Rapid and Efficient Co-Transcriptional Splicing Enhances Mammalian Gene Expression

Kirsten A. Reimer[1], Claudia Mimoso[2], Karen Adelman[2], and Karla M. Neugebauer[1*]

[1] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, 06520, USA
[2] Department of Biological Chemistry and Molecular Pharmacology, Blavatnik Institute, Harvard Medical School, Boston, MA 02115, USA

*Correspondence: karla.neugebauer@yale.edu

ABSTRACT

Pre-mRNA splicing is tightly coordinated with transcription in yeasts, and introns can be removed soon after they emerge from RNA polymerase II (Pol II). To determine if splicing is similarly rapid and efficient in mammalian cells, we performed long read sequencing of nascent RNA during mouse erythropoiesis. Remarkably, 50% of splicing occurred while Pol II was within 150 nucleotides of 3′ splice sites. PRO-seq revealed that Pol II does not pause around splice sites, confirming that mammalian and yeast spliceosomes can act equally rapidly. Two exceptions were observed. First, several hundred introns displayed abundant splicing intermediates, suggesting that the spliceosome can stall after the first catalytic step. Second, some genes – notably globins – displayed poor splicing coupled to readthrough transcription. Remarkably, a patient-derived mutation in β-globin that causes thalassemia improves splicing efficiency and proper termination, revealing co-transcriptional splicing efficiency is a determinant of productive gene output.

Keywords: nascent RNA, erythropoiesis, globin, co-transcriptional splicing, PacBio, long read sequencing

INTRODUCTION

Mammalian genes contain multiple introns of variable lengths that must be removed by a multi-megadalton complex — the spliceosome —for genetic information to be expressed correctly. The spliceosome assembles at the 5′ and 3′ splice sites (SSs) that demarcate intron boundaries and then catalyzes two transesterification reactions to excise introns and ligate exons together (Wilkinson et al., 2019). Splicing is a highly-regulated process; it is influenced by environmental factors, including stress and developmental cues, as well factors in the local pre-messenger RNA (pre-mRNA) environment such as intron structure and RNA-binding protein occupancy (Baralle and Giudice, 2017; Jeong, 2017; Lin et al., 2016; Pai and Luca, 2019). Additionally, splicing is linked to other mRNA processing reactions including 5′ end capping, cleavage, polyadenylation, and nuclear export (reviewed in Bentley, 2014; Herzel et al., 2017).

Across species, tissues, and cell types, splicing occurs co-transcriptionally (Custodio and Carmo-Fonseca, 2016; Neugebauer, 2019). This demands that the constellation of regulatory factors affecting splicing also act co-transcriptionally. However, the spatial and temporal window in which splicing and its regulation occur is poorly defined *in vivo*, particularly in mammals. Recent work has focused on studying the coordination between transcription and splicing in living cells and how this affects gene expression. For example, altering the rate of RNA polymerase II (Pol II) elongation has been shown to affect the outcome of splicing; when Pol II transcribes more slowly than usual, widespread changes in alternative splicing occur (Aslanzadeh et al., 2018; Braberg et al., 2013; Carrillo Oesterreich et al., 2016; de la Mata et al., 2003; Fong et al., 2014; Ip et al., 2011; Jonkers and Lis, 2015; Schor et al., 2013), indicating that the splicing and transcription machinery are finely tuned to one another and that this coordination is a strong determinant of gene expression. The presence of introns increases gene expression through direct effects on transcription and chromatin (Bieberstein et al., 2012; Brinster et al., 1988; Fiszbein et al., 2019; Shaul, 2017), though the underlying mechanisms are still unknown. Finally, in addition to efficient splicing being necessary for RNA export, it has been shown that efficient splicing allows release of RNA from the site of transcription (Custodio et al., 1999).

Previously, our lab has characterized co-transcriptional splicing in budding and fission yeasts using two single-molecule sequencing methods (Carrillo Oesterreich et al., 2016; Herzel et al., 2018). To gain greater biological insight into the role of co-transcriptional splicing as a gene regulatory mechanism, we analyzed nascent RNA transcription and splicing in a mammalian system: murine erythroleukemia (MEL) cells undergoing erythroid differentiation. We have employed two single-molecule sequencing approaches to directly measure co-transcriptional splicing of nascent RNA: (*i*) Long read sequencing (LRS), which enables genome-wide analysis of splicing with respect to Pol II position and (*ii*) Precision Run-On sequencing (PRO-seq), enabling the assessment of Pol II density at these sites. This study determines the spatial window in which co-transcriptional splicing occurs in mammalian cells, co-transcriptional splicing efficiency for thousands of genes, Pol II elongation behavior across splice junctions, and the effects of efficient co-transcriptional splicing on gene output. These findings provide a framework for future work on the coordination between splicing and transcription in mammalian cells. In particular, the demonstration of highly efficient splicing in mammals in the absence of transcriptional pausing causes us to rethink key features of splicing regulation.

RESULTS

PacBio Long Read Sequencing of Nascent RNA Yields High Coverage

Murine erythroleukemia (MEL) cells are a model for murine erythropoiesis which are immortalized at the proerythroblast stage and can be induced to enter terminal erythroid differentiation by treatment with 2% DMSO for five days. Phenotypic changes in the cells include decreased cell volume, increased levels of β-globin expression, and a visual increase in hemoglobinization (**Figures S1A-C**). We used chromatin purification of uninduced and induced MEL cells followed by biochemical depletion of polyA(+) and ribosomal RNA in order to enrich for Pol II-associated RNA (hereafter referred to as "nascent RNA"; **Figure 1A**). Chromatin purification in stringent washing conditions of 0.5 M urea allows release of contaminating RNAs, but retains the ternary complex of nascent RNA and Pol II on chromatin (**Figure S1D**; Wuarin and Schibler, 1994). Importantly, we were unable to detect any changes in splicing of nascent RNA by RT-PCR when we isolated nascent RNA in the presence or absence of the splicing inhibitor Pladienolide B, demonstrating that splicing is not ongoing during chromatin fractionation or RNA isolation (**Figure S2**).

In order to generate libraries for LRS of nascent RNA, we adapted protocols previously established in our lab (Carrillo Oesterreich et al., 2016; Herzel et al., 2018). Briefly, a DNA adapter was ligated to RNA 3′ ends, followed by reverse transcription using a strand-switching enzyme to create both the first and second strand of cDNA (**Figure 1A**). Two biological replicates, each with two technical replicates, were sequenced using PacBio RSII and Sequel flowcells, yielding a total of 1,010,270 mappable reads (**Table S1**). Reads containing a non-templated polyA tail comprised only 1.9% of the total reads (**Table S1**), confirming the stringent biochemical removal of polyA(+) RNA from our samples. The detected polyA(+) RNAs were bioinformatically filtered in all further analyses. Principal component analysis of all replicates showed the largest variance between induced and uninduced samples, as expected (**Figure S1E**). In all replicates, the non-coding RNA 7SK was the most abundant RNA (accounting for 15.8% and 12.5% of the reads in uninduced and induced libraries, respectively), so these reads were removed in all further analyses. Of the remaining 990,966 reads, the average read length was 675 and 550 nucleotides (nt), and the average coverage in reads per gene was 11.9 and 8.7 for uninduced and induced samples, respectively (**Figure 1B-C**). We observed 637 genes with 100 or more reads and 9,519 genes with 10 or more reads in either of the two treatment conditions (**Figure 1C**). While coverage of 5′ ends was focused downstream of annotated transcription start sites (TSSs), 3′ end coverage was distributed more evenly throughout gene bodies, with an increase just upstream of annotated transcription end sites (TESs), and a drop after TESs (**Figure S1F**).

LRS Reveals Rapid and Efficient Co-transcriptional Splicing

Each long read provides two critical pieces of information: the 3′ end reveals the position of Pol II when the RNA was isolated, and the splice junctions reveal if and where splicing has occurred. In previous studies (Carrillo Oesterreich et al., 2016; Herzel et al., 2018), read density per gene was low, and all reads could be easily displayed. In the present study, the enhanced read depth afforded by advancements in PacBio flowcell technology necessitated new methods of data visualization. Here, we present our LRS data in a format that highlights both the 3′ end position and the splicing status (**Figure 2A**; **Figure S3**). Each transcript was categorized and colored according to its splicing status: either "all spliced", "partially spliced", "all unspliced", or "NA" (transcripts that did not span entire introns).

To quantitate the range of splicing observed in our data, we calculated a per-gene co-transcriptional splicing efficiency (coSE) metric, which we define as the number of spliced introns divided by the total number of fully transcribed introns per gene. We only considered genes which were covered by at least 10 reads in both conditions and did not contain any introns longer than 5 kb. We observed a wide range in coSE values across different genes. For example, the gene *Pabpc1* has a high coSE (0.92 in uninduced and 0.86 in induced cells; **Figure 2B**). *Actb* and *Calr* have more moderate coSE values (0.51-0.69) and a greater fraction of partially spliced or all unspliced reads (**Figure 2B**). In addition, we observed examples of lncRNA genes, like *Snhg5*, which have a majority of all unspliced reads (coSE = 0.38 in uninduced and coSE = 0.06 in induced cells; **Figure 2C**). Genome-wide, we found that 89.5% and 83.9% of introns were spliced in uninduced and induced conditions, respectively; thus, the majority of introns are removed co-transcriptionally (**Figure 2D**). We found a similarly high per-gene coSE between uninduced and induced cells which prompted us to merge these two datasets for further splicing analysis. The fraction of introns spliced was slightly lower in the induced cells compared to the uninduced cells, which may be reflective of regulated intron retention that has previously been reported during erythroid differentiation (Parra et al., 2018; Pimentel et al., 2016). Importantly, our data indicate extremely prevalent splicing under both conditions (57% of genes have coSE > 90%; **Figure 2E**).

How soon after Pol II transcribes a 3′SS can splicing occur? Our data enables us to determine the position of Pol II associated with spliced nascent transcripts by measuring the distance in nucleotides between the 3′ end of each read and the nearest spliced junction (**Figure 3A**). In order to consider only 3′ ends which arise from actively elongating transcripts, reads with 3′ ends arising from splicing intermediates were removed in this analysis (discussed below). While the longest distances observed were just over 6 kb, 75% of splice junctions were within 300 nt of Pol II, and the median distance was 141 nt (**Figure 3B**; **Figure S4A**). The median exon size in the mouse genome is 151 nt (Waterston et al., 2002), suggesting that Pol II is typically within or just downstream of a newly transcribed exon when the upstream intron is spliced. This supports a model whereby some introns are spliced by intron definition in mammals. Pol II distance to the nearest splice junction was shorter for protein-coding genes (median = 127 nt) than non-coding genes (median = 193 nt, 227 nt, 242 nt for antisense, pseudogene, and lincRNA, respectively; **Figure 3C**; **Figure S4B**), consistent with previous findings of our two labs (Pai et al., 2017; Wachutka et al., 2019). Additionally, our analysis of recent LRS data of nascent RNA obtained by direct RNA sequencing using Oxford Nanopore from three different cell types shows that the distance from Pol II to the nearest splice junction is surprisingly similar across organisms and cell types (median distance in human BL1184 = 239 nt, human K562 = 282 nt, *Drosophila* S2 = 409 nt; **Figure S4C**; Drexler et al., 2019). Therefore, we conclude that the spliceosome is able to use a 3′SS when Pol II has transcribed a relatively short distance past the end of an intron.

Pol II Does Not Pause to Allow for Splicing to be Completed

One explanation for the relatively short distances observed between splice junctions and Pol II may be that Pol II pauses downstream of a splice site to allow time for splicing to occur. This model has previously been proposed as a mechanism for splicing and transcription to feedback on each other, with pausing as a checkpoint for correct RNA processing (Alexander et al., 2010; Carrillo Oesterreich et al., 2011; Carrillo Oesterreich et al., 2010; Chathoth et al., 2014; Milligan et al., 2017). However, recent work on co-transcriptional splicing has disagreed on the behavior of Pol II elongation near splice junctions, with some studies indicating long-lived pausing at splice sites, and others reporting no significant pausing (Kwak et al., 2013; Mayer et al., 2015; Sheridan et al., 2019).

To directly evaluate changes in Pol II occupancy in relation to splicing, we measured the density of elongating Pol II genome-wide using Precision Run-On sequencing (PRO-seq) in MEL cells. PRO-seq allows for mapping of actively elongating Pol II complexes at single-nucleotide resolution through labeling of nascent RNA by incorporation of a single biotinylated NTP (Mahat et al., 2016). Comparing PRO-seq with LRS is advantageous, because PRO-seq data provide an independent measure of nascent RNA 3′ ends that are being actively transcribed and not originating from other chromatin-associated intermediates. Although RNA is fragmented for PRO-seq, making reads short (average fragment size = 50 nt), a number of PRO-seq reads contained spliced junctions (**Figure 3D**; 396,257 spliced reads out of 289,610,781 total mapped reads). These data confirm that mammalian splicing can occur when actively engaged Pol II is just downstream of the 3′SS. Thus, two complementary methods to probe Pol II position and splicing status indicate that splicing can occur in close proximity to the 3′SS.

We next used PRO-seq to determine if transcription elongation changes across intron-exon boundaries or contributes to co-transcriptional splicing. As expected, metagene plots around active TSSs reveal prominent promoter-proximal pausing (**Figure 3E**; Core and Adelman, 2019). Analyzing PRO-seq signal around splice sites initially revealed a small peak near the 5′SS (**Figure S4D**). However, we were concerned that high PRO-seq density from TSS peaks might be bleeding through to the first 5′SS. Indeed, analyzing first introns independently demonstrated that elevated PRO-seq signal in the vicinity of the 5′SS was only seen at first introns, and only at introns with 5′SS ≤ 250 nt from the TSS (**Figure S4D-E**). Accordingly, after removal of first introns from our analysis, the PRO-seq signal around the 5′SS and 3′SS showed no evidence of pausing and no significant changes in Pol II elongation (**Figure 3E**). To ask whether introns with particularly high or low splicing efficiency might display different elongation profiles, we evaluated PRO-seq signal around splice sites grouped by the co-transcriptional splicing efficiency (coSE) of the adjacent intron in LRS; again, we found no significant changes in Pol II elongation among any group of introns (**Figure S4F**). Taken together, the lack of altered Pol II occupancy across splice junctions indicates that Pol II elongation is not generally impacted by the transcription of splice site sequences or by splicing itself. Therefore, the efficient co-transcriptional splicing observed just downstream of the 3′SS cannot be attributed to transcriptional pausing.

Splicing Intermediates are Readily Observed at a Subset of Introns

Splicing intermediates have previously been observed using other chromatin-associated RNA sequencing methods (Burke et al., 2018; Chen et al., 2018; Churchman and Weissman, 2011; Nojima et al., 2015; Nojima et al., 2018). Intermediates arise from the upstream lariat-exon intermediate that is created after the first catalytic step of splicing (**Figure 4A**), and can be identified by their 3′ ends which align to the last nucleotide of an exon. Genome-wide, we observed 3′ ends of long reads spread relatively evenly across gene bodies (**Figure S1F**). However, we detected a defined peak in LRS 3′ end coverage at the last nucleotide of exons (**Figure 4B**). These splicing intermediate reads were relatively rare, making up 5.7% of the data, and the majority of genes harbored either 0 or 1 splicing intermediate reads (**Figure S5A**). What was striking about our data was the specific distribution of these intermediates. A small number of genes contained a large number of splicing intermediates, and these intermediate reads tended to be aligned at a single intron within the gene (**Figure 4C**). For example, 219 of the 430 reads mapped to *Alas2* had 3′ ends mapped to the end of exon four (**Figure 4D**). Interestingly, several reads (14/430) mapping to *Alas2* were one of the extremely rare instances of support for potential recursive

splicing. In this case, we observed an unannotated splice junction which generated a new 5′SS immediately adjacent, with the junction sequence caglGUAUGU (**Figure 4E**).

To determine what features of specific introns might lead to increased splicing intermediates, we normalized the number of splicing intermediates observed for each intron. The resulting metric, normalized intermediate count (NIC), is defined as the number of splicing intermediate reads at the -1 position relative to each intron divided by the sum of splicing intermediate reads and spliced reads. We binned all unique introns based on their observed NIC value, and calculated the splice site strength of the introns in each bin using the MaxEnt method (Yeo and Burge, 2004). We found that while the 5′SS score remained constant across all bins, introns with the highest NIC value tended to have a lower 3′SS score (**Figure 4F**); this result remained true even when we downsampled our data to include the same number of introns in each bin (**Figure S5B**). We generated sequence logos from the regions used to calculate the 3′SS score and found that introns in the highest NIC value bin had a weaker, non-consensus polypyrimidine tract (**Figure 4G**). Taken together, the data suggest that the transition from step 1 to step 2 of splicing may be occurring more slowly at introns with weak 3′SSs, allowing the detection of splicing intermediate reads. We also separated our PRO-seq data by intron NIC value, then looked at Pol II density around splice sites. We observed no significant differences in Pol II density around splice sites in any NIC category (**Figure 4H**). We conclude that splicing does not feedback on Pol II elongation, regardless of how efficiently an intron is spliced, but that introns with weak 3′SSs may cause a delay between the catalytic steps of splicing.

Erythroid Genes with Poor Splicing Efficiency Display Readthrough Transcription

This dataset gave us the opportunity to look at the expression of endogenous globin genes under the physiologically relevant condition of terminal erythroid differentiation. We observed an increased number of long reads at the β-globin (*Hbb-b1*) locus upon induction, in agreement with an increased level of β-globin mRNA detected by RT-qPCR (**Figure 5A**; **Figure S1C**; 39 reads in uninduced vs. 605 reads in induced conditions). This suggests that LRS provides a semi-quantitative measure of gene expression. To our surprise, a significant fraction of β-globin long reads in the induced condition had their 3′ ends up to 5 kb downstream of the annotated polyA site (PAS), indicating incomplete termination by Pol II and readthrough transcription. We confirmed readthrough transcription past the β-globin PAS by detecting an increased PRO-seq signal in this region (**Figure 5B**). At the same time, the splicing efficiency of β-globin decreased upon induction (coSE = 1.00 in uninduced cells, coSE = 0.56 in induced cells; **Figure 5A**). The α-globin genes (*Hba-a1* and *Hba-a2*) exhibited a similar increase in coverage downstream of the PAS and decrease in splicing efficiency upon induction. However, readthrough transcription at α-globin was limited to the region 1-2 kb downstream of the PAS (**Figure S6**). Together, these data suggest that under physiologically relevant conditions, a significant fraction of globin gene transcription is inefficiently terminated.

The most striking feature of the β-globin readthrough transcripts was that they were predominantly unspliced (**Figure 5A**). We observed similar unspliced readthrough transcripts in the α-globin genes (**Figure S6A**), as well as several other genes, for example *Snhg5*, *Psmb6*, and *Eif1* (**Figure 2C**; **Figure S3**). In order to examine this phenomenon genome-wide, we analyzed read coverage downstream of annotated PASs for reads that were either all unspliced or contained at least one splice junction. We found that coverage of unspliced reads was globally higher in the region 1 kb downstream of a PAS than it was for spliced reads (**Figure 5C**; note that some background coverage is due to intergenic mapping). We then categorized our long reads as being readthrough transcripts if the 5′ end originated within a

gene body and the 3′ end mapped more than 100 nt downstream of the last annotated PAS. In comparison to all transcripts, which are mostly all spliced (76%), readthrough transcripts were mostly unspliced (59%; **Figure 5D**). This decreased splicing efficiency in readthrough transcripts suggests functional links between splicing and Pol II termination.

β-thalassemia Mutation Generates Enhanced Cryptic Splicing and Decreases Readthrough Transcription

To investigate how mutations in splice sites alter co-transcriptional splicing efficiency, we took advantage of a known β-thalassemia allele. A patient-derived G>A mutation in intron 1 of human β-globin (*HBB*) leads to new AG dinucleotide in intron 1 which is used as a cryptic 3′SS 19 nt upstream of the canonical 3′SS (**Figure 6A**). This thalassemia-causing mutation, known as IVS-110, generates an *HBB* mRNA containing an in-frame stop codon, resulting in a 90% reduction in functional HBB protein (Spritz et al., 1981; Vadolas et al., 2006). We utilized two previously reported MEL cell lines: a control cell line expressing an integrated copy of a human β-globin minigene (MEL-*HBB* ^{WT}), and a cell line expressing the human β-globin minigene with the IVS-110 mutation (MEL-*HBB* $^{IVS\text{-}110(G>A)}$; Patsali et al., 2018). We treated the MEL-*HBB* ^{WT} and MEL-*HBB* $^{IVS\text{-}110(G>A)}$ cells with 2% DMSO to induce erythroid differentiation and then generated an LRS library targeting the integrated *HBB* locus in three biological replicates. Specific targeting of this locus during library preparation resulted in an average of 11,636 nascent RNA reads that mapped to the *HBB* gene per replicate (**Table S1**), allowing rigorous statistical analysis. As previously reported, the majority of intron 1 splicing in the MEL-*HBB* $^{IVS\text{-}110(G>A)}$ cell line occurs at the cryptic 3′SS (average 94%; **Figure S7**).

To determine whether use of the upstream cryptic 3′SS would affect the spatial window in which splicing could occur, we measured the distance from this splice junction to Pol II positions. We found no significant difference between the use of the canonical 3′SS in MEL-*HBB* ^{WT} cells, the canonical 3′SS in MEL-*HBB* $^{IVS\text{-}110(G>A)}$ cells, or the cryptic 3′SS in MEL-*HBB* $^{IVS\text{-}110(G>A)}$ cells (**Figure 6B**; significance tested by nested ANOVA). Only reads that contained a splice junction, either at the cryptic or canonical 3′SS, and had a 3′ end before the end of intron 2 were considered to ensure our findings were not confounded by potential effects from splicing of intron 2. These data would suggest that even when the spliceosome is faced with a choice between two splice sites, either can be efficiently used within the same spatial window.

We next asked whether or not the thalassemia-causing mutation caused any changes in overall levels of β-globin splicing. As previously noted, the endogenous mouse β-globin had a relatively low splicing efficiency after induction (**Figure 5A**); this was mirrored in the *HBB* minigene (**Figure 6C**). We found that on average, MEL-*HBB* ^{WT} cells had intron 1 spliced in 12.8% of reads (**Figure 6D**). Surprisingly, MEL-*HBB* $^{IVS\text{-}110(G>A)}$ cells exhibited increased splicing at intron 1, with 19.2% of reads containing either the cryptic or canonical splice junction (**Figure 6D**). At the same time, we observed an even greater difference in splicing of intron 2 between the WT and mutant allele (2.8% in MEL-*HBB* ^{WT} cells; 11.2% in MEL-*HBB* $^{IVS\text{-}110(G>A)}$ cells; **Figure 6E**). This suggests that the increased splicing of intron one has a cooperative effect on the splicing of intron 2.

Finally, we asked whether or not the perturbation we observed in the level of splicing had an effect on the high level of readthrough transcription we noted previously at the endogenous mouse β-globin gene (**Figure 5A-B**). Abundant readthrough transcription was also observed at the integrated *HBB* loci in both MEL-*HBB* ^{WT} and MEL-*HBB* $^{IVS\text{-}110(G>A)}$ cells (**Figure 6C**). However, MEL-*HBB* $^{IVS\text{-}110(G>A)}$ cells exhibited a

significantly lower fraction of readthrough transcripts (**Figure 6F**), indicating that changes in splicing efficiency can impact transcription termination. Thus, a single point mutation observed in β-thalassemia that creates a cryptic 3′SS was found to affect the splicing efficiency of a downstream intron as well as transcription termination. This finding indicates a previously unappreciated level of crosstalk between splicing efficiency within a gene and downstream polyA cleavage.

DISCUSSION

This study highlights a broad range of co-transcriptional splicing efficiencies across mammalian genes and introns through a genome-wide analysis of differentiating mammalian erythroid cells. We visualized transcription and splicing dynamics with unprecedented depth and accuracy through long read sequencing of nascent RNA and PRO-seq. We show that splicing catalysis can occur while Pol II is just 75-300 nt past the 3′SS and find no evidence for transcriptional pausing in this window. Thus, we infer that spliceosome assembly and transition to catalysis can occur in close physical proximity to Pol II. Two striking cases stood out from our observations of splicing. First, introns that contain a weak 3′SS seem to induce a stall in the splicing reaction itself, causing a buildup of splicing intermediates. Second, inefficient splicing was correlated with inefficient transcription termination globally (**Figure 7**). Paradoxically, abundant readthrough transcription occurred at the globin gene loci despite the induction of globin expression during erythroid differentiation. Remarkably, a patient-derived, thalassemia-causing point mutation in β-globin increased splicing efficiency and decreased readthrough transcription. These data show for the first time that co-transcriptional splicing efficiency determines gene expression through impacts on transcriptional readthrough.

These data suggest that despite changes in the complexity of yeast and mammalian spliceosomes, the length and number of introns in mammals, as well as the number of accessory factors that can influence splicing, the mammalian spliceosome is capable of assembling and acting in the same spatial window as the yeast spliceosome. (Carrillo Oesterreich et al., 2016; Herzel et al., 2018). Consistent with these findings, many groups have measured the fraction of splicing that occurs co-transcriptionally using methods such as high-density tiling arrays, short-read sequencing, metabolic labeling, and imaging-based assays (Alpert et al., 2017; Neugebauer, 2019). These measurements made in yeast, fly, mouse, and human cells converge on values of 75-87% of splicing being co-transcriptional, suggesting widely conserved features of transcription and splicing mechanisms (Ameur et al., 2011; Carrillo Oesterreich et al., 2016; Carrillo Oesterreich et al., 2010; Girard et al., 2012; Herzel et al., 2018; Khodor et al., 2011; Tilgner et al., 2012). In the present study, the majority of splicing occurs co-transcriptionally, with an average of 87% of introns being removed before Pol II reaches the gene end. Here, spliced mammalian nascent RNAs can be observed even in short PRO-seq reads.

A recent paper reports that the majority of introns in both human and fly nascent RNA appear unspliced and that Pol II is 2-4 kb downstream of the 3′SS when splicing occurs (Drexler et al., 2019). Such delays between transcription and splicing are unlikely, given the levels of co-transcriptional splicing discussed above. One possible explanation for this discrepancy is that the purification of 4sU-labeled RNA performed in the Drexler *et al.* study may have inadvertently enriched for long, intron-containing RNAs over shorter, spliced RNAs; longer RNAs have a greater probability of containing a labeled U residue, and introns are much more U-rich than exons. It is also unclear whether 4sU incorporation affects spliceosome assembly and catalysis due to changes in base-pairing among U-rich RNA elements in

introns and snRNAs (Testa et al., 1999). However, despite differences in the fraction of unspliced RNAs in these datasets and the conclusions drawn, our analysis of the distance between splice junctions and the RNA 3′ end in the Drexler *et al*. datasets shows that the distances at which the spliceosome can act are in fact very similar to what we find in mouse (**Figure S4C**). Thus, all currently available LRS data show that a large fraction of splicing catalytic events are completed in close physical proximity to Pol II, consistent with the possibility that the spliceosome and Pol II may interact physically to achieve regulation (David et al., 2011; Gu et al., 2013; Harlen et al., 2016; Nojima et al., 2018; Yu and Reed, 2015).

Importantly, PRO-seq – a short read method that specifically and quantitatively detects elongating polymerase molecules (Kwak et al., 2013) – corroborated the efficiency of co-transcriptional splicing. We were able to find spliced reads within the PRO-seq data, validating the observations made with LRS of purified nascent RNA with an independent method. Similarly, mNET-seq data, which is generated by short-read sequencing of nascent RNA from immunoprecipitated Pol II, has revealed examples of spliced reads as well (Nojima et al., 2018). Having observed many examples wherein an RNA 3′ end was only a short distance beyond the 3′SS, we considered the hypothesis that Pol II pausing at or near 3′SSs could provide extra time for splicing (Alexander et al., 2010; Carrillo Oesterreich et al., 2011; Carrillo Oesterreich et al., 2010; Chathoth et al., 2014; Milligan et al., 2017). However, our analysis shows that any detection of a PRO-seq peak at 5′SSs—albeit small in meta-analysis—is caused by bleed-through from promoter-proximal pausing, and we simply do not detect pausing at 3′SSs (**Figure 3E**; **Figure S4D-E**), in agreement with a recent study (Sheridan et al., 2019). Further, we detect no evidence that Pol II elongation is impacted by the splicing efficiency of a transcribed intron (**Figure S4F**). These data strongly support the assumption that Pol II travels uniformly across splice junctions, enabling an estimation of the time elapsed since a junction was transcribed based on the RNA 3′ end position. Using the median distance from splicing events to Pol II in our data (141 nt), and taking into account the range of measured Pol II elongation rates (0.5-6 kb/min; Jonkers and Lis, 2015), we conclude that the mammalian spliceosome is able to recognize the 3′SS and undergo the second step of splicing catalysis within 1.4-17 seconds.

Our LRS data capture precursors, intermediates, and products of the splicing reaction. To our surprise, we detected splicing intermediates that were distributed unevenly throughout introns. Due to the 3′ end chemistry and structure of splicing intermediates, we can only capture the upstream exon portion of the nascent RNA undergoing splicing and cannot observe any of the downstream lariat intermediates. Nevertheless, we were able to identify poor sequence consensus at the downstream 3′SS as a strongly correlating feature of splicing intermediates. This evidence is consistent with a model where modulation of the transition between catalytic steps of splicing can alter splicing fidelity or outcome (Smith et al., 2008). Although it is tempting to associate the weak 3′SS sequences we identified with 3′SS selection that occurs *before* spliceosome assembly and involves recognition by U2AF and/or other factors, the spliceosomes associated with these intermediates have already assembled and undergone step 1 chemistry. During the splicing reaction, the substrates in the catalytic center shift from the branch site adenosine and the 5′SS (first step) to the 3′ end of the first exon and the 3′SS AG (second step), which is typically 30-60 nt downstream of the branch site. Therefore, the only factors that could interact and respond to our identified weak 3′SS sequences must be at the core of the spliceosome. A recent Cryo-EM study of human spliceosomes has identified several spliceosomal components that may be in a position to regulate the transition from step 1 to step 2 (Fica et al., 2019). Future studies of these

enigmatic new players may reveal a role for 3′SS diversity in the regulation of splicing by stalling between catalytic steps.

By using erythroid differentiation as a model system, our LRS data could resolve a mystery shrouding β-globin pre-mRNA splicing. Two previous studies used stably integrated β-globin reporter genes combined with high resolution fluorescence microscopy to track pre-mRNA transcription and splicing in HEK293 and U2OS cells (Coulon et al., 2014; Martin et al., 2013). One study reported data consistent with co-transcriptional splicing, while the other strongly favored post-transcriptional splicing. The LRS data presented here explains that, at least in MEL cells, there are two major populations of globin transcripts: all-spliced and all-unspliced. In any biochemical assay or short read RNA-seq assay that examines populations of pre-mRNA, inefficient splicing would be one explanation for the bulk result. In reality, a fraction of the transcripts is efficiently spliced and productively expressed, because polyA cleavage and termination also occur efficiently. The fraction of efficiently spliced β-globin transcripts increased in the case of the IVS-110 thalassemia allele we studied; although this alternative 3′SS yields an out of frame mRNA that will – like many thalassemia alleles of β-globin – be degraded by nonsense-mediated decay (Kurosaki et al., 2019), improving splicing efficiency could be a general strategy for increasing gene output in a variety of disease settings that feature poor splicing.

Here, we report a striking abundance of readthrough transcription at the β-globin locus, which accompanies inefficient splicing. In *S. pombe,* unspliced readthrough transcripts were likely targeted for degradation by the nuclear exosome (Herzel et al., 2018), suggesting these may be non-productive transcripts. Interestingly, many physiological stresses – such as osmotic stress, heat shock, cancer, aging, and viral infection – cause transcriptional readthrough (Enge et al., 2017; Grosso et al., 2015; Muniz et al., 2017; Vilborg et al., 2015; Vilborg et al., 2017). The molecular mechanism of transcriptional readthrough is currently not understood; here we have implicated defective splicing in transcriptional readthrough by showing that improving splicing efficiency increases proper cleavage and termination. The correlation between inefficient splicing and transcriptional readthrough observed in these cells as well as the direct causative effect detected in the expression of β-globin suggests a mechanism whereby splicing must take place in a spatial and/or temporal “window of opportunity”. If this window is passed and no splicing has occurred, it is likely that polyA cleavage will be suppressed and Pol II will fail to terminate. This strong connection between splicing efficiency, 3′ end formation and transcription termination introduces previously unknown layers of regulation to mammalian gene expression.

ACKNOWLEDGMENTS

We thank Petros Patsali for sharing the MEL-*HBB* ^{WT} and MEL-*HBB* $^{IVS-110(G>A)}$ cell lines, Michael Antoniou for sharing an annotation of the GLOBE vector, and John Conboy for advice on erythroblast fractionation. We thank Hagen Tilgner, Tucker Carrocci, David Phizicky, Tara Alpert, and Telmo Henriques for helpful discussions and comments on the manuscript, and Ethan Brown for help with preparation of LRS figures. This work was initiated through pilot funding from NIDDK under Grant U54DK106857 to the Yale Cooperative Center of Excellence in Hematology (to K.M.N.). It was further supported by the National Institutes of Health (NIH R01 GM112766 to K.M.N) and startup funds provided by Harvard Medical School to K.A. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. K.A.R. is supported by a Postgraduate Scholarship from the Natural Sciences and Engineering Research Council of Canada (NSERC), and C.M. is supported by a National Science Foundation Graduate Research Fellowship (DGE1745303).

AUTHOR CONTRIBUTIONS

Conceptualization, K.M.N. and K.A.R.; Investigation, K.A.R. and C.M.; Data Curation, K.A.R. and C.M.; Writing -- Original Draft, K.A.R. and K.M.N.; Writing -- Review & Editing, K.A.R., C.M., K.A., and K.M.N.; Visualization, K.A.R. and C.M.; Supervision, K.A. and K.M.N.; Funding Acquisition, K.A. and K.M.N.

DECLARATION OF INTERESTS

The authors declare no competing interests.

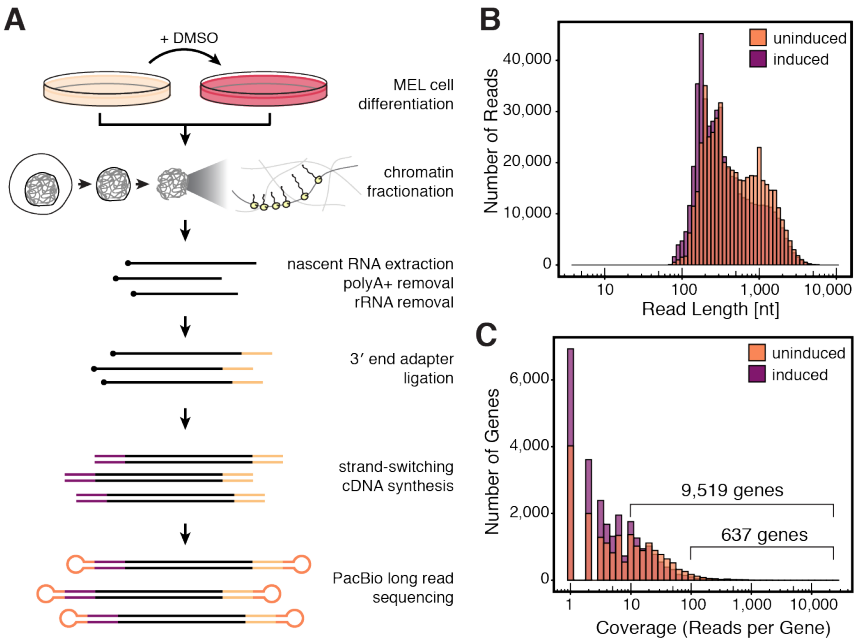

Figure 1. Long read sequencing of nascent RNA with PacBio technology yields high coverage in differentiating mouse erythroblasts

(A) Schematic of nascent RNA isolation and sequencing library generation. MEL cells are treated with 2% DMSO to induce erythroid differentiation, then cells are fractionated to purify chromatin, and chromatin-associated nascent RNA is subsequently depleted of polyadenylated and ribosomal RNAs. An adapter is ligated to the 3′ ends of all remaining RNAs, then a strand-switching reverse transcriptase is used to create double-stranded cDNA that is the input for PacBio library preparation. **(B)** Read length distribution of PacBio long reads. **(C)** Read depth of PacBio long reads. For **(B)** and **(C)**, data represent two biological replicates and two technical replicates combined. See also **Figures S1, Figure S2**, and **Table S1**.

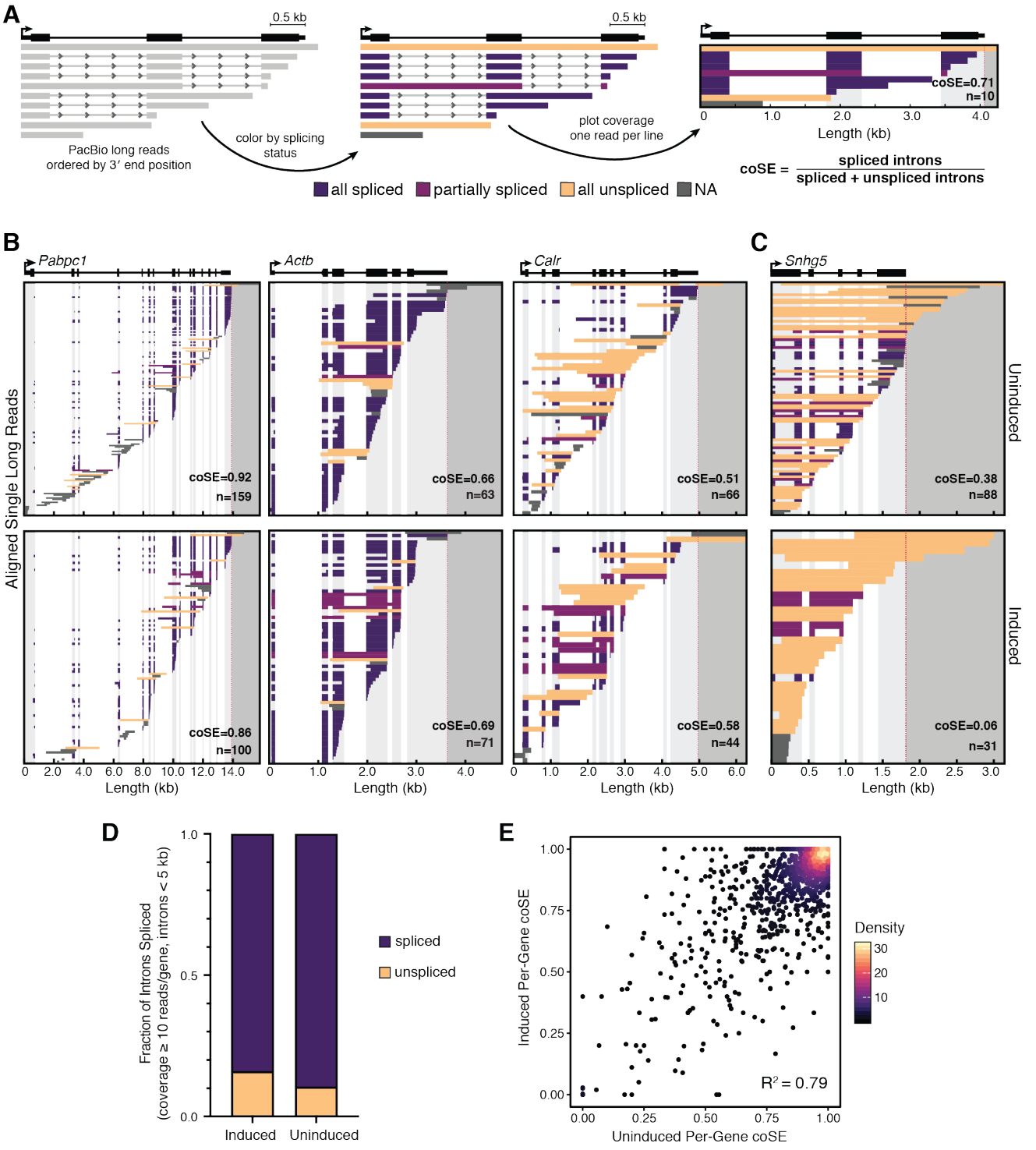

Figure 2. Long read sequencing reveals a range of co-transcriptional splicing efficiencies

(A) Illustration of LRS data. Gene diagram is shown at the top, with the black arrow indicating the TSS. Reads are aligned to the genome and ordered by 3′ end position. Each read is colored by its splicing status: reads that cross one or more introns but contain no splice junction are “all unspliced” (yellow); reads that cross one or more intron and are spliced at every intron are “all spliced” (dark purple); reads that cross two or more introns but are not spliced at all introns are “partially spliced” (light purple); and reads that do not span an entire intron are designated “NA” (gray). Each horizontal row represents one

read. Regions of missing sequence (e.g. spliced introns) are transparent. Light gray shading indicates regions of exons, and dark gray shading indicates the region downstream of the annotated PAS (dotted red line). N is the number of individual long reads aligned to each gene, and coSE is the calculated co-transcriptional splicing efficiency (coSE) for each gene; coSE is defined as the number of spliced introns/(spliced introns + unspliced introns) across all reads aligned to a gene. **(B,C)** LRS data are shown for uninduced (top) and induced (bottom) MEL cells at four representative genes: **(B)** protein-coding genes *Pabpc1*, *Actb*, and *Calr*, and **(C)** the lncRNA *Snhg5*. **(D)** Fraction of introns spliced co-transcriptionally. Only introns from genes with ≥ 10 aligned reads and introns less than 5 kb in uninduced and induced conditions are considered (n = 1,210 genes; 15,911 introns) **(E)** Per-gene coSE in uninduced and induced conditions. Each point represents a single gene as defined in **(D)**. R^2 is Pearson correlation coefficient. For **(B)** - **(E)**, data represent two biological replicates and two technical replicates combined. See also **Figure S3**.

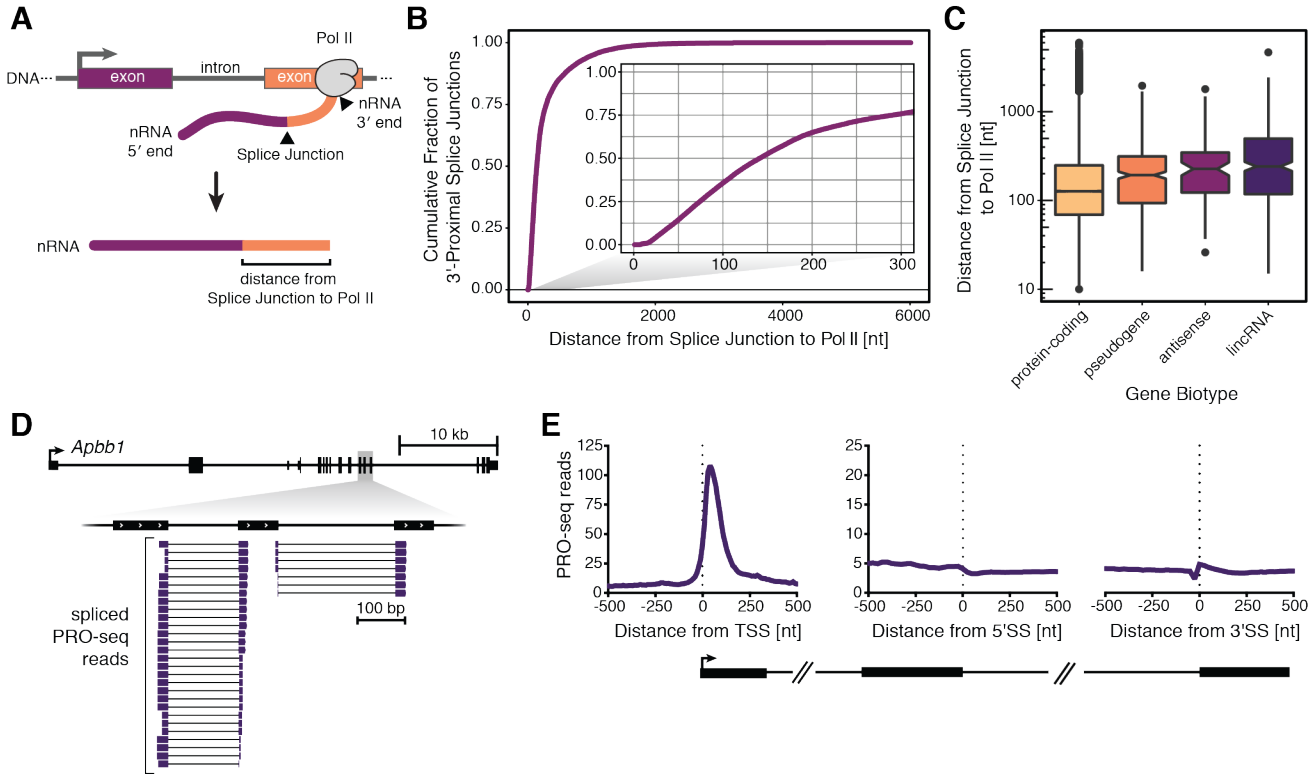

Figure 3. Splicing can occur when elongating Pol II is just downstream of a newly transcribed 3′ splice site

(A) Schematic defining the distance from the 3′ end of a nascent RNA (nRNA) to the most 3′-proximal splice junction. 3′ end sequence reports the position of Pol II when nascent RNA was isolated. **(B)** Distance (nt) from the 3′-most splice junction to Pol II position is shown as a cumulative fraction. Inset is a zoom in on the first 300 nt past the 3′SS (n = 184,456 observations). **(C)** Distance (nt) from the 3′-most splice junction to Pol II position is shown categorized by gene biotype (n = 108,280 protein-coding genes, 440 pseudogenes, 90 antisense genes, 350 lincRNA genes). For **(B)** and **(C)**, data represent two biological replicates and two technical replicates in uninduced and induced cells combined. **(D)** Genome browser view showing spliced PRO-seq reads aligned to the *Apbb1* gene, where 3′ ends of reads represent the position of elongating Pol II. Only spliced reads, filtered from all reads, are shown. **(E)** PRO-seq 3′ end coverage is shown aligned to active transcription start sites (TSS), 5′ splice sites (5′SS), and 3′ splice sites (3′SS). For **(D)** and **(E)**, data represent three biological replicates for uninduced and induced cells combined. See also **Figure S4**.

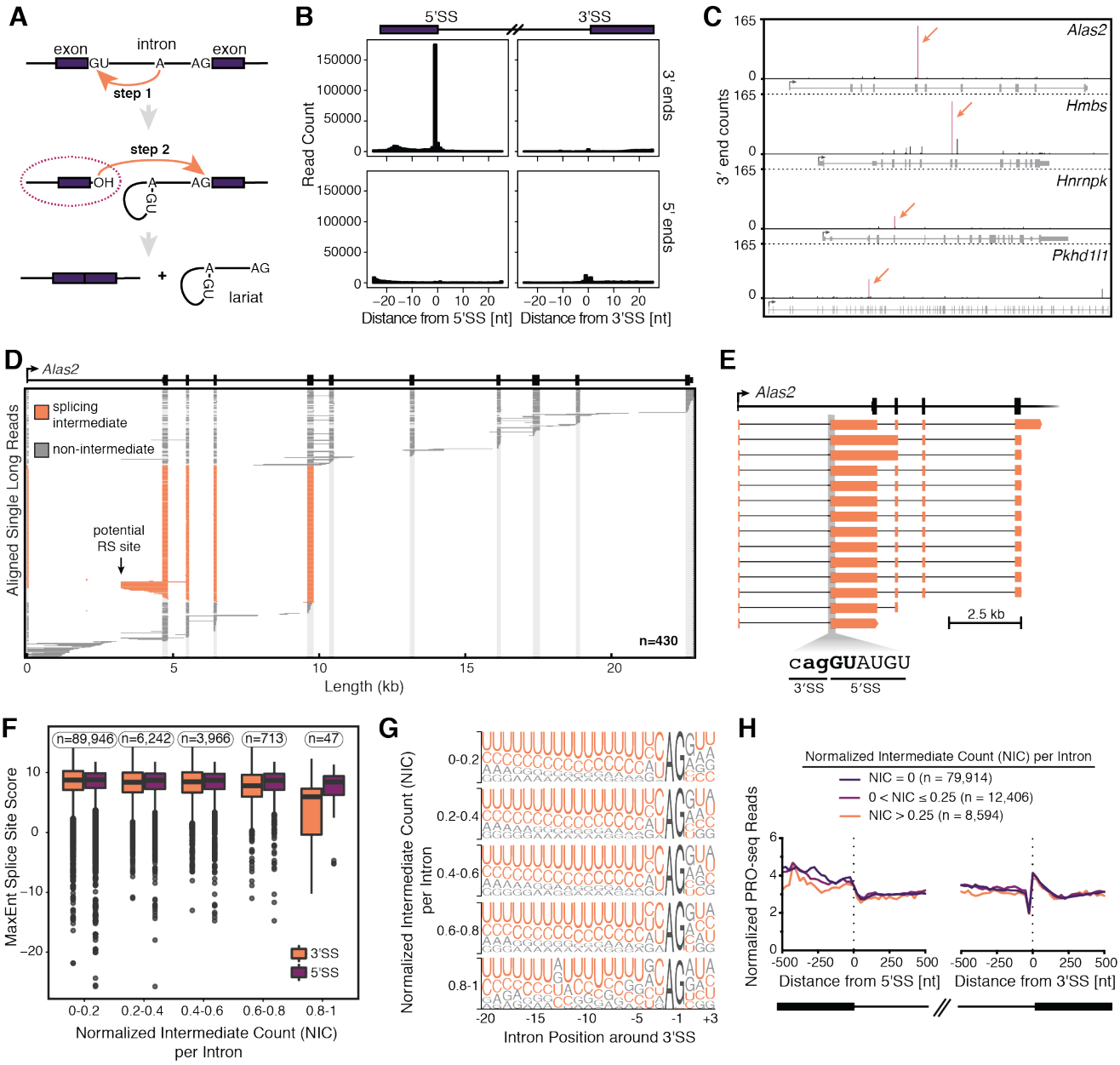

Figure 4. Splicing intermediates are abundant at introns with weak 3′ splice sites

(A) Schematic definition of first step splicing intermediates (dotted red oval), which have undergone the first step of splicing and have a free 3′-OH moiety that can be ligated to the 3′ end DNA adapter. Splicing intermediate reads are characterized by a 3′ end at the -1 position of a 5′SS (last nucleotide of the upstream exon). **(B)** Coverage of long read 3′ ends (top panels) and 5′ ends (bottom panels) aligned to all mm10 5′SSs (left) and 3′SSs (right). **(C)** Coverage of long read 3′ ends across four example genes: *Alas2*, *Hmbs*, *Hnrnpk*, and *Pkhd1l1*. Orange arrows indicate the positions where the most abundant splicing intermediates are observed in each gene. **(D)** Individual long reads are shown for the gene *Alas2*. Diagram is similar to **Figure 2**, except that individual reads are colored depending on whether they are splicing intermediates (orange) or not (gray). Data for uninduced and induced cells are shown combined. Potential recursive splicing site is indicated by an arrow; recursively spliced reads are shown in detail in **(E)**. **(F)** MaxEnt splice site scores for 5′SS (in purple) and 3′SS (in orange) for all introns is shown categorized by the normalized intermediate count (NIC) at each intron (n = the number of introns in each

category). For **(F)** - **(H)**, NIC is defined as the number of splicing intermediate reads divided by the sum of the intermediate reads plus spliced reads at a 5′SS. **(G)** Sequence logos of the -20 to +3 nt region around the 3′SS used to calculate the 3′SS score in **(F)**. 3′SS dinucleotide is shown in dark gray, pyrimidines are shown in orange, and purines are shown in light gray. For **(B)** - **(G)**, data represent two biological replicates and two technical replicates from uninduced and induced cells combined. **(H)** PRO-seq 3′ end coverage aligned to 5′SSs, and 3′SSs for introns in three categories of NIC values (n = the number of introns in each category). Data represent three biological replicates for uninduced and induced cells combined. See also **Figure S5**.

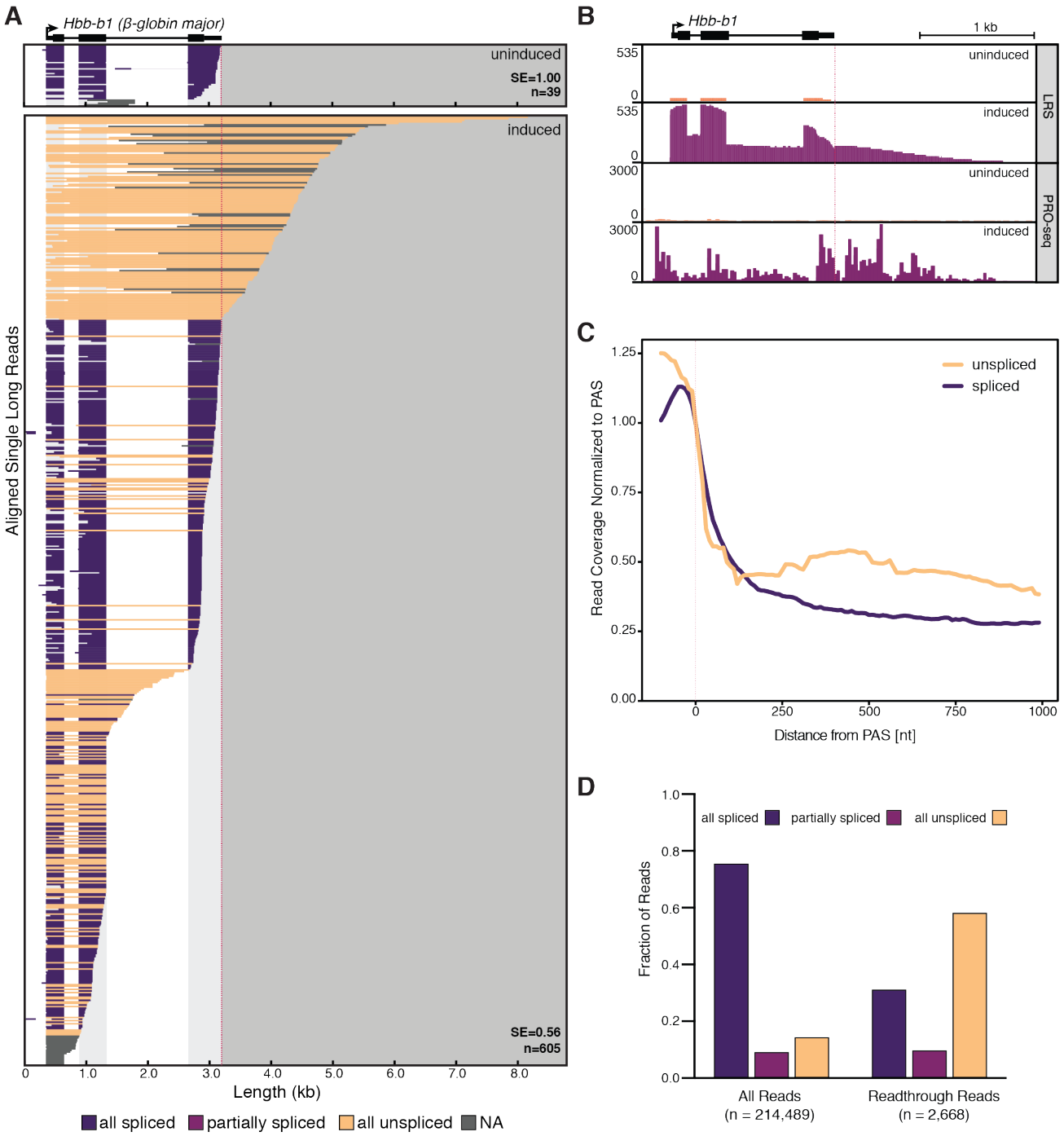

Figure 5. Poor splicing efficiency is associated with readthrough transcription

(A) Individual long reads are shown for the major β-globin gene (*Hbb-b1*). Diagram is as described in **Figure 2**. **(B)** LRS coverage and PRO-seq 3′ end coverage in uninduced and induced cells is shown at the *Hbb-b1* gene. Scale at the left indicates coverage in number of reads, and red dotted line indicates PAS. We note that the duplicated copies of β-globin in the genome (*Hbb-b1* and *Hbb-b2*) impedes unique mapping of short PRO-seq reads in the coding sequence, artificially reducing gene body reads. **(C)** Long read coverage in the region downstream of the last annotated PAS is shown normalized to coverage at the PAS. Unspliced coverage shows reads with no splice junctions (n = 626,130 reads), and spliced coverage shows reads with one or more splice junctions (n = 368,031 reads). Red dotted line indicates

PAS. **(D)** Fraction of all reads (left) and readthrough reads (right) categorized by splicing status (as described in **Figure 2**). A readthrough read is categorized as having a 5′ end within a gene region and a 3′ end greater than 100 nt downstream of the gene end (n = number of reads in each category). For **(A-D)**, LRS data represent two biological replicates and two technical replicates combined, and PRO-seq data represent three biological replicates combined. For **(D-E)**, data from uninduced and induced cells are shown combined. See also **Figure S6**.

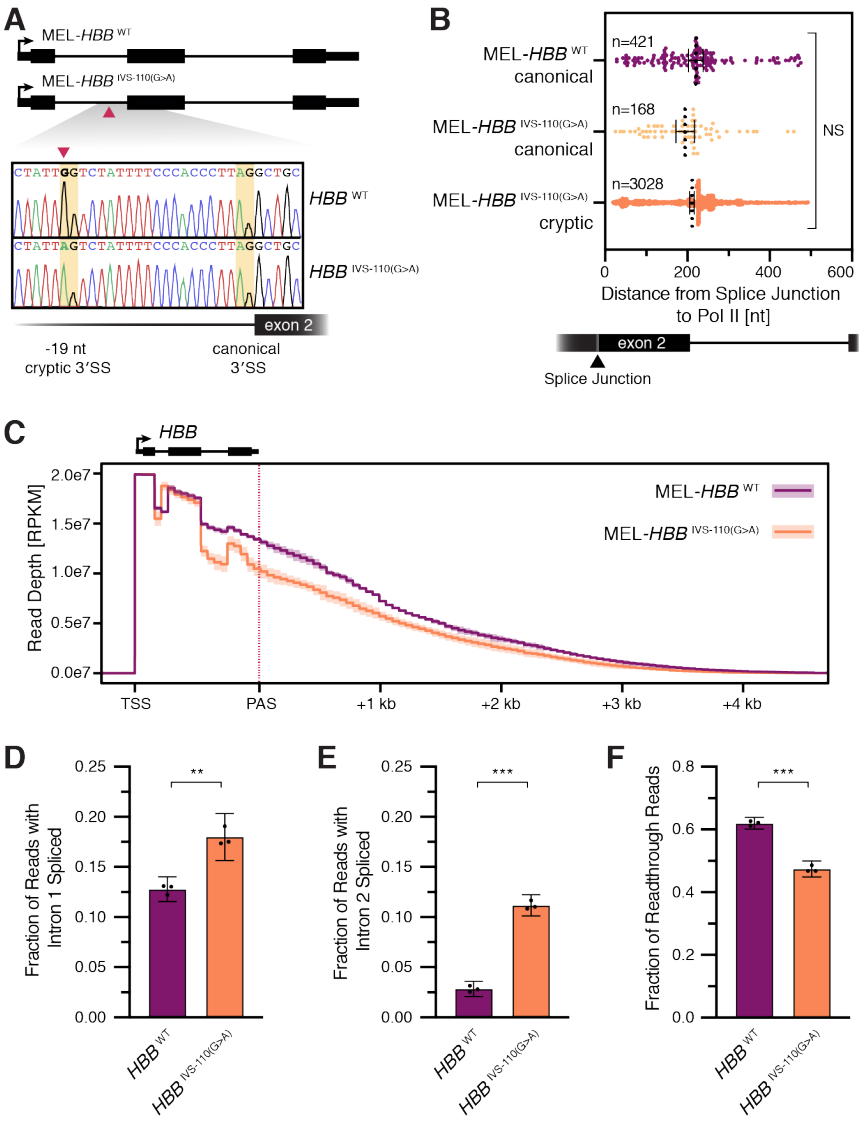

Figure 6. Increased splicing of intron 1 in β-globin thalassemia allele leads to decreased readthrough transcription

(A) Top: schematic describing two engineered MEL cell lines. MEL-*HBB* ^{WT} contains an integrated copy of a wild type human globin minigene. In MEL-*HBB* $^{IVS\text{-}110(G>A)}$, a single point mutation at the +110 position of intron 1 (red triangle) mimics a disease-causing thalassemia allele. Bottom: Sanger sequencing of the *HBB* minigene coding strand shows that a G>A mutation leads to a new AG dinucleotide in the RNA produced from the *HBB* $^{IVS\text{-}110(G>A)}$ gene. This AG is utilized as a cryptic 3′SS and is located 19 nt upstream of the canonical 3′SS. **(B)** Distance from exon1-exon2 splice junctions to 3′ ends for all long reads detected, using targeted LRS of the integrated *HBB* loci. For the MEL-*HBB* $^{IVS\text{-}110(G>A)}$ cell line, data are separated by use of the canonical (yellow) or cryptic (orange) splice site. Only reads that have 3′ ends before the end of intron 2 are considered in this analysis. Data represent three biological replicates combined; dotted black line represents mean of replicates and solid black bars represent 95% confidence intervals, n = number of reads. Significance tested by nested 1-way ANOVA. NS = not significant; p = 0.2726. **(C)** Normalized coverage of long reads mapped to the *HBB* allele in MEL-*HBB* ^{WT} cells (purple) and MEL-*HBB* $^{IVS\text{-}110(G>A)}$ cells (orange). Solid lines are the mean of three biological replicates and shaded windows are 95% confidence intervals. Red dotted line indicates PAS. **(D-F)** The fraction of targeted long

reads from MEL-*HBB* ^{WT} cells (purple) and MEL-*HBB* $^{IVS\text{-}110(G>A)}$ cells (orange) that: **(D)** have *HBB* intron 1 spliced, **(E)** have *HBB* intron 2 spliced, and **(F)** are classified as readthrough transcripts. Reads containing the cryptic and canonical splice junction in the MEL-*HBB* $^{IVS\text{-}110(G>A)}$ cells are shown combined in **(D)**. For **(D-F)** black points represent three biological replicates, bars indicate mean of biological replicates, and error bars represent 95% confidence intervals. Significance tested by unpaired t-test: ** $p=0.0011$, *** $p<0.0001$. See also **Figure S7**.

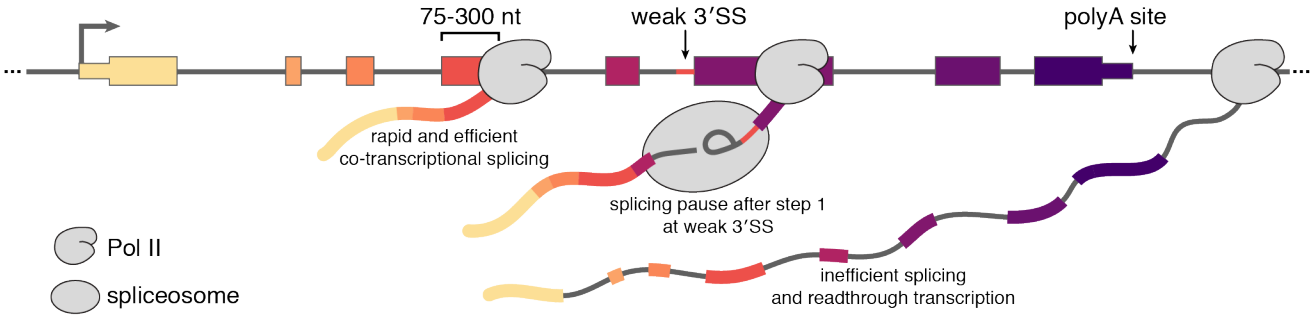

Figure 7. Rapid and efficient co-transcriptional splicing enhances productive gene expression
Model describing the variety of co-transcriptional splicing efficiencies observed during mouse erythropoiesis. When splicing occurs co-transcriptionally, it can occur rapidly in a spatial window where Pol II has transcribed on average 75-300 nt past a 3′SS. However, when a weak 3′SS is encountered, the spliceosome itself may stall between step 1 and step 2 of splicing, allowing splicing intermediates to be detected more readily. Additionally, inefficient splicing is associated with failure to cleave and terminate transcription at the PAS.

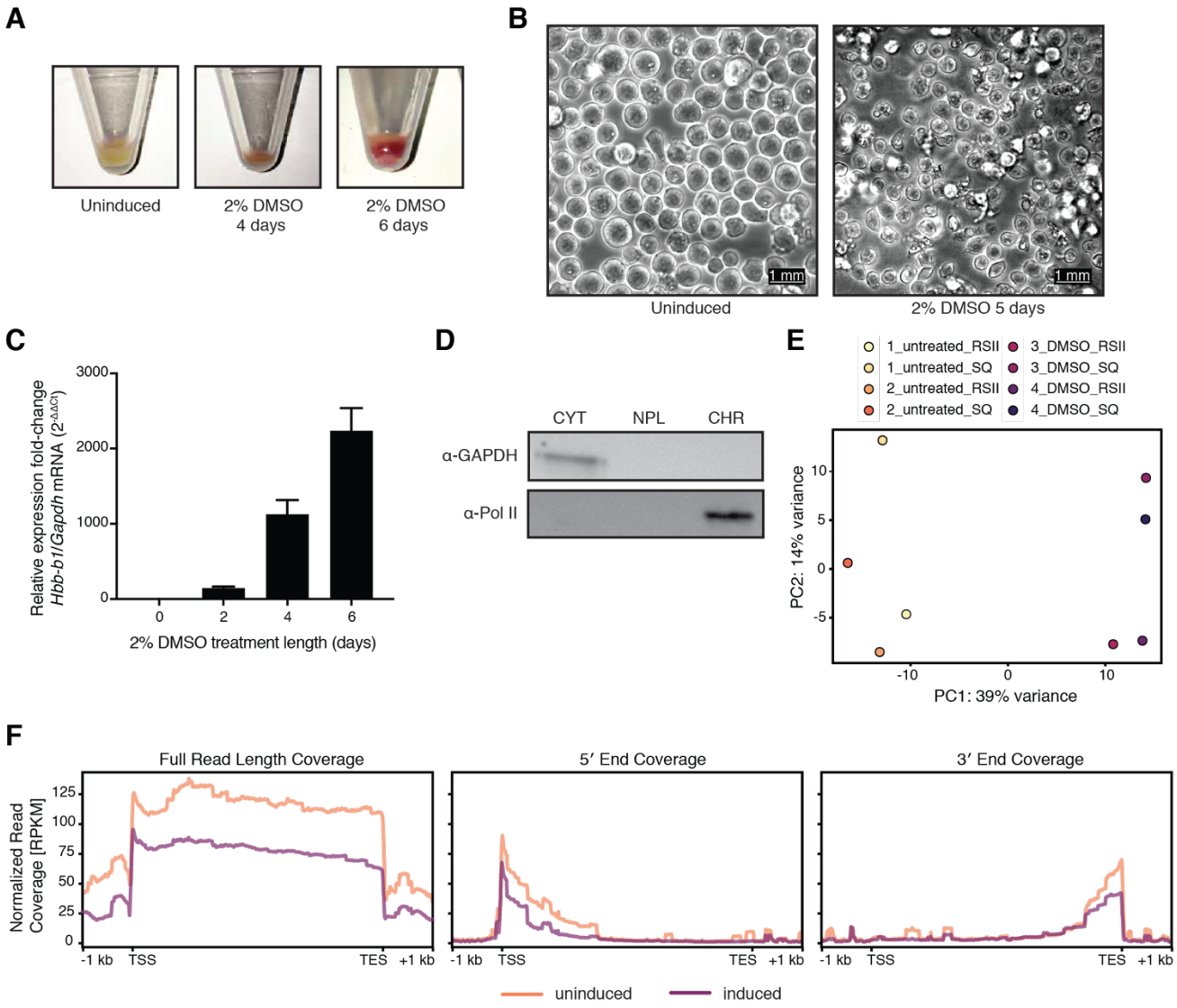

Figure S1. Related to Figure 1. DMSO treatment induces erythroid differentiation

(A) MEL cells after culturing in 2% DMSO for 0, 4, or 6 days. **(B)** Bright field microscopy of MEL cells uninduced (left) and induced for 5 days (right). Scale bar is 1 mm. **(C)** RT-qPCR measurement of *Hbb-b1* (β-globin) mRNA levels relative to *Gapdh* mRNA from total RNA in MEL cells treated with 2% DMSO for 0, 2, 4, or 6 days. Data represent mean of 3 technical replicates, and error bars represent SEM. **(D)** Western blot of subcellular fractions collected during chromatin fractionation (CYT = cytoplasm, NPL = nucleoplasm, CHR = chromatin). **(E)** Principal component analysis of LRS data comparing biological replicates (uninduced vs. induced) and technical replicates (data collected on RSII vs. Sequel [SQ] flowcells). Input data were PacBio long reads quantified using Salmon and analyzed by DESeq2. **(F)** Normalized read coverage (RPKM) of long read 5′ ends (left), full reads (middle), and 3′ ends (right) is shown across a metaplot of all mm10 genes +/- 1 kb (TSS = transcription start site, TES = transcription end site). Data represent two biological replicates and two technical replicates combined.

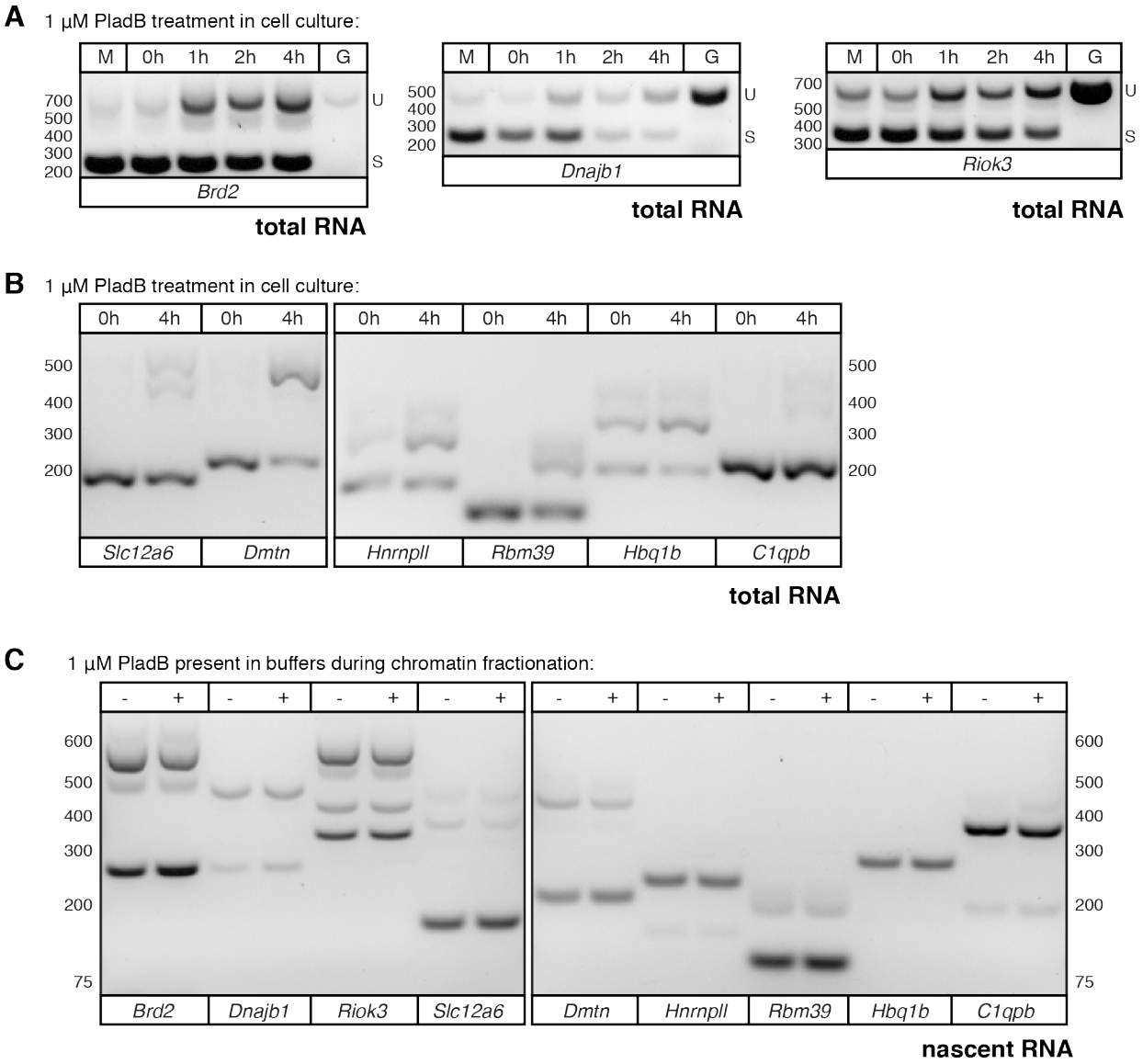

Figure S2. Related to Figure 1. Splicing is not ongoing during chromatin purification and nascent RNA isolation

(A) RT-PCR on total RNA collected from MEL cells treated with 1 μM splicing inhibitor Pladienolide B in cell culture for 0, 1, 2, and 4 hours. Total RNA was reverse transcribed with random hexamers, and PCR primers span a single intron in each gene. Three representative genes are shown (left: *Brd2*, middle: *Dnajb1*, and right: *Riok3*). M indicates mock treatment with DMSO, and G indicates amplification of genomic DNA to determine the size of unspliced RNA. U indicates size of unspliced amplicon, and S indicates size of spliced amplicon. **(B)** RT-qPCR from total RNA (as in **(A)**), showing six additional genes (*Slc12a6*, *Dmtn*, *Hnrnpll*, *Rbm39*, *Hbq1b*, *C1qbp*) after treatment with 1 μM Pladienolide B for 0h and 4h. **(C)** RT-PCR on nascent RNA isolated from chromatin which was fractionated in the absence (-) or presence (+) of 1 μM Pladienolide B. Nascent RNA was reverse transcribed with random hexamers and PCR primers were the same as in **(A)** and **(B)**.

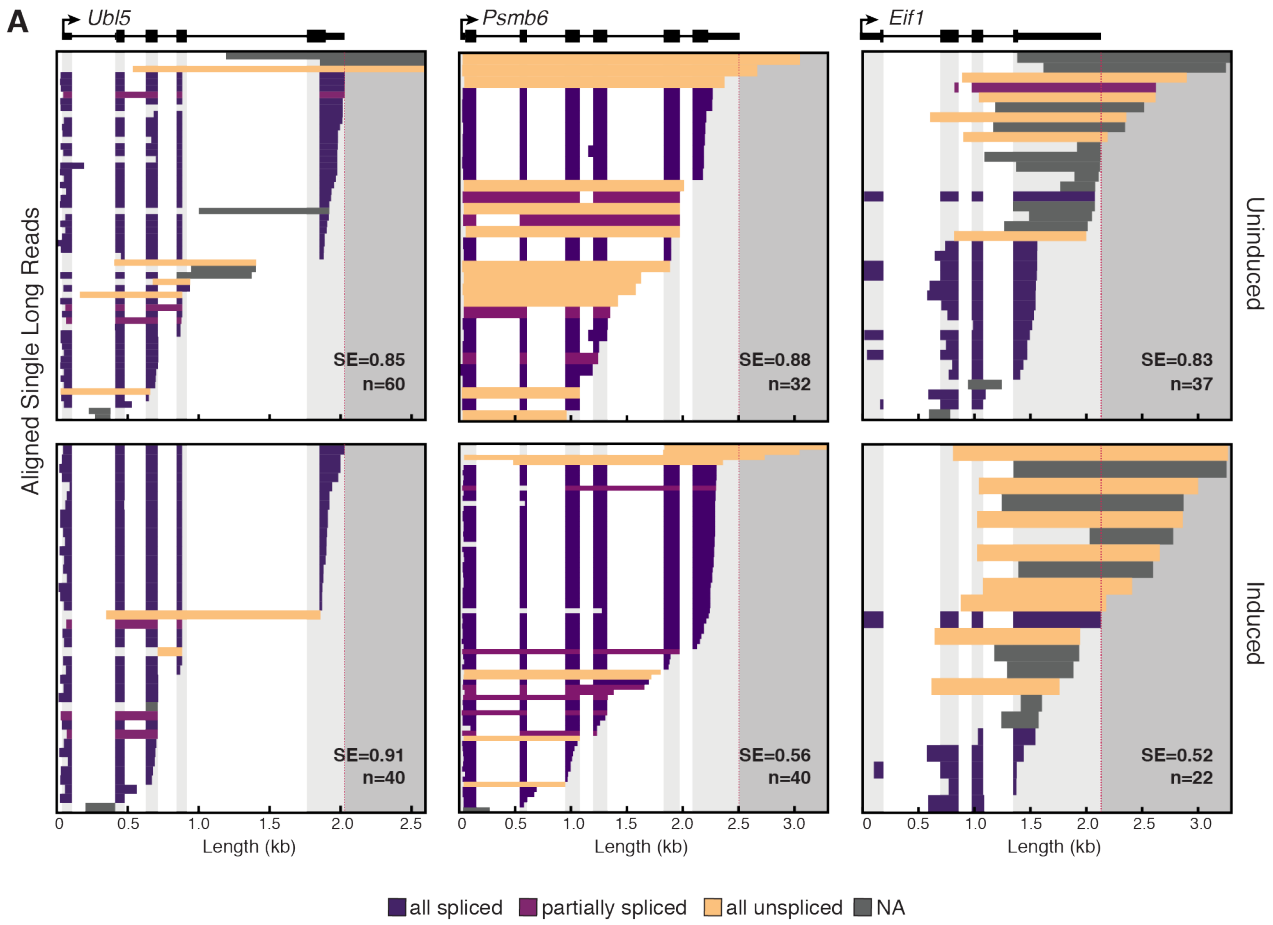

Figure S3. Related to Figure 2. Long read sequencing reveals gene-specific splicing efficiencies
(A) Individual long reads for protein-coding genes *Ubl5*, *Psmb6*, and *Eif1* are shown as described in **Figure 2**. Data represent two biological replicates and two technical replicates combined.

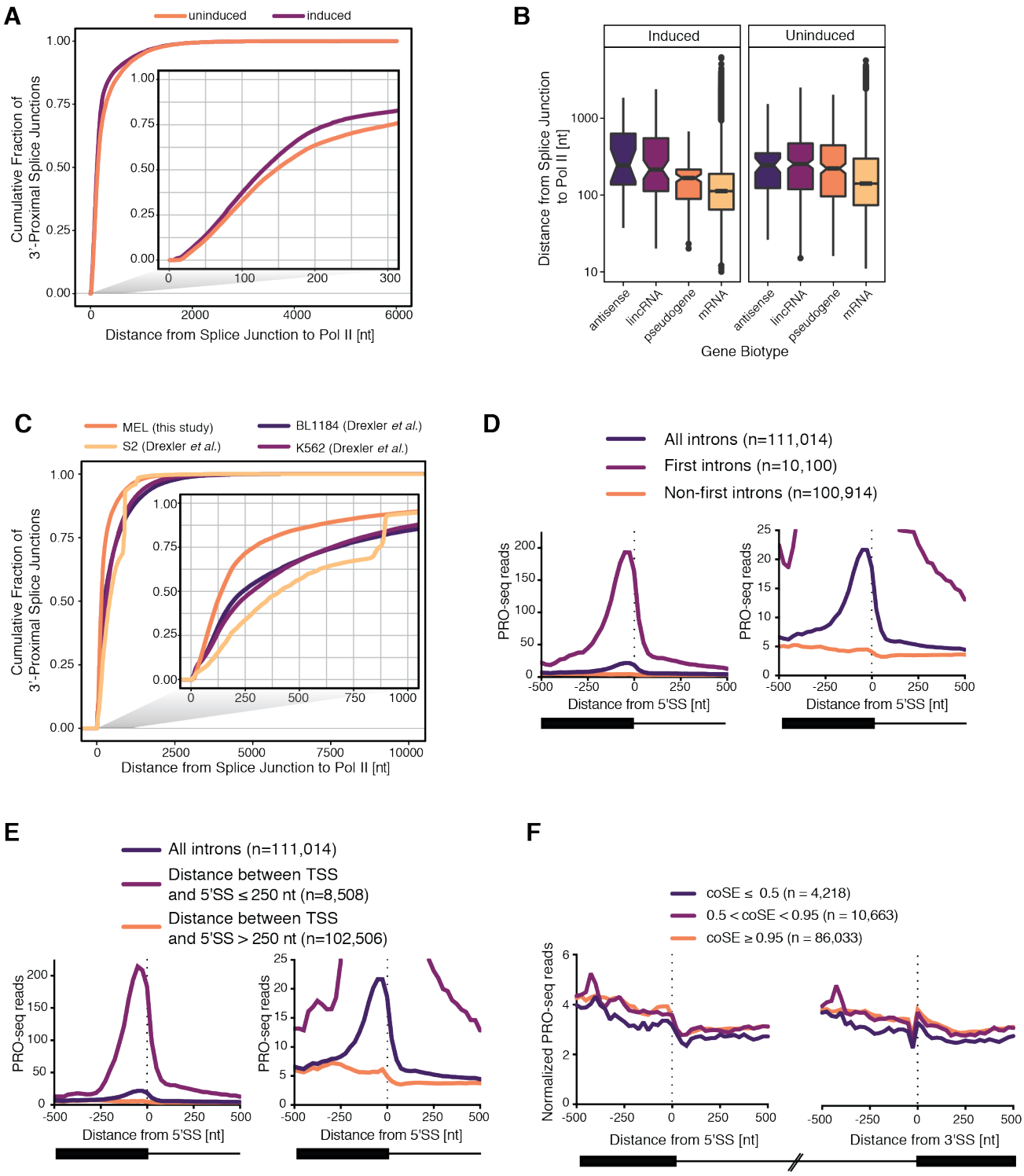

Figure S4. Related to Figure 3. PRO-seq signal at 5′SS is overlapping with TSS signal for short first introns

(A) Distance (nt) from the most 3′-proximal splice junction to Pol II position is shown as a cumulative fraction. Inset is a zoom in on the first 300 nt. Analysis is the same as in **Figure 3B**, but with reads from uninduced and induced cells plotted separately. **(B)** Distance (nt) from the most 3′-proximal splice junction to Pol II position is shown categorized by gene biotype. Analysis is the same as in **Figure 3C**, but with reads from uninduced and induced cells plotted separately. **(C)** Distance (nt) from the most 3′-

proximal splice junction to Pol II position is shown as a cumulative fraction. Inset is a zoom in on the first 1000 nt. Analysis is the same as in **Figure 3B**, but with data from this study and nanoCOP data from Drexler *et al*. **(D)** PRO-seq 3′ end coverage aligned to 5′SSs for all introns from active transcripts (dark purple), first introns only (light purple), and non-first introns (orange). Right panel shows data scaled to show all introns. **(E)** PRO-seq 3′ end coverage aligned to 5′SSs for all introns (dark purple), introns where the distance from the TSS to the 5′SS is ≤ 250 nt (light purple), and introns where the distance to the TSS to the 5′SS is > 250 nt (orange). Right panel shows data scaled to show all introns. **(F)** PRO-seq 3′ end coverage around splice sites separated by per-intron coSE. For each intron, coSE was calculated as the number of reads fully overlapping the intron that were spliced divided by the total number of reads fully overlapping the intron. For **(A-B)**, data represent two biological replicates and two technical replicates combined. For **(C)**, data from this study represent two biological replicates and two technical replicates from uninduced and induced cells combined. Data from Drexler et al. represent two samples in human BL1184 cells, 9 samples in *Drosophila* S2 cells, and 11 samples in human K562 cells. in or **(D-F)** data represent three biological replicates from uninduced and induced cells combined, and n = number of introns introns in each category.

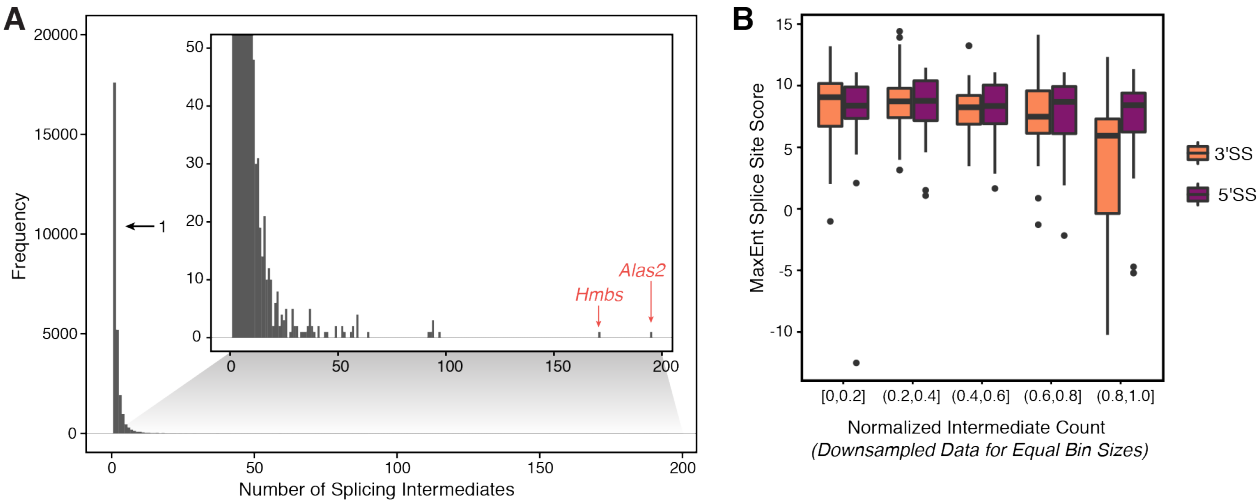

Figure S5. Related to Figure 4. Splicing intermediates are rare at most introns, but highly abundant at a few introns

(A) Histogram showing frequency of splicing intermediates upstream of each unique intron in the mm10 genome. Most introns show 0 or 1 splicing intermediates, while some introns, like those in *Alas2* and *Hmbs*, indicated with orange arrows, exhibit nearly 200 splicing intermediates reads. **(B)** MaxEnt splice site scores for 5′SS (purple) and 3′SS (orange) for a randomly downsampled subset of introns is shown categorized by the normalized intermediate count (NIC) at each intron. The number of introns in each bin is equal to the number of introns in the highest NIC category in **Figure 4F** (n=47).

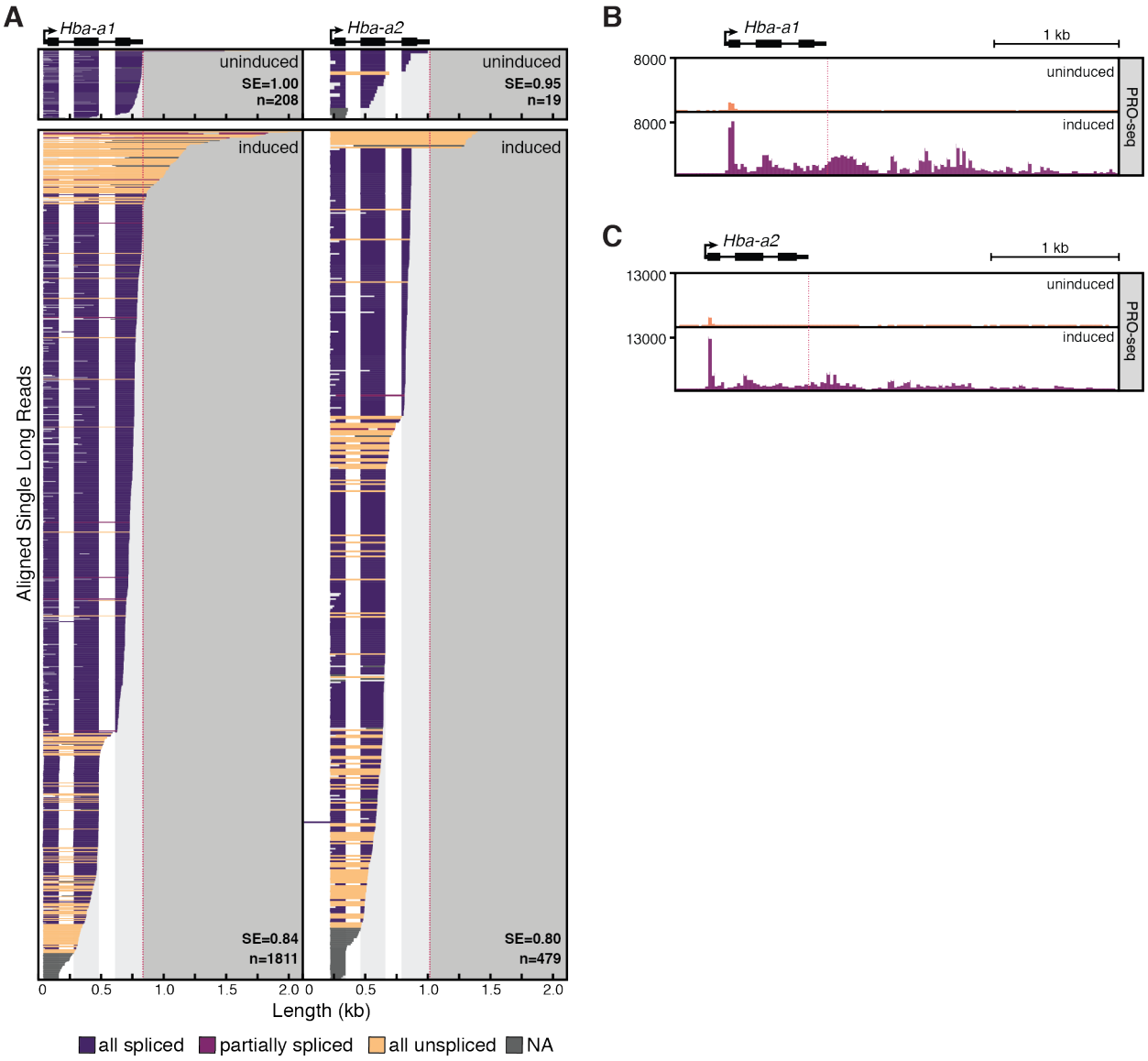

Figure S6. Related to Figure 5. α-globin genes exhibit readthrough transcription

(A) Individual long reads are shown for the α-globin 1 gene (*Hba-a1*) and α-globin 2 gene (*Hba-a2*). Diagrams are as described in **Figure 2**. Data represent two biological replicates and two technical replicates combined. **(B-C)** PRO-seq 3′ end read coverage is shown downstream of the *Hba-a1* **(B)** and *Hba-a2* **(C)** gene loci. We note that the duplicated copies of α-globin in the genome (*Hba-a1* and *Hba-a2*) impedes unique mapping of short PRO-seq reads in the coding sequence, artificially reducing gene body reads. Red dotted line indicates PAS. Data represent three biological replicates combined.

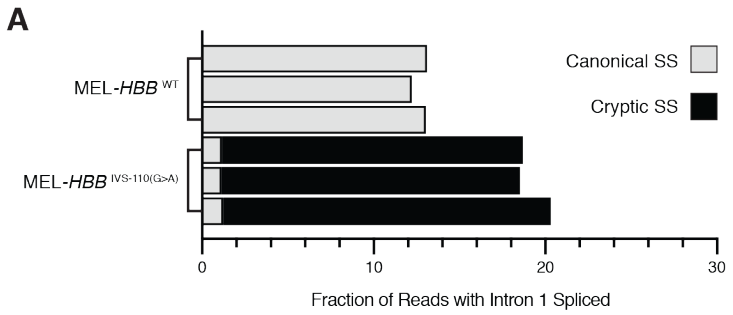

Figure S7. Related to Figure 6. The majority of MEL-*HBB* ^{WT} intron 1 splicing is at the cryptic 3′SS
(A) The fraction of reads spanning intron 1 in MEL-*HBB* ^{WT} or MEL-*HBB* $^{IVS\text{-}110(G>A)}$ that are spliced at either the canonical (gray) or cryptic (black) splice junction. Each bar represents a single biological replicate for each cell type.

METHODS

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Karla Neugebauer (karla.neugebauer@yale.edu), Department of Molecular Biophysics and Biochemistry, Yale University, New Haven CT 06520. This study did not generate new unique reagents.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines, cell culture, and treatments

Murine Erythroleukemia cells (MEL; obtained from Shilpa Hattangadi, Yale School of Medicine) were maintained at 37°C and 5% CO_2 in DMEM + Glutamax medium (GIBCO) containing 100 U/ml penicillin, 100 µg/ml streptomycin (GIBCO), and 10% fetal bovine serum (GIBCO). To induce erythroid differentiation, cells were diluted to 50,000 cells/ml in 10 ml fresh culture medium and incubated as above for 16 hours. DMSO was then added directly to the culture medium to a final concentration of 2% and incubated as above for 5 days. For Pladienolide B treatment, cells were diluted to 50,000 cells/ml in fresh culture medium, then incubated as above for two days until reaching a density of approximately 5 million cells/ml. Pladienolide B (Santa Cruz) dissolved in DMSO was added directly to the culture medium at a final concentration of 1 µM. MEL-*HBB* ^{WT} and MEL-*HBB* $^{IVS\text{-}110(G>A)}$ cell lines are described previously (Patsali et al., 2018), and were maintained and differentiated as above.

METHOD DETAILS

Subcellular Fractionation

Subcellular fractionation was adapted from previously published protocols (Mayer and Churchman, 2017; Pandya-Jones and Black, 2009), with modifications to centrifugation speeds in order to retain intact nuclei. All steps were performed on ice, and all buffers contained 25 uM α-amanitin, 40 U/ml SUPERase.IN, and 1x Roche cOmplete protease inhibitor mix. Briefly, 20 million cells were rinsed once with PBS/1 mM EDTA, then lysed in 250 µl cytoplasmic lysis buffer (10 mM Tris-HCl pH 7.5, 0.05% NP40, 150 mM NaCl) by gently resuspending then incubating on ice for 5 minutes. Lysate was then layered on top of a 500 µl cushion of 24% sucrose in cytoplasmic lysis buffer and spun at 2000 rpm for 10 min at 4°C. The supernatant (cytoplasm fraction) was removed, and the pellet (nuclei) were rinsed once with 500 µl PBS/1 mM EDTA. Nuclei were resuspended in 100 µl nuclear resuspension buffer (20 mM Tris-HCl pH 8.0, 75 mM NaCl, 0.5 mM EDTA, 0.85 mM DTT, 50% glycerol) by gentle flicking, then lysed by the addition of 100 µl nuclear lysis buffer (20 mM HEPES pH 7.5, 1 mM DTT, 7.5 mM $MgCl_2$, 0.2 mM EDTA, 0.3 M NaCl, 1 M Urea, 1% NP-40), vortexed for 2 x 2 seconds, then incubated on ice for 3 min. Chromatin was pelleted by spinning at 14,000 rpm for 2 min at 4°C. The supernatant (nucleoplasm fraction) was removed, and the chromatin was rinsed once with PBS/1 mM EDTA. Chromatin was immediately dissolved in 100 µl PBS and 300 µl TRIzol Reagent (ThermoFisher).

Nascent RNA Isolation

RNA was purified from chromatin pellets in TRIzol Reagent (ThermoFisher) using the RNeasy Mini kit (Qiagen) according to the manufacturer's protocol, including the on-column DNase I digestion. For genome-wide nascent RNA-seq, samples were depleted three times of polyA(+) RNA using the Dynabeads mRNA DIRECT Micro Purification Kit (ThermoFisher), each time keeping the supernatant,

then depleted of ribosomal RNA using the Ribo-Zero Gold rRNA Removal Kit (Illumina). For targeted nascent RNA-seq, polyA(+) and rRNA depletion were omitted.

Western Blotting

Cytoplasm, nucleoplasm, and chromatin fractions from cell fractionation were adjusted to an equal volume with PBS. Nucleoplasm and chromatin fractions were homogenized by sonication, and all samples were spun at 14,000 rpm for 10 min at 4°C before gel loading. Western blots were performed with antibodies against Pol II 4H8 (Santa Cruz Biotechnology) and GAPDH (Santa Cruz Biotechnology).

qPCR

Total RNA was extracted from 10 million cells in TRIzol Reagent (ThermoFisher) according to the manufacturer's protocol after 0, 2, 4, or 6 days of treatment with 2% DMSO as described above. cDNA was generated with SuperScript III Reverse Transcriptase (ThermoFisher) using random hexamer primers (ThermoFisher) according to the manufacturer's protocol. For primers used to amplify *Hbb-b1* and *Gapdh*, see **Table S2**. qPCR reactions were assembled using iQ SYBR Green Supermix (BioRad) and quantified on a Stratagene MX3000P qPCR machine. Expression fold changes were calculated using the ΔΔCt method.

Microscopy

Live cells were imaged in bright field on an Olympus CKX41 microscope.

RT-PCR after Pladienolide B treatment

For total RNA samples, RNA was extracted from approximately 5 million cells treated with Pladienolide B as described above and using TRIzol Reagent (ThermoFisher) according to the manufacturer's protocol. For nascent RNA samples, RNA was extracted from the chromatin pellet after subcellular fraction as described above, except with the addition or not of Pladienolide B to all subcellular fractionation buffers at a final concentration of 1 μM. Poly(A)+ RNA was further depleted from this sample as described above. cDNA was generated from all RNA samples with SSIII RT (ThermoFisher) using random hexamer primers (ThermoFisher) according to the manufacturer's protocol. PCR was performed using Phusion High-Fidelity DNA Polymerase (NEB) according to the manufacturer's protocol. For the list of intron-flanking primers used in these experiments, see **Table S2**.

PacBio Sequencing Library Preparation

<u>Genome-wide nascent RNA sequencing</u>

Nascent RNA was isolated as described above from cells uninduced and treated with 2% DMSO for 5 days. A DNA adapter (**Table S2**) was ligated to 3′ ends of nascent RNA using the T4 RNA ligase kit (NEB) by mixing 50 pmol adapter with 300-600 ng nascent RNA. cDNA was generated from the adapter ligated RNA using the SMARTer PCR cDNA Synthesis Kit (Clontech), replacing the CDS Primer IIA with a custom primer complementary to the 3′ end adapter for first strand synthesis (**Table S2**). cDNA was amplified by 15 cycles of PCR using the Advantage 2 PCR Kit (Clontech), cleaned up using a 1x volume of AMPure beads (Agencourt), then PacBio library preparation was performed by the Yale Center for Genome Analysis using the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences). The library was sequenced on four RSII flowcells and four Sequel 1 flowcells.

<u>HBB targeted nascent RNA sequencing</u>

Nascent RNA was isolated as described above from cells treated with 2% DMSO for 5 days, except that polyA(+) and ribosomal RNA depletion steps were omitted. A DNA adapter was ligated to 3′ ends as above, and custom RT primers were used to add barcodes during reverse transcription with SSIII reverse transcriptase (ThermoFisher; **Table S2**). cDNA was amplified by 26 cycles of PCR using the Advantage 2 PCR Kit (Clontech), but with custom gene-specific forward primers that were complementary to a unique region in the 5′UTR of the human *HBB* gene in combination with the SMARTer IIA primer (Clontech). PCR amplicons were cleaned up with a 2X volume of AMPure beads (Agencourt), and PacBio library preparation was performed at the Icahn School of Medicine at Mt. Sinai Genomics Core Facility using the SMRTbell Template Prep Kit 1.0 (Pacific Biosciences). The library was sequenced on one Sequel 1 flowcell.

PRO-seq Library Preparation

Cell Permeabilization

All buffers were cooled on ice, all steps were performed on ice, and all samples were spun at 300 xg at 4°C unless otherwise noted. MEL cell differentiation was induced as previously described. Uninduced and induced cells were washed with PBS and resuspended in 1 ml Buffer W (10 mM Tris-HCl pH 8.0, 10 mM KCl, 250 mM sucrose, 5 mM $MgCl_2$, 0.5 mM DTT, 10% glycerol), then strained through a 40 µm nylon mesh filter. 9X volume of Buffer P (Buffer W + 0.1% IGEPAL CA-630) was immediately added to each sample, cells were nutated for 2 minutes at room temperature, then spun for 4 minutes. Cells were washed in Buffer F (50 mM Tris-Cl pH 8.3, 40% glycerol, 5 mM $MgCl_2$, 0.5 mM DTT, 1 µL/mL SUPERase.In [ThermoFisher]), then resuspended in Buffer F at a final volume of 1 x 10^6 permeabilized cells per 40 µL. Samples were flash frozen in liquid nitrogen and stored at -80°C.

Library Generation

One million permeabilized uninduced and induced MEL cells were spiked with 5% permeabilized *Drosophila* S2 cells for data normalization and used as input for PRO-seq. Three biological replicates were generated per treatment condition. Nascent RNA was labeled through a biotin-NTP run-on: permeabilized cells was added to an equal volume of a 2X run-on reaction mix (10 mM Tris-HCl pH 8.0, 300 mM KCl, 1% Sarkosyl, 5 mM $MgCl_2$, 1 mM DTT, 200 µM biotin-11-A/C/G/UTP (Perkin-Elmer), 0.8 U/µL SUPERase.In [ThermoFisher]), and incubated at 30°C for 5 mins. RNA was isolated using the Total RNA Purification Kit (Norgen Biotek Corp). Fragmentation of isolated RNA was performed by base hydrolysis with 0.25 N NaOH for 9 minutes on ice, followed by neutralization with 1X volume of 1 M Tris-HCl pH 6.8. To select for nascent RNAs, 48 µL of washed Streptavidin M-280 magnetic beads (ThermoFisher) in binding buffer (300 mM NaCl, 10 mM Tris-HCl pH 7.4, 0.1% Triton X-100) was added to the fragmented RNA, and samples were rotated at room temperature for 20 minutes. The Streptavidin M-280 magnetic beads were washed twice in each of the following three buffers: high salt buffer (2 M NaCl, 50 mM Tris- HCl pH 7.4, 0.5% Triton X-100), binding buffer (above), and low salt buffer (5 mM Tris-HCl pH 7.4, 0.1% Triton X-100). Beads were resuspended in TRIzol Reagent (ThermoFisher) and heated at 65°C for 5 mins twice to elute the RNA from the beads. A subsequent ethanol precipitation was performed for RNA purification. Nascent RNA was resuspended in 10 uM of the VRA3 3′ end adapter (**Table S2**). 3′ end ligation was performed using T4 RNA ligase I (NEB) for 2 hours at room temperature. A second Streptavidin M-280 magnetic bead binding was performed to enrich for ligated nascent RNAs. The beads were subsequently washed twice in high, binding, and low salt buffers, then once in 1X ThermoPol Buffer (NEB). To prepare nascent RNA for 5′ end adapter ligation, the 5′ ends of the RNA were decapped and repaired. 5′ end decapping was performed using RNA 5′ Pyrophosphohydrolase

(NEB) at 37°C for 1 hour. The beads were washed once in high and low salt buffer, then once in 1X T4 PNK Reaction Buffer (NEB). Samples were treated with T4 Polynucleotide Kinase (NEB) for 1 hour at 37°C for 5′-hydroxyl repair. Next, T4 RNA ligase I (NEB) was used to ligate the reverse 5′ RNA adapter VRA5 (**Table S2**) as described previously. Following the 5′ end ligation, beads were washed twice in high, binding, and low salt buffers, then once in 0.25X FS Buffer (ThermoFisher). Reverse transcription was performed using Superscript IV Reverse Transcriptase (ThermoFisher) with 25 pmol of the Illumina TRU-seq RP1 Primer (**Table S2**). The RT product was eluted from the beads by heating the samples twice at 95°C for 30 seconds. All libraries were amplified by 12 cycles of PCR with 12.5 pmol of Illumina TRU-seq RPI-index primers, excess RP1 primer, and Phusion Polymerase (NEB). The amplified library was purified using the ProNex Size-Selective Purification System (Promega) and sequenced using NextSeq 500 machines in a mid-output 150 bp cycle run.

Genome assembly and annotation

For all analyses in this study, except where noted with the *HBB* targeted LRS, the mouse reference genome GRCm38/mm10 genome assembly was used with the GENCODE VM20 annotation. PRO-seq spike-in data were mapped to the dm3 reference genome.

Long Read Sequencing Data Preprocessing

<u>Genome-wide nascent RNA sequencing</u>

Combined consensus sequence (CCS) reads were generated in FASTQ format, and Porechop was used to separate chimeric reads and trim external adapters with the SMRTer IIA sequences AAGCAGTGGTATCAACGCAGAGTAC and GTACTCTGCGTTGATACCACTGCTT with settings --extra_end_trim 0 --extra_middle_trim_good_side 0 --extra_middle_trim_bad_side 0 --min_split_read_size 100. Cutadapt was used to remove the unique 3′ end adapter on all reads in two rounds of filtering. First any reads with the adapter at the 3′ end were trimmer with settings -a CTGTAGGCACCATCAAT -e 0.1 -m 15 --untrimmed-output=untrimmed.fastq, and any reads which did not contain the full adapter were retained and their reverse complement was generated. Then, a second round of filtering with cutadapt using the settings -a CTGTAGGCACCATCAAT -e 0.1 -m 15 --discard-untrimmed was used to remove adapters from the reverse complement reads, and reads without the 3′ adapter were discarded. This ensures that each read contains a successfully ligated 3′ adapter which marks Pol II position, and since sequencing occurs in both forward and reverse orientations randomly, it places all reads in the correct 5′ to 3′ orientation. Reads from the two adapter trimming steps were combined into a single file, then Prinseq-lite was used to remove PCR duplicates with settings -derep 1. Prinseq-lite was used again to trim 6 non-templated nucleotides added at the 5′ end by the strand-switching reverse transcriptase and the 5 nt of the 3′ end adapter UMI with settings -trim_left 6 -trim_right 5. Reads were then mapped to the mm10 genome using minimap2 with settings -ax splice -uf -C5 --secondary=no, and the resulting SAM files were converted to BAM and BED files for downstream analysis using samtools and bedtools. Reads overlapping the 7SK genomic region (chr9:78175302,78175633 in the mm10 genome) were filtered using samtools before all downstream analyses. All data generated using Nanopore sequencing from Drexler *et al.* (GEO: GSE123191) were downloaded in FASTQ format and mapped to either the hg38 or dm6 genome using minimap2 with settings -ax splice -ut -k14, then converted to SAM, BAM, and BED formats as above. All data were visualized in and exported from IGV to generate genome browser figures.

HBB targeted nascent RNA sequencing

Porechop was used on raw FASTQ reads to remove external adapters and separate chimeric reads with the common forward sequence and the SMRTer IIA reverse sequence GACGTGTGCTCTTCCGATCT and GTACTCTGCGTTGATACCACTGCTT (as well as the reverse complement sequences) with settings --extra_end_trim 0 --extra_middle_trim_good_side 0 --extra_middle_trim_bad_side 0 --min_split_read_size 100 --middle_threshold 75. Reads were filtered and trimmed if they contained the 3′ end adapter as described above using the 3′ end adapter sequence plus the barcode sequence (**Table S2**). Prinseq was used to demultiplex and trim reads as above, then cleaned FASTQ files were mapped to a custom annotation of the integrated *HBB* locus, which is based on the GLOBE vector (Miccio et al., 2008).

PolyA Read Filtering

Genome-wide nascent RNA sequencing

Mapped reads in SAM format were filtered to remove reads that contained a polyA tail using a custom script (available on Github). Briefly, mapped reads that had soft-clipped bases at the 3′ end were discarded if the soft-clipped region of the read contained 4 or more A's and the fraction of A's was greater than 0.9. Similarly, reads with soft-clipped bases at the 5′ end (resulting from minus strand reads) containing at least 4 T's and having a fraction of T's greater than 0.9 were discarded.

HBB targeted nascent RNA sequencing

Additional parameters were added to the above criteria for removing polyA-containing reads from targeted data mapped to the *HBB* locus based on empirical observation. Since the *HBB* locus is integrated randomly in the MEL genome, long readthrough transcripts that have coverage past the annotated *HBB* locus read into random genomic regions and cause long stretches of mismatched soft-clipped bases. A custom script was used to filter polyA-containing reads but retain readthrough transcripts (available on Github). Briefly, reads were discarded if: they contained a fraction of A's or T's greater than 0.7 in the soft-clipped region that starts past the end of the *HBB* locus annotation; they contained a fraction of A's or T's greater than 0.7 and 4 or more A's or T's in the soft-clipped region starting within 50 nt of the annotated PAS; they contained a stretch of soft-clipped reads greater than 20 nt that starts within the annotated *HBB* gene.

Long Read Depth and Coverage Calculations

Library depth was calculated using bedtools coverage across a file of collapsed genes in the mm10 genome. A file of all known transcripts in the mm10 genome was downloaded from UCSC, then collapsed to group each transcript by its parent Gene ID. The coordinates for each gene were collapsed to include the longest possible TSS-TES distance. Metagene plots of 5′ end, 3′ end, and entire read coverage across all genes were generated using deepTools. Briefly, coverage was calculated and normalized by RPKM using the bamCoverage function, then coverage was scaled over all genes using the computeMatrix scale-regions function, and plots were generated using the plotProfile function. For coverage downstream of the PAS, input reads were separated using a custom script (available on Github) based on whether or not there was at least one "N" in the CIGAR string indicating a splice junction, then coverage for spliced and unspliced reads was calculated around the last nucleotide of annotated genes (described above) using the computeMatrix reference-point function. Coverage was normalized by RPKM, then scaled based on normalized coverage in the 10-nt window immediately downstream of the PAS. Coverage of 5′ end and 3′ ends of all reads was calculated across a 50-nt window around intron

5′SS and 3′SS using bedtools coverage. For metagene coverage of HBB-targeted data, long readthrough reads were adjusted to convert “S” to “M” in their CIGAR string using a custom script (available on Github). Extremely long reads extended through the end of the integrated HBB locus and into random genomic loci, and thus resulted in long soft-clipped ends that could not be included in calculating coverage. After converting these soft-clipped ends to mapped ends, coverage was computed, normalized by RPKM and scaled to the HBB locus, then plotted as above using deepTools.

Splicing Status Classification and Splicing Efficiency Calculation

Long reads were overlapped with a file of all unique introns in the mm10 genome that was extended by 1 nt on either end, resulting in reads that fully spanned an intron. Then, read junction coordinates were compared to the coordinates of each intron they overlapped to determine if the overlapped intron was spliced in the read. If the junction was not present in the read, a 10 nt window was included in the search for the junction to allow for slight mismatches in alignments. If the junction was not found, the intron was classified as unspliced. To classify splicing status or splicing efficiency, intron splicing status was then grouped by read name, by intron ID, or by gene ID.

Distance from Splice Junction to 3′ End Calculation

All long reads in BAM format were first classified as spliced or unspliced based on whether or not they contained at least one “N” in their CIGAR string using pysam. Splicing intermediates (defined below), were filtered out from the MEL cell data in this analysis, since their 3′ ends do not represent the position of Pol II, but rather an upstream exon currently undergoing splicing. For all remaining reads, data were converted to BED12 format using bedtools, and the last block size, which represents the distance from the most distal splice junction to the 3′ end of the read, was calculated. Coordinates of the last spliced intron were also recorded, and each intron was matched to a transcript and categorized by gene biotype using TxDb.Mmusculus.UCSC.mm10.knownGene and EnsDb.Mmusculus.v79 in R.

Splicing Intermediates Analysis

Long reads were categorized as being splicing intermediates if the 3′ end of the read aligned exactly at the -1 position of an intron (last nucleotide of an exon). Introns considered in this analysis were all non-first introns from active transcripts in the GENCODE VM20 annotation as previously described. The number of intermediates aligned upstream of each intron was counted using bedtools intersect. The Normalized Intermediate Count (NIC), was calculated for each intron by dividing the number of splicing intermediate reads by the sum of splicing intermediate reads and spliced reads. The sequence of the 23 nt region surrounding intron 3′SS (-20:+3) and the 9 nt region surrounding the 5′SS (-3:+6) were extracted using bedtools getfasta, and these sequences were used to calculate 5′ and 3′ splice site scores using MaxEntScan (Yeo and Burge, 2004). Nucleotide sequence logos of the same regions were generated using WebLogo.

Readthrough Transcripts Analysis

Bedtools intersect was used to identify long reads with 5′ ends originating in a gene body (as described above). The set of genes considered in this analysis was further filtered to remove intronless genes, any genes that were overlapping with another gene, and a set of custom curated genes with a significant number of poorly mapped or chimeric reads (ENSMUSG00000023048, ENSMUSG00000075015, ENSMUSG00000075014, and ENSMUSG00000031939). Reads with a mapping score of less than 60 were discarded, and remaining reads were categorized as being readthrough transcripts if their 3′ ends

were greater than 100 nt downstream of the end of the gene which the 5′ end overlapped with. Splicing status classification of readthrough transcripts was carried out as described above.

HBB$^{IVS-110(G>A)}$ Splicing and Readthrough Analysis

To determine the fraction of targeted HBB long reads that were spliced at intron 1, reads were overlapped with a region spanning either the cryptic or canonical intron 1 +/- 1 nt using bedtools intersect. Then from these overlapping reads, only reads containing the exact canonical or cryptic splice junction were identified. To measure the distance from the exon1-exon2 splice junction to the 3′ end, splicing intermediates were filtered out of the data as described above, then reads were overlapped with a region spanning *HBB* intron 1 +/- 1 nt as above. Reads were then filtered to keep only reads that contained either the cryptic or canonical intron 1 splice junction, had a block count of exactly 2, and had their 3′ ends before the end of intron 2. The last block sizes for these reads were calculated as described above.

Gene Expression Quantification

Long reads were quantified using Salmon with settings salmon quant -i transcripts_index -l U –gcBias and principal component analysis was performed using DESeq2.

PRO-seq Data Preprocessing

Cutadapt was used to trim paired-end reads to 40 nt, removing adapter sequence and low quality 3′ ends, and discarding reads that were shorter than 20 nt (-m20 -q 1). Additionally, in order to align reads using Bowtie, 1 nt was removed from the 3′ end of all trimmed reads. Trimmed paired-end reads were first mapped to the *Drosophila* dm3 reference genome using Bowtie, and subsequent uniquely mapped reads to the dm3 genome were used to determine percent spike-in return across all samples. Paired-end reads that failed to align to the dm3 genome were mapped to the mm10 reference genome. Read alignment to the dm3 and mm10 genomes were performed with settings - k1 -v2 –best -X1000 --un. SAM files were sorted using samtools. Read pairs uniquely aligned to the mm10 genome were separated, and strand-specific single nucleotide bedGraphs of the 3′ end mapping positions, corresponding to the biotinylated RNA 3′ end, were generated. Due to the “forward/reverse” orientation of Illuminia paired-end sequencing, “+” and “-“ stranded bedGraph files were switched at the end of the pipeline (Mahat et al., 2016). bedGraph files across replicates in each cell treatment were merged by summing the read counts per nucleotide position. Since the spike-in return was comparable between biological replicates within a treatment type, and no comparisons were made between the two treatment conditions, no further normalizations were performed.

PRO-seq Data Analysis

A list of active transcripts in MEL cells was first generated using PRO-seq signal within a 300 nt window around annotated TSSs in the GENCODE mm10 vM20 annotation. Intron annotations that did not correspond to an actively expressed transcript and had zero spliced read counts, suggesting no evidence of the intron’s usage in MEL cells, were removed. Additionally, if two intron annotations shared a 5′SS or 3′SS, the annotation with the most spliced reads was kept. Finally, first intron (rank = 1) annotations were filtered out from the final list of unique introns to avoid bleed-through PRO-seq signal from the promoter-proximally paused Pol II. A similar result was observed after removing any introns in which the intron’s 5′SS was within 250 nt of any active TSS. Metagene plots around the TSS and splice sites were generated by plotting the average PRO-seq reads at each indicated position with respect to the TSS, 5′SS, or 3′SS respectively. In order to extract PRO-seq reads that were spliced, filtered and trimmed

PRO-seq reads were mapped to the mm10 reference index using STAR with the following changes to default settings: -- outMultimapperOrder Random --outFilterType BySJout –alignSJoverhangMin 8 --outFilterIntronMotifs RemoveNoncanonicalUnannotated. All reads in BAM format were filtered for reads that contained an “N” in their CIGAR string using pysam. Resulting reads were filtered to discard reads with an “N” size > 10,000 using pysam to remove poorly mapped reads or reads mapped across very large introns.

QUANTIFICATION AND STATISTICAL ANALYSIS

All information about statistical testing for individual experiments can be found in figure legends, including statistical tests used, number of replicates, and number of observations.

Table S1. Related to Figure 1. RNA sequencing and mapping statistics

Read counts representing raw reads, mapped reads, and polyA-filtered reads (where applicable) for genome-wide LRS, *HBB*-targeted LRS, and PRO-seq. LRS samples represent combined counts for two biological and two technical replicates in each induction conditions, and PRO-seq samples represent combined counts for three biological replicates in each induction condition.

Sample	**Sequencing Protocol**	**Raw read number**	**Mapped read number**	**PolyA-filtered read number**
MEL_LRS_uninduced	PacBio LRS	583,632	545,477	538,452
MEL_LRS_induced	PacBio LRS	571,997	464,793	452,514
MEL_LRS_HBB_WT	PacBio LRS targeted	68,121	66,909	20,651
MEL_LRS_HBB_IVS110	PacBio LRS targeted	79,472	78,268	32,582
MEL_PROseq_uninduced	PRO-seq	220,070,650	127,803,736	NA
MEL_PROseq_induced	PRO-seq	208,414,166	109,366,055	NA

Table S2. Related to STAR Methods. Oligonucleotides used in this study

Oligonucleotide sequences used in this study for LRS library preparation, qPCR, RT-PCR, and PRO-seq library preparation.

Description	Sequence
3′ end DNA adapter	/5rApp/NNNNNCTGTAGGCACCATCAAT/3ddC/
RT primer for genome-wide first strand synthesis	AAGCAGTGGTATCAACGCAGAGTACATTGATGGTGCCTACAG
RT primer for targeted first strand synthesis barcode 1	AAGCAGTGGTATCAACGCAGAGTACCACATATCAGAGTGCGGATTGATGGTGCCTACAG
RT primer for targeted first strand synthesis barcode 2	AAGCAGTGGTATCAACGCAGAGTACACACACAGACTGTGAGGATTGATGGTGCCTACAG
RT primer for targeted first strand synthesis barcode 3	AAGCAGTGGTATCAACGCAGAGTACACACATCTCGTGAGAGGATTGATGGTGCCTACAG
RT primer for targeted first strand synthesis barcode 4	AAGCAGTGGTATCAACGCAGAGTACCACGCACACACGCGCGGATTGATGGTGCCTACAG
RT primer for targeted first strand synthesis barcode 5	AAGCAGTGGTATCAACGCAGAGTACCATATATATCAGCTGTGATTGATGGTGCCTACAG
RT primer for targeted first strand synthesis barcode 6	AAGCAGTGGTATCAACGCAGAGTACTCTGTATCTCTATGTGGATTGATGGTGCCTACAG
PCR primer for targeted amplification of human *HBB*	GACGTGTGCTCTTCCGATCTCACGACACGACGATGTcaactgtgttcactagcaacct
qPCR primer F *Hbb-b1*	ATGCCAAAGTGAAGGCCCAT
qPCR primer R *Hbb-b1*	CCCAGGAGCCTGAAGTTCTC
qPCR primer F *Gapdh*	AATGTGTCCGTCGTGGATCTGA
qPCR primer R *Gapdh*	GATGCCTGCTTCACCACCTTCT
RT-PCR primer F *Brd2*	GATTATCACAAAATTATAAAACAGCC
RT-PCR primer R *Brd2*	CTGCTAACTTGGCCCCC
RT-PCR primer F *Dnajb1*	CCTTTCCCAAGGAAGGG
RT-PCR primer R *Dnajb1*	GTTTCTCAGGTGTTTTGGG
RT-PCR primer F *Riok3*	TGTTGCTGAAGGACCATTC
RT-PCR primer R *Riok3*	ATTTTCCATTCTTGCTGTGTTC
RT-PCR primer F *Slc12a6*	GACGTGTGCTCTTCCGATCTGGATAACATCATACCTTTCCTTAGG
RT-PCR primer R *Slc12a6*	ATGGAAAGAATTGGGGCC
RT-PCR primer F *Dmtn*	GACGTGTGCTCTTCCGATCTCCACCCATCTACAAACAGAGAG
RT-PCR primer R *Dmtn*	CCACAACGGCCAGCGACG
RT-PCR primer F *Hnrnpll*	GACGTGTGCTCTTCCGATCTTAAAGTGTTTGACGCGAAAG
RT-PCR primer R *Hnrnpll*	TCGGGACTCGTATCTGGTA
RT-PCR primer F *Rbm39*	GACGTGTGCTCTTCCGATCTTGCCTCATAGCATCAAATTAAG
RT-PCR primer R *Rbm39*	CTCACAGGGCTCTTGTCTT
RT-PCR primer F *Hbq1b*	GACGTGTGCTCTTCCGATCTGGACCCTGCTAACTTCCAG
RT-PCR primer R *Hbq1b*	TCAGCGATATTTGGAGACC
RT-PCR primer F *C1qbp*	GACGTGTGCTCTTCCGATCTCACAGATTCCCTGGACTGG

RT-PCR primer R C1qbp	CTACTGGTTCTTGACAAAGCTTT
VRA3 3′ end adapter	/5Phos/rGrArUrCrGrUrCrGrGrArCrUrGrUrArGrArArCrUrCrUrGrArArC /3InvdT/
VRA5 5′ end adapter	rCrCrUrUrGrGrCrArCrCrCrGrArGrArArUrUrCrCrA
RP1 primer	AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACA GTCCGA

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit polyclonal anti-GAPDH	Santa Cruz Biotechnology	FL-335/sc-25778
Mouse monoclonal anti-Pol II	Santa Cruz Biotechnology	CTD4H8/sc-47701
Bacterial and Virus Strains		
Biological Samples		
Chemicals, Peptides, and Recombinant Proteins		
DMEM + GlutaMAX	Gibco	10569-010
Fetal Bovine Serum (FBS)	Gibco	16000-044
Penicillin Streptomycin	Gibco	15140-122
α-Amanitin	Sigma	A2263
SUPERase.In	ThermoFisher	AM2694
1X cOmplete Protease Inhibitor Cocktail	Sigma	11697498001
TRIzol Reagent	ThermoFisher	15596018
RNase-Free DNase Set	Qiagen	79254
Pladienolide B	Santa Cruz Biotechnology	445493-23-2
Random Hexamer Primers	ThermoFisher	SO142
Phusion High-Fidelity DNA Polymerase	NEB	M0530S
Biotin-11-NTPs	Perkin-Elmer	NEL54(2/3/4/5)001
Critical Commercial Assays		
RNeasy Mini Kit	Qiagen	74104
Dynabeads mRNA DIRECT Micro Purification Kit	ThermoFisher	61021
Ribo-Zero Gold rRNA Removal Kit	Illumina	MRZG126
T4 RNA ligase Kit	NEB	M0351L
SMARTer PCR cDNA Synthesis Kit	Clontech	634925
Advantage 2 PCR Kit	Clontech	639201
AMPure XP Beads	Agencourt	A63880
SuperScriptIII Reverse Transcriptase	ThermoFisher	18080044
iQ SYBR Green Supermix	Biorad	1708880
SMRTbell Template Prep Kit 1.0	Pacific Biosciences	100-259-100
Total RNA Purification Kit	Norgen Biotek Corp.	17200
Dynabeads M-280 Streptavidin	ThermoFisher	11205D
T4 RNA Ligase I	NEB	M0204S
ThermoPol Reaction Buffer	NEB	B9004S
RNA 5′ Pyrophosphohydrolase (RppH)	NEB	M0356S
T4 Polynucleotide Kinase	NEB	M0201S
Lysis Buffer, FS	ThermoFisher	4480724
SuperScript IV Reverse Transcriptase	ThermoFisher	18090010
ProNex Size-Selective Purification System	Promega	NG2001

Deposited Data		
Raw and analyzed data	This paper	GEO: GSE144205
Raw image data	This paper	http://dx.doi.org/10.17632/5vrtbpnj4k.1
nanoCOP data from BL1184, K562, and S2 cells	Drexler et al. (2019)	GEO: GSE123191
Experimental Models: Cell Lines		
MEL	Shilpa Hattangadi	N/A
MEL-*HBB* ^{WT}	Patsali et al. (2018)	N/A
MEL-*HBB* $^{IVS-110(G>A)}$	Patsali et al. (2018)	N/A
Experimental Models: Organisms/Strains		
Oligonucleotides		
See Table S2	This paper	N/A
Recombinant DNA		
Software and Algorithms		
Porechop v0.2.4	N/A	https://github.com/rrwick/Porechop
Cutadapt v1.9.1	Martin (2011)	https://cutadapt.readthedocs.io/en/stable/
Bowtie v1.2.2	Langmead et al. (2009)	http://bowtie-bio.sourceforge.net/index.shtml
STAR v2.7.0a	Dobin et al. (2013)	http://code.google.com/p/rna-star/
Prinseq-lite v0.20.4	Schmieder and Edwards (2011)	https://sourceforge.net/projects/prinseq/files/
Minimap2 v2.12-r827	Li (2018)	https://github.com/lh3/minimap2
samtools v1.9	Li et al. (2009)	http://samtools.sourceforge.net/
bedtools v2.27.1	Quinlan and Hall (2010)	https://bedtools.readthedocs.io/en/latest/
MaxEnt Scan	Yeo and Burge (2004)	http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq_acc.html
WebLogo v3.7.1	Crooks et al. (2004)	http://weblogo.threeplusone.com/
deepTools v3.3.0	Ramirez et al. (2016)	https://deeptools.readthedocs.io/en/develop/
Pysam v0.15.0	N/A	https://github.com/pysam-developers/pysam

Salmon v0.14.1	Patro et al. (2017)	https://github.com/COMBINE-lab/salmon
DESeq2	Love et al. (2014)	https://bioconductor.org/packages/release/bioc/html/DESeq2.html

REFERENCES

Alexander, R.D., Innocente, S.A., Barrass, J.D., and Beggs, J.D. (2010). Splicing-dependent RNA polymerase pausing in yeast. Mol Cell *40*, 582-593.

Alpert, T., Herzel, L., and Neugebauer, K.M. (2017). Perfect timing: splicing and transcription rates in living cells. Wiley interdisciplinary reviews RNA *8*.

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., and 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.

Aslanzadeh, V., Huang, Y., Sanguinetti, G., and Beggs, J.D. (2018). Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast. Genome Res *28*, 203-213.

Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nature reviews Molecular cell biology *18*, 437-451.

Bentley, D.L. (2014). Coupling mRNA processing with transcription in time and space. Nature reviews Genetics *15*, 163-175.

Bieberstein, N.I., Carrillo Oesterreich, F., Straube, K., and Neugebauer, K.M. (2012). First exon length controls active chromatin signatures and transcription. Cell Rep *2*, 62-68.

Braberg, H., Jin, H., Moehle, E.A., Chan, Y.A., Wang, S., Shales, M., Benschop, J.J., Morris, J.H., Qiu, C., Hu, F.*, et al.* (2013). From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. Cell *154*, 775-788.

Brinster, R.L., Allen, J.M., Behringer, R.R., Gelinas, R.E., and Palmiter, R.D. (1988). Introns increase transcriptional efficiency in transgenic mice. Proc Natl Acad Sci U S A *85*, 836-840.

Burke, J.E., Longhurst, A.D., Merkurjev, D., Sales-Lee, J., Rao, B., Moresco, J.J., Yates, J.R., 3rd, Li, J.J., and Madhani, H.D. (2018). Spliceosome Profiling Visualizes Operations of a Dynamic RNP at Nucleotide Resolution. Cell *173*, 1014-1030 e1017.

Carrillo Oesterreich, F., Bieberstein, N., and Neugebauer, K.M. (2011). Pause locally, splice globally. Trends Cell Biol *21*, 328-335.

Carrillo Oesterreich, F., Herzel, L., Straube, K., Hujer, K., Howard, J., and Neugebauer, K.M. (2016). Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. Cell *165*, 372-381.

Carrillo Oesterreich, F., Preibisch, S., and Neugebauer, K.M. (2010). Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. Mol Cell *40*, 571-581.

Chathoth, K.T., Barrass, J.D., Webb, S., and Beggs, J.D. (2014). A splicing-dependent transcriptional checkpoint associated with prespliceosome formation. Mol Cell *53*, 779-790.

Chen, W., Moore, J., Ozadam, H., Shulha, H.P., Rhind, N., Weng, Z., and Moore, M.J. (2018). Transcriptome-wide Interrogation of the Functional Intronome by Spliceosome Profiling. Cell *173*, 1031-1044 e1013.

Churchman, L.S., and Weissman, J.S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368-373.

Core, L., and Adelman, K. (2019). Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. Genes & development *33*, 960-982.

Coulon, A., Ferguson, M.L., de Turris, V., Palangat, M., Chow, C.C., and Larson, D.R. (2014). Kinetic competition during the transcription cycle results in stochastic RNA processing. Elife *3*, e1002215.

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

Custodio, N., and Carmo-Fonseca, M. (2016). Co-transcriptional splicing and the CTD code. Critical reviews in biochemistry and molecular biology *51*, 395-411.

Custodio, N., Carmo-Fonseca, M., Geraghty, F., Pereira, H.S., Grosveld, F., and Antoniou, M. (1999). Inefficient processing impairs release of RNA from the site of transcription. EMBO J *18*, 2855-2866.

David, C.J., Boyne, A.R., Millhouse, S.R., and Manley, J.L. (2011). The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex. Genes & development *25*, 972-983.

de la Mata, M., Alonso, C.R., Kadener, S., Fededa, J.P., Blaustein, M., Pelisch, F., Cramer, P., Bentley, D., and Kornblihtt, A.R. (2003). A slow RNA polymerase II affects alternative splicing in vivo. Mol Cell *12*, 525-532.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15-21.

Drexler, H.L., Choquet, K., and Churchman, L.S. (2019). Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. Mol Cell.

Enge, M., Arda, H.E., Mignardi, M., Beausang, J., Bottino, R., Kim, S.K., and Quake, S.R. (2017). Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. Cell *171*, 321-330.e314.

Fica, S.M., Oubridge, C., Wilkinson, M.E., Newman, A.J., and Nagai, K. (2019). A human postcatalytic spliceosome structure reveals essential roles of metazoan factors for exon ligation. Science *363*, 710-714.

Fiszbein, A., Krick, K.S., Begg, B.E., and Burge, C.B. (2019). Exon-Mediated Activation of Transcription Starts. Cell *179*, 1551-1565.e1517.

Fong, N., Kim, H., Zhou, Y., Ji, X., Qiu, J., Saldi, T., Diener, K., Jones, K., Fu, X.D., and Bentley, D.L. (2014). Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. Genes & development *28*, 2663-2676.

Girard, C., Will, C.L., Peng, J., Makarov, E.M., Kastner, B., Lemm, I., Urlaub, H., Hartmuth, K., and Luhrmann, R. (2012). Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. Nat Commun *3*, 994.

Grosso, A.R., Leite, A.P., Carvalho, S., Matos, M.R., Martins, F.B., Vitor, A.C., Desterro, J.M., Carmo-Fonseca, M., and de Almeida, S.F. (2015). Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. Elife *4*.

Gu, B., Eick, D., and Bensaude, O. (2013). CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo. Nucleic Acids Res *41*, 1591-1603.

Harlen, K.M., Trotta, K.L., Smith, E.E., Mosaheb, M.M., Fuchs, S.M., and Churchman, L.S. (2016). Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. Cell Rep *15*, 2147-2158.

Herzel, L., Ottoz, D.S.M., Alpert, T., and Neugebauer, K.M. (2017). Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. Nature reviews Molecular cell biology *18*, 637-650.

Herzel, L., Straube, K., and Neugebauer, K.M. (2018). Long-read sequencing of nascent RNA reveals coupling among RNA processing events. Genome Res *28*, 1008-1019.

Ip, J.Y., Schmidt, D., Pan, Q., Ramani, A.K., Fraser, A.G., Odom, D.T., and Blencowe, B.J. (2011). Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. Genome Res *21*, 390-401.

Jeong, S. (2017). SR Proteins: Binders, Regulators, and Connectors of RNA. Molecules and cells *40*, 1-9.

Jonkers, I., and Lis, J.T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. Nature reviews Molecular cell biology *16*, 167-177.

Khodor, Y.L., Rodriguez, J., Abruzzi, K.C., Tang, C.H., Marr, M.T., 2nd, and Rosbash, M. (2011). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes & development *25*, 2502-2512.

Kurosaki, T., Popp, M.W., and Maquat, L.E. (2019). Quality and quantity control of gene expression by nonsense-mediated mRNA decay. Nature reviews Molecular cell biology *20*, 406-420.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science *339*, 950-953.

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

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094-3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Lin, C.L., Taggart, A.J., and Fairbrother, W.G. (2016). RNA structure in splicing: An evolutionary perspective. RNA Biol *13*, 766-771.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

Mahat, D.B., Kwak, H., Booth, G.T., Jonkers, I.H., Danko, C.G., Patel, R.K., Waters, C.T., Munson, K., Core, L.J., and Lis, J.T. (2016). Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc *11*, 1455-1476.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet *17*, 10.14806/ej.14817.14801.14200.

Martin, R.M., Rino, J., Carvalho, C., Kirchhausen, T., and Carmo-Fonseca, M. (2013). Live-cell visualization of pre-mRNA splicing with single-molecule sensitivity. Cell Rep *4*, 1144-1155.

Mayer, A., and Churchman, L.S. (2017). A Detailed Protocol for Subcellular RNA Sequencing (subRNA-seq). Current protocols in molecular biology *120*, 4.29.21-24.29.18.

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*, 541-554.

Miccio, A., Cesari, R., Lotti, F., Rossi, C., Sanvito, F., Ponzoni, M., Routledge, S.J., Chow, C.M., Antoniou, M.N., and Ferrari, G. (2008). In vivo selection of genetically modified erythroblastic progenitors leads to long-term correction of beta-thalassemia. Proc Natl Acad Sci U S A *105*, 10547-10552.

Milligan, L., Sayou, C., Tuck, A., Auchynnikava, T., Reid, J.E., Alexander, R., Alves, F.L., Allshire, R., Spanos, C., Rappsilber, J.*, et al.* (2017). RNA polymerase II stalling at pre-mRNA splice sites is enforced by ubiquitination of the catalytic subunit. Elife *6*, e27082.

Muniz, L., Deb, M.K., Aguirrebengoa, M., Lazorthes, S., Trouche, D., and Nicolas, E. (2017). Control of Gene Expression in Senescence through Transcriptional Read-Through of Convergent Protein-Coding Genes. Cell Rep *21*, 2433-2446.

Neugebauer, K.M. (2019). Nascent RNA and the Coordination of Splicing with Transcription. Cold Spring Harb Perspect Biol *11*.

Nojima, T., Gomes, T., Grosso, A.R.F., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. Cell *161*, 526-540.

Nojima, T., Rebelo, K., Gomes, T., Grosso, A.R., Proudfoot, N.J., and Carmo-Fonseca, M. (2018). RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing. Mol Cell *72*, 369-379 e364.

Pai, A.A., Henriques, T., McCue, K., Burkholder, A., Adelman, K., and Burge, C.B. (2017). The kinetics of pre-mRNA splicing in the Drosophila genome and the influence of gene architecture. Elife *6*, 1123.

Pai, A.A., and Luca, F. (2019). Environmental influences on RNA processing: Biochemical, molecular and genetic regulators of cellular response. Wiley interdisciplinary reviews RNA *10*, e1503.

Pandya-Jones, A., and Black, D.L. (2009). Co-transcriptional splicing of constitutive and alternative exons. RNA *15*, 1896-1908.

Parra, M., Booth, B., Weiszmann, R., Yee, B., Yeo, G.W., Brown, J.B., Celniker, S.E., and Conboy, J.G. (2018). An important class of intron retention events in human erythroblasts is regulated by cryptic exons proposed to function as splicing decoys. Rna *24*, 1255-1265.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods *14*, 417-419.

Patsali, P., Papasavva, P., Stephanou, C., Christou, S., Sitarou, M., Antoniou, M.N., Lederer, C.W., and Kleanthous, M. (2018). Short-hairpin RNA against aberrant HBB(IVSI-110(G>A)) mRNA restores beta-globin levels in a novel cell model and acts as mono- and combination therapy for beta-thalassemia in primary hematopoietic stem cells. Haematologica *103*, e419-e423.

Pimentel, H., Parra, M., Gee, S.L., Mohandas, N., Pachter, L., and Conboy, J.G. (2016). A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. Nucleic Acids Res *44*, 838-851.

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

Ramirez, F., Ryan, D.P., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dundar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res *44*, W160-165.

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863-864.

Schor, I.E., Fiszbein, A., Petrillo, E., and Kornblihtt, A.R. (2013). Intragenic epigenetic changes modulate NCAM alternative splicing in neuronal differentiation. Embo j *32*, 2264-2274.

Shaul, O. (2017). How introns enhance gene expression. The international journal of biochemistry & cell biology *91*, 145-155.

Sheridan, R.M., Fong, N., D'Alessandro, A., and Bentley, D.L. (2019). Widespread Backtracking by RNA Pol II Is a Major Effector of Gene Activation, 5' Pause Release, Termination, and Transcription Elongation Rate. Mol Cell *73*, 107-118 e104.

Smith, D.J., Query, C.C., and Konarska, M.M. (2008). "Nought may endure but mutability": spliceosome dynamics and the regulation of splicing. Mol Cell *30*, 657-666.

Spritz, R.A., Jagadeeswaran, P., Choudary, P.V., Biro, P.A., Elder, J.T., deRiel, J.K., Manley, J.L., Gefter, M.L., Forget, B.G., and Weissman, S.M. (1981). Base substitution in an intervening sequence of a beta+-thalassemic human globin gene. Proc Natl Acad Sci U S A *78*, 2455-2459.

Testa, S.M., Disney, M.D., Turner, D.H., and Kierzek, R. (1999). Thermodynamics of RNA-RNA duplexes with 2- or 4-thiouridines: implications for antisense design and targeting a group I intron. Biochemistry *38*, 16655-16662.

Tilgner, H., Knowles, D.G., Johnson, R., Davis, C.A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T.R., and Guigo, 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.

Vadolas, J., Nefedov, M., Wardan, H., Mansooriderakshan, S., Voullaire, L., Jamsai, D., Williamson, R., and Ioannou, P.A. (2006). Humanized beta-thalassemia mouse model containing the common IVSI-110 splicing mutation. J Biol Chem *281*, 7399-7405.

Vilborg, A., Passarelli, M.C., Yario, T.A., Tycowski, K.T., and Steitz, J.A. (2015). Widespread Inducible Transcription Downstream of Human Genes. Mol Cell *59*, 449-461.

Vilborg, A., Sabath, N., Wiesel, Y., Nathans, J., Levy-Adam, F., Yario, T.A., Steitz, J.A., and Shalgi, R. (2017). Comparative analysis reveals genomic features of stress-induced transcriptional readthrough. Proc Natl Acad Sci U S A *114*, E8362-e8371.

Wachutka, L., Caizzi, L., Gagneur, J., and Cramer, P. (2019). Global donor and acceptor splicing site kinetics in human cells. Elife *8*.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P.*, et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Wilkinson, M.E., Charenton, C., and Nagai, K. (2019). RNA Splicing by the Spliceosome. Annu Rev Biochem.

Wuarin, J., and Schibler, U. (1994). Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. Mol Cell Biol *14*, 7219-7225.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Journal of computational biology : a journal of computational molecular cell biology *11*, 377-394.

Yu, Y., and Reed, R. (2015). FUS functions in coupling transcription to splicing by mediating an interaction between RNAP II and U1 snRNP. Proc Natl Acad Sci U S A *112*, 8608-8613.