Abstract

Co-Transcriptional Splicing in Murine Erythroblasts

Kirsten Audrey Reimer

2021

Eukaryotic genes contain non-coding sequences called introns. The removal of introns from pre-mRNAs, termed splicing, is carried out by the spliceosome, a multi-megadalton molecular complex of proteins and RNAs. Splicing occurs co-transcriptionally across multiple cell types and species. The Neugebauer lab has developed single molecule nascent RNA sequencing methods—including single molecule intron tracking (SMIT) and long-read sequencing (LRS) of nascent RNA—to visualize the precursors, intermediates, and products of transcription and splicing in budding and fission yeasts. Using these methods, the lab was able to estimate the kinetics of single intron removal in both yeasts by relating the 3′ end of nascent RNA (the position of RNA Polymerase II) to progress of the splicing reaction. In both species of yeast, splicing proceeded rapidly and co-transcriptionally.

In comparison to yeast, mammalian genes are much more complex—on average they contain eight long introns surrounded by short exons. It was unclear how the presence of many more long introns, often with more poorly conserved splice site sequences, would affect how splicing and transcription are coordinated. Thus, I have optimized new methods to isolate nascent RNA and analyze co-transcriptional splicing in mammalian cells.

To determine how splicing is integrated with transcription elongation and 3′ end formation in mammalian cells, I performed long-read sequencing of individual nascent RNAs and PRO-seq during murine erythropoiesis. I chose murine erythroid leukemia (MEL) cells as a model system, as they can be easily differentiated *in vitro*, and they express a subset of erythroid-specific genes at high levels.

Many studies of gene expression have historically been carried out in erythroblasts, and the biogenesis of β-globin mRNA—the most highly expressed transcript in erythroblasts—was the focus of many seminal studies on the mechanisms of pre-mRNA splicing.

I isolated nascent, chromatin-associated RNAs from MEL cells before and after induction of terminal erythroid differentiation and performed long-read sequencing on the Pacific Biosciences Sequel platform. Splicing was not accompanied by transcriptional pausing and was detected when RNA polymerase II (Pol II) was within 75 – 300 nucleotides of 3′ splice sites, often during transcription of the downstream exon. Interestingly, several hundred introns displayed abundant splicing intermediates, suggesting that splicing delays can take place between the two catalytic steps of splicing. Overall, splicing efficiencies were correlated among introns within the same transcript, and intron retention was associated with inefficient 3′ end cleavage. Remarkably, a thalassemia patient-derived mutation introducing a cryptic 3′ splice site improves both splicing and 3′ end cleavage of individual β-globin transcripts, demonstrating functional coupling between the two co-transcriptional processes as a determinant of productive gene output.

Thus, I conclude that highly expressed pre-mRNAs in MEL cells are largely spliced co-transcriptionally, and that the mammalian spliceosome can assemble and act rapidly on this set of pre-mRNAs. A previously unappreciated level of cross-talk between splicing and 3′ end cleavage efficiencies is involved in erythroid development. Together, this work provides a high-resolution description of mammalian gene expression and shows that short-read RNA sequencing of bulk RNA can conceal coordinated behaviours that can only be observed at the level of individual nascent transcripts.

Co-Transcriptional Splicing in Murine Erythroblasts

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Kirsten Audrey Reimer

Dissertation Director: Karla M. Neugebauer

June 2021

Contents

List of Figures

List of Tables

Acknowledgements

I am grateful to every person who has helped me along the path to completing my PhD. It takes a community to get where I am now, and I am humbled to be part of such a supportive community.

Thank you to my teachers and role models for encouraging me and inspiring me to pursue a path in science. My high school math teacher, Mr. Chris Sdoutz, and chemistry teacher, Mr. Greg Woolgar, made learning fun. Thank you for imparting your enthusiasm for learning more about the world and for pushing me to tackle hard questions with confidence. In my undergraduate degree at the University of Northern British Columbia, I was fortunate to cross paths with many generous professors. In particular, Dr. Stephen Rader continues to go above and beyond in advocating for me. Stephen's continuous excitement about science is contagious. Dr. Martha Stark taught me nearly everything I know how to do in a laboratory. Her attention to detail rubbed off on me in a good way. Thank you both for sparking my interest in RNA biochemistry and for helping me accomplish things I didn't know I could.

Thank you to everyone I've met at Yale for contributing in ways large and small to my education. While I've had many enthusiastic mentors along the way, Dr. Karla Neugebauer's excitement for science and for life truly cannot be contained. She has shown me how to navigate issues both personal and professional with grace. She is an amazing of example of success, and it is a joy to try and emulate

someone who has, in her own words, "the best job in the world". Thank you for every opportunity you have given me and for giving me the space to create my own opportunities. Dr. Joan Steitz is a venerable leader in the field of RNA splicing, and it is always an honour to hear her thoughts on my work. Dr. Matthew Simon trusted me with the role of head Teaching Assistant in a class of nearly 150 undergraduate biochemistry students. I learned a lot about biochemistry and about myself. Thank you both for serving on my thesis committee, for providing creative and insightful feedback when I needed it, and importantly for believing in me. The Neugebauer lab has been a seriously fun place to spend the last five years. It never feels like going in to work when you get to see your friends every day. Thank you to all members past and present, and particularly to Dahyana for being hands down the best bay mate, to Tara for always being willing to show me how to get to the next step, to Ed for being a friend from day one, to Korinna for her unending kindness and skill, to Dave for asking *all* the important questions, and to Tucker for being there to talk about science and everything else, and doing so as a proud member of the quiet club.

Thank you to my friends in New Haven for keeping me company in all the hours during a PhD spent outside of classrooms and labs (and many of the hours spent inside of them as well). Jay, Ed, Meg, Josie, Chris, Melanie, and many others in my class: thank you for the all the skiing, hiking, climbing, biking, baking, pizza eating, book clubbing and more. I've been so lucky to be a part of an amazing community of young scientist who I am proud to call colleagues and lifelong friends.

Thank you to my family for their ceaseless support. Coming home during the holidays was always my favourite time of year, and visiting with all my cousins, aunts, uncles, and grandparents gave me a much needed shift in perspective. Thank you for supporting me, even when you weren't really sure what I was doing. My dad and I were doing science experiments at home before I was even in kinder-

garten; he remains the most important scientist in my life. My mom supports me wholeheartedly, and she shows me what it means to be a successful woman, always leading by example. Thank you both for inspiring me to set big goals and for never missing an opportunity to remind me that I can accomplish anything I put my mind to. My sister is my best friend. Thank you for always being there. Finally, thank you to Walker, for absolutely everything.

Abbreviations

4sU 4-thiouridine.

bp basepairs.

BPS branchpoint sequence.

BSA bovine serum albumin.

CCS combined consensus sequence.

cDNA complementary DNA.

CoSE Co-transcriptional Splicing Efficiency.

DMSO dimethyl sulfoxide.

DTT dithiothreitol.

HSPC hematopoietic stem and progenitor cell.

IR intron retention.

LRS long-read sequencing.

MDS myelodysplastic syndromes.

MEL murine erythroid leukemia.

mRNA messenger RNA.

NGS Next Generation Sequencing.

NIC Normalized Intermediate Count.

NMD nonsense-mediated decay.

PAS polyA site.

PCR polymerase chain reaction.

Pol II RNA polymerase II.

polyA+ RNA polyadenylated RNA.

pre-mRNA pre-messenger RNA.

PRO-seq precision run-on sequencing.

PTC premature termination codon.

RBC red blood cell.

RNA-seq RNA sequencing.

rpm revolutions per minute.

rRNA ribosomal RNA.

RT reverse transcriptase.

SMIT single molecule intron tracking.

snRNA small nuclear RNA.

snRNP small nuclear ribonucleoprotein.

SS splice site.

TBS Tris-buffered saline.

TSS transcription start site.

Publications

Research during my PhD has contributed to the following publications:

Key references

1. **Reimer, K. A.**, Mimoso, C., Adelman, K., and Neugebauer, K. M. (2021). Co-transcriptional splicing regulates 3′ end cleavage during mammalian erythropoiesis. Mol Cell *81*,998-1012.e7.

2. Alpert, T. * , **Reimer, K. A.***, Straube, K., and Neugebauer, K. M. (2020). Long-read Sequencing of Nascent RNA from Budding and Fission Yeasts. In Methods in Molecular Biology: Nascent RNA Sequencing Techniques. (Springer: Springer Protocols) *(accepted)*.

3. **Reimer, K. A.** and Neugebauer, K. M. (2020). Preparation of Mammalian Nascent RNA for Long Read Sequencing. Curr Protoc Mol Biol *133*, e128.

4. **Reimer, K. A.** and Neugebauer, K. M. (2018). Blood Relatives: Splicing Mechanisms underlying Erythropoiesis in Health and Disease [version 1; peer review: 3 approved]. F1000Research *7(F1000 Faculty Rev)*:1364.

* . Equal contributions

Author Contributions

Text and figures in this thesis have been adapted from (Reimer et al., 2021), (Reimer and Neugebauer, 2020), and (Reimer and Neugebauer, 2018). With the exception of some PRO-seq library preparation steps and data analysis, I performed all experiments and data analysis as described. I prepared MEL cells for PRO-seq by performing cell differentiation and permeabilization, and Claudia Mimoso performed further PRO-seq library preparation steps. Claudia Mimoso and Dr. Karen Adelman performed most initial PRO-seq data analysis in close discussion with me. I prepared all final figures presented in this thesis.

Chapter 1

Introduction

1.1 Eukaryotic Pre-mRNA Splicing

Eukaryotic genes contain intervening sequences—introns—which must be removed after transcription and before translation of a pre-mRNA to ensure proper gene expression. Introns are recognized and removed by a multi-megadalton molecular machine called the spliceosome, which is comprised of five small nuclear RNAs and hundreds of associated proteins (Papasaikas and Valcarcel, 2016). The spliceosome assembles *de novo* on each intron, recognizing nucleotide sequences called the 5′ and 3′ splice sites (SS) that demarcate intron boundaries. The spliceosome then catalyzes two transesterification reactions to excise the intron and ligate the flanking exons together (Wilkinson et al., 2019).

Decades of biochemistry, genetics, and more recently structural biology, have shed light on the stunningly complex mechanism of pre-mRNA splicing (**Figure 1.1**). The spliceosome is not a pre-assembled enzyme, but rather a dynamic coming and going of protein-RNA complexes called small nuclear ribonucleoproteins (snNRPs). Spliceosome assembly begins with U1 and U2 snRNP binding to the 5′SS and the branchpoint sequence (BPS) of the pre-mRNA substrate respectively.

A helicase protein then unwinds the helix formed between the 5′SS and U1 small nuclear RNA (snRNA), and transfers the 5′SS sequence to U6 snRNA, which is assembled in a complex called the tri-snRNP, consisting of U4, U5, and U6 snRNPs. Next, U1 and U4 snRNPs are released, leaving U2, U5, and U6 snRNPS to perform the catalytic steps of splicing. After the first transesterification step, the 5′-exon and lariat intermediates are rearranged to allow for the second step to occur, resulting in the release of the lariat, and the ligation of the 5′ exon to the 3′ exon. Finally, all the components of the spliceosome must be disassembled and recycled to prepare for assembly and splicing of another intron.

In mammalian cells, genes typically encode pre-mRNAs containing 8-10 introns of variable lengths (ranging from 50 to 500,000 nt), creating a high cellular demand for spliceosomes relative to other cellular machineries, which only act once per transcript. Splicing is also a highly-regulated process; it is influenced by environmental factors, developmental cues, and factors in the local pre-messenger RNA (pre-mRNA) environment, such as RNA secondary structure and RNA-binding protein occupancy (Baralle and Giudice, 2017; Jeong, 2017; Lin et al., 2016; Pai and Luca, 2019). Since the 5′SS, 3′SS and branchpoint sequence are rarely exact consensus sequences, additional regulators are thought to explain the correct recognition of introns. Additionally, trans-acting factors likely play a large role in how constitutive and alternative splice sites are chosen. These working models still largely rely on *in vitro* biochemistry and often do not explain changes in alternative splicing or overall gene expression observed upon experimental perturbation or disease-associated mutations of splicing factors (Joshi et al., 2017; Manning and Cooper, 2017). Thus, despite detailed knowledge of modulatory factors, the mechanisms underlying the gene regulatory potential of pre-mRNA splicing are not fully understood *in vivo*.

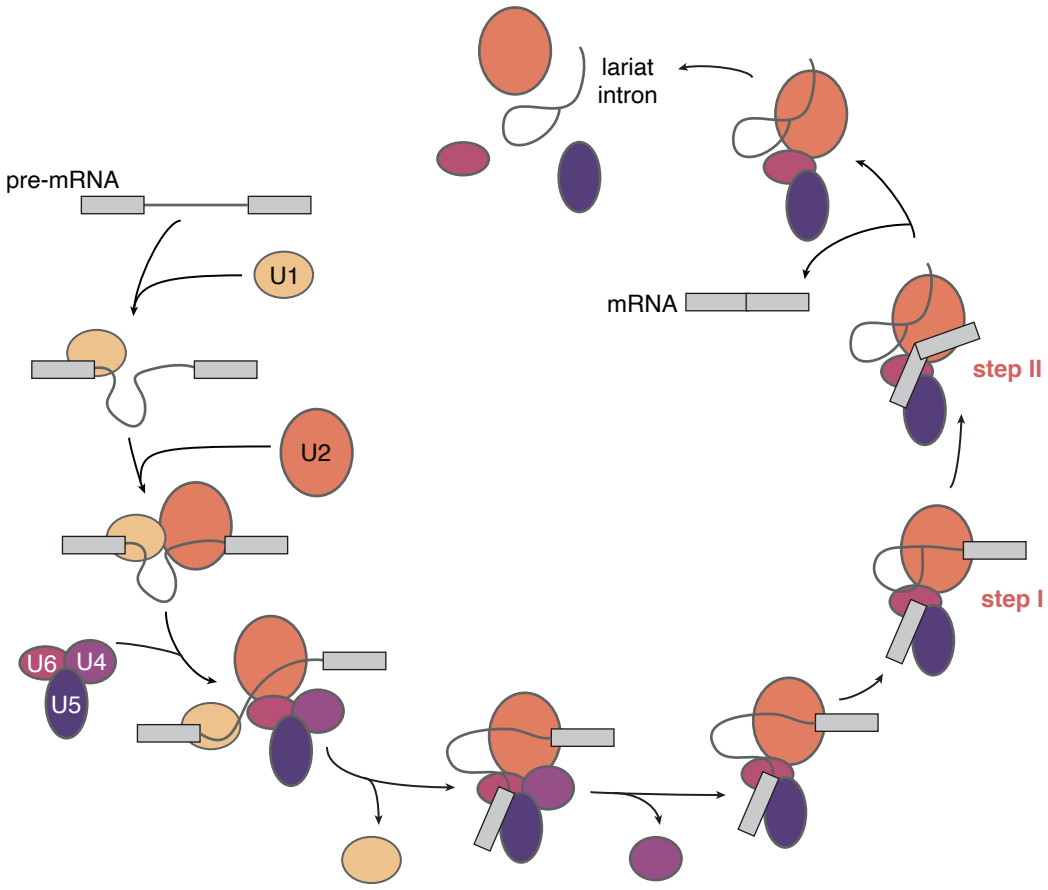

Figure 1.1: Simplified model of the pre-mRNA splicing cycle.
The spliceosome assembles on each intron in pre-mRNA in a step-wise manner. U1, U2, U4, U5, and U6 snNRPs (coloured ovals) facilitate positioning the pre-mRNA (top left, grey) in the spliceosome active site as well as carrying out the catalysis of splicing. Two transesterification reactions (step I and step II, red) result in branching of the lariat and exon ligation. The final mRNA product is released from the spliceosome, and the entire complex is disassembled.

1.2 Co-transcriptionality

Across species, tissues, and cell types, splicing can occur during pre-mRNA synthesis by Pol II (Custodio and Carmo-Fonseca, 2016; Neugebauer, 2019; Carrocci and Neugebauer, 2019), meaning that introns can be excised before Pol II terminates transcription. Thus, spliceosome assembly occurs as the nascent RNA is growing longer and more diverse in sequence and structure. Spliceosomes may not assemble on all introns at the same time, because promoter-proximal introns are synthesized before promoter-distal introns. Thus, the questions of whether introns are spliced in the order they are transcribed and how splicing of individual introns within a given transcript might be coordinated are currently the subject of intense investigation. Co-transcriptional splicing also demands that the constellation of splicing factors capable of regulating a splicing event bind the nascent RNA coordinately with the timing imposed by transcription and in a relevant spatial window. For example, a splicing inhibitor element in a given nascent RNA would only be influential if it were transcribed before the target intron was removed.

Another issue raised by co-transcriptional RNA processing is how splicing is coordinated with other pre-mRNA processing steps (Bentley, 2014; Herzel et al., 2017). In a recent study from the Neugebauer lab in fission yeast (Herzel et al., 2018), "all or none" splicing of individual nascent transcripts was discovered, wherein nascent transcripts were more likely than would be expected by chance to contain multiple introns that were either all spliced or all unspliced. This suggested both positive and negative cooperativity among splicing of neighboring introns. These transcripts appeared to have opposite fates regarding 3′ end formation: all spliced transcripts are efficiently cleaved, whereas unspliced transcripts exhibit inefficient 3′ end cleavage. Indeed, crosstalk among introns was observed in human cells at the same time by others (Kim and Abdel-Wahab, 2017; Tilgner et al., 2018). However, those studies did not explore coupling to 3′ end formation. Cleavage of

the nascent RNA by the cleavage and polyadenylation machinery at polyA sites (PAS) releases the RNA from Pol II and the RNA is subsequently polyadenylated (Kumar et al., 2019). Coupling between splicing and 3′ end cleavage is important, because uncleaved transcripts are degraded by the nuclear exosome in *Schizosaccharomyces pombe* (Herzel et al., 2018; Meola et al., 2016; Zhou et al., 2015). Whether 3′ end cleavage efficiency contributes to gene expression levels in mammalian cells is currently unknown.

Previous studies from the Neugebauer lab have shown that only a small portion of the downstream exon may be needed for 3′SS identification and splicing in yeasts (Carrillo Oesterreich et al., 2016; Herzel et al., 2018; Alpert et al., 2020). Interestingly, altering the rate of Pol II elongation affects splicing outcomes, and can introduce widespread changes in alternative splicing (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). Taken together, these findings suggest that transcription elongation rates may govern the amount of downstream RNA available for *cis* regulation at the time that splicing takes place. This in turn would determine which *trans*-acting regulatory factors could be recruited to the nascent RNA to modulate splicing. To obtain mechanistic insights into these processes, we need to understand how mammalian cells—with many more introns per gene and vastly increased levels of alternative splicing compared to yeast—coordinate co-transcriptional splicing with transcription elongation.

1.3 Methods for Studying Co-transcriptional Splicing

One method for studying the process of co-transcriptional splicing involves examining the sequence of nascent RNA directly. Sequencing of nascent RNA reports two critical pieces of information: the position of RNA Pol II during the process of

RNA synthesis (marked by the nascent RNA's 3′ end, which is present in the catalytic center of Pol II) as well as the sequence of the pre-mRNA substrate acted on by the spliceosome and its processing status (splicing, 3′ end cleavage, polyadenylation, or modification) (Alpert et al., 2017) (**Figure 1.2**). Data from nascent RNA sequencing also provide unique information on multiple co-transcriptional RNA processing steps simultaneously, enabling the identification of coupled reactions. For example, we and others have observed coordination among intron splicing events by using long-read sequencing of nascent RNA (Drexler et al., 2020; Herzel et al., 2018; Reimer et al., 2021; Tilgner et al., 2018). Co-transcriptional RNA folding and nucleobase modifications are topics of current intense investigation that can also be analyzed by long-read sequencing (Ke et al., 2017; Liu et al., 2019; Parker et al., 2020; Saldi et al., 2018). Thus, by performing long-read sequencing on nascent RNA, the researcher has the opportunity to track both the progression of Pol II elongation by the 3′ end of the nascent RNA sequence and relate Pol II position to the progression of RNA processing detected in the internal sequence of the nascent RNA. These approaches rely on two important advancements: first, the ability to isolate nascent RNA, and second the maturation of long-read sequencing technology.

1.3.1 Long-Read Sequencing

Advancements in nucleotide sequencing technology now allow for much longer DNA and RNA molecules to be sequenced. Whereas previous "short-read" Next Generation Sequencing (NGS) libraries consist of 50-300 nucleotide fragments, new technologies, termed "long-read" sequencing (LRS), allow DNA and RNA up to hundreds of thousands of bases in length to be sequenced (van Dijk et al., 2018). This technology is able to determine the nucleotide sequence of large tracts of genomic DNA, making it a powerful tool for *de novo* genome assembly (Sohn and Nam, 2018). Long-read sequencing additionally provides a wealth of information

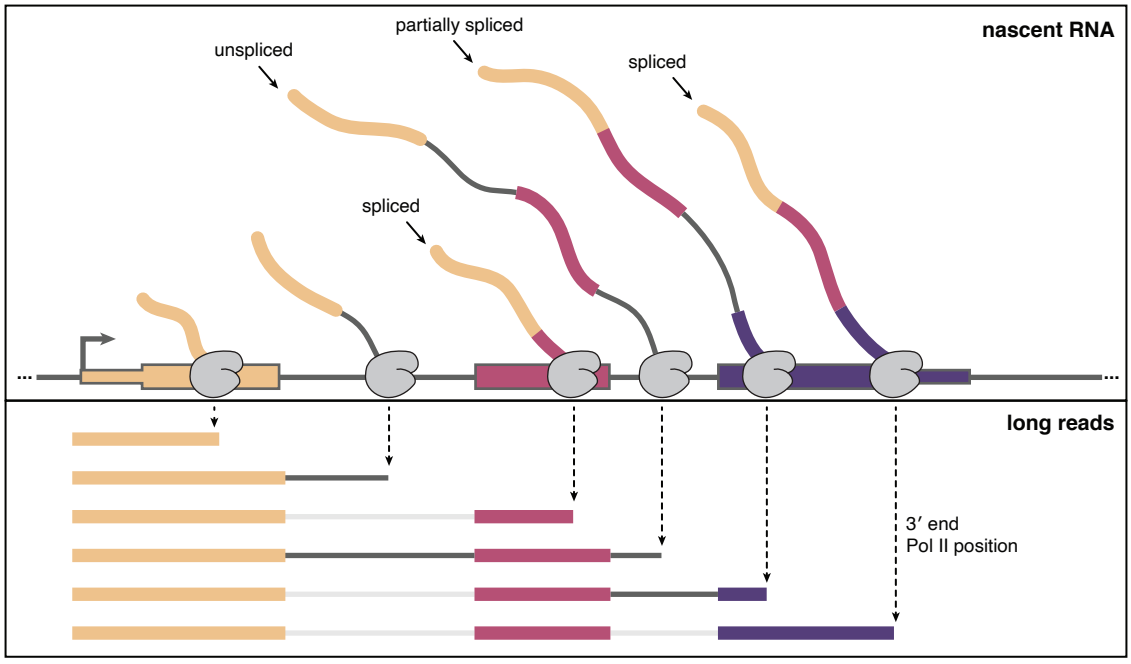

Figure 1.2: Information gained from nascent RNA long-read sequencing. Each nascent RNA molecule contains information about its co-transcriptional splicing status—whether intron sequences are present or not (top panel). Additionally, the 3′ end of each nascent RNA molecule represents the position of Pol II (bottom panel, dotted arrows), as the RNA end is embedded in the elongating Pol II active site at the time of isolating the RNA. Long-read sequencing captures all nucleotide sequence information from 5′ end to 3′ end for each nascent RNA molecule.

about the transcriptome by revealing the identity of full-length mRNA and non-coding RNA transcripts from their 5′ to 3′ ends (Carrillo Oesterreich et al., 2016; Deveson et al., 2018; Hardwick et al., 2019; Herzel et al., 2018; Lagarde et al., 2017; Singh et al., 2019; Tang et al., 2020). Longer reads inherently contain more information, potentially including the unique splicing status, RNA modification status, transcript start site, and polyA cleavage site of each read. Longer reads are less ambiguous when trying to interpret patterns of alternative isoform usage, *i.e.* which exons are ligated together in the same transcript (Tilgner et al., 2015, 2018; Workman et al., 2019). The resulting data thereby characterize the functional genome more completely than short-read NGS data and reveal novel gene products.

The technology behind long-read sequencing was first described in 2003, where a DNA polymerase was used to incorporate fluorescent nucleotides and single molecule sequences could be read out with fluorescent microscopy (Braslavsky et al., 2003). While this initial report only provided "sequence fingerprints" of 5 base pairs in length, the technology has no theoretical maximum on read length, and current read lengths upward of 2 million base pairs have been reported (Payne et al., 2019). Pacific Biosciences and Oxford Nanopore are the leading platforms for long-read sequencing, and both platforms utilize polymerases for reading sequence information, although in slightly different manners (Midha et al., 2019). The biggest advantage of long-read sequencing for RNA is the ability to detect unique isoforms rather than having to infer isoform usage from smaller junctions. With the exception of one study (Kim and Abdel-Wahab, 2017), previous methods for sequencing nascent RNA using short-read sequencing technology have provided information about the splicing status of a single intron, but the reads are not long enough to cover multiple introns, let alone an entire nascent RNA from transcription start site to Pol II active site (Khodor et al., 2011; Tilgner et al., 2012).

1.3.2 Nascent RNA Isolation

Before the advent of biochemical nascent RNA isolation techniques, nascent RNA was first observed directly in a preparation of "chromatin spreads" (Miller and Beatty, 1969). Since nascent RNA makes up such a small fraction of total RNA in the cell, the first experiments tracking nascent RNA molecules were performed on radioactively labeled, highly-abundant species, for example β-globin pre-mRNA, immunoglobin heavy and light chains, and SV40 viral transcripts (Kinniburgh and Ross, 1979; Lai et al., 1978; Schibler et al., 1978). One important finding that led to the ability to biochemically fractionate all nascent RNA from cells was that the ternary complex of Pol II, nascent RNA, and chromatin remains stable even under harsh conditions that compact the chromatin from the surrounding nucleoplasm—up to 2 M urea and 0.3 M NaCl (Wuarin and Schibler, 1994). This allowed chromatin-associated nascent RNA and Pol II to be isolated by relatively low-speed centrifugation. The protocol detailed in Chapter 2 is based on similar approaches for subcellular fractionation using chromatin purification that have been described previously (Pandya-Jones and Black, 2009), and has been adapted previously in the Neugebauer lab for use in budding and fission yeasts (Carrillo Oesterreich et al., 2010; Herzel et al., 2018). In particular, I have further modified this protocol to have a gentler centrifugation speed for collecting nuclei of developing erythroblasts, which have a weaker membrane as they begin the process of enucleation (Pimentel et al., 2016).

Other methods are available for isolating nascent RNA which involve metabolic labeling of newly-transcribed RNAs with a nucleotide analog such as 4-thiouridine (Duffy and Simon, 2016; Garibaldi et al., 2017). While there are some drawbacks to this method, namely difficulty in ensuring unbiased incorporation of the nucleotide analog and possible effects on RNA processing, this method could be considered as an alternative to chromatin-associated RNA purification. Metabolic labeling and

chromatin fractionation are roughly similar in cost and ease of use, however one should be cautioned against using metabolic labeling specifically if the downstream application is to detect splicing, since there is some debate as to how 4-thiouridine incorporation may affect splicing (Testa et al., 1999). Moreover, metabolically labeled RNA is not necessarily nascent, *i.e.* the metabolic label will end up in newly polyadenylated RNA as well. Therefore, some methods of nascent RNA isolation use immunoprecipitation of Pol II as a further enrichment step, and this could be considered if genetic tagging of Pol II is a viable option in your cell type of choice. However, to date, only one lab has reported using this method in conjunction with long-read sequencing (Drexler et al., 2020).

1.3.3 Template-Switching Reverse Transcriptase

A key reagent in the protocol described in Chapter 2 is the template-switching reverse transcriptase (RT) (Zhu et al., 2001). This enzyme is able to synthesize the first strand of cDNA from the nascent RNA template using the blunt ligated adapter at the 3′ end. Then, once the enzyme reaches the 5′ end of the nascent RNA template, it incorporates several nucleotides through non-templated addition, which act as an annealing point for the unique template-switching oligo. The RT can then switch strands and generate the second strand all in the same reaction. Importantly, this means that the resulting cDNA molecule retains the original 5′ and 3′ ends. This is extremely informative for analysis of both Pol II position and transcription start site. The template-switching RT is also more efficient at switching templates in the presence of a 5′-m^{7}G cap, which is installed on mRNAs after only 23 nucleotides of transcription (Rasmussen and Lis, 1993), enriching for full-length cDNA molecules in the final sequencing library (Wulf et al., 2019).

1.4 Erythropoiesis as a Model System in RNA Biology

Erythropoiesis is the developmental pathway by which red blood cells (RBC)—specialized hemoglobin-containing cells that deliver oxygen throughout the body—are produced from hematopoietic stem and progenitor cells (HSPC). Morphologically, erythropoiesis includes the loss of the cell nucleus and acquisition of a characteristic disk-like shape (**Figure 1.3**). Early molecular biologists identified erythropoiesis-associated gene expression patterns such that the globin genes are among the best understood eukaryotic genes. β-globin was among the first proteins to be sequenced and was the first protein to be characterized structurally by using x-ray crystallography. The β-globin gene and mRNA were also among the first to be cloned. These advances facilitated early discoveries in gene regulation, such as the transcriptional control of globin genes by long-range enhancer and repressor elements present in the locus control region (Grosveld et al., 1987). This system of transcriptional regulation is currently being exploited to discover how chromosomal regions interact and how chromatin looping might become a therapeutic target in diseases of RBCs (Krivega and Dean, 2016; Yu and Ren, 2017).

RNA biology is an area in which erythropoiesis, globin gene regulation, and disease mutations have led to fundamental discoveries. Globin pre-mRNA, which contains two introns, was an early model substrate for the investigation of splicing mechanisms (Konkel et al., 1978; Lerner et al., 1980), and mutations in globin genes at the 5′ and 3′ splice sites proved to be the cause of various forms of thalassemias (Maquat et al., 1980). Thalassemias are hemoglobin deficiencies resulting from aberrant globin expression. Some thalassemia mutations cause intron retention (IR) and lead to nonsense-mediated decay (NMD), a major gene regulatory mechanism that degrades mRNA transcripts containing premature termination codons (PTC) present in retained introns (Chang et al., 1979; Maquat et al., 1981). Finally, other thalassemia mutations disrupt the nucleotide sequence that signals 3′ end cleav-

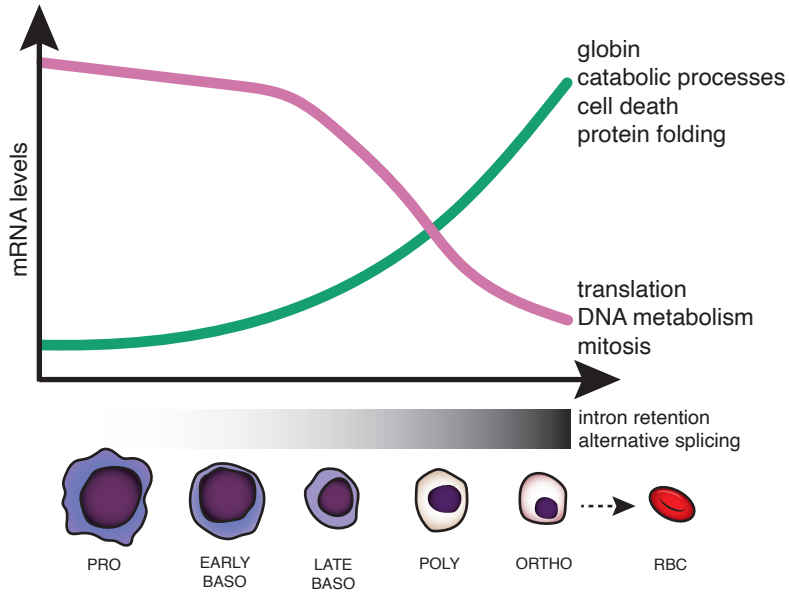

Figure 1.3: Changes in gene expression and splicing occur during terminal erythroid differentiation.
Erythropoiesis is characterized by changes in cell morphology, including nuclear size, color (due to hemoglobinization), and chromatin condensation, which are coordinated with changes in gene expression. During terminal erythroid differentiation, cells progress from proerythroblasts (PRO), to basophilic erythroblasts (EARLY and LATE BASO), to polychromatophilic erythroblasts (POLY), to orthrochromatic erythroblasts (ORTHO), before enucleation to become red blood cells (RBCs) (also called reticulocytes). In human erythroblasts, a subset of genes is downregulated — some top associated Gene Ontology terms are shown to the right (purple line) — while a subset of genes is concurrently upregulated (green line). Changes in splicing occur in the later stages of erythropoiesis (mostly from late baso to ortho), including increased alternative splice site usage and intron retention.

age and polyadenylation of β-globin mRNA, showing the importance of this RNA processing mechanism in health and disease.

1.4.1 Transcriptome-Wide Changes During Erythropoiesis

During erythropoiesis, each cell division is coincident with major changes in gene expression, resulting in daughter cells that are morphologically and transcriptionally distinct from the mother cell (An et al., 2015). Transcriptome-wide profiling using RNA sequencing (RNA-seq) has allowed unbiased dissection of changes that occur along this developmental pathway (An et al., 2014). The greatest number of changes in gene expression in human erythroblasts — either upregulation or downregulation — occurr between the late basophilic to polychromatic and polychromatic to orthochromatic stages, where roughly equal numbers of genes are upregulated as are downregulated. These transcripts are enriched for different annotated functions reflecting cellular events in the differentiation process (**Figure 1.3**), emphasizing the changing transcriptional landscape that underlies massive globin gene expression in terminal stages of erythroid development. In contrast, when mouse erythroblasts are analyzed in the same manner, the overwhelming majority of genes, including key transcription factors, are downregulated. The cause for species-specific differences in transcriptome changes is not immediately clear but likely reflects distinct properties of human and mouse erythrocytes, including differences in size, life span, oxygen-carrying capacity, and metabolism (An et al., 2015). An analysis of chromatin accessibility, histone modifications, and transcription factor binding in a mouse embryonic stem cell model of hematopoiesis has revealed a complex regulatory network that drives changes in the transcriptome during differentiation (Goode et al., 2016). It remains to be seen whether these mechanisms differ between human and mouse erythroblasts, explaining the pronounced differences in transcriptomes.

1.4.2 Splicing Regulation in Normal Erythropoiesis

Splicing can contribute to the regulation of transcript levels by activating cellular programs, such as NMD, to reduce transcript levels. Moreover, alternative splicing leads to the expression of different transcripts and protein products from the same gene (Papasaikas and Valcarcel, 2016). How does splicing regulation contribute to transcriptome diversity in erythroid development? Early work using microarrays to detect changes in splicing during erythropoiesis found altered splicing in known trans-acting splicing factors (for example, SNRP70, HRNPLL, and MBNL2), which are RNA-binding proteins that regulate how the spliceosome assembles on pre-mRNA and how different 5′ and 3′SSs are chosen (Yamamoto et al., 2009). This suggested a regulatory feedback loop, whereby changes in splicing factors could affect the splicing of many downstream genes necessary for development. Subsequent work has focused on identifying stage-specific changes in splicing transcriptome-wide by using RNA-seq, a less biased approach that does not rely on known intron–exon boundaries (An et al., 2015). In addition, mapping the gene expression networks governed by splicing in erythroid differentiation has aided identification of the functional significance of splicing regulation (Conboy, 2017).

One of the first and best-characterized examples of alternative splicing in erythropoiesis is the stage-specific inclusion of exon 16 of the *4.1R* protein-encoding gene, which is crucial for erythrocyte membrane integrity (Yamamoto et al., 2009; Hou et al., 2002). Changes in expression levels and specific binding of the hnRNP A/B protein affect this developmental switch (Hou et al., 2002). Since then, alternative splicing has emerged as a more widespread phenomenon (Baralle and Giudice, 2017; Cheng et al., 2014; Wong et al., 2018). The muscleblind-like protein 1 (MBNL1) is a sequence-specific splicing factor that undergoes extensive alternative splicing during differentiation (Cheng et al., 2014). Cheng *et al*. showed that a specific *Mbnl1* isoform which includes the alternative exon 5 accumulates in the

nucleus in later stages of erythroid differentiation (Cheng et al., 2014). The inclusion *Mbnl1* isoform is responsible for regulating the splicing of downstream genes important for erythroid differentiation, as knockdown of the *Mbnl1* inclusion isoform alone blocked differentiation and caused defects in proliferation. Mirroring the previous findings observed by microarrays, Pimentel *et al*. report a program of highly dynamic alternative isoform switching in late-stage human erythroblasts using RNA-seq (Pimentel et al., 2014). An increase in steady-state levels of transcripts containing PTCs, which likely trigger NMD of these transcripts, was observed in the later stages of differentiation, suggesting that alternative splicing coupled to NMD may be a novel, stage-specific gene regulatory mechanism.

1.4.3 Intron Retention During Hematopoiesis

Intron retention (IR) is a class of alternative splicing wherein an intron is not removed by the spliceosome, potentially introducing PTCs and targeting the transcript for NMD. Alternatively, it is possible that certain intron-retained transcripts remain in the nucleus and undergo splicing with delayed kinetics (Boutz et al., 2015; Bhatt et al., 2012; Boothby et al., 2013; Mauger et al., 2016). IR was only recently recognized as a widespread occurrence (Wong et al., 2013; Braunschweig et al., 2014), and developing erythroid cells exhibit robust IR. Pimentel *et al*. showed that late human erythroblasts accumulate hundreds of transcripts containing retained introns and that the formation of these IR transcripts are enriched for splicing factors and iron-homeostasis factors (Pimentel et al., 2016). These results were corroborated at the single-cell level in human immortalized myelogenous leukemia K562 cells (Abdelmoez et al., 2018). The top three categories of nuclear IR transcripts by gene ontology analysis were RNA metabolism, RNA splicing, and the C complex spliceosome. The retained introns detected in late human erythroblasts were more likely to be found next to alternative exons that contained PTCs (Pimentel

et al., 2016), in line with previous studies suggesting that IR followed by NMD is an important mechanism that regulates levels of splicing factors (Ni et al., 2007).

How is IR triggered during erythropoiesis? Key insights are emerging from studies of transcripts encoding the important core splicing factor SF3B1. SF3B1 expression is also subject to IR during erythroid differentiation, and work suggests that SF3B1 regulation by IR may constitute a regulatory hub leading to the downregulation of transcripts encoding other splicing factors (Pimentel et al., 2016). Indeed, a series of highly conserved cryptic SSs were identified for their activity in promoting IR in SF3B1 transcripts (Parra et al., 2018). The identified intronic sequences are sufficient to promote IR in SF3B1 and can also promote retention when inserted into other introns. The cryptic exons generated by these SSs are proposed to act as splicing decoys, sequestering components of the spliceosome and ultimately preventing productive splicing by blocking the appropriate cross-intron interactions needed to define the intron for splicing. Alternatively, reduced levels of SF3B1 might preferentially affect splicing efficiencies or the half-lives of pre-mRNAs/mRNAs encoding splicing factors or both. Although we presume that most of these instances reflect the downregulation of IR transcripts, the possibility that certain splicing events are delayed remains. Interestingly, delaying gene expression through IR is physiologically relevant in other cell types, including developing spermatocytes, neuronal cells, platelets, granulocytes, and stimulated macrophages (Wong et al., 2013; Yap et al., 2012; Edwards et al., 2016; Denis et al., 2005; Naro et al., 2017). In the case of erythroid differentiation, how introns are retained in a seemingly stage-specific and cell type-specific way remains to be fully understood.

1.4.4 Misregulation of Splicing in β-Thalassemia

Misregulation of splicing underlies a growing number of human diseases (Urbanski et al., 2018; Dvinge et al., 2016; Scotti and Swanson, 2016; Faustino and Cooper, 2003). Generally, mutations either in *cis* or in *trans* can affect splicing outcomes. *Cis* mutations may disrupt the intrinsic sequences that demarcate SSs in a transcript (5′ and 3′SSs). In contrast, mutations in any member of the core spliceosome machinery can produce splicing defects in *trans*, causeing deleterious effects for a large number of downstream splicing substrates. Both types of splicing defects have been characterized in the erythroid lineage (Hahn et al., 2015).

β-thalassemias are a family of disorders defined by mutations in the β-globin gene, causing a reduction of β-globin mRNA, insufficient hemoglobinization of maturing RBCs, and anemia (Thein, 2013). β-thalassemia is one of the most prevalent diseases caused by somatic mutations worldwide, yet currently the only available curative treatment is an allogenic transplant of HSPCs from a matched donor. This option is unavailable for many patients because of the cost of treatment and limited availability of matched donors. Although we possess a quite thorough understanding of the molecular basis and pathophysiology of this disease, better treatments are sorely needed. Many β-thalassemia patients are dependent on transfusions from blood donors to maintain proper levels of healthy, circulating RBCs. However, this therapy often leads to complications related to iron overload, including organ damage. The majority of β-globin alleles that cause thalassemia contain point mutations (**Figure 1.4**). These mutations can affect virtually any step in the correct expression of β-globin mRNA from transcription initiation (**Figure 1.4 A**), to splicing (**Figure 1.4 B, C**), to 3′ end cleavage and polyadenylation (**Figure 1.4 E**). Because of this, β-thalassemia is an attractive target for applying genome-editing tools to correct β-globin mRNA processing and expression, providing a potential cure for β-thalassemia.

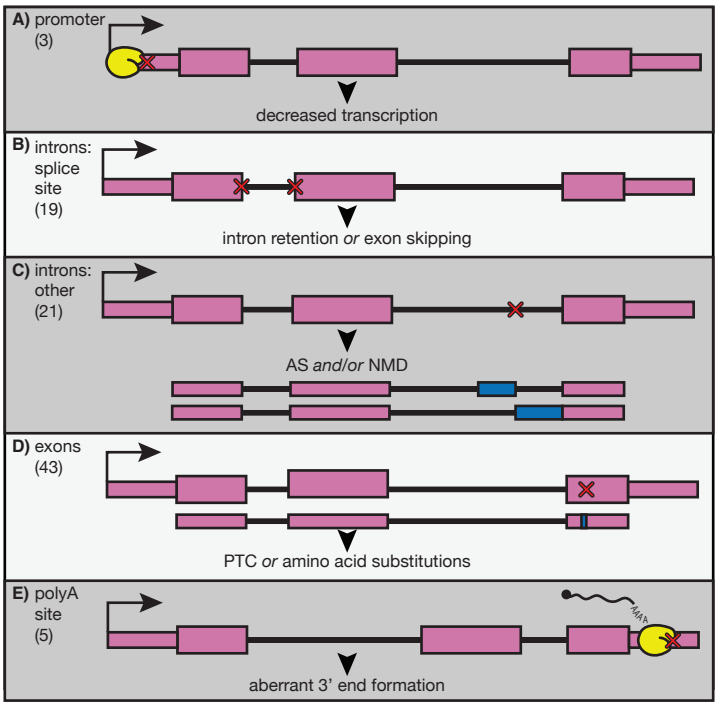

Figure 1.4: Single-nucleotide mutations in key regulatory regions of the β-globin gene disrupt expression in β-thalassemia.
Point mutations in varying regions of the β-globin gene are shown schematically, and the frequencies of these mutations listed in the HbVar database (http://globin.bx.psu.edu/hbvar/menu.html) are shown in brackets at the left. Gene regions are divided into **(A)** promoter, **(B)** splice sites, **(C)** other intronic regions, **(D)** exons, and **(E)** polyadenylation site. These mutations (red X's) have varying effects, illustrated below each example, but all lead to decreased or abolished expression of the β-globin transcript. AS, alternative splicing; NMD, nonsense-mediated decay; PTC, premature termination codon.

1.5 Summary

Erythropoiesis provides an excellent model in which to study RNA splicing in both healthy and diseased states. Mutations affecting genes important for mature RBC function (for example, β-globin) have revealed aberrant splicing which leads to hindered erythropoiesis, often in unexpected ways. In the following chapters, I report an analysis of nascent RNA transcription and splicing in murine erythroid leukemia (MEL) cells undergoing erythroid differentiation, a developmental program that exhibits well-known, drastic changes in gene expression (An et al., 2014; Reimer and Neugebauer, 2018). I 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. I have rigorously determined the spatial window in which co-transcriptional splicing occurs and defined co-transcriptional splicing efficiency for thousands of mouse introns, Pol II elongation behavior across splice junctions, and the effects of efficient co-transcriptional splicing on 3′ end cleavage. Specifically, a patient-derived β-thalassemia allele with a single point mutation at the 3′SS shows an increase in both co-transcriptional splicing efficency and 3′ end cleavage efficiency compared to the wild type allele. These findings identify the pre-mRNA substrates of splicing and show that splicing of multiple introns within individual transcripts is coordinated with 3′ end cleavage. In particular, the demonstration of highly efficient splicing in the absence of transcriptional pausing causes us to rethink key features of splicing regulation in mammalian cells.

Chapter 2

Methods

2.1 Preparation of Nascent RNA for Long-Read Sequencing

In this protocol, chromatin-associated RNA is purified from murine erythroleukemia (MEL) cells by first performing subcellular fractionation to physically separate chromatin (Basic Protocol 1: Subcellular Fractionation). Then, nascent RNA is enriched by depleting polyadenylated RNA (polyA+ RNA) and ribosomal RNA (rRNA), and nascent RNA 3′ ends are ligated to a unique adapter to retain the position of the Pol II active site when RNA was isolated (Basic Protocol 2: Nascent RNA Isolation and Adapter Ligation). Finally, a template-switching reverse transcriptase is used to generate a full-length cDNA copy of the nascent RNA. Minimal polymerase chain reaction (PCR) cycles are used to amplify a cDNA library before long-read sequencing (Basic Protocol 3: cDNA Amplicon Preparation). See **Figure 2.1** for an overview of this protocol.

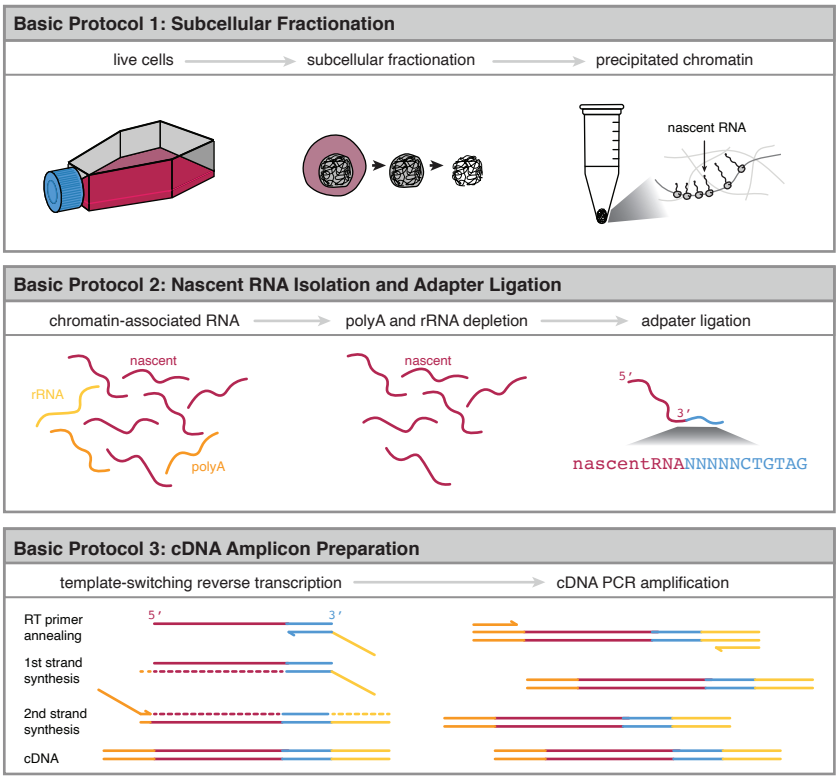

Figure 2.1: Overview of nascent RNA isolation and preparation of long-read sequencing library protocol.

2.1.1 Critical Parameters

Attaining enough nascent RNA from cells is critical for the success of this protocol. The limiting reagent is nascent RNA that has been depleted of polyA+ RNA and rRNA, of which you need 600 ng for the downstream adapter ligation reaction. However, the polyA+ depletion and rRNA depletion steps can have a fairly low yield. The yield for polyA+ depletion is often approximately 70%, and the yield for rRNA depletion can be 5-10%. Thus, I suggest not continuing to Basic Protocol 2 unless you have at least a total of 14 μg of chromatin-associated RNA at the end of Basic Protocol 1. In my experience, 20 million MEL cells yield approximately 5 μg of chromatin-associated RNA, so fractionating 3 x 20 M cells in parallel and combining the chromatin-associated RNA after isolating from Trizol should be sufficient. However, the yield of nascent RNA is variable between cell types, so this should be monitored closely. The quality of the nascent RNA is also extremely important, as this will directly affect the quality of the downstream cDNA library. The RNA 260/280 and 260/230 values should be monitored by Nanodrop where mentioned, and the RNA can be quickly inspected by running it on an agarose gel (**Figure 2.3**). In addition to after collecting the chromatin-associated RNA (Basic Protocol 2, step 22), the user could run the RNA on an agarose gel after polyA+ depletion if they suspected RNA degradation based on the Nanodrop reading.

Subcellular fractionation (Basic Protocol 1) should be optimized for each cell type, and success should be monitored by Western blotting. A successful fractionation is one where Pol II remains in the chromatin fraction, and the cytoplasmic and nucleoplasmic markers are strongly depleted from the chromatin fraction (**Figure 2.2**). When attempting this protocol with another cell type, the main step that will need to be optimized is the subcellular fractionation. Since this protocol is written for MEL cells which are erythroid cells, the centrifugation step to isolate nuclei has been decreased to prevent their slightly weaker nuclear membranes from burst-

ing during centrifugation (Basic Protocol 1, step 7). For most other cell types, this centrifugation speed should be gradually increased and the nuclei and cytoplasm fractions should be observed under a light microscope to observe that nuclei are intact. If cells that grow in monolayer are used, then they should be rinsed quickly with 1X PBS + 1 mM EDTA on the plate before using a cell scraper to harvest them (in place of Basic Protocol 1, steps 1 and 2). The yield of chromatin-associated RNA is variable between different cell types, so this should be monitored closely, and the number of input cells should be adjusted so that the user yields a total of approximately 14 μg of chromatin-associated RNA at the end of Basic Protocol 2, RNA Isolation. However, the subcellular fractionation tends to work best with no more than 20 million cells in the tube, so adjust the number of tubes that are combined rather than adjusting the number of cells in the tube.

Multiple samples can be sequenced on the same flow cell if a barcode is introduced during the reverse transcription step. See **Table 5.1** for a list of barcode sequences that can be inserted in the custom RT primer used for this step. The barcoded RT primer can be ordered from any oligonucleotide synthesis company but should be ordered PAGE-purified.

During the PCR cycle number optimization, the goal is to use as few rounds of PCR amplification as possible to minimize amplification bias in the library. A cycle number of 11-15 is usually chosen, and anything more than 18 cycles should be reconsidered. If this happens, consider adjusting the amount of cDNA input in the PCR reaction.

2.1.2 Strategic Planning

MEL Cell Growth

Before beginning, the user should have MEL cells actively growing. These cells grow in suspension in DMEM + GlutaMAX (Thermo Fischer Scientific, cat. no.

10569-010) supplemented with Fetal Bovine Serum (Thermo Fischer Scientific, cat. no. 26140) and Penicillin-Streptomycin (Thermo Fischer Scientific, cat. no. 15140122) in non-treated T-25 flasks (Thermo Fischer Scientific, cat. no. 169900). For more information on MEL cell growth and induction, see (Antoniou, 1991). Cells should be counted with a hemocytometer to determine their density, and the culture density should be recorded for several days before beginning the protocol to ensure consistent doubling is occurring. Cultures should be diluted back to 5 x 10^4 cells/ml once they reach a density of 1-2 x 10^6 cells/ml. Once cells are in active growth, they should double in number approximately every 12 hours. At this point, the cells are ready to use for this protocol.

Minimum Cell Input

For one replicate to be prepared using this protocol, approximately 80 million cells are needed as input: 60 million cells will be used to isolate nascent RNA and 20 million cells will be used to test the success of subcellular fractionation by Western blot. This number of cells can often be gathered from 4 flasks of cells, each containing 20 million cells (10 ml culture at a density of 2 x 10^6 cells/ml).

2.1.3 Basic Protocol 1: Subcellular Fractionation

Here, the user will perform subcellular fractionation to separate the chromatin from the nucleoplasm and cytoplasm of mammalian cells. This approach is based on the evidence that under harsh conditions that lyse the nuclear membrane and dissociate any non-specifically bound RNA, the ternary complex of Pol II, chromatin, and nascent RNA remains intact and precipitates from solution (Wuarin and Schibler, 1994). After subcellular fractionation, the user will collect a sample from each fraction and check that characteristic marker proteins are in each fraction by Western blot. If the fractionation is successful, the user will then continue to collect nascent

RNA from the chromatin fraction.

Care should be taken to minimize time in between steps, especially after cell lysis and nuclear lysis steps. In order to monitor the success of fractionation, a minimum of two samples should be fractionated in parallel. For one sample, each fraction is kept for use in Western blotting. For the second sample, the chromatin fraction is used to isolate RNA. All buffers used from the point of cell lysis to pelleting the chromatin contain α-amanitin, an inhibitor of Pol II, which prevents Pol II from elongating and allows the position of Pol II to be accurately captured. Note that buffers containing α-amanitin, a potent toxin, should be handled with extreme care, and all buffer waste should be disposed of properly. All buffers should be freshly prepared and chilled on ice before beginning. This protocol is specific for murine erythroleukemia cells but has been easily adapted to other mammalian cell types in our lab. This protocol works best on freshly harvested, not frozen, cells.

Materials

Reagents and Chemicals

- MEL cells in active growth
- 1X PBS (Americanbio, cat. no. AB11072-01000) + 1 mM EDTA (Americanbio, cat. no. AB00502-00100)
- Cell lysis buffer (see recipe in **Reagents and Solutions**, page 54)
- Sucrose buffer (see recipe in **Reagents and Solutions**, page 54)
- Nuclear resuspension buffer (see recipe in **Reagents and Solutions**, page 54)
- Nuclear lysis buffer (see recipe in **Reagents and Solutions**, page 54)
- Trizol Reagent (Thermo Fischer Scientific, cat. no. 15596026)

- NuPAGE 4X LDS Sample Buffer (Thermo Fischer Scientific, cat. no. NP0007)
- 4-12% Bis-Tris gel (Thermo Fischer Scientific, cat. no. NP0322PK2, or poured in house)
- Precision Plus Protein Dual Color Standards (BioRad, cat. no. 1610374)
- NuPAGE MOPS-SDS running buffer (Thermo Fischer Scientific, cat. no. NP0001)
- NuPAGE transfer buffer (Thermo Fischer Scientific, cat. no. NP0006)
- 5% bovine serum albumin (BSA) in 0.1% TBS-T (Tris-buffered saline, 0.1% Tween 20)
- 0.1% TBS-T
- 3% BSA in 0.1% TBS-T
- anti-GAPDH antibody (Santa Cruz, cat. no. sc-25778)
- anti-U1-70K (CB7) antibody (hybridoma supernatant available upon request)
- anti-NXF1 TAP N-19 antibody (Santa Cruz, cat. no. sc-17310)
- anti-Pol II (4H8) antibody (Santa Cruz, cat. no. sc-47701)
- Mouse IgG, HRP-linked whole Ab (Cytiva, cat. no. NA931)
- Rabbit IgG, HRP-linked whole Ab (Cytiva, cat. no. NA934)
- Pierce ECL Western blotting substrate (Thermo Scientific, cat. no. 32106)
- 15-ml conical polypropylene tubes (Sigma, cat. no. CLS430791)

Equipment

- Tabletop centrifuge
- Wide bore P1000 pipette tips (VWR, cat. no. 89049-166)
- Refrigerated microcentrifuge
- 1.5-ml tubes (Dot Scientific, cat. no. RN1700-GMT)
- vortex
- Sonicator (Branson, fitted with a 2 mm microtip probe)
- Heat block at 95°C
- 0.2 μm nitrocellulose membrane (BioRad, cat. no. 1620112)
- Chemiluminescent Imager

Protocol Steps

Cell Fractionation

1. Count actively growing MEL cells using a hemocytometer. For each replicate, aliquot 20 million cells from each flask into a 15-ml conical tube, for a total of 4 tubes with 80 million cells. Centrifuge all tubes 5 min at 1,500 revolutions per minute (rpm), room temperature.

   Note: *As mentioned in the strategic planning section, three of the tubes will be used for nascent RNA isolation, and one tube will be used for confirming the success of subcellular fractionation by western blot. Perform all steps through step 15 on all 15-ml tubes simultaneously, each with 20 million cells.*

2. Gently resuspend the cell pellet in 1 ml ice-cold PBS + 1 mM EDTA by pipetting up and down approximately 5 times while using a wide bore P1000 pipette tip.

3. Centrifuge 5 min at 1,500 rpm, 4°C.

4. Gently resuspend the cell pellet in 250 μl cell lysis buffer.
   Note: *Pipette up and down while using a wide bore P1000 pipette tip just until cells are in a turbid suspension, approximately 10 times up and down. Some small clumps of cells are OK.*

5. Incubate 5 min on ice.

6. Add 500 μl of sucrose buffer to a new 1.5-ml tube and carefully layer the cell lysate on top.

7. Centrifuge the tube 10 min at 2,000 rpm, 4°C in a microcentrifuge.

8. Aspirate the supernatant (this is the cytoplasmic fraction), and transfer to a new 1.5-ml tube. Set aside on ice.
   Note: *Be careful not to disturb the pellet at the bottom of the tube.*

9. Rinse the white nuclear pellet with 500 μl of ice-cold PBS + 1mM EDTA by pipetting the solution down the side of the tube to avoid disturbing the pellet, then aspirating the solution completely.

10. Resuspend the nuclear pellet in 100 μl of nuclear resuspension buffer by pipetting the buffer on top of the pellet, then gently flicking the closed tube.
    Note: *It is easiest to drag the tube across a bumpy surface, such as across the holes of a tube rack several times. The nuclei should easily resuspend into a somewhat turbid suspension.*

11. Add 100 μl of nuclear lysis buffer, then vortex for 5 seconds.

12. Incubate 3 min on ice.

13. Centrifuge for 2 min at 14,000 rpm, 4°C.

14. Aspirate the supernatant (this is the nucleoplasmic fraction), and transfer to a new 1.5-ml tube. Set aside on ice.

15. Rinse the chromatin pellet left in the tube with 500 μl of ice-cold PBS/1 mM EDTA.

    Note: *Make sure to remove as much supernatant as possible to remove nucleoplasmic RNA contamination in chromatin-associated RNA. The chromatin pellet should be very stable.*

16. Add 100 μl ice-cold PBS to one tube (this is the chromatin fraction for Western blot), then set aside on ice. To the other three tubes, add 100 μl ice-cold PBS and 300 μl Trizol, then vortex briefly just until the pellet releases from the bottom of the tube.

    Note: *The chromatin pellet should be very insoluble. It will not dissolve in Trizol.*

17. Either transfer the three tubes containing Trizol to -80°C for up to 1 month or continue immediately to nascent RNA isolation (Basic Protocol 2).

Western Blot

18. Place the 1.5-ml tubes containing the cytoplasmic fraction (from step 8), nucleoplasm fraction (from step 14), and chromatin fraction (from step 16) on ice. Adjust the volume in each tube with PBS so that the three fractions contain an approximately equal volume.

    Note: *The cytoplasmic fraction should be the largest, approximately 750 μl.*

19. For the nucleoplasm and chromatin fractions, sonicate on ice at 30% amplitude for 1 min total, with 10 seconds on followed by 20 seconds off.

20. Centrifuge all three tubes (cytoplasm, nucleoplasm, and chromatin) for 10 min at 14,000 rpm, 4°C.

21. Aliquot 20 μl of the supernatant from each tube into a new 1.5-ml tube.

22. Add 5 μl 4X LDS sample buffer to each tube. Mix by pipetting up and down.
    Note: *Try to avoid creating bubbles when pipetting up and down by keeping the tip submerged.*

23. Incubate 5 min at 95°C.

24. Centrifuge 1 min at 14,000 rpm, room temperature.

25. Load cytoplasm, nucleoplasm, and chromatin samples on a 4-12% Bis-Tris gel alongside 10 μl of prestained ladder. Run the gel in 1X MOPS-SDS running buffer at 180 V until the dye front is just off the bottom of the gel, approximately 50 minutes.
    Note: *Aim to load the prestained ladder in the outermost two lanes to use as a guide for cutting the membrane later. For more details on SDS gel electrophoresis, see (Gallagher, 2012).*

26. Transfer to a 0.2 μm nitrocellulose membrane in 1X NuPAGE transfer buffer for 2 h at 30 V in a cold room at 4°C.
    Note: *For more details on Immunoblotting, see (Ni et al., 2017).*

27. Rinse the nitrocellulose membrane briefly in distilled water.

28. Block the membrane in 5% BSA in 0.1% TBS-T overnight in a cold room at 4°C on a nutator.
    Note: *Incubate the membrane in a closed container to prevent evaporation, and make sure to add enough solution to cover the membrane completely.*

29. Rinse the membrane at least 1 hour in 0.1% TBS-T at room temperature in a nutator.

30. Cut the membrane using a sterile blade and a hard flat edge at the 100 kDa and 50 kDa ladder marks to separate the membrane into three pieces. Incubate the membranes with primary antibodies listed in **Table 2.1** in 3% BSA in 0.1% TBS-T for at least 1 hour at room temperature: incubate the top portion of the blot with the α-Pol II 4H8 antibody, the middle portion of the blot with the α-U1-70K antibody, and the bottom portion of the blot with the α-GAPDH antibody. After incubation, rinse at least 3 times for 10 minutes each with fresh 0.1% TBS-T in a nutator.

    Note: *The antibody concentration and time for incubation will need to be optimized for antibodies for other cell types or antigens. Note that I propose an alternative marker for the nucleoplasm, NXF1, as U1-70K does not work well in all cell types. I typically use antibody dilutions of 1:2,000 for α-Pol II 4H8, 1:2,000 for α-GAPDH, and 1:5 for α-U1-70K.*

Table 2.1: Primary antibodies for verification of subcellular fractionation.

Antibody	Localization	Running Size	Source Organism
α-GAPDH	cytoplasm	37 kDa	rabbit
α-U1-70K	nucleoplasm	70 kDa	mouse
α-NXF1	nucleoplasm	70 kDa	mouse
α-Pol II (4H8)	chromatin	240 kDa	mouse

31. Incubate the membranes with secondary antibodies in 3% BSA in 0.1% TBS-T for at least 1 hour at room temperature: use α-rabbit HRP for the GAPDH membrane, and α-mouse HRP for the U1-70K and Pol II 4H8 membranes.

    Note: *I typically use antibody dilutions of 1:10,000 for α-rabbit HRP, and 1:8,000 for α-mouse HRP.*

32. Rinse the membrane a final 4 times for at least 10 minutes each in fresh 0.1%

TBS-T at room temperature in a nutator after secondary antibody incubation.

33. Cover the membrane in approximately 1 ml prepared ECL Western blotting substrate for approximately 60 seconds and expose the membrane to film or a digital chemiluminescence reader. See **Figure 2.2** for an example of a successful fractionation.

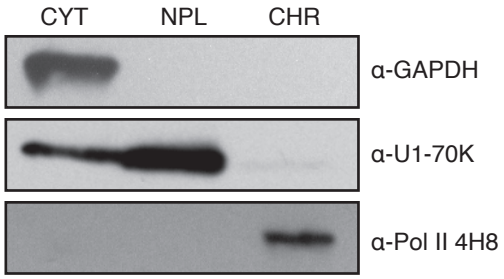

Figure 2.2: Western blot after successful subcellular fractionation of MEL cells. Cytoplasm (CYT), nucleoplasm (NPL), and chromatin (CHR) fractions are loaded from left to right. Primary antibodies for each blot are shown to the right; GAPDH is a marker for the cytoplasm, U1-70K is a marker for the nucleoplasm, and Pol II 4HD is a marker for RNA Pol II, which should be in the chromatin fraction.

2.1.4 Basic Protocol 2: Nascent RNA Isolation and Adapter Ligation

Nascent RNA refers to RNA which is being actively synthesized by Pol II. While subcellular fractionation enriches for RNA that is physically bound to the ternary complex of Pol II on chromatin, further purification is needed to remove contaminating mRNA (*i.e.* not nascent) and rRNA. In this protocol, ribosomal RNA and chromatin-associated polyadenylated RNA are depleted to enrich for nascent RNA. While a number of alternative methods are commercially available for both of these procedures, in my experience, kits with complementary oligos conjugated to magnetic beads provide the fastest and most reliable approach for both methods, and

that approach is described here. After that process, a DNA adapter which is necessary for priming reverse transcription in the downstream protocol is blunt ligated to the 3′ end of each nascent RNA. This adapter sequence is used in data analysis to find the exact position where the nascent RNA 3′-OH was purified from a Pol II active site.

Materials

Reagents and Chemicals

- Chromatin-associated RNA (from Basic Protocol 1 step 17)
- Chloroform
- 100% ethanol, room temperature
- RNeasy Mini kit (Qiagen, cat. no. 74104)
- RNase-Free DNase Set (Qiagen, cat. no. 79254)
- DynaBeads mRNA DIRECT Micro Purification kit (Thermo Fischer Scientific, cat. no. 61021)
- UltraPure Glycogen (Thermo Fischer Scientific, cat. no. 10814010)
- 3 M sodium acetate
- 100% ethanol, ice-cold
- 70% ethanol, ice-cold
- 2X Novex Sample Buffer (Thermo Fischer Scientific, cat. no. LC6876)
- DNA Clean and Concentrator-5 kit (Zymo Research, cat. no. D4013)
- RiboMinus™ Eukaryote System v2 (Thermo Fischer Scientific, cat. no. A15026)

- TAE buffer (40 mM Tris·HCl, 20 mM acetic acid, 1 mM EDTA pH 8.0)
- Agarose (Sigma, cat. no. A9539)
- Lonza Gelstar (Thermo Fischer Scientific, cat. no. 50535)
- GeneRuler 1 kb Plus DNA ladder (Thermo Fischer Scientific, cat. no. SM1331)
- DNA adapter (/5rApp/NNNNNCTGTAGGCACCATCAAT/3ddC/)
- T4 RNA ligase kit (NEB, cat. no. M0351L)

Equipment

- Vortex
- Thermomixer
- Refrigerated microcentrifuge
- 2-ml tube (Dot Scientific, cat. no. RN2000-GMT)
- 1.5-ml tubes (Dot Scientific, cat. no. RN1700-GMT)
- Nanodrop
- UV gel imaging system
- Magnetic 1.5-ml tube rack
- Heat block at 70°C
- Heat block at 37°C
- Heat block at 65°C
- 0.2-ml PCR strip tube (Dot Scientific, cat. no. 415-8PCR)
- Thermal cycler

Protocol Steps

RNA Isolation

1. Thaw three tubes with chromatin pellets frozen in Trizol (from Basic Protocol 1, step 17) at room temperature just until samples are liquid. Vortex briefly to mix. If samples were not frozen, proceed immediately to the next step.
   Note: *For MEL cells, you should plan to isolate nascent RNA from at least 3 chromatin fraction pellets (as described in the Strategic Planning section). You will isolate RNA from each of the three (or more) tubes separately, then the RNA will be combined before the polyA + depletion. Perform all steps through step 21 on all tubes simultaneously.*

2. Incubate samples at 50°C for 10 mins with shaking at 1,400 rpm in Thermomixer.
   Note: *This aids in releasing RNA from the insoluble chromatin and improves the yield of purified RNA.*

3. Add 60 μl chloroform to each tube. Vortex thoroughly (at least 30 seconds).
   Note: *Chloroform should be used in a fume hood. Make sure the tube lid is closed securely before vortexing, for example by wrapping the lid with parafilm, as Trizol is caustic.*

4. Incubate 2 min at room temperature.

5. Centrifuge 15 min at 14,000 rpm, 4°C.

6. Transfer the clear upper aqueous phase to a new labeled 2 ml tube. The aqueous phase should be about 250 μl.
   Note: *Make sure to avoid carrying over any interphase. Using a smaller pipette tip (P200) in multiple aliquots is easier than using a larger tip.*

7. Add 3.5 volumes of RLT buffer (from RNeasy Mini kit) to the tube, then mix by vortexing briefly.
   Note: *For example, if the aqueous phase is 250 μl, add 875 μl RLT buffer.*

8. Add 2.5 volumes of room temperature 100% ethanol to the tube. Mix well by pipetting up and down; do not centrifuge. Spin the tube down briefly (approximately 5 seconds) to collect drops from the lid.

9. Transfer the sample, up to 700 μl at a time, to an RNeasy Mini spin column in a collection tube.

10. Centrifuge for 15 seconds at 14,000 rpm, room temperature. Discard flow-through and repeat this step as necessary to pass the entire sample through the column.

11. Add 350 μl of buffer RW1 (from RNeasy Mini kit) to the RNeasy spin column, centrifuge as in step 10, and discard flow-through.

12. For each sample, prepare 80 μl of "DNAse I incubation mix" by adding 10 μl of DNase I stock solution and 70 μl of Buffer RDD (both from the RNase-free DNase kit) to a new 1.5-ml tube.
    Note: *Make a master mix here for as many tubes as you have.*

13. Mix the DNAse I incubation mix by gently inverting the tube end over end several times. Centrifuge briefly to collect drops from the lid.
    Note: *Do not vortex to mix!*

14. Add the DNase I incubation mix (80 μl) directly to the RNeasy Mini spin column membrane.

15. Incubate 15 min at room temperature.

16. Add 350 μl of Buffer RW1 to the spin column, centrifuge as in step 10, and discard the flow-through.

17. Add 500 μl of Buffer RPE (from RNeasy Mini kit) to the spin column, centrifuge as in step 10, and discard the flow-through.

18. Add 500 μl of Buffer RPE to the spin column, centrifuge 2 min at 14,000 rpm at room temperature, and discard the flow-through.

19. Carefully remove the spin column from the collection tube and transfer the column to a new 1.5-ml tube.

20. Add 30 μl of RNase-free water directly to the center of the column membrane. Centrifuge for 1 min at 10,000 rpm at room temperature to elute the RNA.

21. Measure the concentration of eluted RNA by Nanodrop. Measure the concentration of each sample separately, and if the 260/280 and 260/230 values are acceptable, combine the replicates for each sample into one 1.5-ml tube. ***Note:*** *Users should expect a concentration of 150-200 ng/μl with a 260/280 value around 2.0, and a 260/230 value around 2.1. Note that a minimum of 10 μg of nascent RNA is recommended after pooling and before continuing to polyA+ depletion.*

22. Run 250-500 ng of each pooled nascent RNA sample on a 1% TAE agarose gel to confirm the integrity of the RNA. See **Figure 2.3** for an example of intact nascent RNA.

    (a) Aliquot the nascent RNA into a new 1.5-ml tube. Adjust the volume to 5 μl with sterile water.

    (b) Add 5 μl 2X Novex Sample Buffer to the 1.5-ml tube.

    (c) Incubate the tube in a heat block at 65°C for 5 minutes, then immediately transfer to ice for at least 1 minute or longer.

(d) Load the sample on a 1% agarose TAE gel alongside GeneRuler 1 kb Plus DNA ladder.

(e) Run the gel at 85 V for 45 minutes, then image using a UV gel imaging system.

Note: *Keep the remaining nascent RNA on ice while the gel is running before proceeding to the next step. Optionally, stop here by storing the nascent RNA at -80°C for up to one month.*

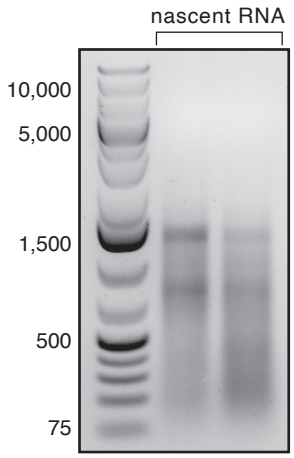

Figure 2.3: Agarose gel showing intact chromatin-associated RNA.
A high-quality, intact sample should appear as a smear of different sizes of RNA, and most should be larger than approximately 100 nt. Some abundant species may cause distinct bands in the nascent RNA smear, and this is OK (Note: lanes show two different unrelated samples).

PolyA + RNA Depletion

Prepare Magnetic Beads

23. Vortex the magnetic beads from the DynaBeads mRNA Direct Micro Purification kit briefly to resuspend.

24. You will need to prepare 50 μl of beads for each sample. Because the depletion will be performed in triplicate, prepare 3 volumes of beads plus 10% for

pipetting error. Aliquot beads into a new 1.5-ml tube.

Note: *E.g. For one sample, aliquot 50 μl x 3 x 10% = 165 μl beads.*

25. Place the tube on a magnetic rack and allow the supernatant to clear (approximately 1 minute).

26. Aspirate the supernatant with a P200 tip and discard. Remove the tube from the magnetic rack.

27. Add 1 volume of lysis/binding buffer (from the DynaBeads mRNA Direct Micro Purification kit) to the tube.

    Note: *E.g. for 165 μl beads, add 165 μl buffer.*

28. Vortex briefly to mix, then centrifuge briefly (approximately 10 seconds) to collect any drops from the lid.

29. Divide the beads equally into 3 1.5-ml tubes. (50 μl in each tube) and set aside.

PolyA+ Depletion

30. Adjust the volume of the input chromatin-associated RNA from step 21 to 300 μl with nuclease-free water.

31. Incubate RNA sample 2 min at 70°C.

32. Add an equal volume (300 μl) of lysis binding buffer (from the DynaBeads mRNA Direct Micro Purification kit) to the RNA. Vortex briefly to mix, then centrifuge briefly to collect any drops in the lid.

33. Pipette the RNA/buffer mixture prepared in step 32 on top of one of the 50 μl aliquots of prepared magnetic beads. Pipette up and down ten times to mix.

34. Incubate the tube 5 min at room temperature.

35. Place the tube on a magnetic rack and allow the supernatant to clear, approximately 1 minute.

36. Carefully aspirate the supernatant and transfer to a clean 1.5-ml tube.
    Note: *This is the polyA-depleted fraction.*

37. Incubate the polyA-depleted fraction 2 min at 70°C.

38. Repeat steps 33-37 two more times (for a total of three rounds of incubation, with fresh beads each time), omitting the final 70°C incubation on the last round.
    Note: *This is why 3 volumes of beads were required on Step 24.*

39. Clean up the polyA-depleted RNA sample using the Clean and Concentrator-5 kit, eluting in 30 μl nuclease-free water.
    Note: *Alternatively, ethanol precipitation can be used to clean up and concentrate RNA.*

40. Measure the concentration of polyA-depleted RNA by Nanodrop.
    Note: *Users should expect a concentration of 350-500 ng/μl with a 260/280 value around 2.0, and a 260/230 value around 2.1. Note that a minimum of 10 μg of polyA+ depleted RNA is recommended before continuing to ribsomal RNA depletion.*

Ribosomal RNA Depletion

Use RiboMinus™ Eukaryote Kit v2 on PolyA-depleted RNA

41. Add the components from the RiboMinus™ Eukaryote Kit in **Table 2.2** to a 1.5-ml tube, in the order listed.
    Note: *The maximum input for the RiboMinus™ kit is 5 μg of RNA. You should have more than this, so you will have to split the sample and do the rRNA depletion in multiple aliquots, then combine all samples at step 60.*

Table 2.2: Ribominus™ probe hybridization reaction components.

Reagent	Amount
2X hybridization buffer	50 μl
RiboMinus™ Eukaryote Probe Mix v2	4 μl
PolyA+ depleted RNA	up to 5 μg
Nuclease-free water	up to 100 μl

42. Mix by briefly vortexing at low speed, then centrifuge briefly to collect drops from the lid.

43. Incubate 10 min in a heat block at 70°C to denature RNA.

44. Immediately transfer the tube to a second heat block at 37°C and allow the sample to cool over a period of 20 minutes. Continue to the next step while the sample is cooling.

Prepare RiboMinus™ Magnetic Beads

45. Resuspend the RiboMinus™ Magnetic Beads (from the RiboMinus™ Eukaryote Kit) by vortexing the bottle briefly.
Note: *Be careful not to confuse the magnetic beads from the RiboMinus™ Eukaryote Kit with the previous magnetic beads from the DynaBeads mRNA Direct Micro Purification Kit (step 23).*

46. For each 5 μg RNA sample, prepare 200 μl of 1X hybridization buffer in a new labelled tube by diluting 2X hybridization buffer (from the RiboMinus™ Eukaryote Kit) with an equal volume of nuclease-free water.

47. For each 5 μg RNA sample, pipette 500 μl of magnetic beads into a new 1.5-ml tube. Place each tube on a magnetic rack to allow the supernatant to clear.

48. Gently aspirate and discard the supernatant.

49. Remove the tubes from the magnetic rack and wash the beads with 500 μl nuclease-free water by pipetting it down the side of the tube where the beads are collected.

50. Place each tube on a magnetic rack to allow the supernatant to clear. Gently aspirate and discard the supernatant.

51. Repeat washing (steps 49-50) one more time (for a total of 2 times).

52. Resuspend the beads in 200 μl of prepared 1X hybridization buffer from step 46.

53. Incubate the tube with beads in a heat block at 37°C for at least 5 minutes, or until the 20 minute incubation of the RNA/probe mix at 37°C is complete.

Capture and Remove rRNA/Probe Complexes

54. Briefly centrifuge the RNA/probe mix to collect the mixture at the bottom of the tube.

55. Add the RNA/probe mix to the prepared RiboMinus™ Magnetic beads from step 53. Mix by pipetting up and down ten times.

56. Incubate the tube in a heat block at 37°C for 5 min. Centrifuge briefly to collect drops.

57. Place the tube on a magnetic rack and allow the supernatant to clear, approximately 1 minute.

58. Transfer the supernatant (approximately 300 μl) to a new 1.5-ml tube.
Note: *This is the rRNA-depleted fraction.*

59. Concentrate and clean up rRNA-depleted RNA by ethanol precipitation:

    Note: *Alternatively, use the magnetic bead clean up kit that comes with some versions of the RiboMinus™ Eukaryote Kit. In my experience, it can be difficult for inexperienced users to work with very small volumes on magnetic beads, therefore, I recommend ethanol precipitation.*

    (a) Add 1 μl glycogen (20 μg/μl), 0.1 volumes 3 M sodium acetate, and 2.5 volumes of ice-cold 100% ethanol to the tube with the RNA.

    Note: *Adding glycogen is optional, but it helps to generate a visible pellet.*

    (b) Mix well by pipetting up and down, then incubate at least 30 minutes at -80°C.

    Note: *Incubate up to overnight at -80°C.*

    (c) Centrifuge 15 min at 14,000 rpm, 4°C.

    (d) Carefully aspirate and discard the supernatant without disturbing the pellet.

    (e) Add 500 μl of ice-cold 70% ethanol to rinse the pellet.

    (f) Centrifuge 5 min at 14,000 rpm, 4°C.

    (g) Carefully aspirate and discard the supernatant without disturbing the pellet.

    (h) Repeat the wash with ice-cold 70% ethanol one more time (for a total of 2 washes).

    Note: *When aspirating the supernatant the second time, be sure to remove as much ethanol as possible. It can help to use a P1000 pipette tip to remove most of the ethanol. Then, centrifuge the tube briefly and use a P10 pipette tip to remove the final few drops.*

    (i) Air dry the pellet with the lid open for 5 min at room temperature.

60. Resuspend the pellet in 7 μl nuclease-free water.

    Note: *Combine multiple samples (divided in step 41) into one 1.5-ml tube at this point. Start by resuspending the first pellet in 7 μl nuclease-free water, then use this same 7 μl to resuspend further pellets.*

61. Measure the concentration of RNA by Nanodrop.

    Note: *Users should expect a concentration of 110-210 ng/μl. Note that it is very important to have at least 600 ng of nascent RNA in a volume of at most 5.5 μl for the next step, or a minimum concentration of 110 ng/μl before proceeding. For more details, see the Critical Parameters section.*

Adapter Ligation

62. Add 600 ng of nascent RNA (in a volume of up to 5.5 μl) and 50 ρmol of DNA adapter (0.5 μl of a 100 ρmol/μl dilution) to a 0.2-ml PCR strip tube. If necessary, adjust the volume to 6 μl with nuclease-free water. Mix by gently flicking the tube, then centrifuge briefly to collect drops.

    Note: *Note that the DNA adapter sequence (/5rApp/NNNNNCTGTAGGCACCATCAAT/3ddC/) includes an activated adenylate group at the 5′ end (5rApp), a sequence of 5 random nucleotides (NNNNN), and a dideoxynucleotide at the 3′ end (3ddC). All three of these custom features must be included when ordering this oligo from a vendor. The adapter should be diluted to 100 ρmol/μl and aliquoted upon arrival, then stored at -20°C.*

63. Incubate 10 min in a thermal cycler at 65°C, then transfer to ice for at least 1 minute.

64. During the incubation of step 63, prepare a master mix (enough for all samples) for the adapter ligation reaction from the components of the T4 RNA

ligase kit, as listed in **Table 2.3**. Add the master mix components to a 1.5-ml tube.

Table 2.3: Adapter ligation reaction components.

Reagent	Amount
10X ligase buffer	2 μl
50% PEG 8000	10 μl
RNase OUT	1 μl
RNA ligase 2 (truncated K227Q)	1 μl

65. Mix *very well* by pipetting up and down.

    Note: *PEG is very viscous, and the success of the adapter ligation reaction can depend on how well it is mixed at this stage. Avoid incorporating bubbles by keeping the pipette tip submerged.*

66. Add 14 μl of adapter ligation master mix to the 6 μl annealed RNA sample (from Step 63) and mix very well again by pipetting up and down. Centrifuge briefly to collect drops.

67. Incubate in a thermal cycler for 12 hours at 16°C, then hold at 4°C.

68. Adjust the volume of the adapter ligation reaction to 100 μl with nuclease-free water.

69. Clean up and concentrate adapter-ligated RNA using the Clean and Concentrator-5 kit, eluting in 10 μl nuclease-free water.

    Note: *Alternatively, ethanol precipitate RNA as described above.*

2.1.5 Basic Protocol 3: cDNA Amplicon Preparation

This library preparation protocol generates a double-stranded cDNA molecule from each nascent RNA that retains the unique nascent RNA 3′ end adapter sequence. First, a template-switching reverse transcriptase generates both the first and second

strand of cDNA in a single step (Zhu et al., 2001), which allows templates with heterogeneous 5′ ends to be incorporated into the library. Importantly, RNAs with a 5′ end that is TMG-capped are incorporated more efficiently than uncapped RNAs (Wulf et al., 2019), which helps enrich for full length RNAs in the library. First, a test is performed to determine the optimal number of PCR cycles to use when amplifying the cDNA. This is to avoid over-amplifying the sample and introducing a size bias in the final library. Then, a final PCR reaction is performed using the optimal number of PCR cycles, and the sample is cleaned up before continuing to generate a long read sequencing library. This protocol is compatible with barcoding.

Materials

Reagents and Chemicals

- Adapter-ligated nascent RNA (from Basic Protocol 2 step 69)
- Custom RT primer, 10 μM (AAGCAGTGGTATCAACGCAGAGTAC**CACATA TCAGAGTGCG**GATTGATGGTGCCTACAG; where the region in bold is a 16-nt barcode, see Critical Parameters section for more details)
- TE buffer (10 mM Tris·HCl, 0.1 mM EDTA, pH 8.0)
- 6X Orange DNA loading dye (Thermo Fischer Scientific, cat. no. R0631)
- SMARTer PCR cDNA Synthesis Kit (Takara/Clontech, cat. no. 634925)
- Advantage 2 PCR kit (Takara/Clontech, cat. no. 639137)
- TAE buffer (40 mM Tris·HCl, 20 mM acetic acid, 1 mM EDTA pH 8.0)
- Agarose (Sigma, cat. no. A9539)
- Lonza Gelstar (Thermo Fischer Scientific, cat. no. 50535)

- GeneRuler 1 kb Plus DNA ladder (Thermo Fischer Scientific, cat. no. SM1331)
- AMPure XP beads (Beckman Coulter, cat. no. A63880)
- 70% ethanol, room temperature
- 10 mM Tris·HCl pH 8.5

Equipment

- 0.2-ml PCR strip tube (Dot Scientific, cat. no. 415-8PCR)
- Thermal cycler
- 1.5-ml tubes (Dot Scientific, cat. no. RN1700-GMT)
- Refrigerated microcentrifuge
- Vortex
- Thermomixer
- 1.5-ml tube magnetic rack
- Nanodrop
- UV gel imaging system

Protocol Steps

Reverse Transcription

1. For each sample, add 3.5 μl adapter ligated RNA (from step Basic Protocol 2 step 69; approximately 210 ng) and 1 μl custom RT primer (10 μM) to a new 0.2-ml tube labeled "+RT". Add the same to a new 0.2-ml tube labeled "-RT". Mix by gently flicking the tubes, then centrifuge briefly to collect drops.

Note: *If multiple samples are to be barcoded, different RT primers with unique barcode sequences should be used for each sample here.*

2. Incubate tubes in a thermal cycler for 3 min at 72°C, then 2 min at 42°C, then hold at 4°C until the next step is prepared.

3. During the incubation of step 2, prepare two master mixes for the reverse transcription reaction from the components of the SMARTer PCR cDNA Synthesis Kit, as listed in **Table 2.4**. Add master mix components to a 1.5-ml tube in the order listed below. Prepare enough master mix for all samples. Make one master mix including the SMARTScribe RT enzyme and one mix without the enzyme, using nuclease-free water in its place.
   Note: *For example, if you are preparing 4 samples, prepare a 5X master mix with RT enzyme, and a 5X master mix without RT enzyme.*

Table 2.4: Reverse transcription reaction components.

Reagent	Amount
5X First Strand buffer	2 μl
100 mM DTT	0.25 μl
10 mM dNTP mix	1 μl
12 μM SMARTer IIA Oligo (stored at -80°C)	1 μl
RNase Inhibitor	0.25 μl
SMARTScribe Reverse Transcriptase (100U)	1 μl

4. Add 5.5 μl of master mix with SMARTScribe RT enzyme to the 0.2-ml tubes from step 1 labeled "+RT", and add 5.5 μl of master mix without SMARTScribe RT enzyme to the 0.2-ml tubes labeled "-RT". Mix by gently flicking the tubes, then centrifuge briefly to collect drops.

5. Incubate in thermal cycler for 90 min at 42°C, then terminate the reaction by incubating 10 min at 70°C, and then hold at 4°C.

6. Dilute each reaction 1:10 with 90 μl TE buffer.

7. Divide each reaction into 10-15 μl aliquots in labeled 0.2-ml PCR strip tubes and freeze immediately at -20°C if stopping here. Store at -20°C for up to three months. Otherwise, keep the tubes on ice and proceed immediately to PCR optimization.

PCR Cycle Number Optimization

8. Assemble a master mix for PCR reactions from the components of the Advantage 2 PCR kit, as listed in **Table 2.5**. For each cDNA sample (both +RT and -RT), assemble enough master mix for one PCR reaction.

Table 2.5: PCR cycle number optimization reaction components.

Reagent	Amount
cDNA	5 μl
10X Advantage 2 buffer	5 μl
50X dNTP mix (10 μM)	1 μl
Primer IIA (12 μM)	1 μl
50X Advantage 2 Polymerase	1 μl
Nuclease-free water	37 μl

9. Add 5 μl of each cDNA sample from step 7 to a new 0.2-ml PCR strip tube, then add 45 μl of the master mix on top. Mix by gently flicking the tube and then centrifuge briefly to collect drops.

10. Place the PCR strip in a thermal cycler with the cycle program indicated in **Table 2.6**. After the initial 8 rounds of amplification, hold the thermal cycler at 4°C, open the thermal cycler lid, and remove a 5 μl aliquot from each reaction. Close the lid and continue cycling for an additional 2 rounds (for a total of 10 rounds), then take out another 5 μl aliquot. Repeat until the sample has gone through a total of 18 cycles of amplification and you have a 5 μl aliquot after every 2 cycles. Keep the aliquots at 4°C until the final round is completed.

Note: *The cycle numbers that are tested may need to be adjusted, but I recommend 8-18 cycles as a starting point.*

Table 2.6: PCR cycle number optimization thermal cycler program.

Temperature (°C)	Time	Cycles
95	1 min	1
95	15 s	
65	30 s	8, 10, 12, 14, 16,18
68	3 min	
4	hold	–

11. Add 1 μl 6X gel loading dye to each aliquot and run all aliquots on a 1% TAE agarose gel at 85 V for 45 min alongside 0.2 μl of GeneRuler 1 kb plus ladder. Load reactions from corresponding +RT and -RT cDNA aliquots at each cycle number for comparison (**Figure 2.4**). Each sample will require 12 lanes plus a ladder. Run the gel at 85 V for 45 minutes, then image using a UV gel imaging system.

12. Select the cycle number to be used for the final PCR amplification. For all samples, the -RT reactions should be empty, and the +RT reactions should show a gradually increasing smear of cDNA with increasing cycle number. The optimal cycle number is one where the cDNA is visible, but before it gets too dark and overloaded.
    Note: *For the example shown in* ***Figure 2.4****, a cycle number of 14 should be chosen.*

Final PCR Amplification

13. For each sample, prepare enough master mix for 8 50 μl PCR reactions using the Advantage 2 PCR kit, according to **Table 2.5**.

14. Add 5 μl of +RT cDNA to 8 0.2-ml PCR strip tubes, then add 45 μl of master mix on top. Mix by gently flicking the tube and then centrifuge briefly to

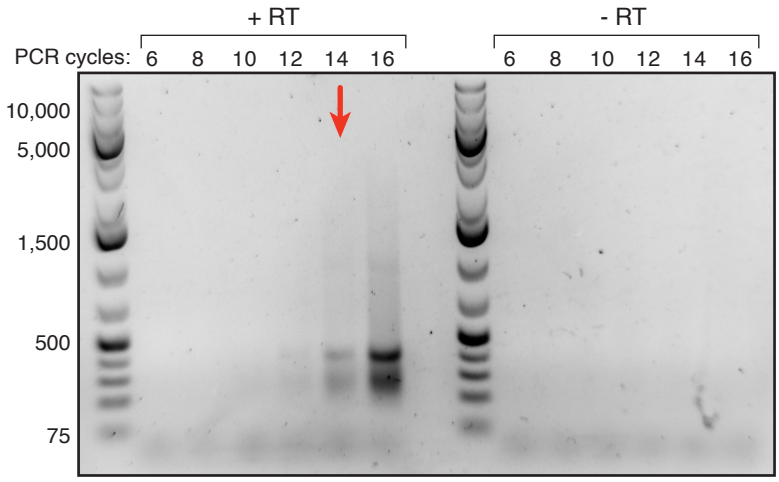

Figure 2.4: Agarose gel showing aliquots from PCR cycle number optimization. On the left, "+RT" samples contain the RT enzyme in the reverse transcription reaction, and on the right, "-RT" control samples that do not contain the RT enzyme. Numbers at the top of the gel indicate the number of PCR amplification cycles. Numbers at the left of the gel indicates size of the DNA ladder in basepairs. The red arrow indicates the optimal cycle number in this example, based on the intensity of the cDNA smear.

collect drops.

15. Place the PCR strip in a thermal cycler with the cycle program indicated in **Table 2.7**.

Table 2.7: Final LRS library PCR thermal cycler program.

Temperature (°C)	Time	Cycles
95	1 min	1
95	15 s	Optimized cycle number
65	30 s	
68	3 min	
68	5 min	1
4	hold	–

16. Combine the 8 PCR reactions for each sample (400 μl total) into a new 1.5-ml tube, and remove a 5 μl aliquot from each sample to run on a gel.

17. Run the 5 μl aliquot from each sample on a 1% TAE agarose gel at 85 V for 45 min to confirm that PCR amplification worked, and that the smear appears

the same as in the cycle optimization test (from step 12).

AMPure Bead PCR Cleanup

18. Mix AMPure beads by gently vortexing the bottle.

19. Add 1X volume of AMPure magnetic beads to each pooled PCR sample (400 μl).

20. Mix the beads/DNA solution thoroughly by vortexing. Centrifuge briefly to collect drops.

21. Allow the DNA to bind to the beads by shaking in a Thermomixer at 1,400 rpm for 10 minutes at room temperature. Centrifuge the tube briefly to collect drops.

22. Place the tube in a magnetic rack and allow the supernatant to clear (at least 1 minute).
    Note: *In my experience, these beads take slightly longer than usual to clear. If the supernatant is still cloudy after 1 minute, wait longer.*

23. With the tube still on the magnetic rack, carefully aspirate and discard the cleared supernatant. Make sure you avoid disturbing the bead pellet.

24. Wash the beads with 500 μl of room temperature 70% ethanol by pipetting it down the side of the tube where the beads are.

25. With the tube still on the magnetic rack, carefully aspirate and discard the supernatant. Repeat once, for a total of 2 ethanol washes.

26. Centrifuge the tube briefly to collect beads and residual ethanol in the bottom of the tube, then place the tube back in the magnetic rack and carefully aspirate any residual ethanol with a P10 pipette tip.

Note: *It is important to remove as much ethanol as possible without disturbing the pellet.*

27. Open the tube lid and air dry for 1 minute at room temperature.

28. Add 40 μl of 10 mM Tris·HCl pH 8.5 to the beads to elute the DNA.

29. Shake the tube in a Thermomixer for 10 minutes at 1,400 rpm at room temperature. Centrifuge the tube briefly to collect drops.

30. Place the tube in the magnetic rack and allow the supernatant to clear (at least 1 minute).

31. Carefully aspirate the supernatant and transfer it to a new 1.5-ml tube.
    Note: *This is the final eluted DNA sample. Be very careful not to carry over any excess magnetic beads during this step. If you are unsure, place the eluted supernatant back on the magnetic rack and repeat.*

32. Determine the concentration of the final DNA sample by Nanodrop.
    Note: *Users should expect a final concentration of 50 – 200 ng/μl, although this will vary depending on the intensity of the cDNA smear that was chosen for the optimal cycle number.*

33. If preparing multiple barcoded samples, pool all barcoded samples together in equal concentrations (*i.e.* put equal ng amount of each in the same tube). The final sample volume will vary depending on the library preparation service you choose, but 50 μl is an approximate volume to aim for.
    Note: *Library preparation should now be completed following manufacturer's instructions for amplicon sequencing on the long-read sequencing platform of your choice.*

2.1.6 Reagents and Solutions

All buffers should be prepared from stock solutions up to 1 week before use and stored at 4°C, excluding α-amanitin, SUPERase.IN, and cOmplete protease inhibitor mix, which should be added immediately before use. All buffers should be chilled on ice before use.

Cell lysis buffer

10 mM Tris·HCl pH 7.5

0.05% NP-40 (Sigma, cat. no. 9016-45-9)

150 mM NaCl

25 μM α-amanitin (Sigma, cat. no. A2263)

40 U/ml SUPERase.IN (Thermo Fischer Scientific, cat. no. AM2694)

1X cOmplete protease inhibitor mix (Sigma, cat. no. 11697498001)

Nuclear Lysis Buffer

20 mM HEPES pH 7.5

1 mM dithiothreitol (DTT)

7.5 mM $MgCl_2$

0.2 mM EDTA

0.3 M NaCl

1 M Urea

1% NP-40 (Sigma, cat. no. 9016-45-9)

25 μM α-amanitin (Sigma, cat. no. A2263)

40 U/ml SUPERase.IN (Thermo Fischer Scientific, cat. no. AM2694)

1x cOmplete protease inhibitor mix (Sigma, cat. no. 11697498001)

Nuclear Resuspesion Buffer

20 mM Tris·HCl pH 8.0

75 mM NaCl

0.5 mM disodium EDTA

0.85 mM dithiothreitol (DTT)

50% (v/v) glycerol

25 μM α-amanitin (Sigma, cat. no. A2263)

40 U/ml SUPERase.IN (Thermo Fischer Scientific, cat. no. AM2694)

1x cOmplete protease inhibitor mix (Sigma, cat. no. 11697498001)

Sucrose Buffer

Cell lysis buffer with 24% (w/v) sucrose

Note: *Add sucrose crystals to prepared cell lysis buffer and mix on a rotary spinner at room temperature until sucrose is dissolved (15 – 30 min).*

2.1.7 Troubleshooting

Fractionation Issues

All fractionation issues should be diagnosed by Western blot of the cytoplasmic, nucleoplasmic, and chromatin fractions. If nucleoplasm or chromatin markers are present in the cytoplasm, the initial centrifugation speed may have been so high that nuclei lysed in the sucrose buffer purification step. Try decreasing centrifugation speed. Additionally, nuclei can be observed under a microscope after this step to ensure they are intact (visibly round). You may also try to decrease the time that cells are incubated with the cell lysis buffer. If the nuclei are difficult to resuspend after pelleting (sticky rather than loosely suspended after flicking the tube), the nuclei may have been prematurely lysed as well. Conversely, if markers for the

nucleoplasm or chromatin are significantly detected in the cytoplasm, the centrifugation speed may not have been high enough to pellet the nuclei, and you may thus wish to increase the centrifugation speed.

RNA Isolation Issues

If at any point during the RNA isolation you end up with less RNA than recommended for the next step, you may either pool multiple samples together to gain enough RNA, or freeze the RNA at -80°C for up to one month while you go back and repeat the previous steps to obtain enough RNA. If the nascent RNA appears degraded on the gel after RNA isolation (a high concentration of small RNAs, no larger RNAs), it is possible that there was RNase contamination. Throw out this RNA sample and repeat from the beginning. Make sure that you are using sterile RNAse-free materials, work quickly while handling the RNA samples, don't touch any tips/tubes accidentally to other surfaces, and keep all tubes that contain RNA on ice unless otherwise noted.

Adapter Ligation Issues

The adapter ligation reaction is usually greater than 90% efficient (Carrillo Oesterreich et al., 2016), but the efficiency can drop if the reaction is not thoroughly mixed before incubation. If you suspect the adapter ligation reaction is not working, for example if the downstream PCR reactions do not yield a product and you have ruled out other issues with RT or PCR (see next point), you can diagnose this problem by running a sample of nascent RNA with and without T4 RNA ligase included in the ligation reaction on a 10% TBE-urea denaturing gel (**Figure 2.5**). In a successful ligation, you should be able to see all the major bands in the nascent RNA sample shift upward when the DNA adapter is present in the reaction.

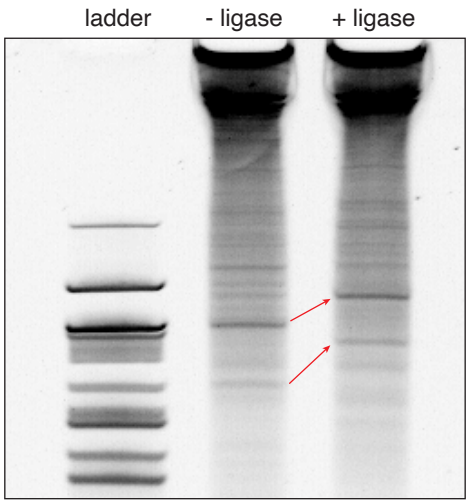

Figure 2.5: Denaturing agarose gel showing successful adapter ligation.
(-) ligase lane contains the adapter ligation reaction mix in the absence of T4 ligase, and (+) ligase lane contains T4 ligase. Successful ligation results in a shift upwards of all major RNA species, as indicated by the red arrows.

RT and PCR Issues

If no PCR product appears even after 18 cycles of amplification, the adapter ligation reaction may not have worked efficiently. Repeat from this step, ensuring the reaction is mixed well. If there is PCR product in the -RT control lanes as well as the +RT lanes, there may be some DNA contamination in the sample. Repeat DNase I treatment if this occurs. If some RNA sequences are very abundant in your cell type, this may appear as a distinct band in the cDNA "smear" (**Figure 2.4**) – this is OK. However, sometimes a high molecular weight smear appears (in the range of 5-20 kb), and this is generally non-specific amplification. A cycle number should be chosen before this high molecular weight smear appears.

2.1.8 Expected Results

Sequencing a cDNA library from 60 million MEL cells on one PacBio Sequel flow cell should yield up to 1 million reads, and typically 68% map uniquely to the mouse genome. Polyadenylated reads can be bioinformatically filtered but should

be minimal (typically around 1%). PCR duplicates can be identified by reads which contain the same barcode in the DNA adapter sequence. However, at this sequencing depth, PCR duplicates are not frequently sequenced (typically less than 1%). The resulting long reads generally have a median length of approximately 720 nucleotides.

2.1.9 Time Considerations

The full protocol described here can be completed in 4 days. On day 1, actively growing MEL cells can be harvested and fractionated, then a diagnostic Western blot can be run. On day 2, nascent RNA can be isolated from the chromatin pellet in Trizol, then polyA+ and rRNA depletion can be completed. Adapter ligation should also be done at the end of this day, since it is the only step that requires an overnight incubation. On day 3, the RT reaction and PCR cycle number optimization can be done. On day 4, the final PCR amplification and clean up can be performed. However, there are flexible stopping points mentioned in the protocol. The most time-sensitive steps are on the first day – it is best to get the live cells harvested and precipitate the chromatin as fast as possible.

2.2 Cell Lines, Cell Culture, and Cell 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. dimethyl sulfoxide (DMSO) was then added directly to the culture medium to a final concentration of 2% and incubated as above for 5 days. For Pla-

dienolide 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.

2.3 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 5.3**. 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 $\Delta\Delta$Ct method.

2.4 Microscopy

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

2.5 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. PolyA+ 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 5.3**.

2.6 *HBB* Targeted Nascent RNA Library Preparation

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 5.3**). 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 either a unique region in the 5′UTR of the human *HBB* gene or the endogenous mouse *Hbb-b1* 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.

2.7 Long-read Sequencing Data Analysis

2.7.1 Genome-wide nascent RNA sequencing data preprocessing

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 trimmed 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. Non-unique reads (reads with the same read name appearing more than once in SAM files) were removed. All data generated using Nanopore sequencing from (Drexler et al., 2020) (GEO accession ID: **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, and non-unique reads were also removed. All data were visualized in and exported from IGV to generate genome browser figures. In all analysis except where noted, LRS data are represented as two biological and two technical replicates combined.

2.7.2 *HBB* targeted nascent RNA sequencing data preprocessing

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 5.3**). 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).

2.7.3 PolyA+ Read Filtering

Genome-wide nascent RNA sequencing data

Mapped reads in SAM format were filtered to remove reads that contained a polyA tail using a custom script (available at https://www.github.com/NeugebauerLab/MEL_LRS). 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 data

Additional parameters were added to the above criteria for removing polyA+ 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 uncleaved 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 uncleaved 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. Uncleaved reads with long stretches of soft-clipped bases that passed this filtering were then recoded to contain a match in the CIGAR string downstream of the PAS in order to include these regions of the long-reads in coverage calculations.

2.7.4 Splicing Status Classification and Co-transcriptional Splicing Efficiency (CoSE) Calculation

The annotation of introns contained in active transcripts (described below for PRO-seq), was first filtered for unique intron start and end coordinates. Additionally, an upstream intron in *Hbb-b1* that was clearly not used in the LRS reads was removed (`ENSMUST00000153218.1_intron_1_0_chr7_103827887_r`). The resulting introns were extended by 1 nt on either end and were overlapped with bed files of long-reads using bedtools intersect in order to get regions of long-reads that spanned entire introns. Spliced junction coordinates in intron-overlapping long-reads 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. Next, reads which did not span the entire intron, but reached at least 35 nt upstream of a 3′SS and were unspliced were counted toward the unspliced count for an intron. To classify splicing status of each read, the number of spliced introns was compared to the total number of introns that was overlapped. To calculate co-transcriptional splicing efficiency (CoSE), the splicing status classification of each intron was recorded as above, and the number of spliced introns and unspliced introns was summed per intron. Introns with identical 5′SS or 3′SS were filtered to keep only the intron with the most total reads. Introns with no spliced reads (no evidence of usage in MEL cells), introns that were longer than 10 kb, and introns covered by fewer than 10 reads were removed. For the remaining introns, CoSE was calculated by dividing the number of spliced reads by the total number of reads.

2.7.5 Distance From Splice Junction to 3′ End Calculation

Splicing intermediates (defined below), were filtered out from the long-read data in this analysis, since their 3′ ends do not represent the position of Pol II, but rather an upstream exon between step I and step II of splicing. For all remaining reads, data were filtered for reads that contained at least 1 splice junction, and then 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 mygene in python. Introns were matched to their corresponding transcript expression level using PRO-seq TSS counts as described below. To determine if certain genes exhibited a longer or shorter distances from 3′ ends to splice junctions, the distance was split into three equal size categories and transcript IDs from each category were entered into the online PANTHER classification system: no significant enrichment was obtained.

2.7.6 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 the same set of introns considered for CoSE as described above. The number of intermediates aligned upstream of each intron was counted using bedtools intersect. The Normalized Intermediate Count (NIC) was calculated for each intron which was covered by at least 10 reads (as above) 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 the 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).

2.7.7 Long-Read Coverage

Transcript coordinates associated with active transcription start sites (as described below) were obtained from the UCSC Table Browser. Transcripts were then grouped by the parent Gene ID, and the largest range of start and end coordinates from the grouped transcripts was kept. Library depth was then calculated using bedtools coverage across this file of collapsed active gene coordinates. Metagene plots of 5′ end, 3′ end, and entire read coverage across the same gene coordinates 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, long-reads were separated by splicing status (see above), then coverage was calculated using bedtools within a window around PASs that corresponded to active TSSs or specifically to a window around the *HBB* PAS. Coverage at all positions was normalized to the coverage at the position 100 nt upstream of the PAS. For coverage of splicing intermediates, bedtools coverage was used to calculate coverage of 5′ ends and 3′ ends across a 50-nt window around 5′SS and 3′SS of introns contained in active transcripts.

2.7.8 Uncleaved Transcripts Analysis

Bedtools intersect was used to identify long-reads with 5′ ends originating in a gene body of active transcripts (as described below). Reads were then categorized as being uncleaved transcripts if their 3′ ends were greater than 50 nt downstream of the PAS of the gene which the 5′ end overlapped with. In the case where a read 5′ end intersected multiple overlapping transcripts, it was only assigned as an uncleaved read if the 3′ end was downstream of all transcript PASs. Splicing status classification of uncleaved transcripts was carried out as described above.

2.7.9 *HBB*- IVS$^{110(G>A)}$ Splicing and 3′ End Cleavage Analysis

For long-reads derived from *HBB*-IVS$^{110(G>A)}$ cells, only reads that were spliced at intron 1 using the cryptic splice site were analyzed, and the rare reads with a splice junction using the canonical splice site were discarded. Splicing status classification, counting of splicing intermediates, and calculating coverage downstream of the PAS were performed as described above but with the custom *HBB* annotation coordinates.

2.8 PRO-seq Library Preparation and Data Analysis

2.8.1 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·HCl pH 8.3, 40% glycerol, 5 mM MgCl2, 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.

2.8.2 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 min. 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 min. 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 min twice to elute the RNA from the beads. A subsequent ethanol precipitation was performed for RNA purification. Nascent RNA was resuspended in 10 μM of the VRA3 3′ end adapter (**Table 5.3**). 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′ Pyrophosphohydro-

lase (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 5.3**) 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 ρmol of the Illumina TRU-seq RP1 Primer (**Table 5.3**). 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 ρmol 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.

2.8.3 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 with settings `-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 Illumina 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.

2.8.4 PRO-seq and Total RNA-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. Additionally, if introns shared both a 5′SS and 3′SS, the intron with the lowest annotated intron number was kept. Finally, first intron annotations were removed for **Figure 3.12** and metagene plots. For all other metagene analyses, introns within 750 nt of a TSS, and introns with fewer than 10 reads were also filtered out from the final list of unique introns to avoid bleed-through PRO-seq signal from the promoter-proximally paused Pol II. Metagene plots around the TSS, splice sites, and PAS were generated by plotting the average PRO-seq reads (of three biological replicates) in uninduced cells at each indicated position with respect to the TSS, 5′SS, 3′SS, or PAS respectively. Violin plots evaluating PRO-seq 3′ end or RNA-seq read coverage were generated by summing the signal at the indicated positions with respect to the 5′SS, 3′SS, or PAS. P-values were calculated using either the Mann-

Whitney or the Wilcoxon matched-pairs signed rank test

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. In all analysis except where noted, PROs-seq data are represented as three biological replicates combined.

2.9 Data and Code Availability

Raw and processed long-read sequencing and PRO-seq data generated in this thesis are deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number **GSE144205**. Raw image data associated with this manuscript are available on Mendeley (http://dx.doi.org/10.17632/5vrtbpnj4k.1). All code supporting the long-read sequencing data analysis is available at https://www.github.com/NeugebauerLab/MEL_LRS. NanoCOP data from (Drexler et al., 2020) analyzed in this thesis can be found at GEO with accession number **GSE123191**, and total RNA-seq from MEL cells analyzed in this thesis can be found at Mouse ENCODE (http://www.mouseencode.org/) with accession number **ENCSR000CWE**.

Chapter 3

Results: Co-transcriptional Splicing Regulates 3′-End Cleavage During Mammalian Erythropoiesis

3.1 PacBio Long-Read Sequencing of Nascent RNA Yields High Read Coverage

Murine erythroleukemia (MEL) cells are immortalized at the proerythroblast stage and can be induced to enter terminal erythroid differentiation by treatment with 2% DMSO for five days (Antoniou, 1991). Phenotypic changes include decreased cell volume, increased levels of β-globin, and visible hemoglobinization (**Figure 3.1 A-C**). I used chromatin purification of uninduced and induced MEL cells to enrich for nascent RNA. Chromatin purification under stringent washing conditions allows release of contaminating RNAs and retains the stable ternary complex formed by elongating Pol II, DNA, and nascent RNA (**Figure 3.1 D**); Wuarin and Schibler (1994). Importantly, spliceosome assembly does not continue during chromatin fractionation or RNA isolation, because the presence of the splicing inhibitor

Pladienolide B throughout the purification process does not change splicing levels (**Figure 3.2**).

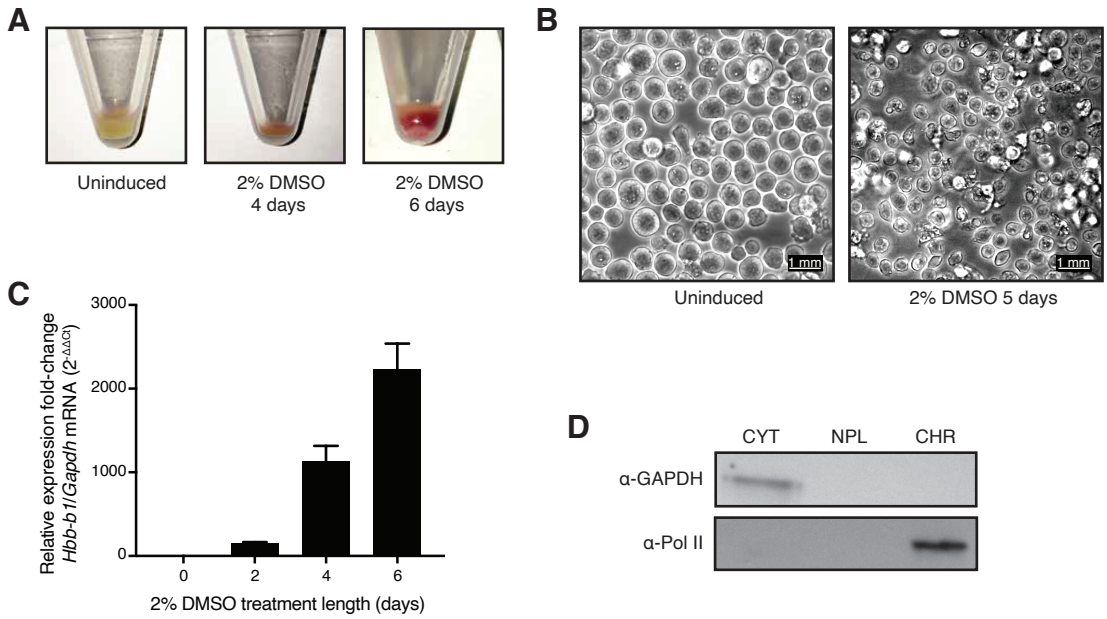

Figure 3.1: DMSO treatment induces erythroid differentiation.
(A) MEL cells after culturing in 2% DMSO for 0, 4, or 6 days. **(B)** Bright field microscopy of MEL cells uninduced (left) and induced for 5 days (right). Scale bar is 1 mm. **(C)** RT-qPCR measurement of *Hbb-b1 (β-globin)* mRNA levels relative to *Gapdh* mRNA from total RNA in MEL cells treated with 2% DMSO for 0, 2, 4, or 6 days. Bar heights 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).

To generate libraries for LRS, I established the protocol outlined in **Figure 3.3**. Two biological replicates, each with two technical replicates, were sequenced using PacBio RSII and Sequel flow cells, yielding a total of 1,155,629 mappable reads (**Table 5.2**). Reads containing a non-templated polyA tail comprised only 1.7% of the total reads (**Table 5.2**) and were removed bioinformatically along with abundant 7SK RNA reads. Of the remaining reads, the average read length was 710 and 733 nucleotides (nt), and the average coverage in reads per gene was 8.4 and 4.8 for uninduced and induced samples, respectively (**Figure 3.4 A-B**). More than 7,500 genes were represented by more than 10 reads per gene in each condition (**Figure**

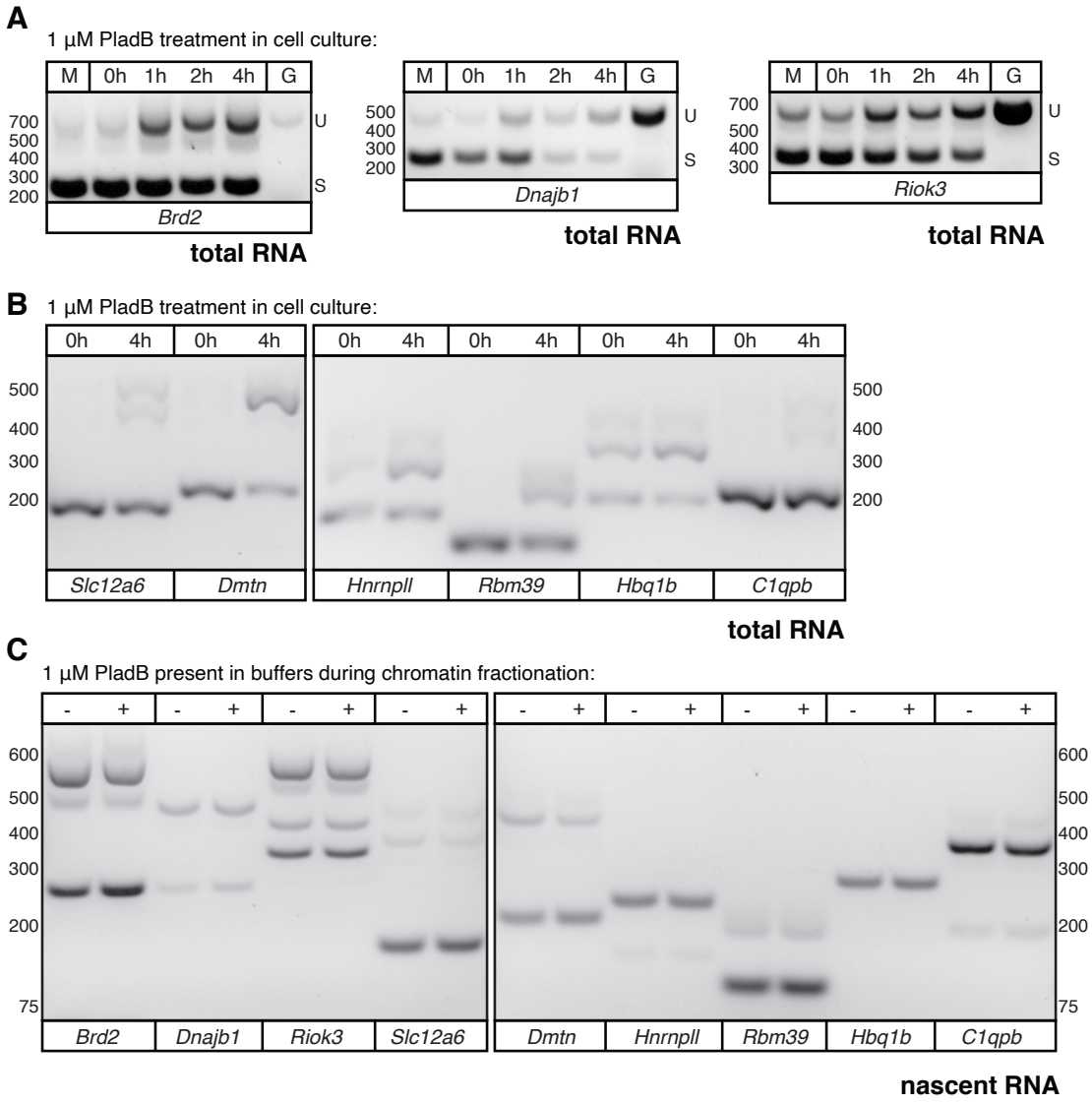

Figure 3.2: Splicing does not continue during chromatin purification and nascent RNA isolation.

(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-PCR from total RNA as in **(A)**, showing six additional genes (*Slc12a6*, *Dmtn*, *Hnrnpll*, *Rbm39*, *Hbq1b*, *C1qbp*) after treatment with 1 μM Pladienolide B for 0 h and 4 h. **(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)**.

3.4 B). Coverage of 5′ ends was focused at annotated transcription start sites (TSSs), with 18.3% of 5′ ends within 50 bp of an active TSS across all samples. As expected, 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 3.5**).

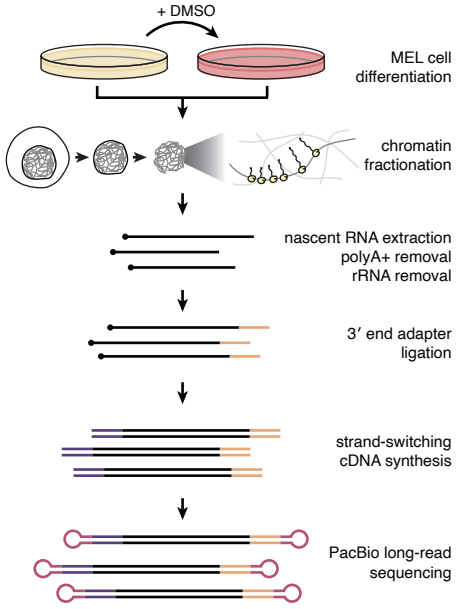

Figure 3.3: Long-read sequencing library preparation workflow.
Schematic of nascent RNA isolation and sequencing library generation. MEL cells are treated with 2% DMSO to induce erythroid differentiation, cells are fractionated to purify chromatin, and chromatin-associated nascent RNA is depleted of polyadenylated and ribosomal RNAs. An adapter is ligated to the 3′ ends of remaining RNAs, then a strand-switching reverse transcriptase is used to create double-stranded cDNA that is the input for PacBio library preparation.

3.2 LRS Reveals Widespread 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; the splice junctions reveal if splicing has occurred and which splice sites were chosen. Here, I present my LRS data in

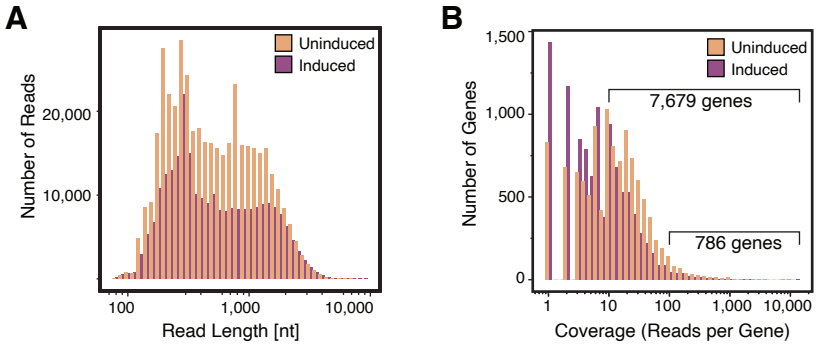

Figure 3.4: Long-read sequencing library length and depth distributions.
(A) Read length (nucleotides) and **(B)** read depth (reads per gene) distributions of PacBio long-reads.

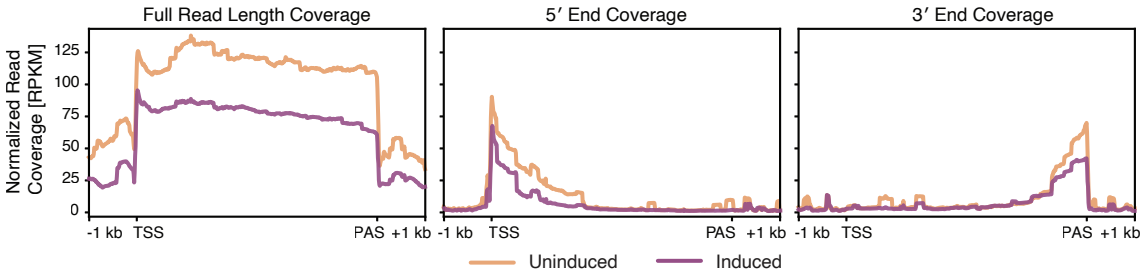

Figure 3.5: Long-read sequencing library metagene coverage.
Normalized read coverage (RPKM) of long-read 5′ ends (left), full reads (middle), and 3′ ends (right) is shown across a metagene plot of all mm10 genes +/- 1 kb (TSS = transcription start site, TES = transcription end site).

a format that highlights 3′ end position and the associated splicing status (**Figure 3.6 A**). Each transcript was categorized and coloured according to its splicing status, which can be either "all spliced", "partially spliced", "all unspliced", or "NA" (transcripts that did not span an entire intron or a 3′SS). For each gene, I calculated the fraction of long-reads that were all spliced, partially spliced, or all unspliced (**Figure 3.6 A**; bar plot far right), enabling a survey of splicing behaviors within individual transcripts (Alpert et al., 2020; Herzel et al., 2018; Kim and Abdel-Wahab, 2017).

Splicing status of individual transcripts varied from gene to gene. For example, the gene *Actb* had mostly all spliced reads (78% and 75% of reads in uninduced and induced cells respectively), while *Calr* and *Eif1* had a greater fraction of all unspliced reads (**Figure 3.6 B**). Genome-wide, the majority of long-reads were all spliced (**Figure 3.7 A**; 68.0% and 73.8% for uninduced and induced cells, respectively), with an average of 88% of all introns being spliced. Therefore, the majority of introns are removed co-transcriptionally. To validate this finding, I examined the read length distribution for reads of each splicing status (**Figure 3.7 B**). As expected, partially spliced and all unspliced reads were on average longer than all spliced reads due to the presence of introns, suggesting that the efficient shortening of nascent RNA due to splicing limits the lengths of long-reads.

To quantify co-transcriptional splicing for each intron detected by at least 10 long-reads, I defined a metric termed the Co-transcriptional Splicing Efficiency (CoSE), tabulated as the number of spliced reads that span the intron divided by the total number of reads (spliced + unspliced) that span the intron (**Figure 3.8 A**). A higher CoSE value indicates a higher fraction of co-transcriptional splicing. To validate this metric, I analyzed an independently generated total RNA-seq dataset in uninduced MEL cells (downloaded from ENCODE; (Davis et al., 2018)). Although nascent RNA is rare in total RNA, the density of reads mapping to a given intron

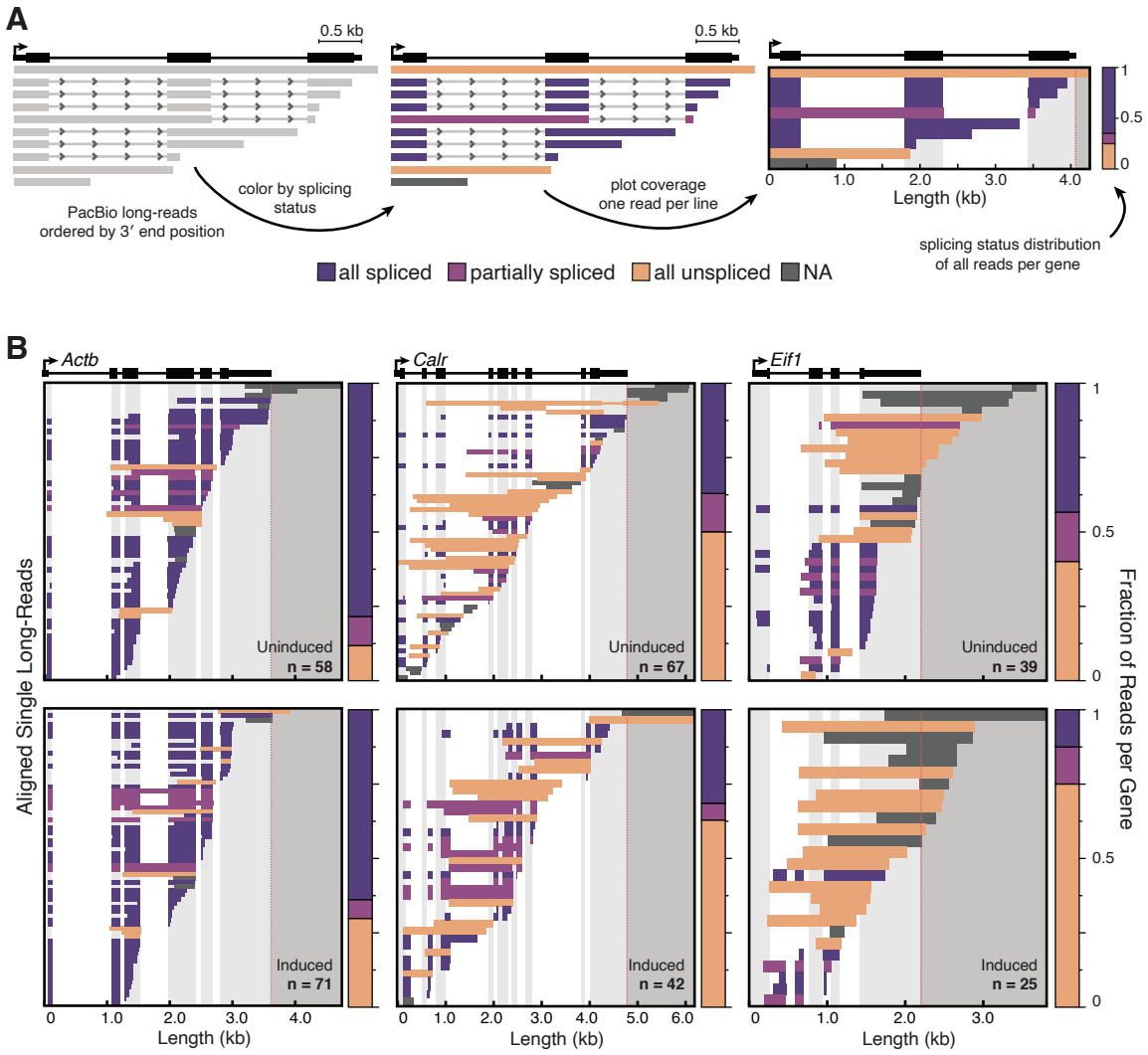

Figure 3.6: Nascent RNA long-read sequencing reveals widespread co-transcriptional splicing.

(A) LRS data visualization for analysis of co-transcriptional splicing. 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. Colour code indicates the splicing status of each transcript. Each horizontal row represents one read. Panels at far right and below: 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). The number of individual long-reads aligned to each gene (n) is indicated. Bar graph at the far right of each plot indicates the fraction of reads that are all spliced (dark purple), partially spliced (light purple), or all unspliced (yellow) for that gene. **(B)** LRS data are shown for uninduced (top) and induced (bottom) MEL cells for three representative genes: *Actb*, *Calr*, and *Eif1*.

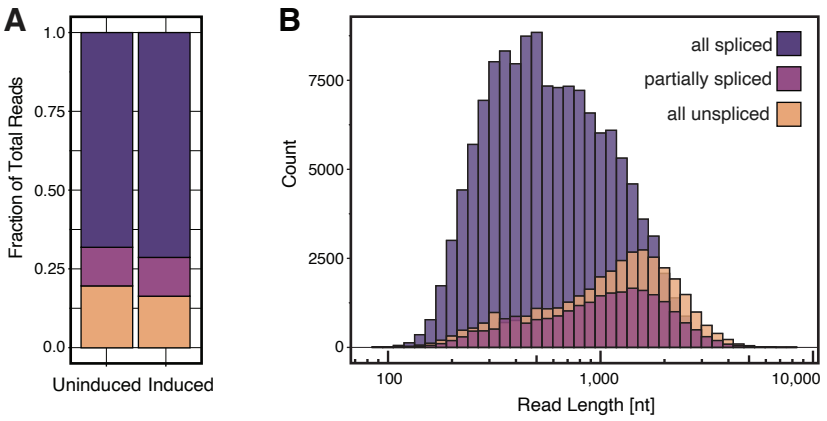

Figure 3.7: Distributions of all spliced, partially spliced, and all unspliced long-reads.
(A) Fraction of long-reads genome-wide that are all spliced, partially spliced, or all unspliced (n = 120,143 reads uninduced, n = 71,639 reads induced). **(B)** Long-read length distribution separated by splicing status (n = 134,857 all spliced reads, n = 23,833 partially spliced reads, n = 35,957 all unspliced reads).

is expected to be inversely proportional to splicing efficiency. The ratio of intron-mapping reads relative to the flanking exon-mapping reads was calculated for each intron and compared to CoSE levels. As expected, higher CoSE corresponded to lower relative intron coverage in the total-RNA seq data (**Figure 3.9 A**). Thus, this independent data set validates the CoSE metric. CoSE values also remained stable across all levels of read coverage (**Figure 3.9 B**).

To determine if intron splicing events are coordinated within the same transcript, I asked how similar CoSE values were between introns in the same transcript. To do so, transcripts containing at least 3 introns with recorded CoSE values (n = 2,028) were compiled. I found that the variance in CoSE between introns within the same transcript was significantly smaller than the variance in CoSE for the same number of randomly assorted introns (**Figure 3.8 B**). Taken together, these results suggest that most introns are well-spliced co-transcriptionally, and that splicing is coordinated in mammalian multi-intron transcripts expressed by both uninduced and induced MEL cells.

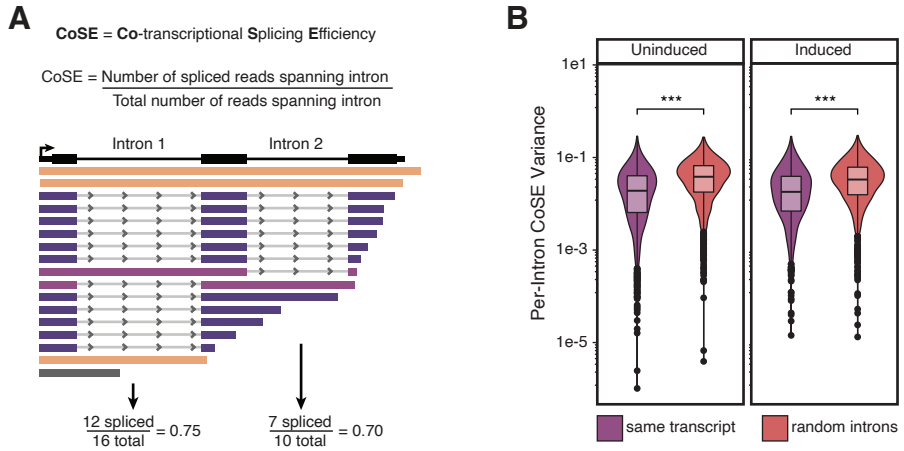

Figure 3.8: Individual mammalian nascent RNA sequences reveal coordination of co-transcriptional splicing.
(A) For each intron that is covered by 10 or more reads, CoSE is defined as the number of reads that are spliced divided by the total number of reads that span the intron. **(B)** Variance in CoSE for transcripts that include 3 or more introns covered by at least 10 reads (n = 1,240 transcripts uninduced, n = 788 transcripts induced) compared to the variance in CoSE for a randomly selected group of introns. Significance tested by Mann Whitney U-test; *** represents p-value < 0.001.

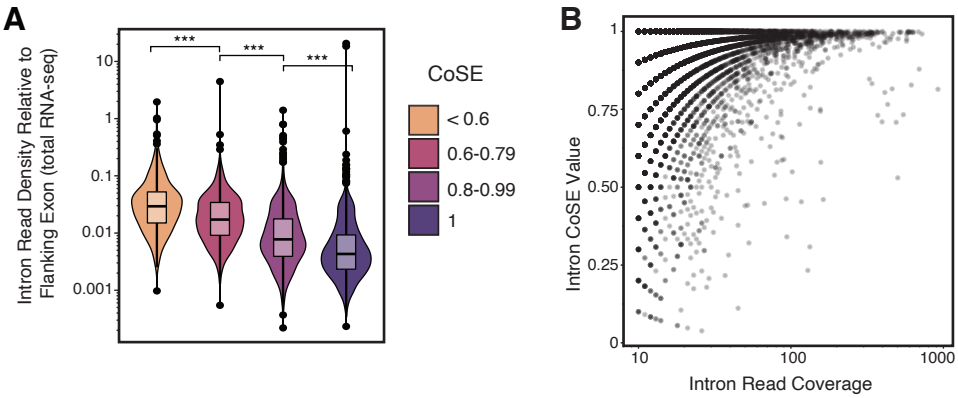

Figure 3.9: CoSE values agree with total RNA-seq and remain stable across intron coverage levels.
(A) Intron coverage relative to flanking exon coverage from untreated MEL total RNA-seq Illumina data (ENCODE) is shown as an independent indicator of unspliced RNA levels. Introns that were covered by at least 10 long-reads in the untreated condition were split into bins based on the CoSE metric calculated from long-read sequencing data. Significance tested by Mann Whitney U-test; *** represents p-value ¡0.001. **(B)** CoSE values as a function of the number of long-reads covering each intron. Each point represents an intron for which a CoSE value was calculated, with uninduced and induced data shown combined (n = 14,422 introns).

3.3 Co-transcriptional Splicing Occurs Rapidly After Intron Transcription

The frequency of all-spliced nascent transcripts implies that splicing in mammalian cells is rapid enough to match the rate of transcription. A direct way to address this is to measure the position of Pol II on nascent RNA when ligated exons are observed. Observing Pol II downstream of a spliced junction indicates that the active spliceosome has assembled and catalyzed splicing in the time it took for Pol II to translocate the measured distance. Therefore, I determined the distance in nucleotides between the 3′ end of each read and the nearest spliced exon-exon junction (**Figure 3.10 A**). To eliminate 3′ ends that arise from splicing intermediates and not from active transcription, reads with 3′ ends mapping precisely to the last nt of exons were removed from this analysis. Although the longest distances between splice junctions and elongating Pol II were just over 6 kb, these were rare. Instead, 75% of splice junctions were within 300 nt of a 3′ end, and the median distance was 154 nt in uninduced cells and 128 nt in induced cells (**Figure 3.10 B**). Therefore, changes in the gene expression program during erythropoiesis did not alter the dynamic relationship between transcription and splicing. Consistent with this, CoSE values were similar when comparing induced to uninduced cells (**Figure 3.10 C**; Spearman's rho = 0.56). In fact, only 66 introns with improved splicing, and 42 introns with reduced splicing displayed $>$ 2-fold change in CoSE upon induction. Taken together, these results show that although global changes in gene expression occur between these two timepoints, the relationship between transcription and splicing remains the same. Overall, these two measurements do not support major changes in splicing efficiency during erythroid differentiation. Moreover, the distance from Pol II to the nearest splice junction was independent of GO category or intron length (**Figure 3.11 B**; GO analysis not shown). Because median exon size

in the mouse genome is 151 nt (Waterston et al., 2002), my data indicate that active spliceosomes can be fully assembled and functional when Pol II is within or just downstream of the next transcribed exon. Recent direct sequencing of nascent RNA seemed to reveal less rapid splicing (Drexler et al., 2020). However, when I analyzed this dataset in the same manner as my own, the cumulative distance from Pol II to the nearest splice junction is similarly close across organisms and cell types (median distance in human BL1184 = 244 nt, human K562 = 310 nt, Drosophila S2 = 335 nt; **Figure 3.11 A**). Thus, I conclude that efficient and coordinated splicing are a general property of metazoan gene expression.

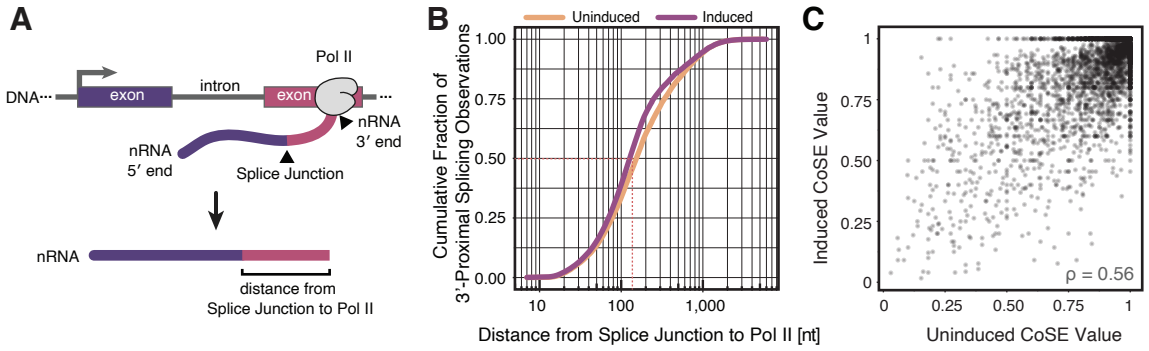

Figure 3.10: Spliceosome-Pol II proximity is unchanged by differentiation. **(A)** Schematic definition of the distance from the 3′ end of a nascent RNA (nRNA) to the most 3′-proximal splice junction. 3′ end sequence reports the position of Pol II when nascent RNA was isolated. **(B)** Distance (nt) from the 3′-most splice junction to Pol II position is shown as a cumulative fraction for uninduced and induced cells (n = 101,911 observations uninduced, n = 66,656 induced). **(C)** CoSE in induced and uninduced conditions. Each point represents a single intron which is covered by at least 10 long-reads in both induced and uninduced conditions. Spearman's rho = 0.56, n = 4,170 introns.

3.4 Pol II Does Not Pause at Splice Sites for Splicing to Complete

Note: PRO-seq experiments and analysis in sections 3.4 - 3.6 were performed in collaboration with Claudia Mimoso and Karen Adelman (Harvard Medical School).

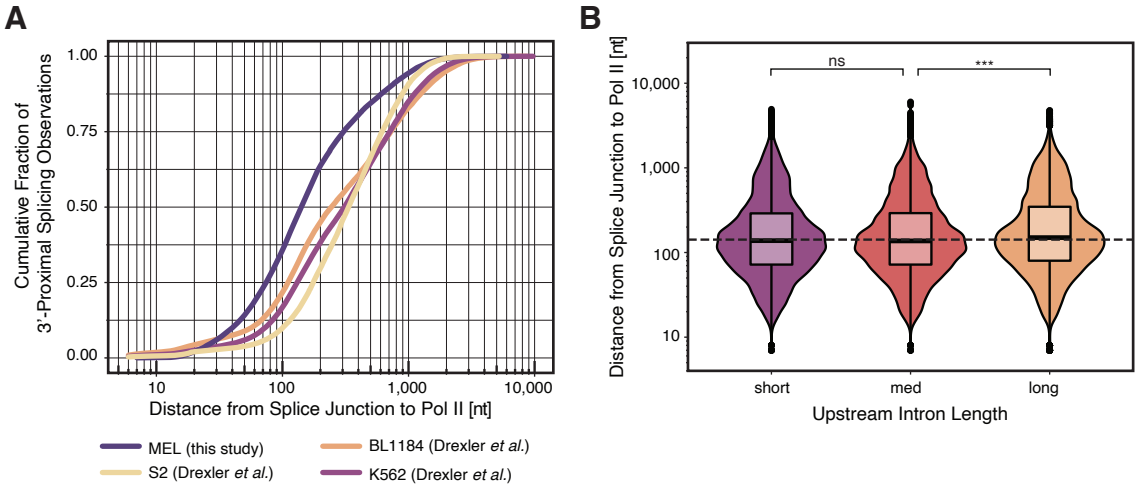

Figure 3.11: Pol II is detected near splice junctions in multiple cell types.
(A) Distance (nt) from the most 3′-proximal splice junction to Pol II position is shown as a cumulative fraction. Analysis is the same as in **Figure 3.10**, but with data from this study and nanoCOP data from Drexler et al. (2020) (n = 39,397 reads in BL1184 cells, n = 183,829 reads in K562 cells, n = 68,956 reads in S2 cells, and n = 168,567 reads in MEL cells). **(B)** Distance (nt) from the most 3′-proximal splice junction to Pol II position is shown categorized by upstream intron length in three equal-sized bins (short = 30 nt – 443 nt, med = 443 nt – 1,738 nt, long >= 1,738 nt). Significance tested by Mann Whitney U-test: *** represents p-value < 0.001, ns represents p-value > 0.05.

One explanation for the relatively short distances observed between splice junctions and Pol II may be that Pol II pauses just downstream of an intron, allowing time for splicing to occur before elongation continues. Alternatively, Pol II could pause at the end of the downstream exon. This model has been proposed as a mechanism for splicing and transcription to feedback on each other, with pausing providing a possible checkpoint for correct RNA processing (Alexander et al., 2010b; Carrillo Oesterreich et al., 2011; Chathoth et al., 2014; Milligan et al., 2017). However, the conclusions of recent work have 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 resolve this controversy, we measured changes in elongating Pol II density genome-wide using Precision Run-On sequencing (PRO-seq) in MEL cells. PRO-seq maps actively elongating Pol II complexes at single-nucleotide resolution by incorporating 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 undergoing active elongation; these 3′ ends cannot originate from other chromatin-associated intermediates, such as splicing intermediates.

We analyzed our PRO-seq data to determine if transcription elongation behavior changes across intron-exon boundaries. Because both induced and uninduced LRS datasets showed an overlapping distribution of Pol II when spliced products are observed, we initially combined the PRO-seq datasets. As expected, metagene plots around active TSSs revealed prominent promoter-proximal pausing (**Figure 3.12 A**); (Core and Adelman, 2019). Analyzing PRO-seq signal around splice sites initially revealed a small peak near the 5′SS. To control for the possibility that high PRO-seq density from TSS peaks might bleed through to the first 5′SS, first introns were independently analyzed. Indeed, elevated PRO-seq signal in the vicinity of

5′SSs was only seen at first introns, and only at introns with 5′SS <= 250 nt from the TSS (**Figure 3.13 A**). Interestingly, middle introns showed a similar profile to terminal introns, both lacking increased PRO-seq signal around splice sites (**Figure 3.13 B**). Accordingly, after removal of first introns from our analysis, PRO-seq signal showed only minor fluctuations around 5′SSs and 3′SSs (**Figure 3.12 A**).

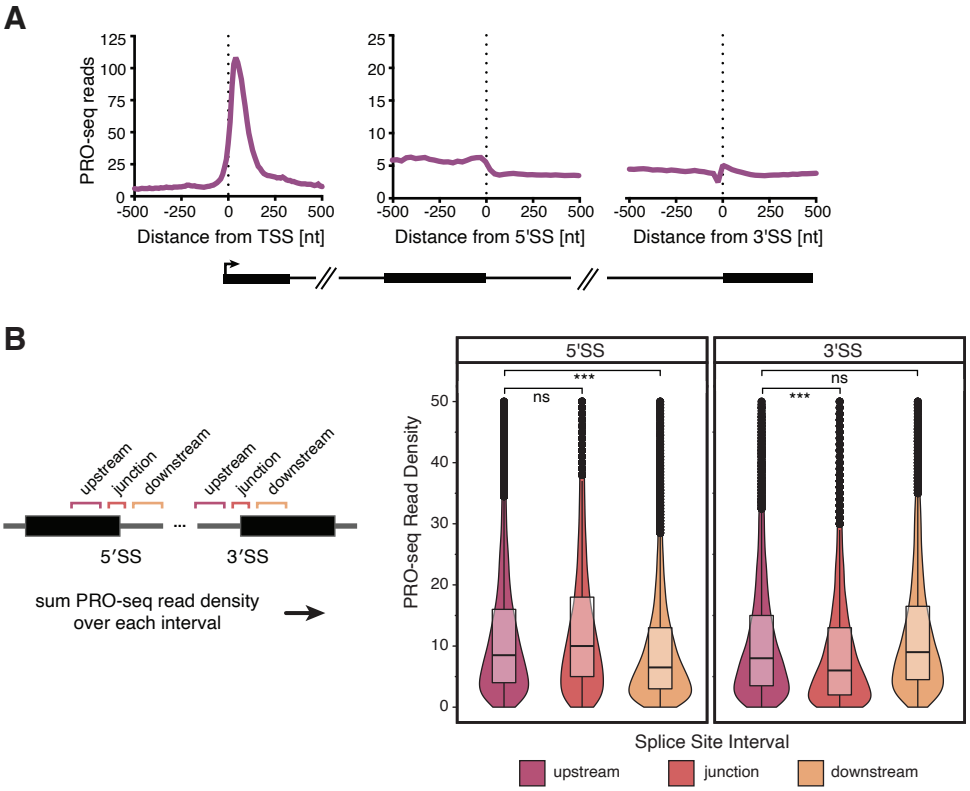

Figure 3.12: Pol II does not pause at 5′ or 3′ splice sites.
(A) PRO-seq 3′ end coverage is shown aligned to active transcription start sites (TSS), 5′ splice sites (5′SS), and 3′ splice sites (3′SS). **(B)** Left: Schematic illustrating the use of colour-coded intervals to quantify PRO-seq reads around each 5′SS and 3′SS to test for significance of pausing. Right: PRO-seq read density summed in each of the intervals indicated at left around 5′SSs (left) and 3′SSs (right) from introns with at least 10 reads in uninduced conditions (n = 3,505). Significance tested by paired t-test; *** represents p-value < 0.001, ns represents p-value > 0.05.

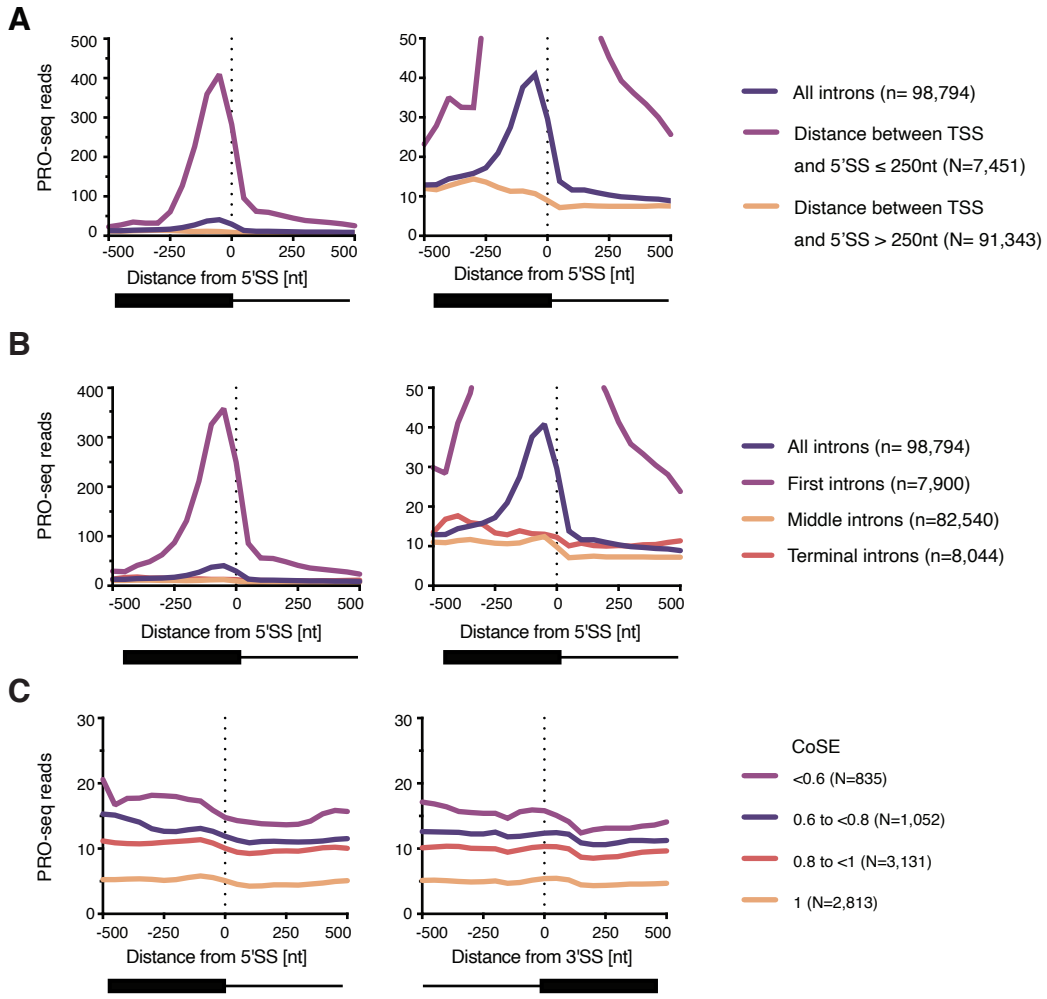

Figure 3.13: PRO-seq reveals no Pol II pause at any subset of introns.
(A) PRO-seq 3′ end coverage aligned to 5′SSs for all introns (dark purple), introns where the distance from the TSS to the 5′SS is <= 250 nt (light purple), and introns where the distance to the TSS to the 5′SS is > 250 nt (light orange). Right panel shows data scaled to show all introns. **(B)** PRO-seq 3′ end coverage aligned to 5′SSs for all introns from active transcripts (dark purple), first introns only (light purple), middle introns (light orange), and terminal intron (dark orange). Right panel shows data scaled to show all introns. **(C)** Average PRO-seq 3′ end coverage is shown around 5′SS (left) and 3′SS (right) for introns in each CoSE category as indicated. N = number of introns in each category.

3.5 Statistical Methods for Evaluating Pol II Pausing

The depth of our PRO-seq libraries enabled us to determine if these minor fluctuations in PRO-seq signal were statistically significant. To do so, we compared summed PRO-seq read counts in three windows surrounding each intron/exon junction for uninduced cells only (**Figure 3.12 B; left**): -150 to -50 nt upstream of the junction, -40 to +10 nt spanning the junction, and +50 to +150 nt downstream of the junction. Comparing the signal between the upstream window and 5′SS-spanning window indicated no statistically significant increase in PRO-seq read density (paired t test $p > 0.05$). In contrast, a significant decrease in PRO-seq signal is observed as Pol II moves from the exon into the intron ($p < 0.0001$; comparing upstream to downstream window at the 5′SS), consistent with data showing faster transcription rates within introns (Jonkers et al., 2014; Veloso et al., 2014). Interestingly, we observe a dip in PRO-seq signal right before the 3′SS ($p < 0.0001$), instead of a peak of Pol II consistent with pausing. Although the cause of this dip remains to be determined, it is unlikely to represent Pol II arrest and/or termination near this junction because signal downstream of the 3′SS does not decrease. In summary, we can detect significant changes in Pol II elongation behavior as it moves from exon to intron and vice versa, but we find no evidence for significant pausing at splice junctions.

To rigorously compare splicing efficiency to Pol II elongation behavior, we evaluated PRO-seq signals around introns binned by CoSE values from our LRS data. Again, we observed no significant differences in Pol II profile around splice sites within any group of introns (**Figure 3.13 C**). Interestingly, the overall level of PRO-seq coverage is lower in transcripts with higher CoSE values, suggesting that lowly expressed transcripts might be more efficiently spliced. Finally, a number of PRO-seq reads contained spliced junctions despite the short read lengths (**Figure 3.14**; 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, within a median distance of 128-154 nt. Taken together, two complementary methods to probe Pol II position and splicing status indicate that splicing can occur when Pol II is in close proximity to the 3′SS and in the absence of transcriptional pausing.

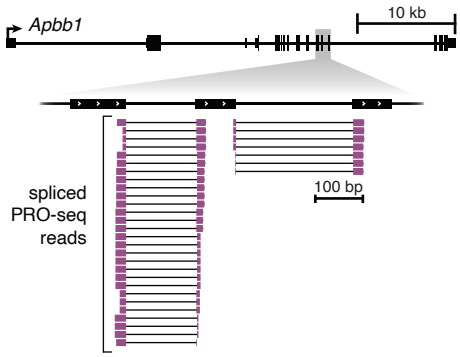

Figure 3.14: Spliced PRO-seq reads confirm co-transcriptional splicing.
Genome browser view showing spliced PRO-seq reads aligned to the *Apbb1* gene, where 3′ ends of reads represent the position of elongating Pol II. Only spliced reads, filtered from all reads, are shown.

3.6 Splicing Intermediates Are Abundant for a Subset of Introns

I expected to readily observe transient splicing intermediates in my nascent RNA libraries, because the two-step transesterification reaction yields a 3′-OH at the end of the upstream exon after step I (**Figure 3.15 A**). Indeed, 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, 2018). As expected, I observed elevated 3′ end coverage precisely at the last nucleotide of exons (**Figure 3.15 B**), with these first step splicing intermediate reads accounting for 7.0% of the data (**Table 5.2**). The rarity of splicing intermediates detected agrees with my finding that splicing does not continue

during chromatin fractionation or RNA isolation (**Figure 3.2**).

Nevertheless, a small number of genes contained a large number of splicing intermediates at a single intron within the gene (**Figure 3.15 C and D**). For example, 216 of the 433 reads mapped to *Alas2* had 3′ ends mapped to the end of exon four (**Figure 3.15 D** and **Figure 3.16 A**). I note that several reads (16/433) mapping to *Alas2* were one of the extremely rare instances of potential recursive splicing. In this case, I observed an unannotated splice junction which generated a new 5′SS immediately adjacent, with the junction sequence cagGUAUGU (**Figure 3.16 B**). Although I observed no other compelling instances of recursive splicing in my dataset, I note that recursive splicing has been previously characterized in extremely long introns, which are difficult to detect using my methods (Pai et al., 2018; Sibley et al., 2015). Nevertheless, I conclude that recursive splicing could occur co-transcriptionally.

To determine what features of specific introns might lead to increased splicing intermediates, I counted and normalized the number of splicing intermediates observed for each intron. The normalized intermediate count (NIC), is defined as the number of splicing intermediate reads at the last nt of the exon divided by the sum of splicing intermediate reads and spliced reads. This metric reports the fraction of long-reads that are captured between step I and step II of splicing. All unique introns which were covered by at least 10 long-reads were binned based on their observed NIC value, and the splice site strength of the introns in each bin was calculated using the MaxEnt algorithm (Yeo and Burge, 2004). While the 5′SS score was relatively constant, introns with the highest NIC value tended to have lower 5′SS and 3′SS scores (**Figure 3.17 A**). Intron length or GC-content showed no similar trend (**Figure 3.17 B and C**).

Finally, I tested whether or not spliceosomal stalling between steps I and II could be associated with Pol II pausing, by analyzing PRO-seq signals downstream of the

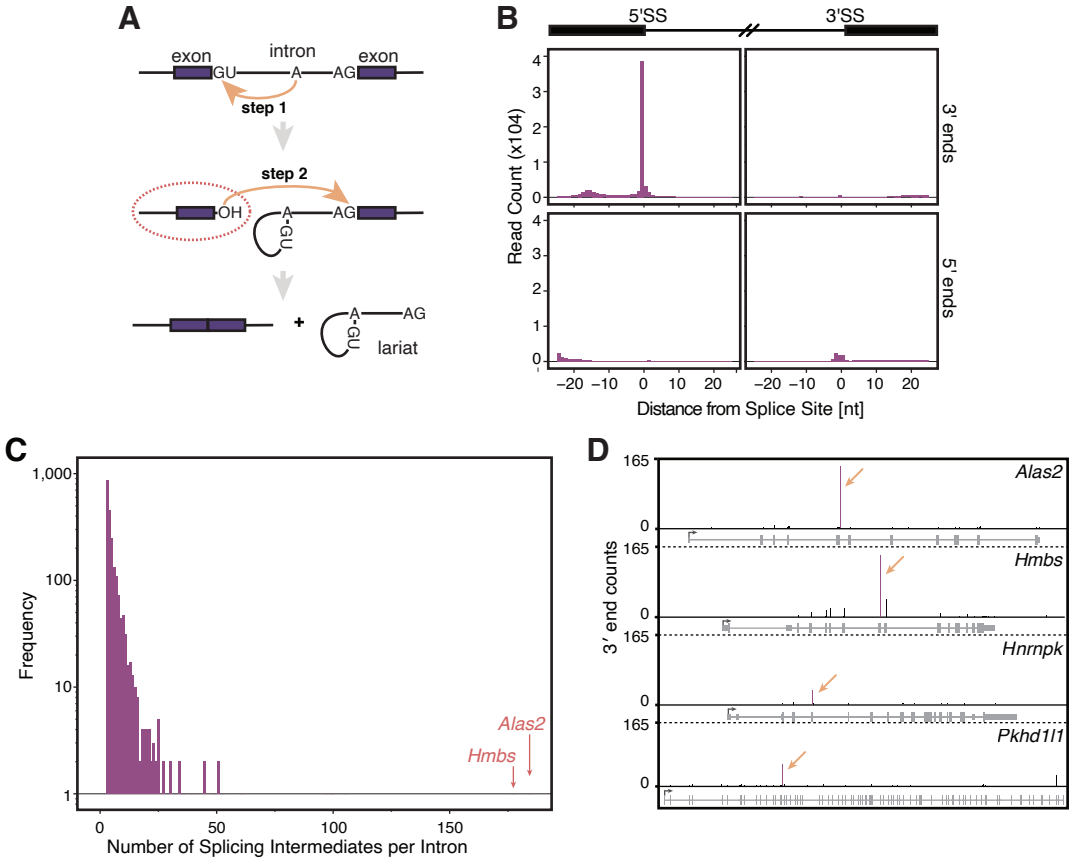

Figure 3.15: Splicing intermediates are detected at a subset of introns.
(A) Schematic definition of first step splicing intermediates (dotted red oval), which have undergone the first step of splicing and have a free 3′-OH that can be ligated to the 3′ end DNA adapter. Splicing intermediate reads are characterized by a 3′ end at the last nucleotide of the upstream exon. **(B)** Coverage of long-read 3′ ends (top panels) and 5′ ends (bottom panels) aligned to 5′SSs (left) and 3′SSs (right) of introns. **(C)** Histogram showing frequency of splicing intermediates upstream of each unique intron from active transcripts in MEL cells. Most introns show 0 or 1 splicing intermediates, while some introns, like those in *Alas2* and *Hmbs*, indicated with orange arrows, exhibit over 150 splicing intermediates reads. **(D)** Coverage of long-read 3′ ends across four example genes. Arrows indicate the positions where the most abundant splicing intermediates are observed.

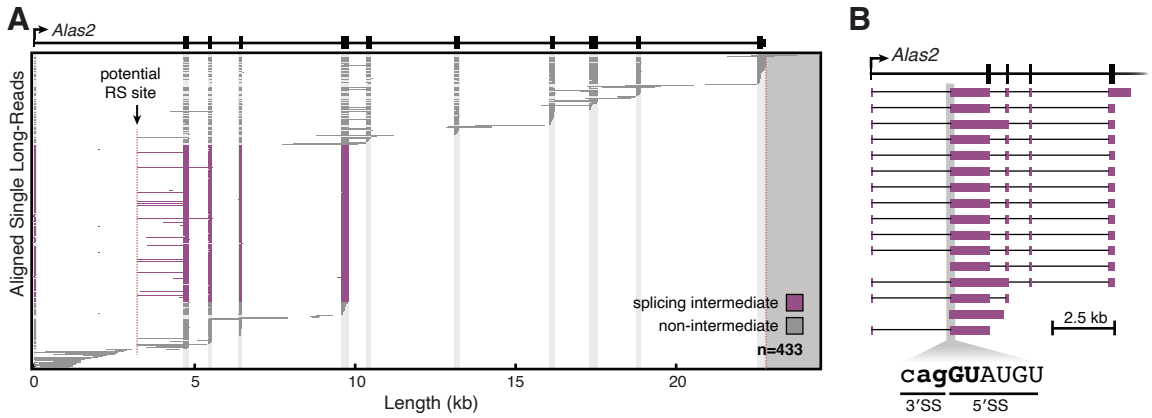

Figure 3.16: Instance of recursive splicing at *Alas2*.
(A) Individual long-reads are shown for the gene *Alas2*. Diagram is similar to **Figure 3.6**, but individual reads are coloured depending on whether they are splicing intermediates (purple) or not (gray). Data for uninduced and induced cells are shown combined. Potential recursive splicing site is indicated by an arrow and dotted line. **(B)** Recursively spliced reads from **(A)** are shown in detail.

associated 3′SSs. Note that PRO-seq signal was universally higher in introns with 1 or more splicing intermediates as was total RNA-seq density in flanking exons, indicating that these genes were more highly expressed (**Figure 3.18 B and C**). However, no differences in Pol II density were detected around 3′SS between introns with NIC = 0 and NIC $>$ 0 (**Figure 3.18 A**). Based on these data, we suggest that introns with weak 3′SSs experience a delay between the catalytic steps of splicing without an associated delay in transcription.

3.7 Unspliced Transcripts Display Poor Cleavage at Gene Ends

Consistent with physiological terminal erythroid differentiation, my induced MEL cells shifted to maximal expression of α- and β-globin genes, each containing two introns. Markedly increased numbers of long-reads mapped to the β-globin (*Hbb-b1*) locus were detected, in agreement with increased β-globin mRNA levels (**Figure**

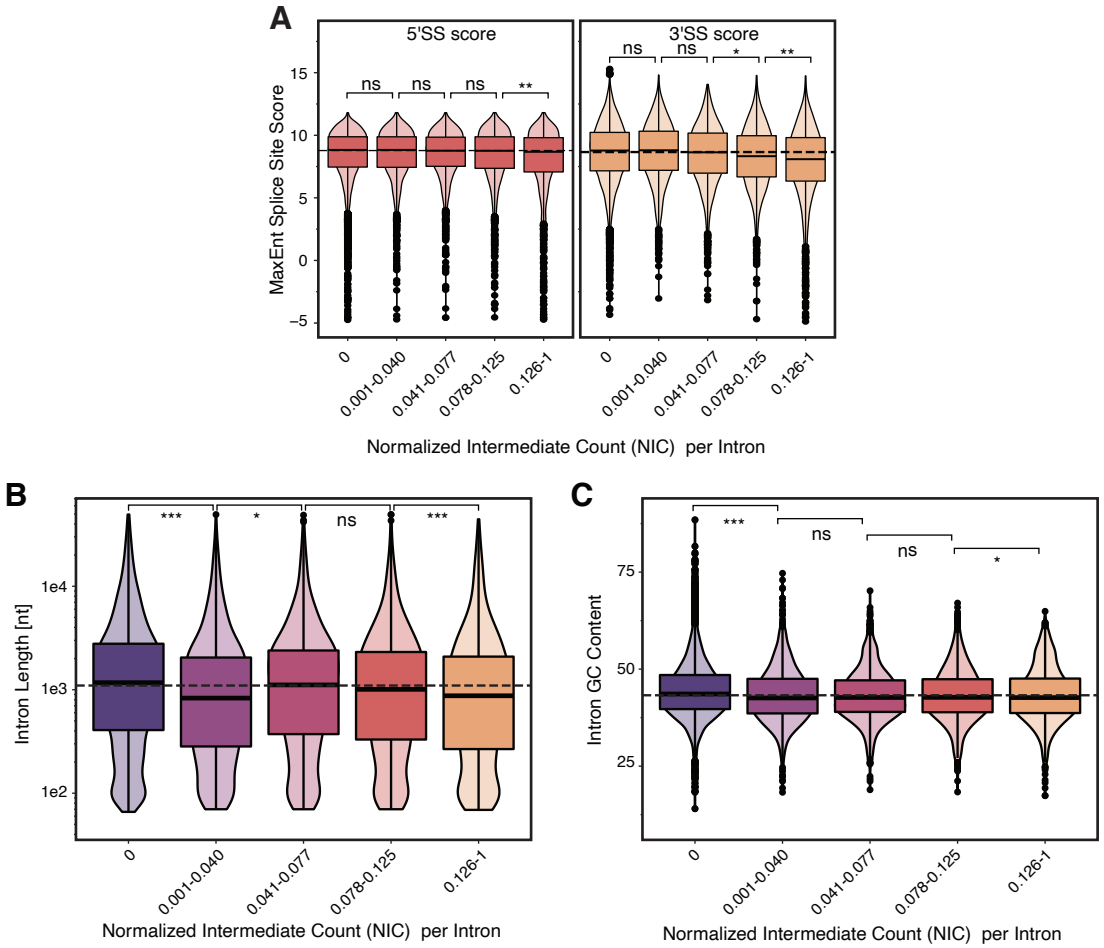

Figure 3.17: Introns with weak 3′SSs accumulate splicing intermediates.
(A) MaxEnt splice site scores for 5′SS (left) and 3′SS (right) for introns with a coverage of at least 10 long-reads is shown categorized by the normalized intermediate count (NIC) at each intron. Introns with NIC = 0 (n = 3,890) are shown separately, and all other introns with NIC $>$ 0 (n = 2,647) are separated in quartiles with NIC values shown. **(B)** Intron length (nt) and **(C)** intron GC-content for unique introns from active transcripts categorized by NIC value. Introns with NIC = 0 (n = 3,890) are shown separately, and all other introns with NIC $>$ 0 (n = 2,647) are separated in quartiles with NIC values shown.

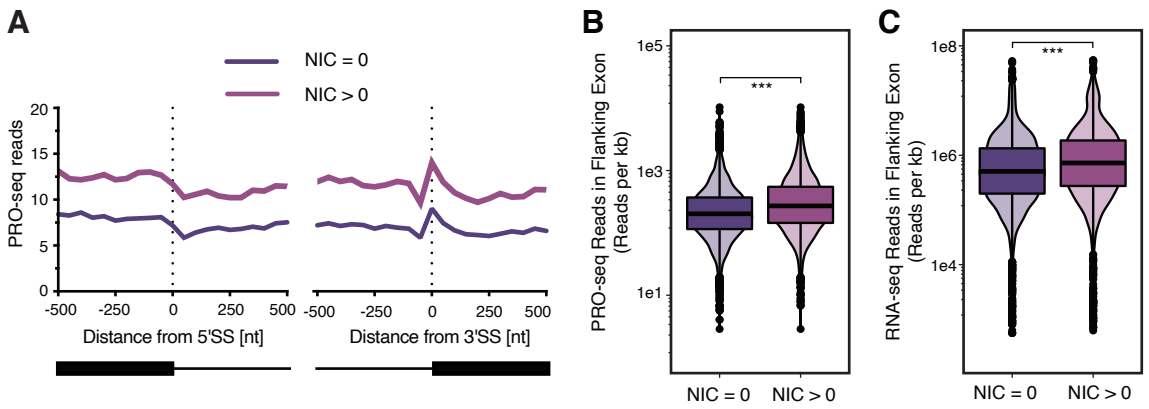

Figure 3.18: Pol II pausing is not assoicated with a delay in catalytic steps of splicing.
(A) Raw PRO-seq 3′ end coverage from uninduced cells aligned to 5′SSs, and 3′SSs for introns with NIC = 0 (n = 4,402), or NIC $>$ 0 (n = 3,427). **(B)** PRO-seq and **(C)** Total RNA-seq (ENCODE) reads summed within the upstream and downstream exons surrounding each intron containing 10 or more long-reads in the uninduced condition and normalized by exon length for introns with NIC = 0 (dark purple) or NIC $>$ 0 (light purple).

3.19 A; **Figure 3.1 C**). To my surprise, a large fraction of individual β-globin long-reads in the induced condition had 3′ ends that mapped up to 2.5 kb downstream of the annotated polyA site (PAS), indicating that these transcripts failed to undergo 3′ end cleavage at the PAS. Pol II occupancy past the β-globin PAS was confirmed by PRO-seq (**Figure 3.19 B**). Notably, PRO-seq reads are commonly detected well past the gene 3′ ends due to transcription termination (Core et al., 2008). However, my LRS data indicate significant transcription past the polyA cleavage site in the absence of 3′ end cleavage, which cannot be revealed by PRO-seq. Remarkably, the β-globin transcripts that escaped 3′ end cleavage were almost uniformly unspliced (**Figure 3.19 A**). The α-globin genes (*Hba-a1* and *Hba-a2*) displayed a similar phenomenon (**Figure 3.20 A-C**). Thus, a significant fraction of nascent globin RNAs undergo "all or none" RNA processing under these conditions of erythroid differentiation: either both introns are efficiently spliced and the nascent RNA is cleaved at the 3′ end or, conversely, both introns are retained and the nascent RNA is inefficiently cleaved.

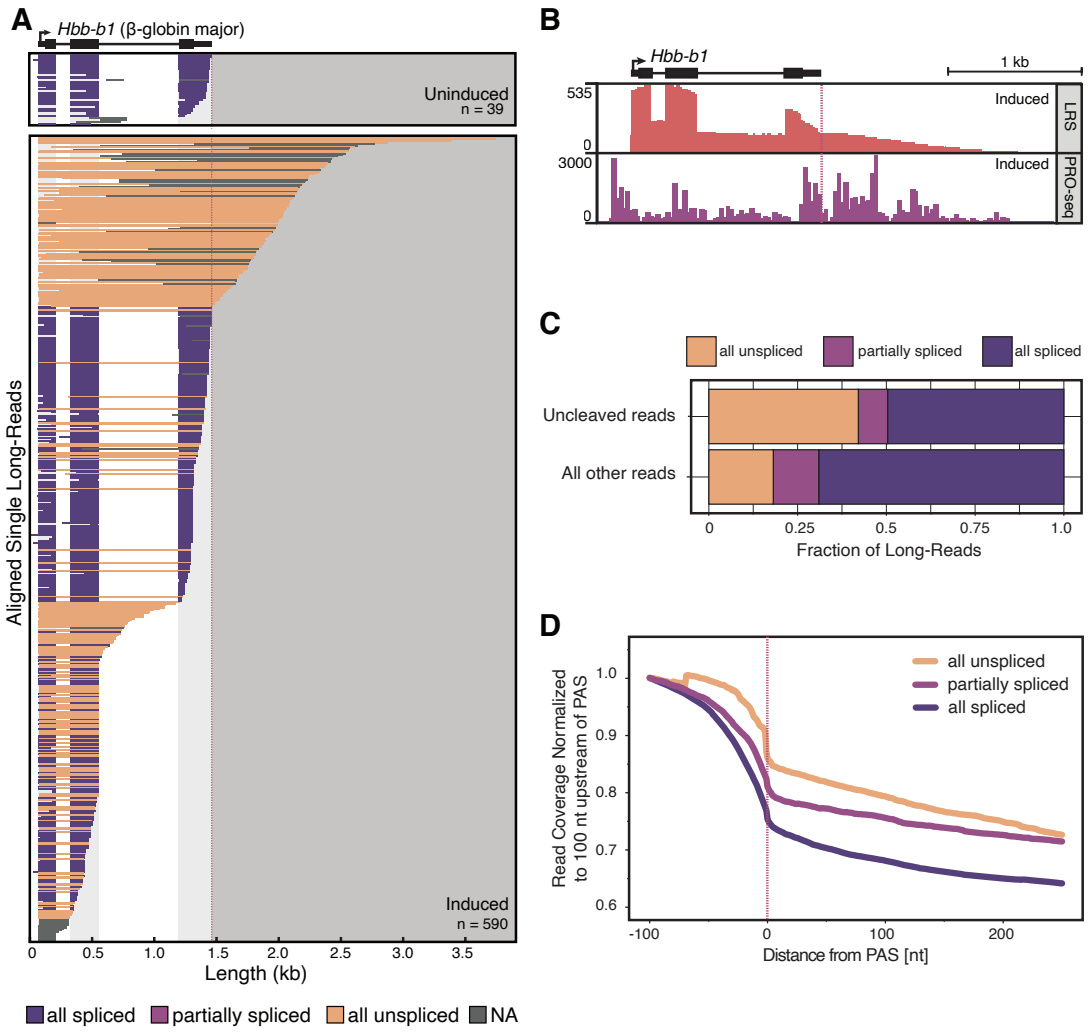

Figure 3.19: Poor splicing efficiency is associated with inefficient 3′ end cleavage.

(A) Individual long-reads are shown for the major β-globin gene (*Hbb-b1*). Diagram is as described in **Figure 3.6**. **(B)** LRS coverage (orange) and PRO-seq 3′ end coverage (purple) in 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. 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)** Fraction of uncleaved long-reads (top) and all other long-reads (bottom) categorized by splicing status (as described in **Figure 3.6**). Uncleaved reads have a 5′ end within an actively transcribed gene region and a 3′ end greater than 50 nt downstream of the PAS (n = 5,694 uncleaved long-reads, and n = 172,612 other long-reads). **(D)** Long-read coverage in the region downstream of PASs is shown for long-reads separated by their splicing status. Coverage is normalized to the position 100 nt upstream of each PAS (n = 35,982 all unspliced reads, n = 24,102 partially spliced reads, and n = 134,581 all spliced reads). Red dotted line indicates PAS position.

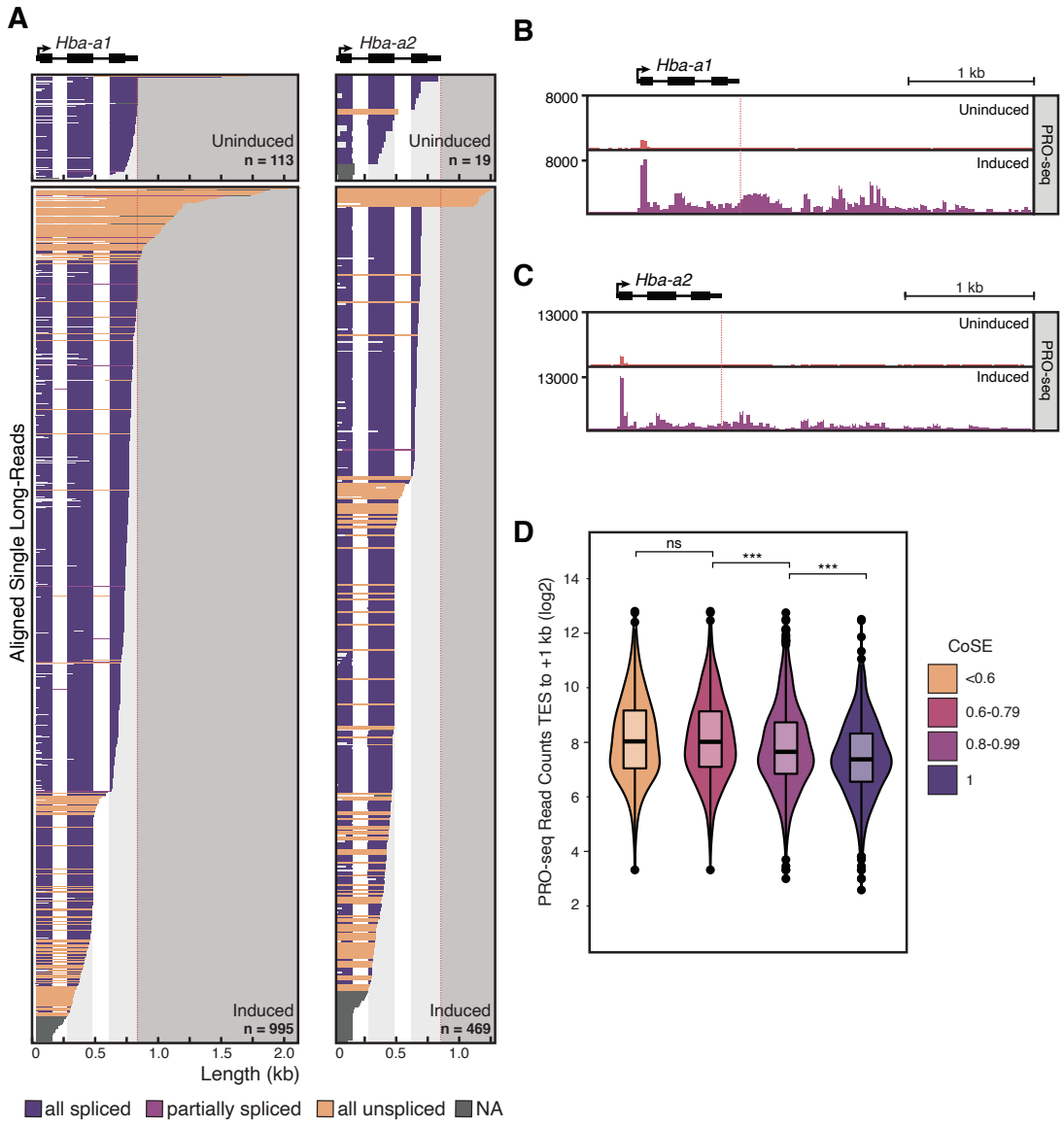

Figure 3.20: α-globin genes exhibit unspliced transcripts that are uncleaved and extend past the PAS.

(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 3.6**. **(B-C)** PRO-seq 3′ end read coverage is shown downstream of the *Hba-a1* **(B)** and *Hba-a2* **(C)** gene loci. 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 the density of gene body reads. Red dotted line indicates PAS. Data represent three biological replicates combined. **(D)** Summed PRO-seq signal in the region 1 kb downstream of the PAS in uninduced cells. Introns were binned by CoSE values, then the PRO-seq read coverage was calculated for each intron and counted only once per bin. Significance tested by Mann-Whitney U-test; *** represents p-value ¡ 0.001, ns represents p-value $>$ 0.05

Because other genes showed evidence of all-or-none splicing (**Figure 3.6 B**), a correlation between splicing and cleavage at the PAS was examined globally. To do so, I categorized long-reads as uncleaved if the 5′ end originated within a gene body and the 3′ end mapped more than 50 nt downstream of an annotated PAS. In comparison to all other long-reads, uncleaved long-reads were 2.5-fold more likely to be unspliced (**Figure 3.19 C**). Next, I analyzed long-read coverage downstream of annotated PASs for reads with different splicing statuses. Coverage of all unspliced reads was globally higher in the region downstream of a PAS than it was for partially spliced or all spliced reads (**Figure 3.19 D**). PRO-seq data support this observation, with significantly less PRO-seq signal observed in the region downstream of the PAS for transcripts harboring introns with the highest CoSE values (**Figure 3.20 D**). This genome-wide decrease in splicing efficiency associated with impaired 3′ end cleavage confirmed the coordination between splicing and 3′ end processing prominently observed in the globin genes.

3.8 A β-thalassemia Mutation Enhances Splicing and 3′ End Cleavage Efficiencies

To investigate how mutations in splice sites alter co-transcriptional splicing efficiency, I 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, creating a cryptic 3′SS 19 nt upstream of the canonical 3′SS (**Figure 3.21 A**). This thalassemia-causing mutation, known as IVSI-110, generates an *HBB* mRNA with an in-frame stop codon, resulting in a 90% reduction in functional HBB protein through nonsense-mediated decay (Spritz et al., 1981; Vadolas et al., 2006). I utilized two MEL cell lines expressing either an integrated copy of a human β-globin minigene (MEL-HBB^{WT}) or the human β-globin minigene with the IVSI-110

mutation (MEL-*HBB*$^{\text{IVS-110(G>A)}}$) (Patsali et al., 2018). Specific targeting of these integrated human *HBB* loci during library preparation resulted in an average of 24,970 nascent RNA long-reads that mapped to the *HBB* gene for each of 3 biological replicates (**Table 5.2**), allowing rigorous statistical analysis. As previously reported, the majority (94%) of intron 1 splicing in the MEL-*HBB*$^{\text{IVS-110(G>A)}}$ cell line occurred at the cryptic 3′SS.

MEL-*HBB*$^{\text{IVS-110(G>A)}}$ cells exhibited a significant increase in the fraction of long-reads that were all spliced and a corresponding decrease in long-reads that were all unspliced (**Figure 3.21 B**). The low co-transcriptional splicing efficiency detected for endogenous mouse β-globin was mirrored in the stably integrated *HBB* minigene, where 80-92% of long-reads were all unspliced (compare **Figures 3.19 A** and **3.21 B**). Long-reads with only intron 1 spliced were present at a higher ratio in the MEL-*HBB*$^{\text{WT}}$ cells, whereas long-reads with only intron 2 spliced were more frequent in the MEL-*HBB*$^{\text{IVS-110(G>A)}}$ cells. This distribution suggests that the cryptic intron 1 is spliced more efficiently than the WT intron 1, leading to a coordinated increase in splicing of intron 2 and a shift from all unspliced to all spliced reads. This interpretation agrees with the coordinated co-transcriptional splicing efficiency I observed in multi-intron transcripts genome-wide (**Figure 3.8 B**). Significantly fewer splicing intermediates were detected for both introns in the MEL-*HBB*$^{\text{IVS-110(G>A)}}$ cells compared to MEL-*HBB*$^{\text{WT}}$ cells, suggesting that inefficient splicing can lead to spliceosomal pausing between catalytic steps I and II (**Figure 3.21 C**).

To rigorously test the possibility that changes in co-transcriptional splicing efficiency determine 3′ end cleavage, read coverage downstream of the *HBB* PAS was used to detect uncleaved long-reads for each category of splicing status (**Figure 3.21 D**). All-unspliced *HBB* reads were detected up to 4 kb past the PAS, similar to endogenous mouse globin genes. When only intron 2 was spliced, cleavage in MEL-*HBB*$^{\text{WT}}$ and MEL-*HBB*$^{\text{IVS-110(G>A)}}$ cells was similar (**Figure 3.21 D**; center right).

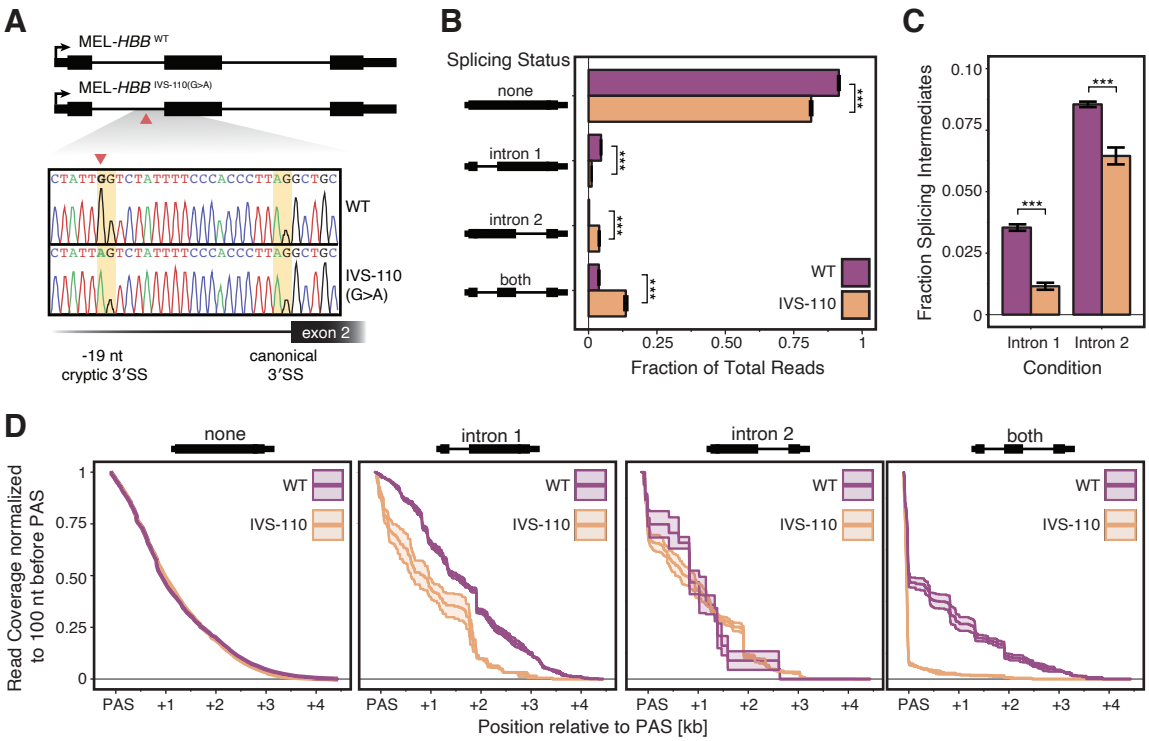

Figure 3.21: Efficient splicing promotes 3′ end cleavage.
(A) Top: schematic describing two engineered MEL cell lines. MEL-*HBB*^{WT} contains an integrated copy of a wild type human globin minigene. In MEL-*HBB*$^{IVS\text{-}110(G>A)}$, a single point mutation (red triangle) mimics a disease-causing thalassemia allele. Bottom: Sanger sequencing of the *HBB* minigene coding strand shows that a G>A mutation leads to a cryptic 3′SS at the AG dinucleotide 19 nt upstream of the canonical 3′SS. **(B)** Distribution of *HBB* long-reads in MEL-*HBB*^{WT} cells (purple) and MEL-*HBB*$^{IVS\text{-}110(G>A)}$ cells (orange) separated by splicing status of intron 1 and intron 2 and measured as a fraction of total reads mapped to the *HBB* gene (n = 20,395 reads in MEL-*HBB*^{WT} cells, and n = 26,244 reads in MEL-*HBB*$^{IVS\text{-}110(G>A)}$ cells). **(C)** Fraction of splicing intermediates at intron 1 and intron 2 in MEL-*HBB*^{WT} cells (purple) and MEL-*HBB*$^{IVS\text{-}110(G>A)}$ cells (orange) measured as a fraction of total reads mapped to the *HBB* gene. For **(B-C)**, significance tested by Mann Whitney U-test; *** represents p-value < 0.001, bar height represents the mean of three biological replicates, and error bars represent standard error of the mean. **(D)** Read coverage in the region downstream of the *HBB* PAS is shown for long-reads separated by their splicing status from MEL-*HBB*^{WT} cells (purple) and MEL-*HBB*$^{IVS\text{-}110(G>A)}$ cells (orange). Coverage is normalized to the position 100 nt upstream of the PAS. Solid line represents the mean coverage of three biological replicates, and shaded windows represent standard error of the mean.

However, a notable decrease in uncleaved reads past the PAS was detected among transcripts spliced at the cryptic 3′SS as compared to the canonical 3′SS (**Figure 3.21 D**; center left). When both introns were spliced, there was an even more dramatic shift to proper cleavage at the PAS in the MEL-*HBB*$^{\text{IVS-110(G>A)}}$ cells (**Figure 3.21 D**; far right). Together, these findings confirm my demonstration of functional coupling between splicing and 3′ end formation at the individual transcript level and highlight the regulatory potential of just a single point mutation. Thus, a previously unappreciated level of crosstalk between splicing and 3′ end cleavage efficiencies is involved in erythroid development.

Chapter 4

Discussion

4.1 Summary

This study reveals functional relationships between co-transcriptional RNA processing events derived from genome-wide analysis of individual nascent transcripts purified from differentiating mammalian erythroid cells. Transcription and splicing dynamics were visualized with unprecedented depth and accuracy through long-read sequencing of nascent RNA and PRO-seq. I conclude that splicing catalysis can occur when Pol II is just 75-300 nt past the intron without transcriptional pausing at the splice sites. Thus, spliceosome assembly and the transition to catalysis often occur when the spliceosome is physically close to Pol II. Two striking cases stood out from my observations of splicing. First, introns that contain a weak 3′SS seem to induce stalling between steps I and II of the splicing reaction itself, causing a buildup of splicing intermediates. Second, inefficient splicing was globally correlated with inefficient 3′ end cleavage on both the population and single transcript level (**Figure 4.1**). I pursued this second phenomenon further in the context of globin gene expression, wherein all two-intron globin genes (two α and two β in mouse) displayed "all-or-none" splicing behavior. Approximately 20% of endoge-

nous nascent *Hbb-b1* transcripts retained both introns and were inefficiently cleaved at the PAS. Remarkably, a patient-derived, thalassemia-causing point mutation in β-globin increased splicing efficiency and 3′ end cleavage. These data show that co-transcriptional splicing efficiency determines 3′ end processing efficiency, as discussed below.

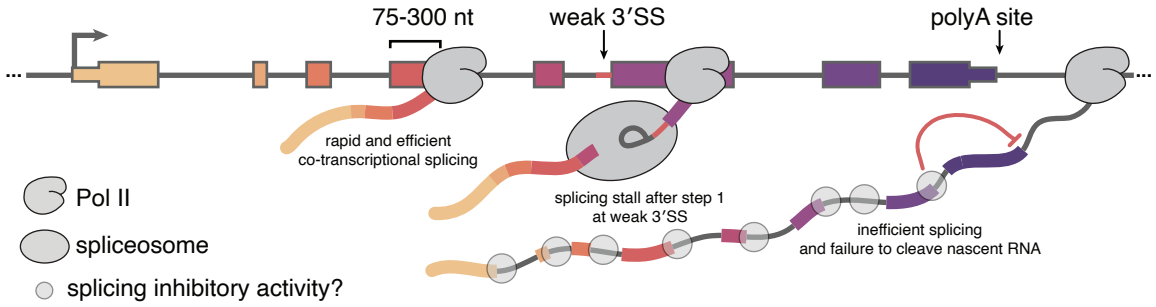

Figure 4.1: Model describing the variety of co-transcriptional splicing phenomena observed during murine erythropoiesis.

4.2 Conservation of Co-transcriptional Splicing Mechanisms

The data presented here indicate that the mammalian spliceosome is capable of assembling and acting on nascent RNA substrates in the same spatial window of transcription as the yeast spliceosome (Carrillo Oesterreich et al., 2016; Herzel et al., 2018). This is 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. My high resolution determination of the fraction of splicing that occurs co-transcriptionally (88%) matches data from alternative methods (e.g. short-read sequencing, metabolic labeling, and imaging) showing that 75-87% of splicing is co-transcriptional in yeast, fly, mouse, and human cells (reviewed in Neugebauer (2019)). These findings suggest widely conserved features of transcription and splicing mechanisms.

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., 2020), which is at odds with my findings. One possible explanation for this discrepancy is that the purification of 4-thiouridine (4sU)-labeled RNA inadvertently enriched for long, intron-containing RNAs that have a greater probability of containing a labeled U residue (introns are U-rich). 4sU incorporation may also impede splicing due to changes in base-pairing among U-rich RNA elements in introns and snRNAs (Testa et al., 1999). Indeed, direct comparison of splicing efficiencies with vs. without 4sU-labelling indicates that 4sU-lableled RNA is less frequently spliced (Drexler et al., 2020). In this context, I note that the principal difference between the data from our two labs is the fraction of unspliced RNAs observed. However, when I analyzed spliced RNAs in the Drexler *et al*. data to determine the distance between splice junctions and the RNA 3′ end, I obtained similar results to my own (**Figure 3.11 A**). I conclude that splicing more typically occurs when Pol is close to the intron.

4.3 Pol II Does Not Pause at Splice Junctions

Importantly, PRO-seq corroborated the efficiency of co-transcriptional splicing. We identified spliced reads within the PRO-seq data, validating the observations made with LRS of purified nascent RNA by an independent method. Similarly, mNET-seq data, which are generated by short-read sequencing of nascent RNA from immunoprecipitated Pol II, has revealed examples of spliced reads (Nojima et al., 2018). Having observed many examples wherein an RNA 3′ end was only a short distance beyond the 3′SS, I considered the hypothesis that Pol II pausing at or near 3′SSs could provide extra time for splicing (Alexander et al., 2010b; Chathoth et al., 2014; Milligan et al., 2017). However, my 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 I do not detect statistically significant pausing at 5′ or 3′SSs (**Figure 3.12 A, B**), in agreement with a recent study using mNET-seq (Sheridan et al., 2019). Importantly, Pol II elongation was not detectably impacted by the splicing efficiency of introns (**Figure 3.13 C**), with no significant differences in PRO-seq signal across splice junctions. These data strongly support the assumption that Pol II travels at a uniform rate across splice junctions. Using the median distance from splicing events to Pol II position in my combined data (142 nt) and taking into account the 0.5-6 kb/min range of measured Pol II elongation rates (Jonkers and Lis, 2015), I calculate that the splicing events detected in my data occurred within 1.4-17 seconds of intron transcription. These rates are similar to those obtained in budding and fission yeasts (Alexander et al., 2010a; Alpert et al., 2020; Carrillo Oesterreich et al., 2016; Eser et al., 2016; Herzel et al., 2018).

4.4 Mechanism of Splicing Catalysis Delay

Due to the 3′ end chemistry and structure of splicing intermediates, I can capture the step I intermediates of splicing with long-read sequencing. Remarkably, splicing intermediates were distributed unevenly among introns and were associated with poor sequence consensus at the downstream 3′SS. 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). Because the spliceosomes associated with these intermediates have already assembled and undergone step I chemistry, the accumulation of intermediates cannot be attributed to defective intron recognition during spliceosome assembly. Instead, the catalytic center of the spliceosome shifts from the branch site (step I) to the 3′SS AG (step II), typically 30-60 nt downstream of the branch site. The mechanism underlying 3′SS choice

during this transition is likely to be influenced by spliceosomal proteins outside of the catalytic core. 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 I to step II (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.

4.5 Resolution of β-globin Splicing Measurements

My LRS data 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" RNAs and "all unspliced" RNAs (**Figure 3.19 A, Figure 3.20 A, Figure 3.21 B**). Although previous studies also linked the splicing of β-globin exon 2 with 3′ end cleavage (Antoniou et al., 1998; Dye and Proudfoot, 1999), here I show all-or-none behavior in the splicing and cleavage decisions made by both α- and β-globins. In any biochemical and/or short-read RNA-seq assay that examines populations of pre-mRNA, inefficient splicing would be one explanation for the bulk result. In reality, one population is unspliced and uncleaved at their 3′ ends. The other population of globin transcripts is efficiently spliced and productively expressed, because polyA cleavage and subsequent Pol II termination also occur efficiently for these RNAs.

4.6 Model for Mechanistic Link Between Splicing and 3′ End Cleavage

The fraction of efficiently spliced β-globin transcripts increased in the thalassemia allele I studied, even though the cryptic 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). Accordingly, a decrease in uncleaved reads extending beyond the PAS in the MEL-*HBB*$^{\text{IVS-110(G>A)}}$ cells was evident compared to the MEL-*HBB*$^{\text{WT}}$ cells. This indicates that splicing efficiency is a determinant of 3′ end cleavage. Several possible mechanisms could be involved. Less efficient splicing can inhibit 3′ end cleavage (Cooke et al., 1999; Davidson and West, 2013; Martins et al., 2011), suggesting that introns retained in transcripts that display readthrough harbor an inhibitory activity that represses 3′ end cleavage (**Figure 4.1**). Candidate inhibitory factors include U1 snRNP, PTB and hnRNP C proteins, each of which binds introns promiscuously *in vivo* (Deng et al., 2020; König et al., 2010). In particular, U1 snRNP binding to introns is known to repress premature 3′ end cleavage and cleavage at PASs (Berg et al., 2012; So et al., 2019; Vagner et al., 2000). I speculate that this inhibitory activity persists longer on inefficiently spliced transcripts, potentially binding and inactivating 3′ end cleavage factors (Deng et al., 2020; So et al., 2019). Additionally, the deposition of stimulatory factors—namely the exon junction complex (EJC) and SR proteins—on exon-exon junctions after splicing may stimulate splicing of the next intron (Singh et al., 2012), potentially contributing to the all-spliced phenomenon. An added stimulus to 3′ end cleavage may also be afforded by loss of U1 snRNP and accumulation of SR proteins, since the RS domain in Fip1 promotes cleavage (Zhu et al., 2018); more generally, SR proteins are associated with polyA site choice (Müller-McNicoll et al., 2016). Thus, more efficient splicing in the thalassemia mutant likely enables 3′ end cleavage by more quickly

removing inhibitory activities and/or recruiting positive effectors. Investigation of these mechanisms awaits future studies that would afford single transcript evaluation of the residence time of intron-bound inhibitory factors (*e.g.* U1 snRNP) coupled with splicing and cleavage outcome.

It is tempting to speculate that improving splicing efficiency could be a general strategy for increasing gene output in a variety of disease settings. My findings substantiate an earlier proposal based on experiments on β-globin that efficient splicing and 3′ end cleavage contribute to gene expression output (Lu and Cullen, 2003), by suggesting that nascent transcripts are earmarked as productive or unproductive during their biogenesis. Moreover, failure of 3′ end cleavage when splicing is impaired would explain why splicing efficiency was previously associated with release of RNA from the site of transcription (Antoniou et al., 1998; Custodio et al., 1999; Dye and Proudfoot, 1999). Importantly, many physiological stresses—such as osmotic stress, heat shock, cancer, aging, and viral infection—cause a failure to cleave at annotated PASs and the production of very long non-coding RNAs (Enge et al., 2017; Grosso et al., 2015; Muniz et al., 2017; Vilborg et al., 2015, 2017). This strong connection between splicing efficiency, 3′ end formation, and transcription termination introduces previously unknown layers of regulation to mammalian gene expression in a variety of physiological contexts.

Chapter 5

Outlook

5.1 Limitations

Several limitations to this study remain to be investigated further. First, the length of long-reads are dependent on reverse transcriptase processivity when copying RNA into cDNA. While I have taken steps to enrich for full-length transcripts in my library generation, some RNAs are likely not fully reverse transcribed and captured in this dataset. Advancements in strand-switching RT enzyme chemistry may improve this in the future (Guo et al., 2020). The length of reads in this study also precludes analysis of complex alternative splicing events, such as cassette exon usage, simply because it is difficult to capture longer reads. Second, I have not addressed directly what the ultimate fate of unspliced and uncleaved nascent RNA is in these cells. While in other studies, the Neugebauer lab found these transcripts were degraded by the nuclear exosome (Herzel et al., 2018), it remains to be tested directly. Finally, a more rigorous test of my proposed mechanism linking splicing and 3′ end cleavage would require tools to probe inhibition of both processes. While chemical inhibitors can be used to block spliceosome assembly globally, these drugs also induce a general stress response in cells (Castillo-Guzman et al., 2020), including

changes in transcription and 3′ end cleavage. Thus, future studies probing this mechanism await specific reagents to test directly the link between splicing and cleavage.

5.2 Future Directions

The experiments described in this thesis provide a detailed view of the coordination between transcription and splicing that occurs during mammalian erythropoiesis. They also raise new hypotheses to be tested. While MEL cells are a useful model system, the field of erythropoiesis is increasingly moving toward using more biologically relevant cell types. It would be interesting to observe nascent RNA processing in mammalian hematopoietic stem and progenitor cells (HSPCs) induced to undergo erythroid differentiation. As discussed in Chapter 1, it is known that human and mouse erythroid cells undergo distinct transcriptomic changes during terminal differentiation (An et al., 2014). I note a distinct lack of intron retention in my LRS libraries after induction of erythropoiesis, and this may be in part because intron retention is less pronounced in mouse. Indeed, the characterized examples of intron retention during erythropoiesis have been observed in human cells (Pimentel et al., 2016). Therefore, it would be extremely informative to study co-transcriptional splicing in a human HSPCs. Additionally, we know that erythroid cells undergo specific stresses, such as oxidative stress (Lee et al., 2020). Due to the documented link between stress and readthrough transcription (Castillo-Guzman et al., 2020), HPSCs would provide an ideal system for inducing a biologically relevant stress and monitoring the effects on nascent RNA 3′ end formation and splicing.

The intriguing connection between splicing and 3′ end formation discussed in this thesis was uncovered using a β-globin mini gene which contained a thalassemia-

causing point mutation. In fact, even in the endogenous context, the β-globin gene is one of the most striking examples of genes harboring all unspliced transcripts which fails to undergo 3′ end cleavage. Therefore, the β-globin minigene-containing MEL cell lines provide ample opportunity for further exploration of the mechanistic relationship between splicing and cleavage. For example, many naturally-occurring mutations in the β-globin splice sites could be inserted into the minigene, and the effect on 3′-end formation could be monitored by long-read sequencing. Additionally, this could provide a useful model system for identifying *trans*-acting factors which may link the two processes. For example, I speculate that U1 snRNP could play a role in binding and suppressing the usage of spurious splice sites (Berg et al., 2012), but also in suppressing polyA cleavage if it is not efficently removed from the pre-mRNA via splicing. It could be suggested then that modulating levels of U1 snRNP (for example with U1 antisense morpholinos (Kaida et al., 2010)) could affect the efficiency of 3′ end formation. To this end, the level of other splicing accessory proteins, such as SR and hnRNP proteins, could also be modulated to test their involvement in the process. Together, the molecular biology reagents and methods developed in this thesis present opportunities to further dissect the mechanistic link between transcription elongation, splicing, and 3′ end formation in an erythroid system.

Appendix

Table 5.1: Barcode sequences for custom RT primer.

Barcode Name	Sequence	Purification
RT_BC_01	CACATATCAGAGTGCG	PAGE
RT_BC_02	ACACACAGACTGTGAG	PAGE
RT_BC_03	ACACATCTCGTGAGAG	PAGE
RT_BC_04	CACGCACACACGCGCG	PAGE
RT_BC_05	CACTCGACTCTCGCGT	PAGE
RT_BC_06	CATATATATCAGCTGT	PAGE
RT_BC_07	TCTGTATCTCTATGTG	PAGE
RT_BC_08	ACAGTCGAGCGCTGCG	PAGE
RT_BC_09	ACACACGCGAGACAGA	PAGE
RT_BC_10	ACGCGCTATCTCAGAG	PAGE
RT_BC_11	CTATACGTATATCTAT	PAGE
RT_BC_12	ACACTAGATCGCGTGT	PAGE
RT_BC_13	CTCTCGCATACGCGAG	PAGE
RT_BC_14	CTCACTACGCGCGCGT	PAGE
RT_BC_15	CGCATGACACGTGTGT	PAGE
RT_BC_16	CATAGAGAGATAGTAT	PAGE
RT_BC_17	CACACGCGCGCTATAT	PAGE
RT_BC_18	TCACGTGCTCACTGTG	PAGE
RT_BC_19	ACACACTCTATCAGAT	PAGE
RT_BC_20	CACGACACGACGATGT	PAGE
RT_BC_21	CTATACATAGTGATGT	PAGE
RT_BC_22	CACTCACGTGTGATAT	PAGE
RT_BC_23	CAGAGAGATATCTCTG	PAGE
RT_BC_24	CATGTAGAGCAGAGAG	PAGE

Sequences that can be used as a barcode during the reverse transcription step of long-read sequencing library preparation.

Table 5.2: RNA sequencing read counts

Sample	Protocol	Raw	Mapped	Non PolyA	Non 7SK	Unique	Non Intermediates
Uninduced	LRS	1,036,691	583,632	576,607	453,009	429,099	402,316
Induced	LRS	720,882	571,997	559,718	395,793	255,182	233,942
HBB^{WT}	targeted LRS	68,784	70,871	37,642	NA	37,275	33,571
$HBB^{IVS\text{-}110(G>A)}$	targeted LRS	79,833	78,948	32,172	NA	31,885	28,323
Uninduced	PRO-seq	220,070,650	127,803,736	N/A	N/A	N/A	N/A
Induced	PRO-seq	208,414,166	109,366,055	N/A	N/A	N/A	N/A

Read counts representing raw reads, mapped reads, polyA-filtered reads, 7SK-filtered reads, unique reads, and non-splicing intermediate 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 condition, and PRO-seq samples represent combined counts for three biological replicates in each induction condition.

Table 5.3: Oligonucleotide sequences used in this thesis.

Description	Sequence
3′ end DNA adapter	/5rApp/NNNNNCTGTAGGCACCATCAAT/3ddC/
genome-wide RT	AAGCAGTGGTATCAACGCAGAGTACATTGATGGTGCCTACAG
targeted RT barcode 1	AAGCAGTGGTATCAACGCAGAGTACCACATATCAGAGTGCGGATTGATGGTGCCTACAG
targeted RT barcode 2	AAGCAGTGGTATCAACGCAGAGTACACACACAGACTGTGAGGATTGATGGTGCCTACAG
targeted RT barcode 3	AAGCAGTGGTATCAACGCAGAGTACACACATCTCGTGAGAGGATTGATGGTGCCTACAG
targeted RT barcode 4	AAGCAGTGGTATCAACGCAGAGTACCACGCACACACGCGCGGATTGATGGTGCCTACAG
targeted RT barcode 5	AAGCAGTGGTATCAACGCAGAGTACCATATATATCAGCTGTGATTGATGGTGCCTACAG
targeted RT barcode 6	AAGCAGTGGTATCAACGCAGAGTACTCTGTATCTCTATGTGGATTGATGGTGCCTACAG
targeted PCR human *HBB*	GACGTGTGCTCTTCCGATCTCACGACACGACGATGTcaactgtgttcactagcaacct
qPCR F *Hbb-b1*	ATGCCAAAGTGAAGGCCCAT
qPCR R *Hbb-b1*	CCCAGGAGCCTGAAGTTCTC
qPCR F *Gapdh*	AATGTGTCCGTCGTGGATCTGA
qPCR R *Gapdh*	GATGCCTGCTTCACCACCTTCT
RT-PCR F *Brd2*	GATTATCACAAAATTATAAAACAGCC
RT-PCR R *Brd2*	CTGCTAACTTGGCCCCC
RT-PCR F *Dnajb1*	CCTTTCCCAAGGAAGGG
RT-PCR R *Dnajb1*	GTTTCTCAGGTGTTTTGGG
RT-PCR F *Riok3*	TGTTGCTGAAGGACCATTC
RT-PCR R *Riok3*	ATTTTCCATTCTTGCTGTGTTC
RT-PCR F *Slc12a6*	GACGTGTGCTCTTCCGATCTGGATAACATCATACCTTTCCTTAGG
RT-PCR R *Slc12a6*	ATGGAAAGAATTGGGGCC
RT-PCR F *Dmtn*	GACGTGTGCTCTTCCGATCTCCACCCATCTACAAACAGAGAG
RT-PCR *Dmtn*	CCACAACGGCCAGCGACG
RT-PCR F *Hnrnpll*	GACGTGTGCTCTTCCGATCTTAAAGTGTTTGACGCGAAAG
RT-PCR R *Hnrnpll*	TCGGGACTCGTATCTGGTA
RT-PCR F *Rbm39*	GACGTGTGCTCTTCCGATCTTGCCTCATAGCATCAAATTAAG
RT-PCR R *Rbm39*	CTCACAGGGCTCTTGTCTT
RT-PCR F *Hbq1b*	GACGTGTGCTCTTCCGATCTGGACCCTGCTAACTTCCAG
RT-PCR R *Hbq1b*	TCAGCGATATTTGGAGACC
RT-PCR F *C1qbp*	GACGTGTGCTCTTCCGATCTCACAGATTCCCTGGACTGG
RT-PCR R *C1qbp*	CTACTGGTTCTTGACAAAGCTTT
VRA3 3′ end adapter	/5Phos/rGrArUrCrGrUrCrGrGrArCrUrGrUrArGrArArCrUrCrUrGrArArC/3InvdT/
VRA5 5′ end adapter	rCrCrUrUrGrGrCrArCrCrCrGrArGrArArUrUrCrCrA
RP1 primer	AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA

/5rApp/ = activated adenylate; /3ddC/ = dideoxycytosine; /5Phos/ = phosphorylated; /3InvdT/ = inverted base; rN = ribonucleotide base

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

Table 5.4: Key Resources

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
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		

continued on next page...

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Raw and analyzed data	This thesis	GEO: GSE144205
Raw image data	This thesis	http://dx.doi.org/10.17632/5vrtbpnj4k.1
nanoCOP data from BL1184, K562, and S2 cells	(Drexler et al., 2020)	GEO: GSE123191
total RNA-seq from MEL cells	Mouse ENCODE (Stamatoyannopoulos et al., 2012)	ENCSR000CWE
Code for LRS data analysis	This thesis	https://github.com/NeugebauerLab/MEL_LRS
Experimental Models: Cell Lines		
MEL	Shilpa Hattangadi	N/A
MEL-*HBB*^{WT}	(Patsali et al., 2018)	N/A
MEL-*HBB*$^{IVS\text{-}110(G>A}$	(Patsali et al., 2018)	N/A
Oligonucleotides		
See **Table 5.3**	This thesis	N/A
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
deepTools v3.3.0	(Ramirez et al., 2016)	https://deeptools.readthedocs.io/en/develop/

continued on next page...

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Pysam v0.15.0	N/A	https://github.com/pysam-developers/pysam
mygene	N/A	https://pypi.org/project/mygene/

References

Abdelmoez, M. N., Iida, K., Oguchi, Y., Nishikii, H., Yokokawa, R., Kotera, H., Uemura, S., Santiago, J. G. and Shintaku, H. (2018). SINC-seq: correlation of transient gene expressions between nucleus and cytoplasm reflects single-cell physiology. Genome Biol *19*, 66.

Alexander, R. D., Barrass, J. D., Dichtl, B., Kos, M., Obtulowicz, T., Robert, M. C., Koper, M., Karkusiewicz, I., Mariconti, L., Tollervey, D., Dichtl, B., Kufel, J., Bertrand, E. and Beggs, J. D. (2010a). RiboSys, a high-resolution, quantitative approach to measure the in vivo kinetics of pre-mRNA splicing and 3'-end processing in Saccharomyces cerevisiae. RNA *16*, 2570–80.

Alexander, R. D., Innocente, S. A., Barrass, J. D. and Beggs, J. D. (2010b). Splicing-dependent RNA polymerase pausing in yeast. Mol Cell *40*, 582–93.

Alpert, T., Herzel, L. and Neugebauer, K. M. (2017). Perfect timing: splicing and transcription rates in living cells. Wiley Interdiscip Rev RNA *8*, 10.1002/wrna.1401.

Alpert, T., Straube, K., Carrillo Oesterreich, F. and Neugebauer, K. M. (2020). Widespread Transcriptional Readthrough Caused by Nab2 Depletion Leads to Chimeric Transcripts with Retained Introns. Cell Rep *33*, 108496.

An, X., Schulz, V. P., Li, J., Wu, K., Liu, J., Xue, F., Hu, J., Mohandas, N. and Gallagher, P. G. (2014). Global transcriptome analyses of human and murine terminal erythroid differentiation. Blood *123*, 3466–77.

An, X., Schulz, V. P., Mohandas, N. and Gallagher, P. G. (2015). Human and murine erythropoiesis. Curr Opin Hematol *22*, 206–11.

Antoniou, M. (1991). Induction of Erythroid-Specific Expression in Murine Erythroleukemia (MEL) Cell Lines. Methods Mol Biol *7*, 421–34.

Antoniou, M., Geraghty, F., Hurst, J. and Grosveld, F. (1998). Efficient 3'-end formation of human beta-globin mRNA in vivo requires sequences within the last intron but occurs independently of the splicing reaction. Nucleic Acids Res *26*, 721–9.

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. Nat Rev Mol Cell Biol *18*, 437–451.

Bentley, D. L. (2014). Coupling mRNA processing with transcription in time and space. Nat Rev Genet *15*, 163–75.

Berg, M. G., Singh, L. N., Younis, I., Liu, Q., Pinto, A. M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L. and Dreyfuss, G. (2012). U1 snRNP determines mRNA length and regulates isoform expression. Cell *150*, 53–64.

Bhatt, D. M., Pandya-Jones, A., Tong, A. J., Barozzi, I., Lissner, M. M., Natoli, G., Black, D. L. and Smale, S. T. (2012). Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. Cell *150*, 279–90.

Boothby, T. C., Zipper, R. S., van der Weele, C. M. and Wolniak, S. M. (2013). Removal of retained introns regulates translation in the rapidly developing gametophyte of Marsilea vestita. Dev Cell *24*, 517–29.

Boutz, P. L., Bhutkar, A. and Sharp, P. A. (2015). Detained introns are a novel, widespread class of post-transcriptionally spliced introns. Genes Dev *29*, 63–80.

Braberg, H., Jin, H., Moehle, E. A., Chan, Y. A., Wang, S., Shales, M., Benschop, J. J., Morris, J. H., Qiu, C., Hu, F., Tang, L. K., Fraser, J. S., Holstege, F. C., Hieter, P., Guthrie, C., Kaplan, C. D. and Krogan, N. J. (2013). From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. Cell *154*, 775–88.

Braslavsky, I., Hebert, B., Kartalov, E. and Quake, S. R. (2003). Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci U S A *100*, 3960–4.

Braunschweig, U., Barbosa-Morais, N. L., Pan, Q., Nachman, E. N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M. and Blencowe, B. J. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res *24*, 1774–86.

Burke, J. E., Longhurst, A. D., Merkurjev, D., Sales-Lee, J., Rao, B., Moresco, J. J., Yates, J. R., r., Li, J. J. and Madhani, H. D. (2018). Spliceosome Profiling Visualizes Operations of a Dynamic RNP at Nucleotide Resolution. Cell *173*, 1014–1030 e17.

Carrillo Oesterreich, F., Bieberstein, N. and Neugebauer, K. M. (2011). Pause locally, splice globally. Trends Cell Biol *21*, 328–35.

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*, 571–81.

Carrocci, T. J. and Neugebauer, K. M. (2019). Pre-mRNA Splicing in the Nuclear Landscape. Cold Spring Harb Symp Quant Biol *84*, 11–20.

Castillo-Guzman, D., Hartono, S. R., Sanz, L. A. and Chédin, F. (2020). SF3B1-targeted Splicing Inhibition Triggers Global Alterations in Transcriptional Dynamics and R-Loop Metabolism. 10.1101/2020.06.08.130583. bioRxiv.

Chang, J. C., Temple, G. F., Trecartin, R. F. and Kan, Y. W. (1979). Suppression of the nonsense mutation in homozygous beta 0 thalassaemia. Nature *281*, 602–3.

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*, 779–90.

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 e13.

Cheng, A. W., Shi, J., Wong, P., Luo, K. L., Trepman, P., Wang, E. T., Choi, H., Burge, C. B. and Lodish, H. F. (2014). Muscleblind-like 1 (Mbnl1) regulates pre-mRNA alternative splicing during terminal erythropoiesis. Blood *124*, 598–610.

Churchman, L. S. and Weissman, J. S. (2011). Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature *469*, 368–73.

Conboy, J. G. (2017). RNA splicing during terminal erythropoiesis. Curr Opin Hematol *24*, 215–221.

Cooke, C., Hans, H. and Alwine, J. C. (1999). Utilization of splicing elements and polyadenylation signal elements in the coupling of polyadenylation and last-intron removal. Mol Cell Biol *19*, 4971–9.

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

Core, L. J., Waterfall, J. J. and Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–8.

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.

Custodio, N. and Carmo-Fonseca, M. (2016). Co-transcriptional splicing and the CTD code. Crit Rev Biochem Mol Biol *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*, 2855–66.

Davidson, L. and West, S. (2013). Splicing-coupled 3′ end formation requires a terminal splice acceptor site, but not intron excision. Nucleic Acids Res *41*, 7101–14.

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y. and Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res *46*, D794–d801.

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*, 525–32.

Deng, Y., Shi, J., Ran, Y., Xiang, A. P. and Yao, C. (2020). A potential mechanism underlying U1 snRNP inhibition of the cleavage step of mRNA 3′ processing. Biochem Biophys Res Commun *530*, 196–202.

Denis, M. M., Tolley, N. D., Bunting, M., Schwertz, H., Jiang, H., Lindemann, S., Yost, C. C., Rubner, F. J., Albertine, K. H., Swoboda, K. J., Fratto, C. M., Tolley, E., Kraiss, L. W., McIntyre, T. M., Zimmerman, G. A. and Weyrich, A. S. (2005). Escaping the nuclear confines: signal-dependent pre-mRNA splicing in anucleate platelets. Cell *122*, 379–91.

Deveson, I. W., Brunck, M. E., Blackburn, J., Tseng, E., Hon, T., Clark, T. A., Clark, M. B., Crawford, J., Dinger, M. E., Nielsen, L. K., Mattick, J. S. and Mercer, T. R. (2018). Universal Alternative Splicing of Noncoding Exons. Cell Syst *6*, 245–255.e5.

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. (2020). Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. Mol Cell *77*, 985–998 e8.

Duffy, E. E. and Simon, M. D. (2016). Enriching s(4) U-RNA Using Methane Thiosulfonate (MTS) Chemistry. Curr Protoc Chem Biol *8*, 234–250.

Dvinge, H., Kim, E., Abdel-Wahab, O. and Bradley, R. K. (2016). RNA splicing factors as oncoproteins and tumour suppressors. Nat Rev Cancer *16*, 413–30.

Dye, M. J. and Proudfoot, N. J. (1999). Terminal exon definition occurs cotranscriptionally and promotes termination of RNA polymerase II. Mol Cell *3*, 371–8.

Edwards, C. R., Ritchie, W., Wong, J. J., Schmitz, U., Middleton, R., An, X., Mohandas, N., Rasko, J. E. and Blobel, G. A. (2016). A dynamic intron retention program in the mammalian megakaryocyte and erythrocyte lineages. Blood *127*, e24–e34.

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.e14.

Eser, P., Wachutka, L., Maier, K. C., Demel, C., Boroni, M., Iyer, S., Cramer, P. and Gagneur, J. (2016). Determinants of RNA metabolism in the Schizosaccharomyces pombe genome. Mol Syst Biol *12*, 857.

Faustino, N. A. and Cooper, T. A. (2003). Pre-mRNA splicing and human disease. Genes Dev *17*, 419–37.

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.

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 Dev *28*, 2663–76.

Gallagher, S. R. (2012). One-dimensional SDS gel electrophoresis of proteins. Curr Protoc Mol Biol *Chapter 10*, Unit 10.2A.

Garibaldi, A., Carranza, F. and Hertel, K. J. (2017). Isolation of Newly Transcribed RNA Using the Metabolic Label 4-Thiouridine. Methods Mol Biol *1648*, 169–176.

Goode, D. K., Obier, N., Vijayabaskar, M. S., Lie, A. L. M., Lilly, A. J., Hannah, R., Lichtinger, M., Batta, K., Florkowska, M., Patel, R., Challinor, M., Wallace, K., Gilmour, J., Assi, S. A., Cauchy, P., Hoogenkamp, M., Westhead, D. R., Lacaud, G., Kouskoff, V., Gottgens, B. and Bonifer, C. (2016). Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. Dev Cell *36*, 572–87.

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*, e09214.

Grosveld, F., van Assendelft, G. B., Greaves, D. R. and Kollias, G. (1987). Position-independent, high-level expression of the human beta-globin gene in transgenic mice. Cell *51*, 975–85.

Guo, L. T., Adams, R. L., Wan, H., Huston, N. C., Potapova, O., Olson, S., Gallardo, C. M., Graveley, B. R., Torbett, B. E. and Pyle, A. M. (2020). Sequencing and Structure Probing of Long RNAs Using MarathonRT: A Next-Generation Reverse Transcriptase. J Mol Biol *432*, 3338–3352.

Hahn, C. N., Venugopal, P., Scott, H. S. and Hiwase, D. K. (2015). Splice factor mutations and alternative splicing as drivers of hematopoietic malignancy. Immunol Rev *263*, 257–78.

Hardwick, S. A., Bassett, S. D., Kaczorowski, D., Blackburn, J., Barton, K., Bartonicek, N., Carswell, S. L., Tilgner, H. U., Loy, C., Halliday, G., Mercer, T. R., Smith, M. A. and Mattick, J. S. (2019). Targeted, High-Resolution RNA Sequencing of Non-coding Genomic Regions Associated With Neuropsychiatric Functions. Front Genet *10*, 309.

Herzel, L., Ottoz, D. S. M., Alpert, T. and Neugebauer, K. M. (2017). Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. Nat Rev Mol Cell Biol *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.

Hou, V. C., Lersch, R., Gee, S. L., Ponthier, J. L., Lo, A. J., Wu, M., Turck, C. W., Koury, M., Krainer, A. R., Mayeda, A. and Conboy, J. G. (2002). Decrease in hnRNP A/B expression during erythropoiesis mediates a pre-mRNA splicing switch. EMBO J *21*, 6195–204.

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. Mol Cells *40*, 1–9.

Jonkers, I., Kwak, H. and Lis, J. T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife *3*, e02407.

Jonkers, I. and Lis, J. T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. Nat Rev Mol Cell Biol *16*, 167–77.

Joshi, P., Halene, S. and Abdel-Wahab, O. (2017). How do messenger RNA splicing alterations drive myelodysplasia? Blood *129*, 2465–2470.

Kaida, D., Berg, M. G., Younis, I., Kasim, M., Singh, L. N., Wan, L. and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. Nature *468*, 664–8.

Ke, S., Pandya-Jones, A., Saito, Y., Fak, J. J., Vagbo, C. B., Geula, S., Hanna, J. H., Black, D. L., Darnell, J. E., J. and Darnell, R. B. (2017). m(6)A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. Genes Dev *31*, 990–1006.

Khodor, Y. L., Rodriguez, J., Abruzzi, K. C., Tang, C. H., Marr II, M. T. and Rosbash, M. (2011). Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. Genes Dev *25*, 2502–12.

Kim, Y. J. and Abdel-Wahab, O. (2017). Therapeutic targeting of RNA splicing in myelodysplasia. Semin Hematol *54*, 167–173.

Kinniburgh, A. J. and Ross, J. (1979). Processing of the mouse beta-globin mRNA precursor: at least two cleavage-ligation reactions are necessary to excise the larger intervening sequence. Cell *17*, 915–21.

Konkel, D. A., Tilghman, S. M. and Leder, P. (1978). The sequence of the chromosomal mouse beta-globin major gene: homologies in capping, splicing and poly(A) sites. Cell *15*, 1125–32.

Krivega, I. and Dean, A. (2016). Chromatin looping as a target for altering erythroid gene expression. Ann N Y Acad Sci *1368*, 31–9.

Kumar, A., Clerici, M., Muckenfuss, L. M., Passmore, L. A. and Jinek, M. (2019). Mechanistic insights into mRNA 3'-end processing. Curr Opin Struct Biol *59*, 143–150.

Kurosaki, T., Popp, M. W. and Maquat, L. E. (2019). Quality and quantity control of gene expression by nonsense-mediated mRNA decay. Nat Rev Mol Cell Biol *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*, 950–3.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M. and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol *17*, 909–15.

Lagarde, J., Uszczynska-Ratajczak, B., Carbonell, S., Pérez-Lluch, S., Abad, A., Davis, C., Gingeras, T. R., Frankish, A., Harrow, J., Guigo, R. and Johnson, R. (2017). High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. Nat Genet *49*, 1731–1740.

Lai, C. J., Dhar, R. and Khoury, G. (1978). Mapping the spliced and unspliced late lytic SV40 RNAs. Cell *14*, 971–82.

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.

Lee, P., Chandel, N. S. and Simon, M. C. (2020). Cellular adaptation to hypoxia through hypoxia inducible factors and beyond. Nat Rev Mol Cell Biol *21*, 268–283.

Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L. and Steitz, J. A. (1980). Are snRNPs involved in splicing? Nature *283*, 220–224.

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*, 2078–9.

Lin, C. L., Taggart, A. J. and Fairbrother, W. G. (2016). RNA structure in splicing: An evolutionary perspective. RNA Biol *13*, 766–71.

Liu, H., Begik, O., Lucas, M. C., Ramirez, J. M., Mason, C. E., Wiener, D., Schwartz, S., Mattick, J. S., Smith, M. A. and Novoa, E. M. (2019). Accurate detection of m(6)A RNA modifications in native RNA sequences. Nat Commun *10*, 4079.

Lu, S. and Cullen, B. R. (2003). Analysis of the stimulatory effect of splicing on mRNA production and utilization in mammalian cells. RNA *9*, 618–30.

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*, 1455–76.

Manning, K. S. and Cooper, T. A. (2017). The roles of RNA processing in translating genotype to phenotype. Nat Rev Mol Cell Biol *18*, 102–114.

Maquat, L. E., Kinniburgh, A. J., Beach, L. R., Honig, G. R., Lazerson, J., Ershler, W. B. and Ross, J. (1980). Processing of human beta-globin mRNA precursor to mRNA is defective in three patients with beta+-thalassemia. Proc Natl Acad Sci U S A *77*, 4287–91.

Maquat, L. E., Kinniburgh, A. J., Rachmilewitz, E. A. and Ross, J. (1981). Unstable beta-globin mRNA in mRNA-deficient beta o thalassemia. Cell *27*, 543–53.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet *17*, 10.14806/ej.17.1.200.

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*, 1144–55.

Martins, S. B., Rino, J., Carvalho, T., Carvalho, C., Yoshida, M., Klose, J. M., de Almeida, S. F. and Carmo-Fonseca, M. (2011). Spliceosome assembly is coupled to RNA polymerase II dynamics at the 3′ end of human genes. Nat Struct Mol Biol *18*, 1115–23.

Mauger, O., Lemoine, F. and Scheiffele, P. (2016). Targeted Intron Retention and Excision for Rapid Gene Regulation in Response to Neuronal Activity. Neuron *92*, 1266–1278.

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.

Meola, N., Domanski, M., Karadoulama, E., Chen, Y., Gentil, C., Pultz, D., Vitting-Seerup, K., Lykke-Andersen, S., Andersen, J. S., Sandelin, A. and Jensen, T. H. (2016). Identification of a Nuclear Exosome Decay Pathway for Processed Transcripts. Mol Cell *64*, 520–533.

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*, 10547–52.

Midha, M. K., Wu, M. and Chiu, K. P. (2019). Long-read sequencing in deciphering human genetics to a greater depth. Hum Genet *138*, 1201–1215.

Miller, O. L., J. and Beatty, B. R. (1969). Visualization of nucleolar genes. Science *164*, 955–7.

Milligan, L., Sayou, C., Tuck, A., Auchynnikava, T., Reid, J. E., Alexander, R., Alves, F. L., Allshire, R., Spanos, C., Rappsilber, J., Beggs, J. D., Kudla, G. and Tollervey, D. (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.

Müller-McNicoll, M., Botti, V., de Jesus Domingues, A. M., Brandl, H., Schwich, O. D., Steiner, M. C., Curk, T., Poser, I., Zarnack, K. and Neugebauer, K. M. (2016). SR proteins are NXF1 adaptors that link alternative RNA processing to mRNA export. Genes Dev *30*, 553–66.

Naro, C., Jolly, A., Di Persio, S., Bielli, P., Setterblad, N., Alberdi, A. J., Vicini, E., Geremia, R., De la Grange, P. and Sette, C. (2017). An Orchestrated Intron Retention Program in Meiosis Controls Timely Usage of Transcripts during Germ Cell Differentiation. Dev Cell *41*, 82–93 e4.

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

Ni, D., Xu, P. and Gallagher, S. (2017). Immunoblotting and Immunodetection. Curr Protoc Protein Sci *88*, 10.10.1–10.10.37.

Ni, J. Z., Grate, L., Donohue, J. P., Preston, C., Nobida, N., O'Brien, G., Shiue, L., Clark, T. A., Blume, J. E. and Ares, M., J. (2007). Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. Genes Dev *21*, 708–18.

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 e4.

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

Pai, A. A., Paggi, J. M., Yan, P., Adelman, K. and Burge, C. B. (2018). Numerous recursive sites contribute to accuracy of splicing in long introns in flies. PLoS Genet *14*, e1007588.

Pandya-Jones, A. and Black, D. L. (2009). Co-transcriptional splicing of constitutive and alternative exons. RNA *15*, 1896–908.

Papasaikas, P. and Valcarcel, J. (2016). The Spliceosome: The Ultimate RNA Chaperone and Sculptor. Trends Biochem Sci *41*, 33–45.

Parker, M. T., Knop, K., Sherwood, A. V., Schurch, N. J., Mackinnon, K., Gould, P. D., Hall, A. J., Barton, G. J. and Simpson, G. G. (2020). Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m(6)A modification. Elife *9*.

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.

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.

Payne, A., Holmes, N., Rakyan, V. and Loose, M. (2019). BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. Bioinformatics *35*, 2193–2198.

Pimentel, H., Parra, M., Gee, S., Ghanem, D., An, X., Li, J., Mohandas, N., Pachter, L. and Conboy, J. G. (2014). A dynamic alternative splicing program regulates gene expression during terminal erythropoiesis. Nucleic Acids Res *42*, 4031–42.

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*, 838–51.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–2.

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*, W160–5.

Rasmussen, E. B. and Lis, J. T. (1993). In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. Proc Natl Acad Sci U S A *90*, 7923–7.

Reimer, K. A., Mimoso, C., Adelman, K. and Neugebauer, K. M. (2021). Co-transcriptional splicing regulates 3' end cleavage during mammalian erythropoiesis. Mol Cell *81*, 10.1016/j.molcel.2020.12.018.

Reimer, K. A. and Neugebauer, K. M. (2018). Blood Relatives: Splicing Mechanisms underlying Erythropoiesis in Health and Disease [version 1; peer review: 3 approved]. F1000Res *7*, 1364.

Reimer, K. A. and Neugebauer, K. M. (2020). Preparation of Mammalian Nascent RNA for Long Read Sequencing. Curr Protoc Mol Biol *133*, e128.

Saldi, T., Fong, N. and Bentley, D. L. (2018). Transcription elongation rate affects nascent histone pre-mRNA folding and 3' end processing. Genes Dev *32*, 297–308.

Schibler, U., Marcu, K. B. and Perry, R. P. (1978). The synthesis and processing of the messenger RNAs specifying heavy and light chain immunoglobulins in MPC-11 cells. Cell *15*, 1495–509.

Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–4.

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*, 2264–74.

Scotti, M. M. and Swanson, M. S. (2016). RNA mis-splicing in disease. Nat Rev Genet *17*, 19–32.

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 e4.

Sibley, C. R., Emmett, W., Blazquez, L., Faro, A., Haberman, N., Briese, M., Trabzuni, D., Ryten, M., Weale, M. E., Hardy, J., Modic, M., Curk, T., Wilson, S. W., Plagnol, V. and Ule, J. (2015). Recursive splicing in long vertebrate genes. Nature *521*, 371–375.

Singh, G., Kucukural, A., Cenik, C., Leszyk, J. D., Shaffer, S. A., Weng, Z. and Moore, M. J. (2012). The Cellular EJC Interactome Reveals Higher-Order mRNP Structure and an EJC-SR Protein Nexus. Cell *151*, 915–916.

Singh, M., Al-Eryani, G., Carswell, S., Ferguson, J. M., Blackburn, J., Barton, K., Roden, D., Luciani, F., Giang Phan, T., Junankar, S., Jackson, K., Goodnow, C. C., Smith, M. A. and Swarbrick, A. (2019). High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat Commun *10*, 3120.

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*, 657–66.

So, B. R., Di, C., Cai, Z., Venters, C. C., Guo, J., Oh, J. M., Arai, C. and Dreyfuss, G. (2019). A Complex of U1 snRNP with Cleavage and Polyadenylation Factors Controls Telescripting, Regulating mRNA Transcription in Human Cells. Mol Cell *76*, 590–599.e4.

Sohn, J. I. and Nam, J. W. (2018). The present and future of de novo whole-genome assembly. Brief Bioinform *19*, 23–40.

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*, 2455–9.

Stamatoyannopoulos, J. A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D. M., Groudine, M., Bender, M., Kaul, R., Canfield, T., Giste, E., Johnson, A., Zhang, M., Balasundaram, G., Byron, R., Roach, V., Sabo, P. J., Sandstrom, R., Stehling, A. S., Thurman, R. E., Weissman, S. M., Cayting, P., Hariharan, M., Lian, J., Cheng, Y., Landt, S. G., Ma, Z., Wold, B. J., Dekker, J., Crawford, G. E., Keller, C. A., Wu, W., Morrissey, C., Kumar, S. A., Mishra, T., Jain, D., Byrska-Bishop, M., Blankenberg, D., Lajoie, B. R., Jain, G., Sanyal, A., Chen, K.-B., Denas, O., Taylor, J., Blobel, G. A., Weiss, M. J., Pimkin, M., Deng, W., Marinov, G. K., Williams, B. A., Fisher-Aylor, K. I., Desalvo, G., Kiralusha, A., Trout, D., Amrhein, H., Mortazavi, A., Edsall, L., McCleary, D., Kuan, S., Shen, Y., Yue, F., Ye, Z., Davis, C. A., Zaleski, C., Jha, S., Xue, C., Dobin, A., Lin, W., Fastuca, M., Wang, H., Guigo, R., Djebali, S., Lagarde, J., Ryba, T., Sasaki, T., Malladi, V. S., Cline, M. S., Kirkup, V. M., Learned, K., Rosenbloom, K. R., Kent, W. J., Feingold, E. A., Good, P. J., Pazin, M., Lowdon, R. F., Adams, L. B. and Mouse, E. C. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biology *13*, 418.

Tang, A. D., Soulette, C. M., van Baren, M. J., Hart, K., Hrabeta-Robinson, E., Wu, C. J. and Brooks, A. N. (2020). Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun *11*, 1438.

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*, 16655–62.

Thein, S. L. (2013). The molecular basis of beta-thalassemia. Cold Spring Harb Perspect Med *3*, a011700.

Tilgner, H., Jahanbani, F., Blauwkamp, T., Moshrefi, A., Jaeger, E., Chen, F., Harel, I., Bustamante, C. D., Rasmussen, M. and Snyder, M. P. (2015). Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. Nat Biotechnol *33*, 736–42.

Tilgner, H., Jahanbani, F., Gupta, I., Collier, P., Wei, E., Rasmussen, M. and Snyder, M. (2018). Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. Genome Res *28*, 231–242.

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*, 1616–25.

Urbanski, L. M., Leclair, N. and Anczukow, O. (2018). Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. Wiley Interdiscip Rev RNA *9*, e1476.

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*, 7399–405.

Vagner, S., Rüegsegger, U., Gunderson, S. I., Keller, W. and Mattaj, I. W. (2000). Position-dependent inhibition of the cleavage step of pre-mRNA 3'-end processing by U1 snRNP. RNA *6*, 178–88.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D. and Thermes, C. (2018). The Third Revolution in Sequencing Technology. Trends Genet *34*, 666–681.

Veloso, A., Kirkconnell, K. S., Magnuson, B., Biewen, B., Paulsen, M. T., Wilson, T. E. and Ljungman, M. (2014). Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. Genome Res *24*, 896–905.

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*, 449–61.

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.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M.,

Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K. et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–62.

Wilkinson, M. E., Charenton, C. and Nagai, K. (2019). RNA Splicing by the Spliceosome. Annu Rev Biochem *89*, 359–388.

Wong, A. C. H., Rasko, J. E. J. and Wong, J. J. (2018). We skip to work: alternative splicing in normal and malignant myelopoiesis. Leukemia *32*, 1081–1093.

Wong, J. J., Ritchie, W., Ebner, O. A., Selbach, M., Wong, J. W., Huang, Y., Gao, D., Pinello, N., Gonzalez, M., Baidya, K., Thoeng, A., Khoo, T. L., Bailey, C. G., Holst, J. and Rasko, J. E. (2013). Orchestrated intron retention regulates normal granulocyte differentiation. Cell *154*, 583–95.

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., Zuzarte, P. C., Gilpatrick, T., Payne, A., Quick, J., Sadowski, N., Holmes, N., de Jesus, J. G., Jones, K. L., Soulette, C. M., Snutch, T. P., Loman, N., Paten, B., Loose, M., Simpson, J. T., Olsen, H. E., Brooks, A. N., Akeson, M. and Timp, W. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat Methods *16*, 1297–1305.

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*, 7219–25.

Wulf, M. G., Maguire, S., Humbert, P., Dai, N., Bei, Y., Nichols, N. M., Corrêa, I. R., J. and Guan, S. (2019). Non-templated addition and template switching by Moloney murine leukemia virus (MMLV)-based reverse transcriptases co-occur and compete with each other. J Biol Chem *294*, 18220–18231.

Yamamoto, M. L., Clark, T. A., Gee, S. L., Kang, J. A., Schweitzer, A. C., Wickrema, A. and Conboy, J. G. (2009). Alternative pre-mRNA splicing switches modulate gene expression in late erythropoiesis. Blood *113*, 3363–70.

Yap, K., Lim, Z. Q., Khandelia, P., Friedman, B. and Makeyev, E. V. (2012). Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention. Genes Dev *26*, 1209–23.

Yeo, G. and Burge, C. B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol *11*, 377–94.

Yu, M. and Ren, B. (2017). The Three-Dimensional Organization of Mammalian Genomes. Annu Rev Cell Dev Biol *33*, 265–289.

Zhou, Y., Zhu, J., Schermann, G., Ohle, C., Bendrin, K., Sugioka-Sugiyama, R., Sugiyama, T. and Fischer, T. (2015). The fission yeast MTREC complex targets CUTs and unspliced pre-mRNAs to the nuclear exosome. Nat Commun *6*, 7050.

Zhu, Y., Wang, X., Forouzmand, E., Jeong, J., Qiao, F., Sowd, G. A., Engelman, A. N., Xie, X., Hertel, K. J. and Shi, Y. (2018). Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation. Mol Cell *69*, 62–74.e4.

Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. and Siebert, P. D. (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. Biotechniques *30*, 892–7.