we uncover another mechanism controlling the developmental specific expression of a neuronal protein. The *Gabbr1* RNA is abundant in mESC but its splicing is incomplete and its transcript remains in the chromatin compartment.

*Gabbr1* is expressed as multiple isoforms (Kaupmann et al. 1997). The long *Gabbr1a* isoform comes from a promoter active in all three cell types studied here. *Gabbr1b*, which lacks N-terminal sushi domains, arises from an alternative promoter within intron 5 active in neurons (Vigot et al. 2006). There is also a short transcript derived from an alternative polyadenylation site in intron 4. A microexon 5 between these two introns adds a linker into the 1a isoform (Vigot et al. 2006). This complex intron 4-5 region is largely unprocessed in mESC cells and becomes processed in neurons with the production of cytoplasmic mRNA including exon 5. The depletion of *Ptbp1* from mESC leads to multiple changes in *Gabbr1* splicing including activation of microexon 5 and downstream exon 15. This leads to some expression of neuronal mRNA isoforms but very limited protein expression. Much of the RNA remains nuclear indicating that additional factors prevent its mobilization. Instead of regulation at the level of transcription or mRNA stability, incomplete *Gabbr1* splicing and sequestration of its RNA on chromatin are modulated to control gene output over development.

The *Gabbr1* transcript is extensively bound by PTBP1. Studies have shown that when binding RNA at high stoichiometry, PTBP1 can cause the condensation of RNA/protein liquid droplets *in vitro* (Lin et al. 2015). Extensive PTBP1 binding to the long non-coding RNA *Xist* is required for *Xist* condensation onto the X Chromosome during X inactivation (Pandya-Jones et al. 2020). PTBP1 also drives the condensation of the long non-coding RNA *PNCTR* in the perinucleolar compartment, and a similar mechanism may be involved in its interaction with LINE RNAs (Attig et al. 2018; Yap et al. 2018). It will be interesting to examine whether PTBP1 might create a nuclear condensate of *Gabbr1* RNA. Although *Ptbp1* knockdown led to increased splicing and increased mRNA in the nucleoplasm and cytoplasm, it did not eliminate the enrichment of the unspliced RNA in the chromatin. This may be due to the partial depletion of *Ptbp1* by RNAi, but it seems likely that other proteins will also contribute to the sequestration of *Gabbr1* RNA,

as is seen with *Xist*. If the chromatin enrichment of protein-coding transcripts like *Gabbr1* involve similar mechanisms to those controlling lncRNA function, they may also have similar effects on chromatin condensation and gene expression.

## METHODS

### **Subcellular fractionation, RNA isolation, and library construction**

Total RNA was isolated from mESCs, mNPCs, and cortical neurons (mCtx) that were fractionated into cytoplasmic, soluble nuclear, and chromatin pellet compartments as described previously (Pandya-Jones 2009, Wuariin Schibler 1994, Yeom Damianov 2017, Yeom 2018). After checking RNA quantity and integrity, RNAs longer than 200 nt (long RNA) and shorter than 200 nt (short RNA) were separated using RNeasy MinElute Cleanup Kit (Qiagen). Long RNAs were used for total and poly(A)+ libraries, and short RNAs were used for small RNA library construction. See also Supplemental Materials.

### **Calculation of chromatin partition indices and biotype analysis**

To analyze differential compartmentalization of RNAs, genes were selected that had chromatin expression greater or equal to the median TPM reported by kallisto (2.13 TPM), and had read counts greater than 0 in the cytoplasmic fractions as measured by FeatureCount. This returned 13,036 genes for analysis. DESeq2 was used to measure fold change in read counts between the chromatin-associated and the cytoplasmic poly(A)+ RNA by calculating the average read count among replicates of the chromatin fraction divided by the average read counts of the cytoplasmic fraction. The chromatin partition index was defined as the  $\log_2$  of this ratio (Fig. 1D).

Biotypes were retrieved from Ensembl annotation (V.91). Of the 13,036 genes, 400 genes (3.1 %) were analyzed in each of three ranges of the distribution. Partition indices were from -4.2 to -2.6 for region L, -0.1 to 0.1 for region M, and 4.1 to 8.6 for region R.

### **Measurement of intron retention**

We developed SIRI (Systematic Investigation of Retained Introns), a tool to stringently quantify unspliced introns by deep sequencing (<https://github.com/Xinglab/siri>). In this tool, we first retrieved all introns from Ensembl gene transfer format (GTF) version 91 for the mouse mm10 genome (Hunt et al. 2018). The number of reads mapping to each exon-exon (EE), exon-intron (EI), and intron-exon (IE) junctions were counted to determine the FI (Fraction of intron Inclusion) value of each intron. We selected only introns with a unique intron annotation (U introns) that are not involved in other alternative processing events (Supplemental Fig. S4B). Introns subjected to FI measurement were also required have an intron length greater than and equal to 60 and have a sum of EE + EI + IE reads be greater than and equal to 20 (Supplemental Table S6). From this set, IR events with EE reads no fewer than 2 in at least one cell compartment in one cell type were then kept for downstream analysis.

### **X-means clustering of IR events**

X-means clustering was performed using the Pyclustering tool (Novikov 2019) applied to the FI values determined in all three compartments of each cell type (Fig. 3A), with the maximum number of clusters set at 6. The distance matrix for X-means clustering is based on the Dynamic Time Warping (DTW) algorithm (Berndt and Clifford 1994) for the purpose of investigating directional changes of FI values from chromatin to nucleoplasm to cytoplasm. The Circos plot (Krzywinski et al. 2009) showing the intron group changes from one cell type to another cell type was produced using R (R Core Team 2020) package circlize (version 0.4.4) (Gu et al. 2014) (Fig. 5A).

### **Predicting intron retention patterns by deep learning**

To apply deep learning to IR group prediction, we constructed a compendium of 1,387 intron features of five types: sequence motifs, transcript features, RNA secondary

structure, nucleosome positioning, and conservation (Supplemental Table S9).

Sequence motif features included splice site consensus sequences, position-specific matrices of RNA-binding proteins, dinucleotide and trinucleotide frequencies of introns and flanking exons. Transcript features included the lengths of upstream exon (E1), downstream exon (E2), and intron (I) and intron number in the host gene. The translatability of E1, E2, E1 + E2, I and E1 + I + E2 were defined by confirming the absence of a stop codon in one of the three reading frames. To predict RNA secondary structure, RNA sequences from the regions from -20 to +20 nt relative to each splice site were examined. Sequence intervals from 1 to 70 nt, 70 to 140 nt, 140 to 210 nt from the 5' portion of the intron, and from -210 to -140 nt, -140 to -70 nt and -70 to -1 nt from the 3' portion of the intron were also examined. We computed the free energy of folding for each region with RNAfold (2.2.10) (Kerpedjiev et al. 2015) and used the free energy of unfolding for each region as features for the deep learning. The nucleosome positioning was predicted by NuPoP (version 1.0, set to the mouse model) (Xi et al. 2010) on the last 50 nt of the upstream exon, the first 100 nt of 5' intron region, the last 100 nt of 3' intron region, and the first 50 nt of downstream exon. The training dataset included introns that had grouping information in at least two cell types and excluded U11/U12 introns and other introns lacking GT or AG splice sites. We trained a Deep Neural Network (DNN) (LeCun et al. 2015) with these 1,387 features to predict whether introns belong to group A, B, C, and D for each cell type (Fig. 3A). The training was done with fivefold cross-validation with area under the ROC curves on data held-out during training reported for performance evaluation (Pounraja et al. 2019). To evaluate the strengths of individual features, we assessed the decrease of AUC on held-out data when the values of each feature were substituted by its median.

## DATA ACCESS

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>) under accession numbers GSE159944 for total RNA, GSE159919 for poly(A)+ RNA, GSE159971 for small RNA, and GSE159993 for poly(A)+ RNA in *Ptbp* knockdown experiments in Figs. 5, 6.

Links to the data displayed on the UCSC Genome Browser are here: ([https://genome.ucsc.edu/s/Chiaho/Kay\\_fraction\\_total\\_hub\\_10202020](https://genome.ucsc.edu/s/Chiaho/Kay_fraction_total_hub_10202020) for total RNA, and [https://genome.ucsc.edu/s/Chiaho/Kay\\_fraction\\_polyA%2B\\_hub\\_10202020](https://genome.ucsc.edu/s/Chiaho/Kay_fraction_polyA%2B_hub_10202020) for poly(A)+ RNA).

The source code of data analysis is available in <https://github.com/Xinglab/intron-retention-paper>, as well as in Supplemental Code files. The data resources used to reproduce the analysis are available in <https://doi.org/10.5281/zenodo.4540589>.

## ACKNOWLEDGMENTS

We thank Grigori Enikolopov (Cold Spring Harbor Laboratory) for the Nestin-GFP mouse line; Celine Vuong, Amy Pandya-Jones, Kathrin Plath, and members of the Black lab for help, discussions and comments on the manuscript. Financial support was provided by W.M. Keck Foundation and NIH R01 MH109166 grants to DLB and YX, NIH R01 GM088342 to YX, and R35 GM136426, R01 GM049662 and funding from the David Geffen School of Medicine and Division of Life Sciences at UCLA to DLB. K-HY received a postdoctoral fellowship from Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA. HL was supported by Whitcome Warsaw Family fellowships from UCLA.

## AUTHOR CONTRIBUTIONS

K-HY, ZP, XY and DLB: conception and study design. K-HY: all experiments and data generation, except *Ptbp* knockdown experiments and microscopy performed by HL and WX. ZP, C-HL and K-HY: data processing and analysis. ZP: defining unique introns, clustering analyses and deep learning analyses. K-HY and DLB wrote the manuscript with input from ZP and XY. DLB and XY: supervision and funding.

#### DISCLOSURE DECLARATION

The authors declare no competing interests.

## FIGURE LEGENDS

**Figure 1.** RNA partitioning between subcellular compartments.

**A.** Workflow used in this study. **B.** Genome browser tracks of the *Gapdh* locus in mESC. GENCODE annotated isoforms (M11) are diagrammed at the top. Poly(A)+ RNA (open box), total RNA (grey box), and peak RPM are noted on the left. RNA from chromatin (Chr), nucleoplasmic (Nuc), and cytoplasmic (Cyto) fractions are labeled at the right. The fixed Y scale (RPM) shows the strong enrichment of *Gapdh* RNA in the cytoplasm. The bottom tracks show chromatin RNA with an extended Y scale to observe the intron reads. **C.** Genome browser tracks of the *Xist/Tsix* locus in female mNPCs show strong chromatin enrichment of *Xist* RNA. **D.** Distribution of chromatin partition indices. The Chromatin / Cytoplasm ratio [Chr\_Poly(A)+ / Cyto\_Poly(A)+] of the averaged read counts of each gene are plotted as a distribution along the  $\log_2$  scale, with partition indices of representative genes indicated below. Biotypes of the 400 genes from bottom (left (L), blue bar), peak (middle (M), green bar), and top (right (R), red bar) of the distribution are presented in the bar graph below.

**Figure 2.** Cotranscriptional and posttranscriptional intron excision.

**A.** Genome browser tracks of the *Sorbs1* locus in mESC. Total chromatin RNA (grey box) shows intron reads, but the poly(A)+ RNA (open box) shows primarily exon reads except one posttranscriptional intron. **B.** Genome browser tracks of chromatin RNA at the *Wdr55* and *Vegfb* loci in mESC. Total (grey box) and poly(A)+ (open box) are shown, with cotranscriptionally and posttranscriptionally spliced introns are highlighted in green and red respectively. Yellow highlighted introns were not analyzable due to multiple processing patterns. **C.** Proportions of co- and posttranscriptional splicing for 49,692 U introns in mESC, using criteria described in Supplemental Figs. S4C – S4E. Introns

upstream (2,779) and downstream (2,744) from simple cassette exons were similarly analyzed.

**Figure 3.** Intron Groups defined by their retention level and fractionation behavior.

**A.** X-means clustering was applied to intron FI values and fraction enrichment in mESC, mNPC, and mCtx. The FI distribution for introns in each subcellular fraction and group is shown. **B.** Genome browser tracks (top) and RT-PCR validation (bottom) of representative transcripts in mESC. Validated introns are indicated by a blue highlight and a bracket below. Gel images are one of 3 biological replicates. **C.** The proportion of introns containing a PTC in frame with the upstream sequence is shown for each cluster and cell type. **D.** Percent of introns in each group for U introns from mESC and for detained introns within the U intron set (Boutz et al. 2015).

**Figure 4.** Deep Learning Analysis of Intron Groups.

**A.** Flow diagram for training the deep neural network. **B.** Performance of the model in distinguishing introns of different groups. ROC curves were plotted for individual pairwise comparisons with AUC values shown in parentheses. **C.** t-SNE plots of the 15 genomic features most predictive for distinguishing intron groups. Features distinguishing Group A from Groups C and D are shown above and those distinguishing Group C from Group D below. Features colored blue or red indicate the group for which they are positively correlated.

**Figure 5.** Regulation of intron retention and chromatin association during neuronal development.

**A.** Circos plot (Krzywinski et al. 2009; Gu et al. 2014) of intron group changes between cell types (mESC, mNPC, and mCtx). Introns not changing groups are on the left.

Introns switching groups between cell types on the right. **B.** Genome browser tracks of *Med22* during neuronal differentiation (left three panels) and after *Ptbp* knockdown in mESC (right three panels). Dashed boxes indicate U introns with measured FI values (introns 1 and 3) under each track. Group classification of intron 3 is at the bottom. **C.** Scatter plots of FI change between mESC and neurons (mCtx) plotted for each fraction against FI change after *Ptbp1* knockdown in mESC. Introns with delta FI less than -0.1 in both conditions are in red, and with delta FI greater than 0.1 in blue. Number of introns showing these changes with the number carrying PTBP1 iCLIP tags in parentheses, are above and below (Linares et al. 2015). Intron 3 of *Med22* is circled in green.

**Figure 6.** Chromatin enrichment and PTBP1 regulation of *Gabbr1* transcripts.

**A.** Genome browser tracks of *Gabbr1* in mESC and mCtx. PTBP1 iCLIP tags in mESC are plotted above in pink. Y axis indicates the maximum RPM in each cell type. The green box and bracket mark the intron 4-5 region expanded in panel B. PTBP1 responsive exons 5 and 15 are marked with arrowheads. **B.** Sashimi plots of the *Gabbr1* intron 4-5 region in mESC, mCtx, and after *Ptbp* knockdown in mESC. RPKM is plotted on the Y axis. PTBP1 responsive exon 5 is marked with an arrowhead. Exons encoding the two sushi domains and the 6-aa linker are marked on the conservation track below. **C.** Immunoblot showing expression of GABBR1 protein relative to other proteins in mESC and cortical neurons.

## REFERENCES

- Ameur A, Zaghlool A, Halvardson J, Wetterbom A, Gyllensten U, Cavelier L, Feuk L. 2011. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat Struct Mol Biol* **18**: 1435–1440.
- Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol* **11**: R106.
- Attig J, Agostini F, Gooding C, Chakrabarti AM, Singh A, Haberman N, Zagalak JA, Emmett W, Smith CWJ, Luscombe NM, et al. 2018. Heteromeric RNP Assembly at LINEs Controls Lineage-Specific RNA Processing. *Cell* **174**: 1067-1081.e17.
- Berndt DJ, Clifford J. 1994. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining, AAAIWS'94*, pp. 359–370, AAAI Press, Seattle, WA.
- Bhatt DM, Pandya-Jones A, Tong A-J, Barozzi I, Lissner MM, Natoli G, Black DL, Smale ST. 2012. Transcript Dynamics of Proinflammatory Genes Revealed by Sequence Analysis of Subcellular RNA Fractions. *Cell* **150**: 279–290.
- Boutz PL, Bhutkar A, Sharp PA. 2015. Detained introns are a novel, widespread class of post-transcriptionally spliced introns. *Genes Dev* **29**: 63–80.
- Boutz PL, Stoilov P, Li Q, Lin C-H, Chawla G, Ostrow K, Shiue L, Ares M, Black DL. 2007. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes Dev* **21**: 1636–1652.
- Braun CJ, Stanciu M, Boutz PL, Patterson JC, Calligaris D, Higuchi F, Neupane R, Fenoglio S, Cahill DP, Wakimoto H, et al. 2017. Coordinated Splicing of Regulatory Detained Introns within Oncogenic Transcripts Creates an Exploitable Vulnerability in Malignant Glioma. *Cancer Cell* **32**: 411-426.e11.
- Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. 2014. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res* **24**: 1774–1786.
- Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* **34**: 525–527.
- Brockdorff N, McCabe M, Norris P, Cooper J, Swift S, Kay F. 1992. The Product of the Mouse Xist Gene Is a 15 kb Inactive X-Specific Transcript Containing No Conserved ORF and Located in the Nucleus. **12**.
- Broseus L, Ritchie W. 2020. Challenges in detecting and quantifying intron retention from next generation sequencing data. *Comput Struct Biotechnol J* **18**: 501–508.
- Coulon A, Ferguson ML, de Turreis V, Palangat M, Chow CC, Larson DR. 2014. Kinetic competition during the transcription cycle results in stochastic RNA processing ed. D.L. Black. *eLife* **3**: e03939.

- Davidson L, Kerr A, West S. 2012. Co-transcriptional degradation of aberrant pre-mRNA by Xrn2. *EMBO J* **31**: 2566–2578.
- Duff MO, Olson S, Wei X, Garrett SC, Osman A, Bolisetty M, Plocik A, Celniker SE, Graveley BR. 2015. Genome-wide identification of zero nucleotide recursive splicing in *Drosophila*. *Nature* **521**: 376–379.
- Dvinge H, Bradley RK. 2015. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med* **7**: 45.
- Edfors F, Danielsson F, Hallström BM, Käll L, Lundberg E, Pontén F, Forsström B, Uhlén M. 2016. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Mol Syst Biol* **12**: 883.
- Edwards CR, Ritchie W, Wong JJ-L, Schmitz U, Middleton R, An X, Mohandas N, Rasko JEJ, Blobel GA. 2016. A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. *Blood* **127**: e24–e34.
- Fei J, Jadaliha M, Harmon TS, Li ITS, Hua B, Hao Q, Holehouse AS, Reyer M, Sun Q, Freier SM, et al. 2017. Quantitative analysis of multilayer organization of proteins and RNA in nuclear speckles at super resolution. *J Cell Sci* **130**: 4180–4192.
- Flynn RA, Almada AE, Zamudio JR, Sharp PA. 2011. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc Natl Acad Sci* **108**: 10460–10465.
- Frankiw L, Baltimore D, Li G. 2019a. Alternative mRNA splicing in cancer immunotherapy. *Nat Rev Immunol* **19**: 675–687.
- Frankiw L, Majumdar D, Burns C, Vlach L, Moradian A, Sweredoski MJ, Baltimore D. 2019b. BUD13 Promotes a Type I Interferon Response by Countering Intron Retention in *Irf7*. *Mol Cell* **73**: 803-814.e6.
- Garland W, Jensen TH. 2020. Nuclear sorting of RNA. *WIREs RNA* **11**: e1572.
- Girard C, Will CL, Peng J, Makarov EM, Kastner B, Lemm I, Urlaub H, Hartmuth K, Lührmann R. 2012. Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. *Nat Commun* **3**: 994.
- Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**: 2811–2812.
- Hao S, Baltimore D. 2013. RNA splicing regulates the temporal order of TNF-induced gene expression. *Proc Natl Acad Sci* **110**: 11934–11939.
- Hautbergue GM. 2017. RNA Nuclear Export: From Neurological Disorders to Cancer. In *Personalised Medicine: Lessons from Neurodegeneration to Cancer* (ed. S. El-Khamisy), *Advances in Experimental Medicine and Biology*, pp. 89–109, Springer International Publishing, Cham [https://doi.org/10.1007/978-3-319-60733-7\\_6](https://doi.org/10.1007/978-3-319-60733-7_6) (Accessed September 1, 2020).

- Herzel L, Neugebauer KM. 2015. Quantification of co-transcriptional splicing from RNA-Seq data. *Methods* **85**: 36–43.
- Herzel L, Ottoz DSM, Alpert T, Neugebauer KM. 2017. Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nat Rev Mol Cell Biol* **18**: 637–650.
- Horan L, Yasuhara JC, Kohlstaedt LA, Rio DC. 2015. Biochemical identification of new proteins involved in splicing repression at the *Drosophila* P-element exonic splicing silencer. *Genes Dev* **29**: 2298–2311.
- Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, Parton A, Armean IM, Trevanion SJ, Flicek P, et al. 2018. Ensembl variation resources. *Database* **2018**. <https://doi.org/10.1093/database/bay119> (Accessed January 26, 2021).
- Hurt JA, Robertson AD, Burge CB. 2013. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res* **23**: 1636–1650.
- Hutchinson JN, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB, Chess A. 2007. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* **8**: 39.
- Jacob AG, Smith CWJ. 2017. Intron retention as a component of regulated gene expression programs. *Hum Genet* **136**: 1043–1057.
- Jaillon O, Bouhouche K, Gout J-F, Aury J-M, Noel B, Saudeumont B, Nowacki M, Serrano V, Porcel BM, Ségurens B, et al. 2008. Translational control of intron splicing in eukaryotes. *Nature* **451**: 359–362.
- Kaupmann K, Huggel K, Heid J, Flor PJ, Bischoff S, Mickel SJ, McMaster G, Angst C, Bittiger H, Froestl W, et al. 1997. Expression cloning of GABA B receptors uncovers similarity to metabotropic glutamate receptors. *Nature* **386**: 239–246.
- Keppetipola N, Sharma S, Li Q, Black DL. 2012. Neuronal regulation of pre-mRNA splicing by polypyrimidine tract binding proteins, PTBP1 and PTBP2. *Crit Rev Biochem Mol Biol* **47**: 360–378.
- Kerpedjiev P, Hammer S, Hofacker IL. 2015. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics* **31**: 3377–3379.
- Khodor YL, Menet JS, Tolan M, Rosbash M. 2012. Cotranscriptional splicing efficiency differs dramatically between *Drosophila* and mouse. *RNA* **18**: 2174–2186.
- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. *Genome Res* **19**: 1639–1645.
- LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* **521**: 436–444.

- Li Y, Bor Y-C, Misawa Y, Xue Y, Rekosh D, Hammarskjöld M-L. 2006. An intron with a constitutive transport element is retained in a Tap messenger RNA. *Nature* **443**: 234–237.
- Lin Y, Protter DSW, Rosen MK, Parker R. 2015. Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. *Mol Cell* **60**: 208–219.
- Linares AJ, Lin C-H, Damianov A, Adams KL, Novitsch BG, Black DL. 2015. The splicing regulator PTBP1 controls the activity of the transcription factor Pbx1 during neuronal differentiation ed. B.J. Blencowe. *eLife* **4**: e09268.
- Liu H, Liang C, Kollipara RK, Matsui M, Ke X, Jeong B-C, Wang Z, Yoo KS, Yadav GP, Kinch LN, et al. 2016a. HP1BP3, a Chromatin Retention Factor for Co-transcriptional MicroRNA Processing. *Mol Cell* **63**: 420–432.
- Liu Y, Beyer A, Aebersold R. 2016b. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**: 535–550.
- Makeyev EV, Zhang J, Carrasco MA, Maniatis T. 2007. The MicroRNA miR-124 Promotes Neuronal Differentiation by Triggering Brain-Specific Alternative Pre-mRNA Splicing. *Mol Cell* **27**: 435–448.
- Mauger O, Lemoine F, Scheiffele P. 2016. Targeted Intron Retention and Excision for Rapid Gene Regulation in Response to Neuronal Activity. *Neuron* **92**: 1266–1278.
- Naftelberg S, Schor IE, Ast G, Kornblihtt AR. 2015. Regulation of alternative splicing through coupling with transcription and chromatin structure. *Annu Rev Biochem* **84**: 165–198.
- Naganuma T, Nakagawa S, Tanigawa A, Sasaki YF, Goshima N, Hirose T. 2012. Alternative 3'-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. *EMBO J* **31**: 4020–4034.
- Naro C, Jolly A, Di Persio S, Bielli P, Setterblad N, Alberdi AJ, Vicini E, Geremia R, De la Grange P, Sette C. 2017. An Orchestrated Intron Retention Program in Meiosis Controls Timely Usage of Transcripts during Germ Cell Differentiation. *Dev Cell* **41**: 82-93.e4.
- Ninomiya K, Kataoka N, Hagiwara M. 2011. Stress-responsive maturation of Clk1/4 pre-mRNAs promotes phosphorylation of SR splicing factor. *J Cell Biol* **195**: 27–40.
- Novikov AV. 2019. PyClustering: Data Mining Library. *J Open Source Softw* **4**: 1230.
- Pandya-Jones A, Bhatt DM, Lin C-H, Tong A-J, Smale ST, Black DL. 2013. Splicing kinetics and transcript release from the chromatin compartment limit the rate of Lipid A-induced gene expression. *RNA* **19**: 811–827.
- Pandya-Jones A, Black DL. 2009. Co-transcriptional splicing of constitutive and alternative exons. *RNA* **15**: 1896–1908.

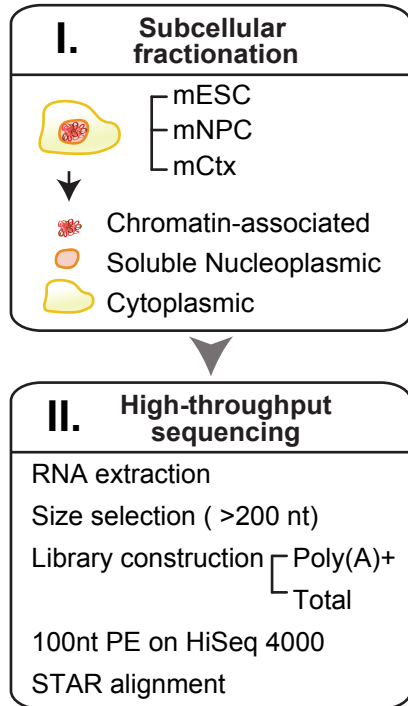
- Pandya-Jones A, Markaki Y, Serizay J, Chitiashvili T, Mancina Leon WR, Damianov A, Chronis C, Papp B, Chen C-K, McKee R, et al. 2020. A protein assembly mediates Xist localization and gene silencing. *Nature* 1–7.
- Parra M, Booth BW, Weiszmann R, Yee B, Yeo GW, Brown JB, Celniker SE, Conboy JG. 2018. An important class of intron retention events in human erythroblasts is regulated by cryptic exons proposed to function as splicing decoys. *RNA N Y N* **24**: 1255–1265.
- Pawlicki JM, Steitz JA. 2008. Primary microRNA transcript retention at sites of transcription leads to enhanced microRNA production. *J Cell Biol* **182**: 61–76.
- Pelleg D, Moore A. 2000. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *In Proceedings of the 17th International Conf. on Machine Learning*, pp. 727–734, Morgan Kaufmann.
- Pendleton KE, Chen B, Liu K, Hunter OV, Xie Y, Tu BP, Conrad NK. 2017. The U6 snRNA m6A Methyltransferase METTL16 Regulates SAM Synthetase Intron Retention. *Cell* **169**: 824-835.e14.
- Pimentel H, Parra M, Gee SL, Mohandas N, Pachter L, Conboy JG. 2016. A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res* **44**: 838–851.
- Popp MW-L, Maquat LE. 2013. Organizing Principles of Mammalian Nonsense-Mediated mRNA Decay. *Annu Rev Genet* **47**: 139–165.
- Pounraja VK, Jayakar G, Jensen M, Kelkar N, Girirajan S. 2019. A machine-learning approach for accurate detection of copy-number variants from exome sequencing. *Genome Res* gr.245928.118.
- Quinn JJ, Chang HY. 2016. Unique features of long non-coding RNA biogenesis and function. *Nat Rev Genet* **17**: 47–62.
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org>.
- Sakabe NJ, de Souza SJ. 2007. Sequence features responsible for intron retention in human. *BMC Genomics* **8**: 59.
- Saldi T, Cortazar MA, Sheridan RM, Bentley DL. 2016. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J Mol Biol* **428**: 2623–2635.
- Schmid M, Jensen TH. 2018. Controlling nuclear RNA levels. *Nat Rev Genet* **19**: 518–529.
- Schmitz U, Pinello N, Jia F, Alasmari S, Ritchie W, Keightley M-C, Shini S, Lieschke GJ, Wong JJ-L, Rasko JEJ. 2017. Intron retention enhances gene regulatory complexity in vertebrates. *Genome Biol* **18**: 216.

- Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. 2008. Divergent Transcription from Active Promoters. *Science* **322**: 1849–1851.
- Sibley CR, Emmett W, Blazquez L, Faro A, Haberman N, Briese M, Trabzuni D, Ryten M, Weale ME, Hardy J, et al. 2015. Recursive splicing in long vertebrate genes. *Nature* **521**: 371–375.
- Spellman R, Llorian M, Smith CWJ. 2007. Crossregulation and Functional Redundancy between the Splicing Regulator PTB and Its Paralogs nPTB and ROD1. *Mol Cell* **27**: 420–434.
- Stewart M. 2019. Polyadenylation and nuclear export of mRNAs. *J Biol Chem* **294**: 2977–2987.
- Tilgner H, Knowles DG, Johnson R, Davis CA, Chakraborty S, Djebali S, Curado J, Snyder M, Gingeras TR, Guigó R. 2012. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res* **22**: 1616–1625.
- Vargas DY, Shah K, Batish M, Levandoski M, Sinha S, Marras SAE, Schedl P, Tyagi S. 2011. Single-Molecule Imaging of Transcriptionally Coupled and Uncoupled Splicing. *Cell* **147**: 1054–1065.
- Vigot R, Barbieri S, Bräuner-Osborne H, Turecek R, Shigemoto R, Zhang Y-P, Luján R, Jacobson LH, Biermann B, Fritschy J-M, et al. 2006. Differential Compartmentalization and Distinct Functions of GABAB Receptor Variants. *Neuron* **50**: 589–601.
- Vilborg A, Steitz JA. 2017. Readthrough transcription: How are DoGs made and what do they do? *RNA Biol* **14**: 632–636.
- Vuong CK, Black DL, Zheng S. 2016. The neurogenetics of alternative splicing. *Nat Rev Neurosci* **17**: 265–281.
- Wang Q, Rio DC. 2018. JUM is a computational method for comprehensive annotation-free analysis of alternative pre-mRNA splicing patterns. *Proc Natl Acad Sci U S A* **115**: E8181–E8190.
- Wegener M, Müller-McNicoll M. 2018. Nuclear retention of mRNAs - quality control, gene regulation and human disease. *Semin Cell Dev Biol* **79**: 131–142.
- Windhager L, Bonfert T, Burger K, Ruzsics Z, Krebs S, Kaufmann S, Malterer G, L'Hernault A, Schilhabel M, Schreiber S, et al. 2012. Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution. *Genome Res* **22**: 2031–2042.
- Wong JJ-L, Ritchie W, Ebner OA, Selbach M, Wong JWH, Huang Y, Gao D, Pinello N, Gonzalez M, Baidya K, et al. 2013. Orchestrated intron retention regulates normal granulocyte differentiation. *Cell* **154**: 583–595.

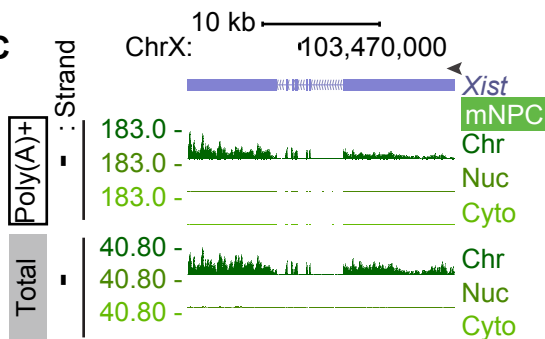
- Wuarin J, Schibler U. 1994. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol Cell Biol* **14**: 7219–7225.
- Xi L, Fondufe-Mittendorf Y, Xia L, Flatow J, Widom J, Wang J-P. 2010. Predicting nucleosome positioning using a duration Hidden Markov Model. *BMC Bioinformatics* **11**: 346.
- Yap K, Lim ZQ, Khandelia P, Friedman B, Makeyev EV. 2012. Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. *Genes Dev* **26**: 1209–1223.
- Yap K, Mukhina S, Zhang G, Tan JSC, Ong HS, Makeyev EV. 2018. A Short Tandem Repeat-Enriched RNA Assembles a Nuclear Compartment to Control Alternative Splicing and Promote Cell Survival. *Mol Cell* **72**: 525-540.e13.
- Yeom K-H, Damianov A. 2017. Methods for Extraction of RNA, Proteins, or Protein Complexes from Subcellular Compartments of Eukaryotic Cells. In *mRNA Processing: Methods and Protocols* (ed. Y. Shi), *Methods in Molecular Biology*, pp. 155–167, Springer, New York, NY [https://doi.org/10.1007/978-1-4939-7204-3\\_12](https://doi.org/10.1007/978-1-4939-7204-3_12) (Accessed August 21, 2020).
- Yeom K-H, Mitchell S, Linares AJ, Zheng S, Lin C-H, Wang X-J, Hoffmann A, Black DL. 2018. Polypyrimidine tract-binding protein blocks miRNA-124 biogenesis to enforce its neuronal-specific expression in the mouse. *Proc Natl Acad Sci* **115**: E11061–E11070.
- Zheng S, Gray EE, Chawla G, Porse BT, O'Dell TJ, Black DL. 2012. PSD-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2. *Nat Neurosci* **15**: 381–388.

Yeom\_Fig 1

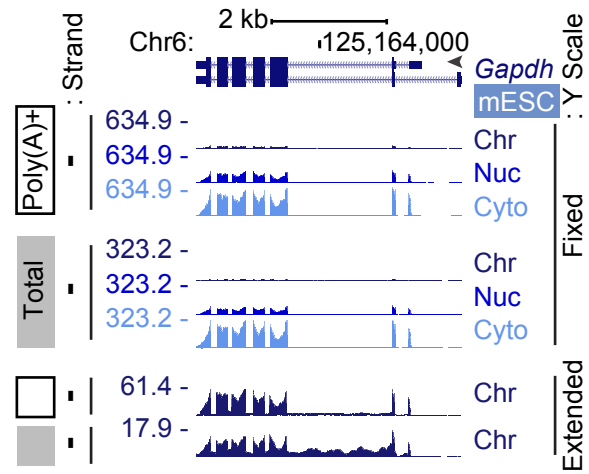
**A**



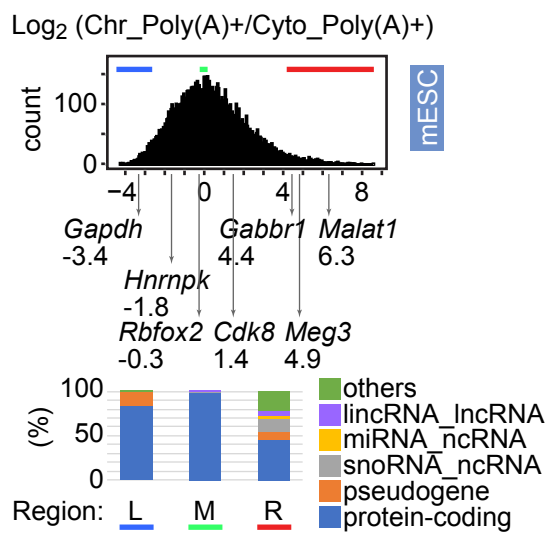
**C**



**B**



**D**



Yeom\_Fig 2

