U1 SNRNP TELESCRIPTING: A TRANSCRIPTIONAL REGULATION MECHANISM THAT HAS DRIVEN INTRON SIZE EXPANSION ACROSS EVOLUTION

Christopher Conrad Venters

A DISSERTATION

in

Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

_____

Gideon Dreyfuss, Ph.D.
Isaac Norris Professor, Biochemistry & Biophysics, HHMI investigator

Graduate Group Chairperson

_____

Dejian Ren, Ph.D.
Professor of Biology

Dissertation Committee

Brian D. Gregory, Ph.D. (Assoc. Prof. of Biology)

Junhyong Kim, Ph.D. (Patricia M. Williams Term Prof. and Chair of Biology)

Scott Poethig, Ph.D. (John H. and Margaret B. Fassitt Prof. of Biology)

Stephen A. Liebhaber, M.D (Prof. of Genetics)

U1 SNRNP TELESCRIPTING: A TRANSCRIPTIONAL REGULATION MECHANISM THAT HAS DRIVEN INTRON SIZE EXPANSION ACROSS EVOLUTION

I dedicate this work to my family. To my mother, father, and sister, thank you for nurturing me into the scientist and person that I have become today. This would not have been possible without your constant support and love; I hope that this work reflects the effort and motivation that I have learned from all of you over the years. To my wife Lauren, I cannot express enough how important you have been to the completion of this project. Your positivity and stability have kept me sane and happy. I thank and love all of you.

ACKNOWLEDGMENTS

I would like to thank all the members of the Dreyfuss lab, both past and present, for helping me and being supportive of my growth over the years. This thesis would not be possible without your contributions. Gideon has taught me much about scientific scholarship and how to synthesize multiple, and often very disparate, pieces of information into exciting new conclusions that drive research forward. He has lead from the fore by challenging convention, presenting thoughtful and rigorous conclusions, and, most of all, not missing the forest for the trees. I could not imagine the Dreyfuss lab without all of the scientists, colleagues, and friends that I have known along the way. Ihab Younis, Pilong Li, and Mike Berg were very helpful post-doctoral mentors from whom I learned many laboratory techniques and critical thinking skills. I was very happy to also be able to work with Jung-Min Oh, Chie Arai, Ranny So, Chao Di, Jingqi Duan, and Zhiqiang Cai on experiments and in writing manuscripts; their diverse research backgrounds helped me to become a more rounded scientist in all regards. I enjoyed getting to know the undergraduate researchers that passed through the lab as well, especially Ethan Fein, Aris Mourelatos, Shawn Foley, and Rediet Mersha. Lastly from the lab, but far from least, I would like to acknowledge Maura Jones and Eric Babiash. They have both been not only wonderful colleagues, but also true friends and life coaches. Whether helping on my projects or sharing a laugh in the break room, Maura and Eric have made this lab feel like a second home. Spending eight years in Philadelphia has also allowed me to meet amazing people along the way. I would especially like to thank Tommy and Krystal Ferguson, Zach Corse, Alex Bennett, Travis Conrad, Caitlin Ferguson, Dan Browne, John Coggins, Jenn Abrams, and Alli Blansfield who have all helped me enjoy my life in the city, around campus, or playing games at summer league. The strong foundation, support, and love given to me

by my parents, Ron Venters and Jennifer Conrad, and my sister, Katie Venters, has also been influential in my time here. I could not imagine completing this process without you. Thank you also to my wife, Lauren DeRuyter. You have done much more for me than I could ever put into words.

ABSTRACT

U1 SNRNP TELESCRIPTING: A TRANSCRIPTIONAL REGULATION MECHANISM THAT HAS DRIVEN INTRON SIZE EXPANSION ACROSS EVOLUTION

Christopher Conrad Venters

Gideon Dreyfuss

U1 snRNP (U1) functions in controlling transcription through both the splicing of introns and the suppression of premature cleavage and polyadenylation. The latter, termed telescripting, is the critical process that allows for the creation of full-length pre-mRNA. Reducing the level of available U1 in cells relative to pre-mRNA, either through global transcription up-regulation or functional U1 inhibition, causes widespread premature cleavage and polyadenylation (PCPA) from cryptic polyadenylation signals in introns. Through the development of novel analytical programs for large sequencing datasets, I identified the gene features that increase sensitivity to U1 inhibition and examined the size-function polarization of genomes that has occurred through intron expansion as a result of telescripting. By conducting a series of high-throughput sequencing experiments for both chromatin immunoprecipitation (ChIPseq) and RNA (RNAseq), I demonstrated that PCPA is co-transcriptional and that it is an evolutionarily conserved transcription regulation mechanism. I have shown that U1 inhibition caused PCPA selectively in large genes (median 39 kb in human), which were enriched for developmental and differentiation functions, while small genes were enriched for acute cell survival and stress response function are up-regulated under the same conditions. Importantly, I proved that PCPA susceptibility is evolutionarily conserved and has been a major driving factor in vertebrate intron size expansion. I demonstrated that this gene size-function polarization allows for large genes to sacrifice transcription during cell stress and activation, supplying small genes necessary for cell survival with RNA processing proteins that boost their

mRNA productivity. Lastly, I have shown that minor U1 inhibition caused general and previously unannotated multi-exon skipping events that can sometimes occur between two, adjacent and same-stranded genes as a part of read through transcription including in disease relevant genes. These results highlight the importance of U1, not only as a mechanism of transcription regulation in all metazoans, but also as a primary contributor to the evolution of gene and genome structure.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1: INTRODUCTION

RNA Transcription

The cellular function most crucial to life is the conversion of genetic information into molecular machinery. Aptly described as the central dogma of molecular biology, permanently stored sequence information, in the form of deoxyribonucleic acid (DNA), is transformed into functional cellular machinery, mainly in the form of protein. At the heart of this transfer is a molecule called ribonucleic acid (RNA), which is generated from copying the genetic information stored in DNA through a process known as transcription. The creation of protein coding messenger RNA (mRNA) from transcriptional units, called genes, is carried out by a large, 12-subunit complex called RNA polymerase II (pol II) (Cramer, Bushnell, and Kornberg 2001; Cramer et al. 2008). There are three definitive stages to transcription: initiation, where a start site is selected, the DNA is unwound, and pol II forms the beginning of the RNA strand; productive elongation of RNA as pol II moves through the gene from 5' to 3'; and finally, termination as pol II is released from the DNA (Figure 1.1) (Holstege, Fiedler, and Timmers 1997).

The main subunit of pol II, Rbp1, has a C-terminal domain (CTD) that is unique to pol II and is critical for regulating the various stages of transcription through phosphorylation of various peptide residues (Shilatifard, Conaway, and Conaway 2003; Bowman and Kelly 2014). This CTD is made up of multiple heptad repeats, the consensus sequence being Tyr1–Ser2–Pro3–Thr4–Ser5–Pro6–Ser7. Near to the transcription start site (TSS), Ser5 is heavily phosphorylated (Ser5p), but this is slowly replaced by Ser2p over the length of the gene during elongation (Komarnitsky, Cho, and Buratowski 2000; Mayer et al. 2010; Odawara et al. 2011). Ser2p addition is also known to help the polymerase overcome

promoter proximal pausing (PPP), a vital transcription checkpoint. These different phosphorylation states help the recruitment of proteins to the CTD that are involved in the various stages of transcription, such as export factors or RNA processing factors (Egloff and Murphy 2008; Heidemann et al. 2013). Much research has been done on the kinetics of transcription, as pol II's variable transcription speed and newly discovered transcriptional checkpoints have been shown to be important regulators in termination site usage and proper co-transcriptional processing of RNA (Darzacq et al. 2007; Wada et al. 2009; David et al. 2011; Jonkers, Kwak, and Lis 2014; Yang et al. 2016).

RNA Splicing and Its Machinery

During the early evolutionary stages of eukaryotes, and still found today in prokaryotes, transcription was a one-to-one process where the entirety of a gene's sequence information is transcribed into utilized RNA. This information will, in turn, be converted by the ribosome into a final peptide strand in a process called translation. More recent evolutionary changes to transcription and, more importantly, RNA post-processing, has allowed for selection of specific sequences to include in the final mRNA from the initially produced RNA strand, pre-mRNA. The first iteration of sequence selection simply involved pieces that were either removed, dubbed introns, or kept as coding sequences, dubbed exons (Chorev and Carmel 2012). The mechanism of removing these intronic pieces of RNA and subsequent ligation of exons to form the final, contiguous mRNA is called splicing.

Modern eukaryotes take advantage of this intron-exon mosaicking for a variety of outcomes, for example to include regulatory sequences or functional RNAs mid-gene, while ultimately not affecting the translated protein product (Chorev and Carmel 2012). Additionally, more complex eukaryotes can choose to include or exclude exons to

generate distinct yet related mRNAs and proteins from a common gene; these exons are known most broadly as alternative exons (Figure 1.2). This is commonly seen with variants of a single gene, called isoforms, that are functionally similar and yet specific to either one tissue or cell type within a complex organism or are necessary only during certain stages of development (E. T. Wang et al. 2008; Baralle and Giudice 2017; Iñiguez and Hernández 2017). The modularity allowed by alternative splicing thus decreases the number of individual genes required to produce the increasingly diverse set of proteins needed for organismal function in higher eukaryotes. As such, over the course of evolution, alternative splicing of transcripts has allowed for an increasingly diverse proteome.

Only after splicing and other processing steps, i.e. 5' end capping and 3' polyadenylation, can the mature mRNA be translated into protein by the ribosome. As this is such a critical process, it should be no surprise that splicing is carried out by a multisubunit, megadalton complex called the spliceosome. This molecular machine is comprised of dozens of proteins and functional RNAs, at the core of which are the U-type small nuclear ribonucleoproteins (snRNPs). These snRNPs were first discovered in the 1960's (Hodnett and Busch 1968; Weinberg and Penman 1968), but, without a physiological foundation in which to place them, it took until 1979 before snRNPs were recognized to be involved in splicing. This discovery was that one family member, U1 snRNP (U1), formed an RNA:RNA base-pair with the 5' splice site (5'ss) of an intron, which is the critical first step in the splicing reaction (Lerner and Steitz 1979; Lerner et al. 1980; Mount et al. 1983; Padgett et al. 1983). The major spliceosome, responsible for processing nearly all introns in eukaryotes, is comprised of U1, U2, U4, U6, while U11, U12, U4atac, and U6atac make up the minor version of this complex seen only in multicellular eukaryotes, with U5 being the only snRNP shared between the two (Hall and Padgett 1994; Patel and Steitz 2003; Wahl, Will, and Lührmann 2009).

The splicing reaction, which is often co-transcriptional, involves many steps of snRNP engagement, rearrangement, and departure (Figure 1.3) (Wahl, Will, and Lührmann 2009; Will and Lührmann 2011). Simply, it begins with U1 snRNA's 5' end recognizing and binding to the 5'ss in order to define the upstream exon, thereby forming the E Complex. U2, aided by the auxiliary splicing factors U2AFs and SF1, then identifies and binds to the branch point adenosine at the 3' end of the intron near the 3' splice site (3'ss). The pre-spliceosomal A Complex is formed by recruitment of additional proteins and the association of U1 and U2 through Prp5 and Sub2. The catalytic component of the spliceosome, the U4/U6.U5 tri-snRNP, is then recruited to form the B Complex as U6 displaces U1 from the intron. The subsequent release of U4 then allows for the C Complex to catalyze the transesterification reaction from the branch site adenosine onto the 5'ss, cleaving the 5' exon junction. A second transesterification ligates the 5' exon onto the 3' exon, resulting in the release of an intron lariat bound to U2, U5, and U6. The snRNPs are then recycled for further rounds of splicing.

The snRNPs themselves are made up of a uridine rich small nuclear RNA (snRNA), ranging in size from 100-300 nucleotides (nt), and a common heptameric ring made of Sm or related Lsm proteins (Sm/Lsm core) (Figure 1.4) (Guthrie and Patterson 1988). U6 is unique in that it is the only snRNP to utilize Lsm proteins that recognize its Lsm site (AU_5) rather than the more common Sm site and proteins. Each snRNP also possesses many accessory proteins associated with specific proteins to aid in their function (Krämer et al. 1995; C. Wang et al. 1998; Pomeranz Krummel et al. 2009). The most well characterized of these accessory proteins are on U1 snRNA, where stem-loop 1 is bound by U1-70K which helps recruit it to the SMN complex (Battle et al. 2006; Lau, Bachorik, and Dreyfuss 2009; So et al. 2016). The remaining two U1 specific proteins may join at any time after this with U1A binding to stem-loop 2, while U1C interacts at the stem near the 3' and 5'

snRNA arms and aids in the critical RNA:RNA base pairing in splicing (Du and Rosbash 2002; Pomeranz Krummel et al. 2009).

Prior to assembly, the Sm proteins congregate into two heterodimers, SmB/D3 and SmD1/D2, and a heterotrimer of SmF/E/G. Experiments *in vitro* showed that they are able to spontaneously assemble into an Sm core around most uracil-rich RNA (Pellizzoni et al. 2002). Given that this protein ring is held tightly together and stable when bound to RNA, even under high salt and detergent conditions, it would be highly deleterious if allowed to run free within a cell as they could form on an off-target RNA and inhibit its canonical function (Yong, Wan, and Dreyfuss 2004). As such, the assembly of the Sm core around the snRNA's Sm site ($AU_{2-4}G$/UUG) is highly regulated by the survival of motor neurons (SMN) complex (U. Fischer, Liu, and Dreyfuss 1997; Liu et al. 1997; Meister et al. 2001; Yong et al. 2004; Cauchi 2010; Utz Fischer, Englbrecht, and Chari 2011; So et al. 2016). The SMN complex is made up of the SMN protein, associated Gemin proteins (2-8), and unrip. Studies into SMN deficiencies, which are known to cause the genetic wasting condition spinal muscular atrophy (SMA), have intriguingly shown a tissue-specific altered snRNP repertoire, rather than uniform decreases across the entire family (Gabanella et al. 2007; Z. Zhang et al. 2008; Workman et al. 2009). This fact, combined with the knowledge that snRNP abundance in cells is highly variable despite their 1:1 stoichiometry in the splicing reaction (Baserga and Steitz 1993), led us to target each snRNP individually for experimental knockdown in order to assess the transcriptomic changes in tissue cultures.

Functional snRNP Knockdown and U1 Telescripting

We directly examined the individual activity of each of these snRNPs through functional knockdown experiments in HeLa cells. We relied on the transfection of antisense

morpholino oligonucleotides (AMOs) to alter the functional snRNP repertoire for our studies rather than small-interfering RNAs (siRNAs) because the former are easier to control, do not destroy the target RNP, and act much more quickly (4-8 hours versus 24-48 hours, respectively (Kaida et al. 2010)). These 25-mer oligos were targeted to a specific sequence of the snRNA both to block their binding to pre-mRNA, and because U1 and U12 AMOs had been shown to inhibit splicing in a few tested introns (Figure 1.4) (Matter and König 2005). U1 AMO, specifically, was of interest due to U1's importance in the first steps of splicing and its high abundance, at roughly 1,000,000 copies in cells; this is several fold more than U4 and U6.

Dose response of the AMO functional knockdowns was tested both with fluorescent *in situ* hybridization and RNase H protection assays (Kaida et al. 2010). The former utilized fluorescent LNA probes complementary to U1's 5' snRNA end, while the latter utilizes RNase H and a DNA probe, similar to the LNA probe, so that bound snRNA gets cleaved and degraded. A decrease in snRNA cleavage was seen after U1 AMO treatments, as it inhibited DNA:RNA base-pairing and RNAseq H cleavage, and allowed us to determine optimal doses for near complete functional inhibition of U1. Initial results with *in vitro* target introns also showed decreased splicing, further confirming these results.

Global testing of these U1 and U2 AMO treatments were initially performed using high density genomic tiling arrays (Figure 1.5) (Kaida et al. 2010). Results from these experiments were puzzling, as RNA reads in many genes were concentrated towards the 5' end, with signal tapering out almost to nothing within a few kilobases (kb) of the transcription start site (TSS) in introns. It should be stated that, despite these novel results, intron retention was apparent in many cases upstream of the point at which the read signal terminated in these introns, confirming U1's role in splicing. In comparison, however, U2

AMO and treatment with spliceostatin A (SSA), which inhibits splicing by suppressing the activity of the critical U2 protein SF3B1, showed global increases in intron retention as seen by increases in splice site and intronic reads (Kaida et al. 2007; Kotake et al. 2007). There was no detectable 5' aggregation of signal with U2 AMO or SSA, indicating that the phenomenon was specific to U1 base-pairing inhibition.

There were multiple possible mechanistic explanations behind the production of these RNAs near the 5' end of genes. For example, the pol II could be paused yet still engaged, similar to those found near the promoter; alternatively, it is possible that full-length mRNAs are actively being degraded from their 3' end, leaving 5' sequences; lastly, pol II transcription could be terminating early within the gene. In order to further study this observation, we turned to other techniques paired with our original AMO knockdown treatments. We targeted a select number of these apparent 3' termination sites with primers to produce complementary DNA (cDNA) for both PCR amplification and sequencing (Scotto-Lavino, Du, and Frohman 2006). This 3' rapid amplification of cDNA ends (3'-RACE) showed that the truncated transcripts were polyadenylated similar to the canonical ends of genes (Kaida et al. 2010). In fact, these stretches of non-genomic polyadenylation (poly(A)) were also found to have consensus polyadenylation signals (PAS; AAUAAA or other variants; (N. J. Proudfoot and Brownlee 1976; Magana-Mora, Kalkatawi, and Bajic 2017)) 20-60nt upstream. These are the signals typically utilized by the cleavage and polyadenylation (CPA) machinery to terminate mRNA transcription at a genes' 3' end (Shi and Manley 2015; Tian and Manley 2017). Utilizing a mutated and inactivated PAS in reporter genes demonstrated that this premature CPA (PCPA) is dependent on the upstream PAS similar to canonical 3' gene ends. PCPA products were still produced from nearby, downstream PAS after mutation, although not near to the

inactivated site. Taken together, these results indicated that this phenomenon is not hinged on a single PAS and is directional from 5' to 3', similar to transcription.

We more precisely identified these PCPA locations using a high throughput sequencing strategy of differentially expressed transcripts (HIDE-seq) (Berg et al. 2012). This protocol utilized subtractive hybridization (Diatchenko et al. 1996; Gurskaya et al. 1996) of cDNA fragments from either control or U1 AMO treated poly(A) transcripts in order to eliminate unchanged sequences (Figure 1.6). PCR amplification of the remaining pool results in a sequencing library only enriched in differentially expressed portions of the genome. HIDE-seq employed on human, mouse, and fruit fly cell lines showed that PCPA was a conserved phenomenon due to U1 base-pairing inhibition across metazoans. Furthermore, we found that PCPA typically occurred within 1kb from the first 5'ss. Assuming that U1 bound to the 5'ss could also function to prevent PCPA, an activity we termed telescripting, this would suggest that a single U1 snRNP could protect roughly 1kb of transcription. This result was confirmed through mutation of several 5'ss, which resulted in PCPA from a PAS ~1kb downstream. Furthermore, the use of a synthetic U1 snRNA, complementary to the now mutated 5'ss, resulted in the restoration of telescripting and, thus, downstream transcription.

There were cases with U1 AMO, however, where transcription extended for 10's of kb before premature termination (Berg et al. 2012). This was suggestive of a mechanism where U1 may bind to a 5'ss for the purpose of splicing and could also suppress a nearby PAS; however, downstream sites may be unprotected by U1. This would be problematic in genes with introns larger than 1kb, a feature found very often in vertebrates, as introns are replete with cryptic PASs. This suggests that U1 should be required for telescripting not only at the 5'ss, but within almost all intronic sequences in genes. Supporting this

conclusion is the fact that U1's base pairing is also known to be degenerate and influenced by other RNA binding proteins (RBPs); in fact, recent studies have shown that U1 does in fact bind throughout introns rather than only at the 5'ss (Engreitz et al. 2014). When combined, these observations are indicative of a model where all PAS usage is under the control of U1, and any PAS could potentially be recognized, and acted upon, by the CPA machinery (Figure 1.7).

While these high levels of U1 AMO knockdowns inhibited ~95-99% of U1, we also used HIDE-seq to examine the effect that significantly lower doses have on transcription efficiency, reasoning that this would be more similar to physiological occurrences (Berg et al. 2012). Blocking 15% of available U1 resulted in much more distal PCPA, best characterized as 3' untranslated region (3'UTR) shortening or alternative 3' end-processing. This phenomenon has been well characterized, as proximal PAS usage in 3'UTRs through alternative polyadenylation (APA) is associated with activated cell states such as immune cell or neuron stimulation, cell proliferation, and cancer (Niibori et al. 2007; Flavell et al. 2008; Sandberg et al. 2008; Mayr and Bartel 2009; Lianoglou et al. 2013). Here, shortened 3'UTRs often result in increased protein translation as there are many binding sites for regulatory factors, such as microRNAs (miRNAs), in the distal part of these regions (Mayr 2017). The loss of these sites can result in changes in mRNA stability, translation efficiency, and mRNA localization (Figure 1.8). There were also cases where low U1 AMO doses induced alternative last exon usage, resulting in shortened, but often annotated and functional, mRNA isoforms. A frequently cited example of this is in the case of *homer-1*, a scaffolding protein that is necessary for synaptogenesis and in the postsynaptic density (PSD) (Niibori et al. 2007). The long form, homer-1l, is constitutively expressed and localized in the PSD, while neuronal stimulation results in an isoform switch to the short version, homer-1s. Without its carboxyl-terminal portion of the protein, homer-

1s antagonizes homer-1l and can cause epilepsy through over-stimulation. Several low U1 AMO concentrations mimicked the physiological isoform switching in *homer-1* in a dose-dependent manner (Berg et al. 2012).

Into the Modern Era

With both the decreasing costs and increasing data quality from high-through sequencing, both for RNA (RNAseq) and target protein immunoprecipitation (various IPseq), we taken advantage of this technology to thoroughly study U1's role in telescripting. In order to more specifically target active transcription after transfection, we also relied on metabolic RNA labeling with 4-thiouridine (4sU). This labeling incorporates 4sU into elongating pre-mRNA, allowing for isolation of nascent transcripts. Combining these two advancements has drastically increased the resolution of transcriptomic changes, elucidating more of the U1 telescripting story and mechanism. I have worked to create innovative analytical work to address the novelty of the discovery of PCPA and availability of such extensive data, as well as flesh out the mechanism of telescripting through critical U1 AMO based experiments.

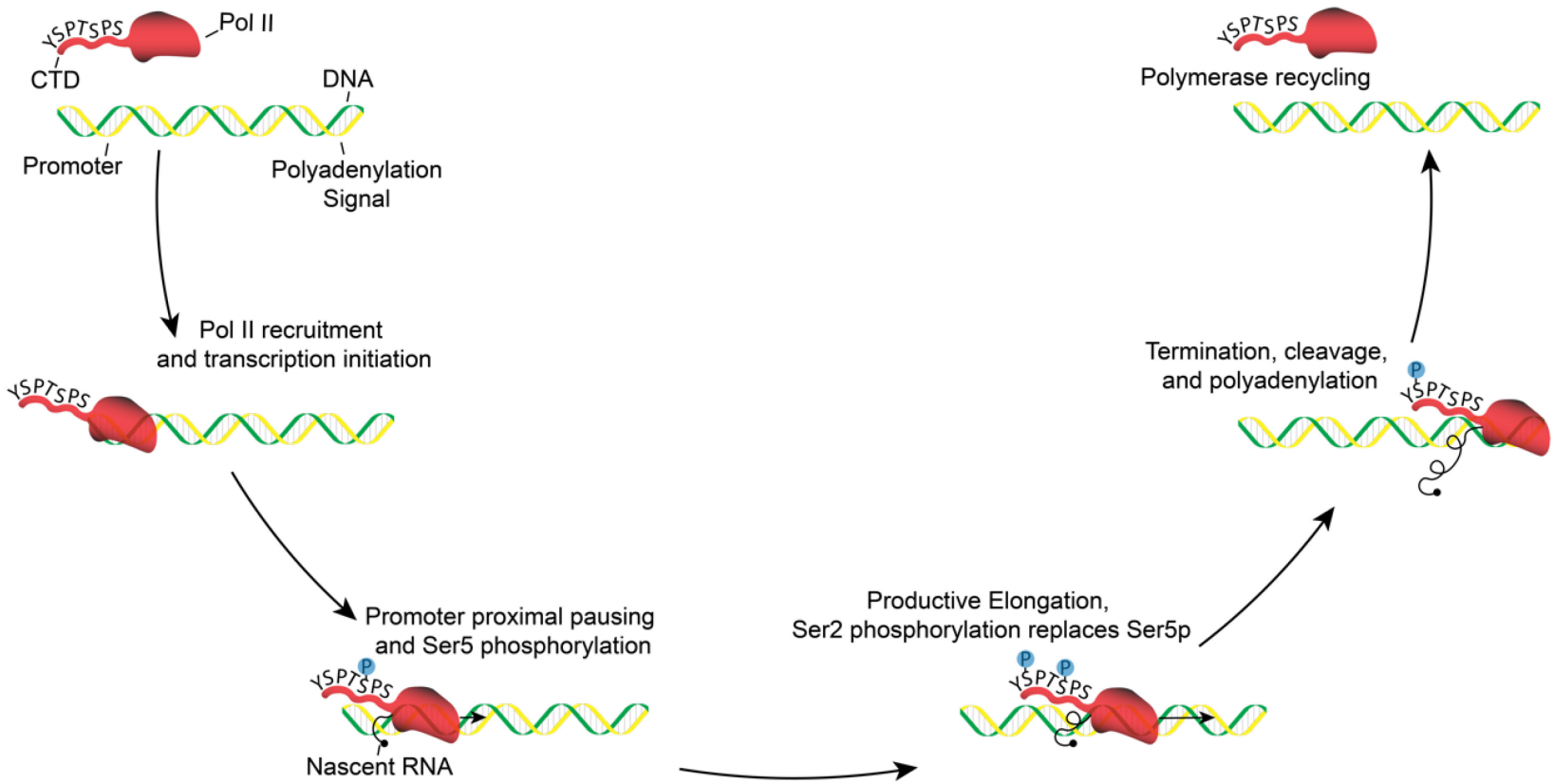

YSPTSPS
Pol II
CTD
DNA
Promoter
Polyadenylation
Signal
Pol II recruitment
and transcription initiation
Promoter proximal pausing
and Ser5 phosphorylation
Nascent RNA
Productive Elongation,
Ser2 phosphorylation replaces Ser5p
Termination, cleavage,
and polyadenylation
Polymerase recycling

Figure 1.1 RNA transcription by RNA polymerase II

A stepwise model of RNA transcription from pol II and the various CTD phosphorylation states associated with each phase. The red line extending from the red pol II body shows the CTD heptad repeat in amino acid code (Y = Tyrosine, S = Serine, P = Proline, T = Threonine). The “P” inset in the blue circle represents the phosphoryl group used for regulation of transcription. The black line and ball represent the nascent pre-mRNA and 5’ cap, respectively.

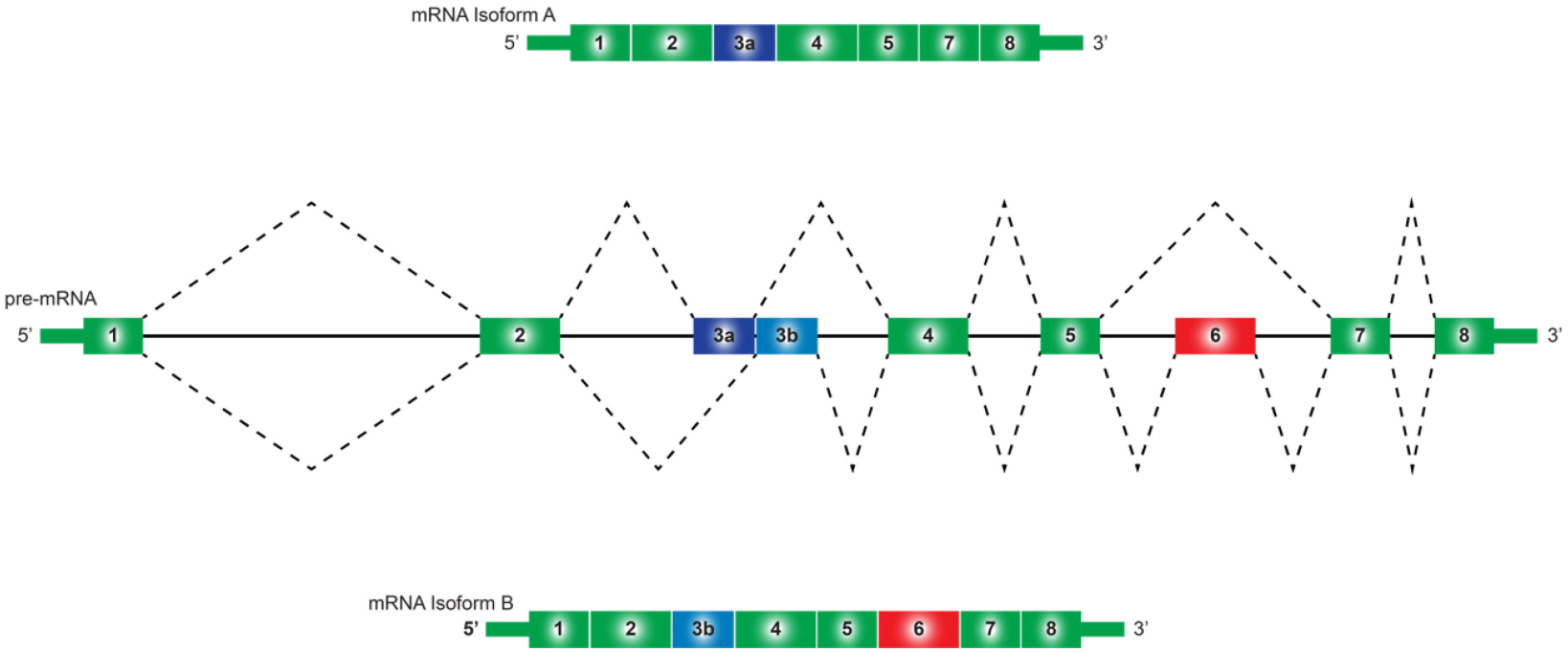

mRNA Isoform A
5'
1
2
3a
4
5
7
8
3'
pre-mRNA
5'
1
2
3a
3b
4
5
6
7
8
3'
mRNA Isoform B
5'
1
2
3b
4
5
6
7
8
3'

Figure 1.2 Structure of pre-mRNA and possible mRNA splicing isoforms

Pre-mRNA, in the middle of the scheme, is comprised of introns (solid black line) and exons (colored and numbered boxes). Exon to exon splicing is depicted by the dashed black lines above and below the pre-mRNA, which creates the associated mRNA products. Exons are differentiated to be included in both mRNA transcripts (green), alternatively spliced (blue), or as cassette exons (red). Thinner green boxes depict UTR regions. Transcription direction is indicated as 5' to 3'.

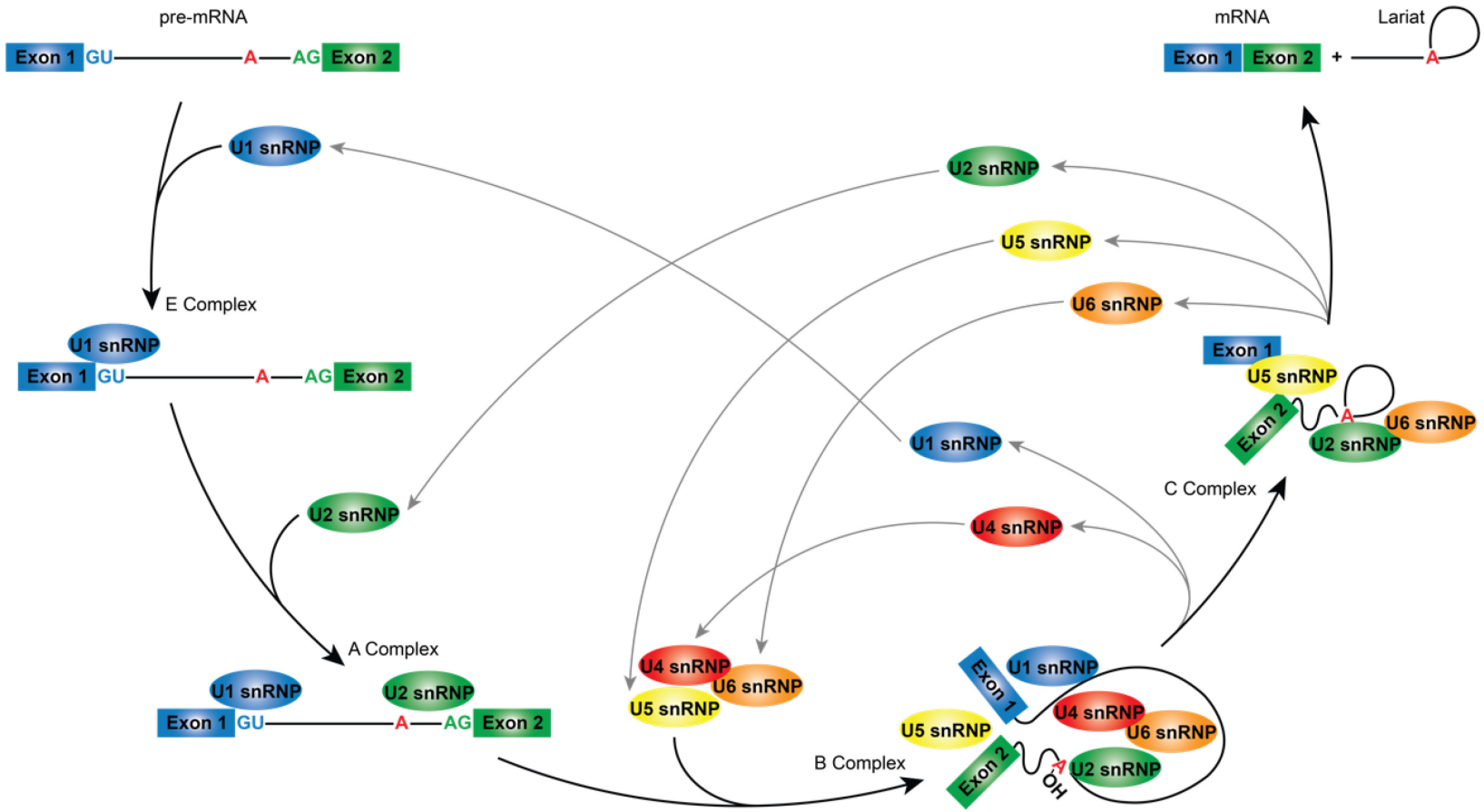

pre-mRNA
Exon 1
GU
A
AG
Exon 2
U1 snRNP
E Complex
U2 snRNP
A Complex
U4 snRNP
U6 snRNP
U5 snRNP
B Complex
OH
C Complex
mRNA
Lariat
+

Figure 1.3 Pre-mRNA splicing mediated by the major spliceosome

A stepwise model of the canonical assembly and disassembly of the spliceosome during the intron splicing process. Additions to the spliceosome are shown as black arrows, while departures are shown as grey arrows. The snRNPs are depicted as colored ovals, exons as colored boxes, and intron as the black line connecting exons. The splicing relevant nucleotide sequences, are shown as colored letters in the intron; invariant dinucleotides in blue and green and branch point adenosine in red. For simplicity, no accessory spliceosomal proteins are shown. Figure adapted from Will and Lührmann (Will and Lührmann 2011).

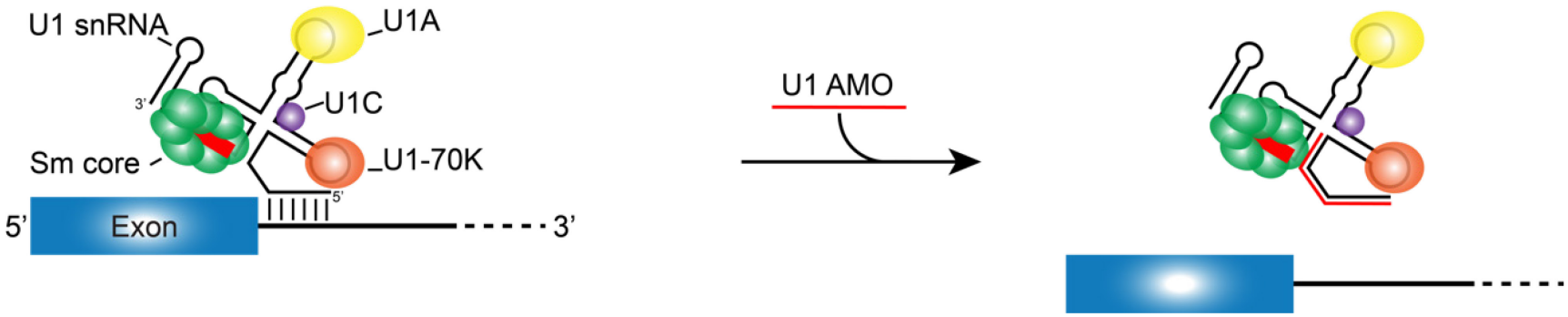

U1 snRNA
U1A
3'
U1C
Sm core
U1-70K
5'
5'
Exon
3'
U1 AMO

Figure 1.4 snRNP structure and mechanism of action of AMO inhibition

Left, U1 snRNP, including the snRNA (black line), U1 specific proteins (colored circles) and Sm core (green circles around the red Sm site), is shown bound to the 5' splice site by base-pairing interactions (vertical black lines). Right, U1 AMO, red line, binds to the 5' sequence of U1 snRNA, inhibiting U1 from binding 5' splice sites and other, similar sequences in introns. Figure adapted from Venters *et al*. 2019 (Venters et al. 2019).

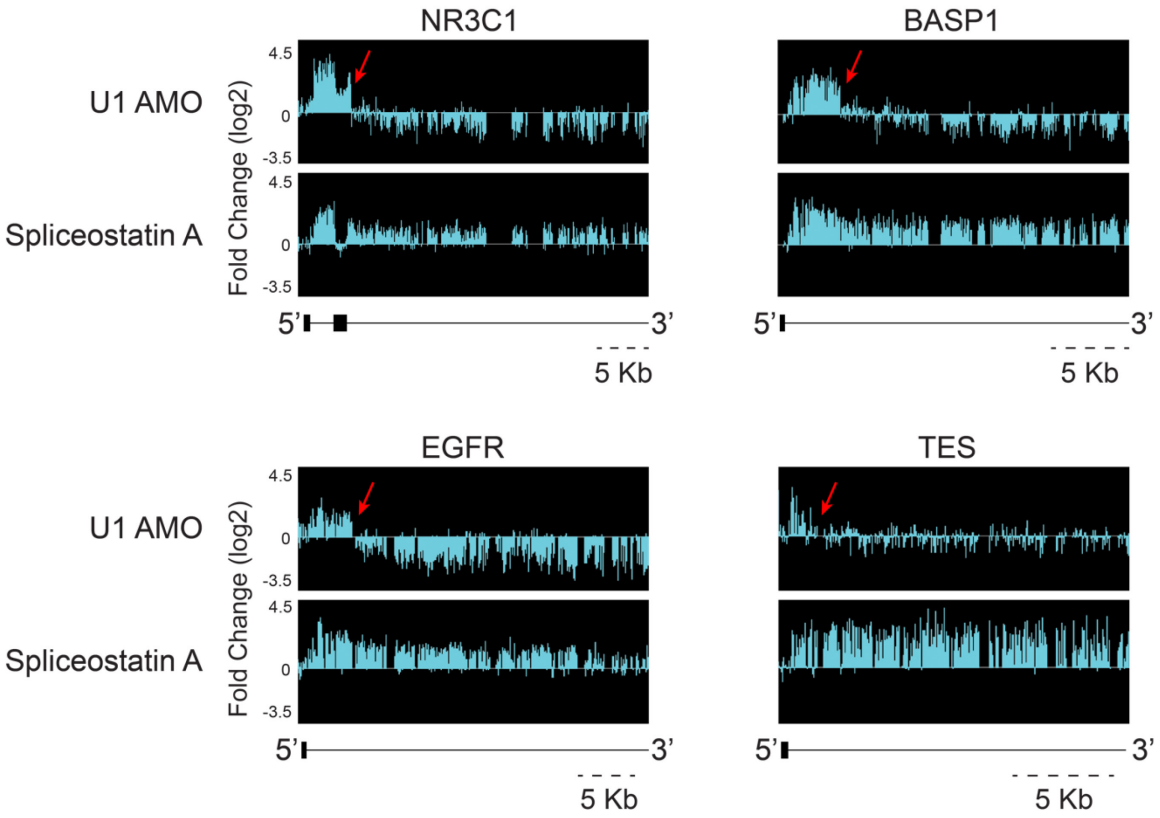
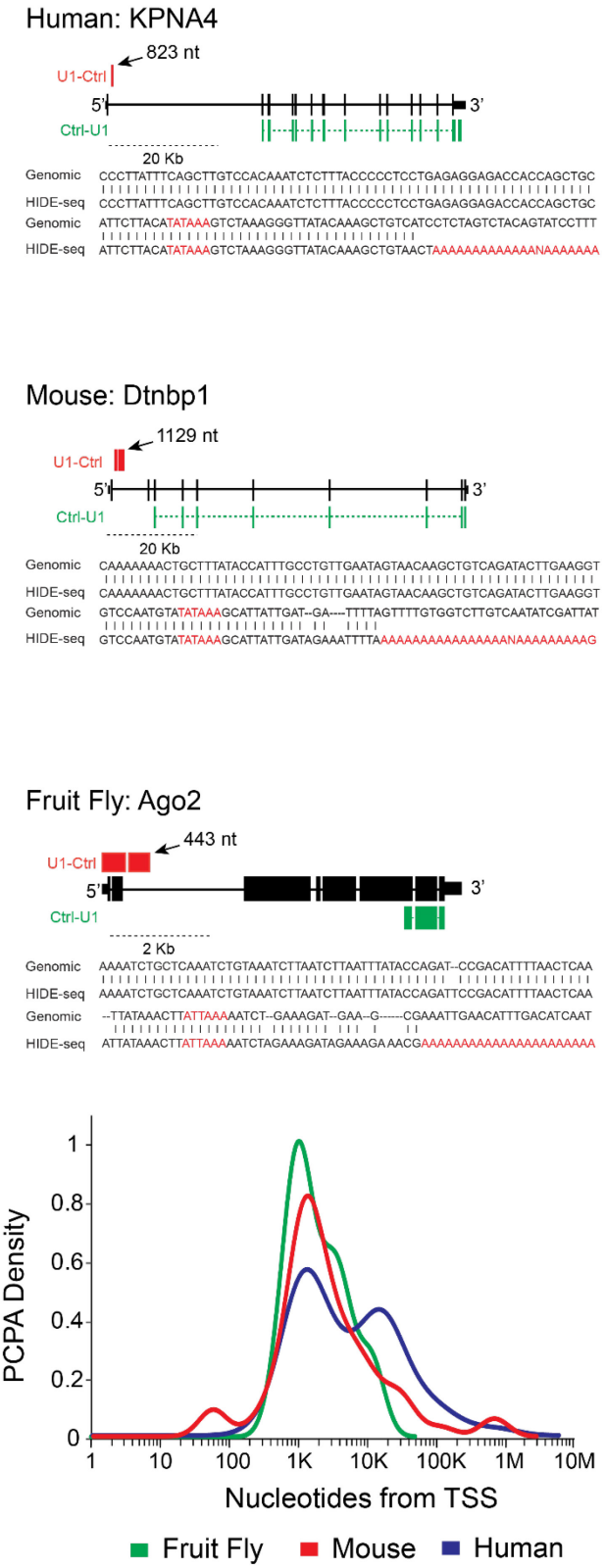

NR3C1
BASP1
EGFR
TES
U1 AMO
Spliceostatin A
Fold Change (log2)
4.5
0
-3.5
5'
3'
5 Kb

Figure 1.5 Genomic tiling array identifies PCPA and splicing inhibition

RNA from control, U1 AMO transfected (8hr, 7.5 μM), or SSA treated (8hr, 100 ng/mL) HeLa cells were analyzed using genomic tiling arrays. Signal intensity $\log_2$ fold change of treated cells versus control cells is shown in light blue above the gene structure shown as black lines for introns and black boxes for exons. Red arrows note the point of signal drop after U1 AMO transfections indicative of PCPA. Genomic distances are shown as dashed black lines. Figure adapted from Kaida *et al.* (Kaida et al. 2010).

a


Human: KPNA4
823 nt
U1-Ctrl
5'
3'
Ctrl-U1
20 Kb
Genomic CCCTTATTTCAGCTTGTCCACAAATCTCTTTACCCCCTCCTGAGAGGAGACCACCAGCTGC
HIDE-seq CCCTTATTTCAGCTTGTCCACAAATCTCTTTACCCCCTCCTGAGAGGAGACCACCAGCTGC
Genomic ATTCTTACATATAAAGTCTAAAGGGTTATACAAAGCTGTCATCCTCTAGTCTACAGTATCCTTT
HIDE-seq ATTCTTACATATAAAGTCTAAAGGGTTATACAAAGCTGTAACTAAAAAAAAAAAAAANAAAAAAA
Mouse: Dtnbp1
1129 nt
U1-Ctrl
5'
3'
Ctrl-U1
20 Kb
Genomic CAAAAAAACTGCTTTATACCATTTGCCTGTTGAATAGTAACAAGCTGTCAGATACTTGAAGGT
HIDE-seq CAAAAAAACTGCTTTATACCATTTGCCTGTTGAATAGTAACAAGCTGTCAGATACTTGAAGGT
Genomic GTCCAATGTATATAAAGCATTATTGAT--GA----TTTTAGTTTTGTGGTCTTGTCAATATCGATTAT
HIDE-seq GTCCAATGTATATAAAGCATTATTGATAGAAATTTTAAAAAAAAAAAAAAAAANAAAAAAAAAG
Fruit Fly: Ago2
443 nt
U1-Ctrl
5'
3'
Ctrl-U1
2 Kb
Genomic AAAATCTGCTCAAATCTGTAAATCTTAATCTTAATTTATACCAGAT--CCGACATTTTAACTCAA
HIDE-seq AAAATCTGCTCAAATCTGTAAATCTTAATCTTAATTTATACCAGATTCCGACATTTTAACTCAA
Genomic --TTATAAACTTATTAAAAATCT--GAAAGAT--GAA--G------CGAAATTGAACATTTGACATCAAT
HIDE-seq ATTATAAACTTATTAAAAATCTAGAAAGATAGAAAGA AACGAAAAAAAAAAAAAAAAAAAAAA

b

PCPA Density
1
0.8
0.6
0.4
0.2
0
1
10
100
1K
10K
100K
1M
10M
Nucleotides from TSS
Fruit Fly
Mouse
Human

Figure 1.6 HIDE-seq enriches for transcriptome changes and PCPA location

(a) Enrichment of transcriptome changes from control transfected or U1 AMO transfected cells is shown from subtractive hybridization of U1 minus control (red, U1-Ctrl) or control minus U1 (green, Ctrl-U1) either above or below gene structures, respectively. Data is from HeLa (Human), NIH-3T3 (Mouse), or S2 (Fruit Fly) cells. Black lines represent introns, black boxes represent exons, and colored dashed lines represent spliced intron signal. Black arrows indicate reads with non-genomic poly(A) tails. Full sequences of HIDE-seq reads are shown below the genomic DNA from human, mouse, or fly. Genomic distances are shown as dashed black lines. (b) Density graphs of reads with poly(A) tails present after U1 AMO transfection in human, mouse, or fly cells relative to their distance from the TSS are plotted. Figure adapted from Berg *et al*. (Berg et al. 2012).

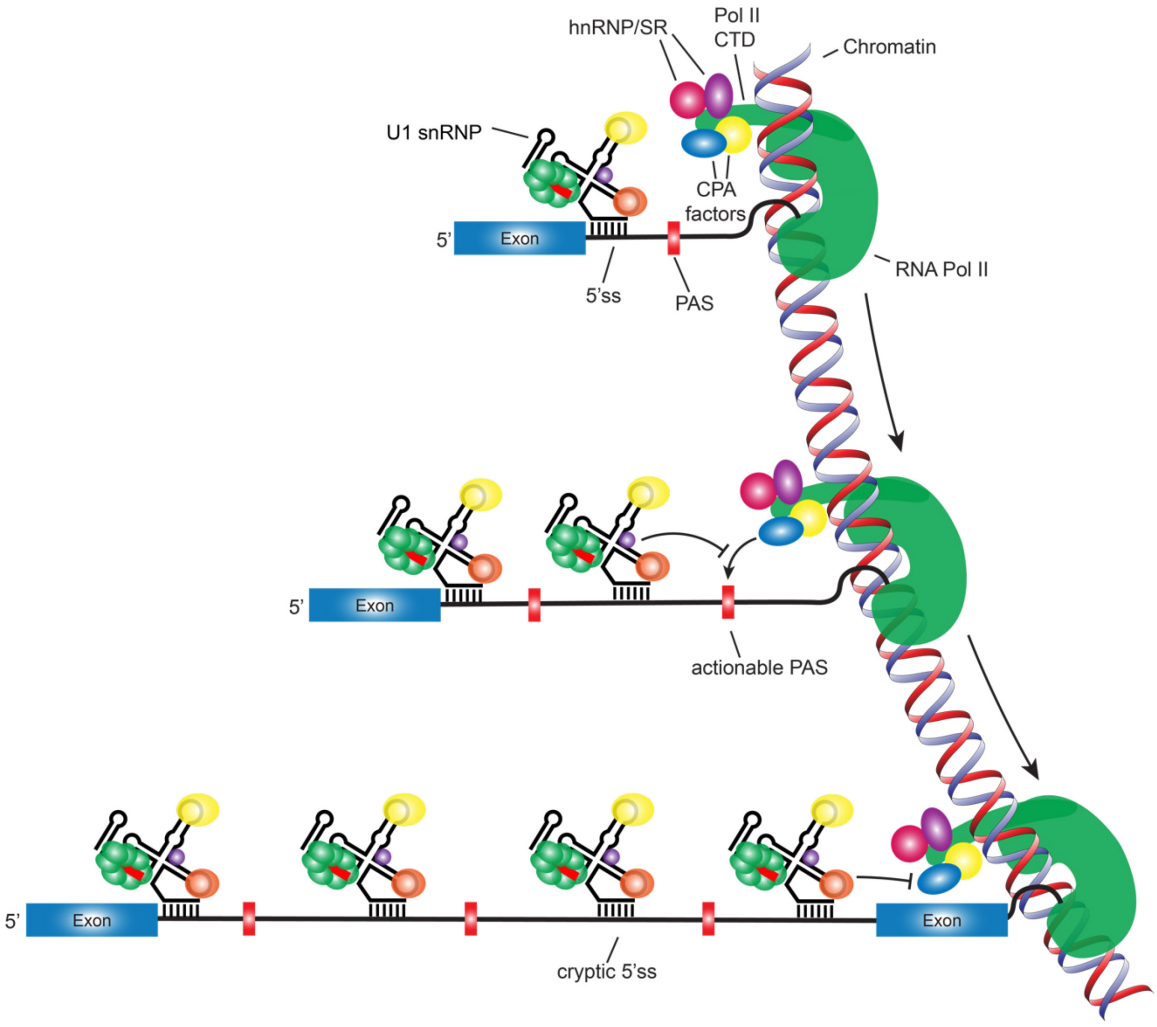
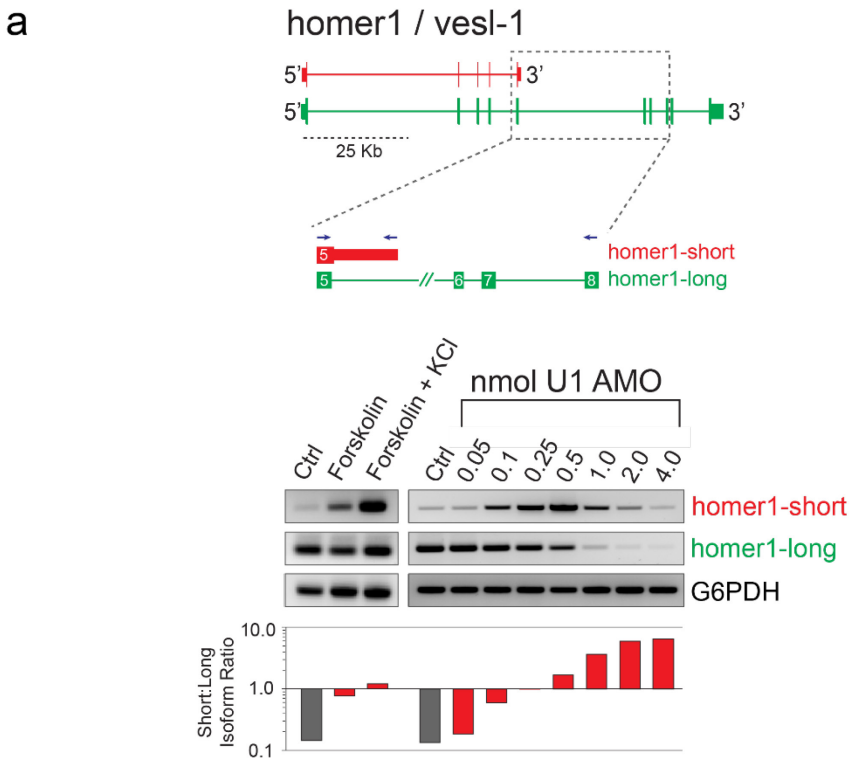

hnRNP/SR
Pol II
CTD
Chromatin
U1 snRNP
CPA
factors
5'
Exon
5'ss
PAS
RNA Pol II
5'
Exon
actionable PAS
5'
Exon
Exon
cryptic 5'ss

Figure 1.7 A proposed model of U1 telescripting

A cartoon model depicting the process of telescripting. Pol II CTD, green line, associates early in transcription with cleavage and polyadenylation (CPA) factors, shown as colored ovals. U1 snRNP, bound to 5' splice sites or intronic U1 binding sites, prevents CPA factors from activating cryptic polyadenylation signals and eliciting premature CPA (PCPA). U1 can protect nascent transcripts from PCPA for distances of roughly 1kb. Exons are shown as blue boxes, PAS as red boxes, and introns as black lines. Figure adapted from Berg *et al*. (Berg et al. 2012).


a
homer1 / vesl-1
5'
3'
25 Kb
homer1-short
homer1-long
Ctrl
Forskolin
Forskolin + KCl
nmol U1 AMO
0.05
0.1
0.25
0.5
1.0
2.0
4.0
homer1-short
homer1-long
G6PDH
Short:Long Isoform Ratio
10.0
1.0
0.1

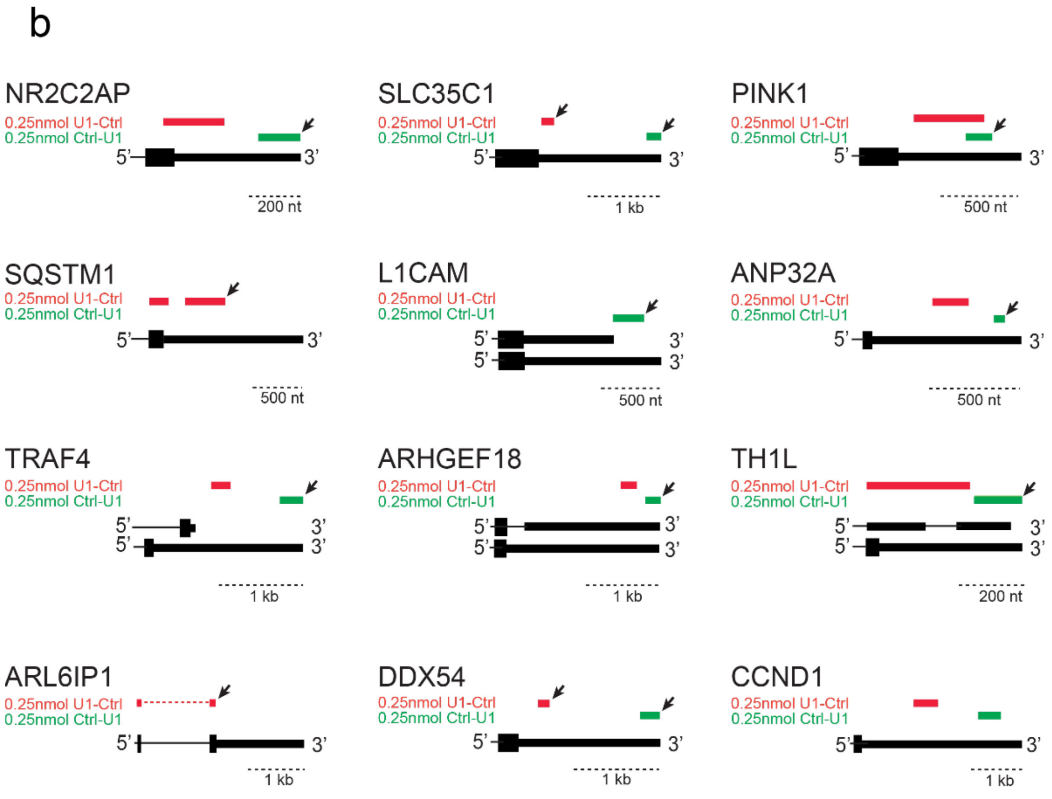

b
NR2C2AP
SLC35C1
PINK1
SQSTM1
L1CAM
ANP32A
TRAF4
ARHGEF18
TH1L
ARL6IP1
DDX54
CCND1
0.25nmol U1-Ctrl
0.25nmol Ctrl-U1
200 nt
1 kb
500 nt

Figure 1.8 Mild U1 base pairing inhibition recapitulates mRNA isoform switching and 3'UTR shortening seen in neuronal activation

(a) Rat PC12 neurons were treated for 3 hours with forskolin or forskolin/KCl, or separately transfected with increasing doses of U1 AMO. RT-PCR results with primers for either the short or long form of *homer-1* show the shift from long to short isoform with U1 AMO or forskolin. Gene structures are shown above, with inset highlighting the short (red) or long (green) isoform structure and arrows for the primer locations. (b) HIDE-seq results from HeLa cells for U1 AMO transfected versus control (red, U1-Ctrl) or vice versa (green, Ctrl-U1) in 3'UTRs. Black arrows indicate reads with non-genomic poly(A) tails. For gene structure lines depict introns, boxes depict exons, and thinner boxes depict UTRs. Genomic distances are shown as dashed black lines. Figure adapted from Berg *et al*. (Berg et al. 2012).

CHAPTER 2: METHODS

Experimental Methods

AMO Transfection

Cells are grown in full media, DMEM supplemented with 10% fetal bovine serum (FBS) and 1% Penicillin and Streptomycin (P/S), until they are 70-80% confluent. If the cell line is adherent, *Danio rerio* ZF4 or *Mus musculus* 3T3 cells, the media is removed and the cells are washed with phosphate buffered saline (PBS) and 0.25% Trypsin-EDTA (ThermoFisher) is used to detach them. Non-adherent cells, *Drosophila melanogaster* Schneider 2 (S2), are spun down at 2,000g for 2 minutes at room temperature (RT), the supernatant is removed and the pellet is washed with PBS. Cells are resuspended in media with 10% FBS, then counted using a NucleoCounter (ChemoMetec). $1x10^6$ cells are used per reaction. AMO transfection of the desired concentration is carried out via electroporation using either the Neon Transfection System (ThermoFisher) or the Amaxa Nucleofector System (Lonza) using buffers and pulse programs optimized for each cell line from the manufacturer. Once electroporated, the cells are immediately resuspended in full media and allowed to grow for the desired experiment time.

4sU Labeling and Purification

30 minutes prior to harvesting cells, 4sU (Sigma) is added to a final concentration of 200μM. The media is then removed and TRIzol (ThermoFisher) is used to harvest and homogenize cells as per the manufacturer's instructions. Total RNA is extracted from the TRIzol with chloroform and precipitated with ethanol.

RNA is biotinylated by shaking at 1,000 RPM for 1.5 hours at RT with 0.2mg/mL EZ-Link HPDP-Biotin (ThermoFisher) in binding buffer (10mM Tris pH 7.5, 1mM EDTA) to a final

concentration of 100ng/mL. Equal volume of chloroform:isoamyl alcohol (24:1) is added and RNA is extracted using a silicone phase divider gel (polydimethylsiloxane and silicone dioxide, Dow Corning). The biotinylated RNA is precipitated using isopropanol, then resuspended in water and captured on Dynabeads MyOne Streptavidin C1 (Invitrogen) as follows. A 100µL aliquot of beads is first prepared over three wash phases: three times with 1mL of a first wash buffer (5mM Tris pH 7.5, 0.5mM EDTA, 1M NaCl), twice with 1mL of a second buffer (0.1M NaOH, 0.05M NaCl), once with 1mL of a third buffer (0.1M NaCl), and then resuspended in 100µL of 10mM Tris pH 7.5, 1mM EDTA, 2M NaCl. The RNA is bound to the beads at equal volume over 15 minutes at RT, then separated from the solution on a magnet. The beads are washed three times with 1mL of wash buffer (100mM Tris pH 7.5, 10mM EDTA, 1M NaCl, 0.1% Tween20) at 65°C and then 3x again with the same wash buffer at RT. The purified RNA is recovered from the beads with two elutions of 100mM DTT and precipitated using ethanol. Final RNA concentration and quality is tested on a Bioanalyzer (Agilent).

Ribosomal RNA (rRNA) Depletion

Depletion of rRNA is done using the Ribo-Zero rRNA Removal Kit (Illumina) following the manufacturer's instructions. For non-human, mouse, or rat samples, the rRNA level is checked on a Bioanalyzer after one pass. If residual, unwanted rRNA remains, a second run with the kit is required. Of note, the Ribo-Zero includes an upper limit to the input RNA concentration. This, combined with the low concentrations of labeled RNA from the short (30 minute) 4sU label, means that 4sU selection should be carried out first. Final RNA concentration and quality is tested on a Bioanalyzer (Agilent).

RNAseq Library Preparation

Strand-specific RNAseq sequencing libraries are created using the KAPA Stranded RNA-Seq Library Preparation Kit (Kapa Biosystems) per the manufacturer's instructions. The SeqCap Adapter Kit A is used (Roche). Final library concentration and quality is tested on a Bioanalyzer (Agilent).

Pol II ChIPseq

Cells, grown as described above, are detached with Trypsin and washed with PBS; 1% formaldehyde in PBS is then used to crosslink protein and DNA at RT for 15 minutes. Excess formaldehyde is quenched using glycine to a final concentration of 125mM. The cells are spun (2,000g for 5 minutes) and resuspended in lysis buffer (5mM HEPES K pH 7.9, 85mM KCl, 0.5% NP-40, cOmplete protease inhibitor from Roche). Nuclei are spun at 3,900g for 5 minutes at 4°C and resuspended in nuclei lysis buffer (50mM Tris-HCl pH 8.0, 1mM EDTA pH 8.0, 1% SDS, cOmplete protease inhibitor from Roche). The chromatin is sonicated using a Covaris ultrasonicator for 5 minutes with the following settings: 10% duty cycle, 5 intensity, and 200 cycle/burst. Chromatin is diluted in 16.7mM Tris-HCl pH 8.0, 1.2mM EDTA pH 8.0, 167mM NaCl, 0.01% SDS, 1.1% Triton X-100, and cOmplete protease inhibitor from Roche and then pre-cleared with Protein A Dynabeads for 1-2 hours. Protein A Dynabeads are bound to either pol II (Abcam ab5131 for Ser5p, ab5095 for Ser2p, or Santa Cruz sc-899X for N-terminus) or control IgG antibodies for 1 hour, then washed 2x with 500µL PBS plus 5% bovine serum albumin (BSA). Diluted chromatin is then bound to the antibodies overnight.

Once bound to beads, the sample is washed 2x each with a low salt buffer (20mM Tris-HCl pH8.0, 2mM EDTA pH 8.0, 150mM NaCl, 0.1% SDS, and 1% Triton X-100), a high salt buffer (20mM Tris-HCl pH8.0, 2mM EDTA pH 8.0, 500mM NaCl, 0.1% SDS, and 1%

Triton X-100), and a LiCl buffer (10mM Tris-HCl pH8.0, 1mM EDTA pH 8.0, 250mM LiCl, 1% NP-40, and 1% deoxycholic acid). The cleaned beads are resuspended in 50mM Tris-HCl pH 8.0, 10mM EDTA pH 8.0, and 0.1% SDS and the crosslinks are reversed overnight. Proteinase K is used to digest the chromatin associated protein, and the DNA is extracted with phenol-chloroform-isoamyl alcohol and precipitated with ethanol. A 2% agarose gel is used to verify the crosslinked DNA, which should be between 200-400nt in length. Sequencing libraries are made using the TruSeq ChIP Sample Preparation Kit (Illumina) as per the manufacturer's instructions.

Analytical Methods

PCPA Detection

Our discovery of PCPA with tiling arrays and HIDE-seq necessitated the development of novel RNAseq analysis tools, as no existing ones could be easily adapted to detect the phenomenon. After visualizing the RNAseq reads from U1 AMO treated samples on a genome browser, the difference between PCPA and other transcriptomic changes was clear. The next step was for me to translate these changes into a computational pipeline that could be easily run for any sample. I structured the analysis with the following parameters in mind: 1) detect as much of the PCPA as possible, especially the sets of genes that we had already established as good markers from previous studies; 2) keep the computational workload to a minimum due to the number of samples and sequencing depth we were using; 3) reduce the number of false positives, arising most likely from intron retention or transcriptional down-regulation; and 4) present the data in a format and style usable by other lab or community members.

I first targeted the most unique and prominent feature of PCPA, which is the aggregation of reads in the 5' side of a gene, seen mostly in the first or second intron, followed by a

decrease in signal over the rest of the gene body (Figure 2.1). I tested multiple methods for detecting the site of read decrease at the PCPA point, including cumulative distributions, inflection points, and sliding windows. These methods were either too computationally intensive for the amount of data I was working with (50-200 million paired-end reads) or were not sufficiently sensitive to detect the cases we were expecting. As a consequence, I decided on a simpler approach of splitting the target regions — initially introns — into a specific number of bins (i.e. allowing for variable length), which would allow for play within the detection parameters. It was important to pick an applicable number of bins per feature; if the number is too high, the bins' small size would mean that minor read changes could skew the results of any statistical tests. Alternatively, if too few bins are used, their large regions could include too much interfering signal against the read drop across a PCPA site and, consequently, reduce the detection rate. Ultimately, trial and error showed that dividing intron one and two into four equal size bins was the simplest approach that resulted in an acceptable number of data points needed for further calculations. As there were some genes that did not show a single PCPA site, but instead showed transcription degradation across the entire gene body, this process was adapted by the postdoc Chao Di to utilize the entire gene length rather than each intron separately. To avoid any genes expressed at low levels, any gene with <10 reads in the first quarter of the target intron or gene body was removed.

Another important feature of PCPA that is included in the calculations is the decrease in transcriptional output from the gene. That is to say, the full-length, spliced mRNA must be decreased. To target this, I extracted multiple metrics for gene expression including exon-exon junctions both across the entire gene and across our target exons (first, second, and third); exon reads, especially terminal exon and coding sequence (CDS) reads; 3'UTR reads; and general FPKM (fragments per kilobase per million mapped reads)

measurements. Many of these metrics resulted in skewed results due to heavy read shifts as a result of PCPA and splicing inhibition from U1 AMO. For example, FPKM measurements are unreliable as a sizeable aggregation of reads in the first 5' exon, due to PCPA, could still result in the gene maintaining a high FPKM value despite not producing functional mRNA. Similarly, the number of exon-exon junction reads, which are valuable as a measure of fully processed mRNA, can be reduced due to splicing inhibition from the U1 AMO and not PCPA. The simplest solution, therefore, is to compare the terminal CDS to the first exon, which allows for direct assessment of the level of full-length transcription to initiation, respectively. After U1 AMO, almost any increase in the signal in 5' exons without a corresponding increase in terminal CDS would be due to PCPA, as even partial transcription termination should produce more 5' transcription at the expense of full-length.

With the PCPA parameters defined as stated above, I then compiled the read values and associated both the decrease in transcriptional output and fewer reads late in the gene as compared with the initial exons. Comparing the values from the U1 AMO treated sample to the control AMO treated sample allowed us to test for statistical significance in the change across all genes in the samples using a Fisher's exact test followed by Benjamini-Hochberg multiple testing, with a cutoff of an adjusted P-value ≤ 0.01 (Benjamini and Hochberg 1995). I worked with a postdoc in our lab, Chao Di, to apply these statistical testing methods to the extracted values.

Intron Size Expansion

With our discovery that PCPA sensitivity was highly correlated with larger gene size, I wanted to examine whether these genes most affected by U1 base-pairing inhibition were also those that experienced the greatest intron expansion across vertebrate evolution (Oh

et al. 2017). I decided to use gene ortholog data for this from Ensembl's Biomart (https://www.ensembl.org/biomart/martview) due to the inclusion of a large number of organisms, variability in data features available, and the up-to-date reference genomes it contains. With the initial RNAseq data from HeLa cells, I acquired each target organism's homology data in reference to the human genome, e.g. I downloaded feature data for all genes in *Drosophila melanogaster* that were known to have a homolog in human. To encompass multiple data points, I used fruit fly, *Drosophila melanogaster*; pufferfish, *Takifugu rubripes;* zebrafish, *Danio rerio*; chicken, *Gallus gallus*; and mouse, *Mus musculus*. Due to the many differences in curation and experimental knowledge on each organism, I had to remove many duplicate gene and isoform entries. After cleaning the data, I kept the longest isoform in order to maintain consistency with our previous analyses. The finalized homology lists were then used to extract gene sizes for target groups.

Novel Splice Site Usage and Multi-Exon Skipping

Despite the fact that much of our results with U1 AMO demonstrated high levels of PCPA genome wide, we also anticipated there to be many splicing changes due to its role in exon definition. As such, I decided to globally analyze splicing changes within our U1 AMO datasets. Existing tools for measuring sample level splicing changes rely on previous gene annotations, such as MISO (Katz et al. 2010), or are computationally intensive or slow, such as JUM (Q. Wang and Rio 2018) or MAJIQ (Vaquero-Garcia et al. 2016); therefore, I sought to develop a simple, gene-based tool for detecting splicing changes for all exon junctions regardless of annotation. To reduce calculation time, rather than defining novel splice junctions from scratch, I utilized TopHat's (Trapnell, Pachter, and Salzberg 2009) or STAR's (Dobin et al. 2013) *de novo* splice junction algorithm via the spliced read files

they output from alignment. To more easily work with the data, I started by removing all reads that align canonically to consecutive exon-exon junctions in any isoform (i.e. a spliced read that spans from exon 1 to exon 2). This is done simply by using genomic coordinates of the splice location in both the BED entry and all introns in the genome. In general, this removes >80% of spliced reads in most samples. The rest of the reads are then filtered to reduce potential false positives due to sequencer or aligner error by removing all junction sites containing >1-5 reads; this threshold depends highly on sequencing depth. Once cleaned, the reads are processed into either splice junctions that utilize both a known 5' and 3'ss, and those that do not, again using genomic coordinates. The former, if the 5' and 3'ss are found to be within one gene, are considered multi-exon skipping, otherwise they are labeled as run-on transcription and trans-gene splicing. The latter are considered *de novo* splicing, and require further analysis to differentiate them from what could be background signal or true novel splice site usage.

For comparison of a treatment versus control, all read values are normalized to the sequencing depth. Then, splicing is compared on a gene by gene basis to determine which genes experience more exon skipping or run-on transcription.

To examine splice site sequences for skipped or included exons, I extracted 9 nucleotides from each 5'ss, three in the upstream exon and six in the intron, and 23 nucleotides from the 3'ss, three in the downstream exon and 20 in the intron. To determine skipped exons, I looked for exons that were entirely overlapped by the inter-splice distance, i.e. the space between the start and end of the junction, of any *de novo* splice junction as called from TopHat or STAR. For each organism, I grouped the exons into those that were always included in the samples (control and any treatment) and those that were skipped, even if this was in one sample but not others. Each splice site sequence was then scored into a

position weight matrix and represented graphically into a logo using Biostrings and seqLogo from Bioconductor in R (Bembom 2014; Huber et al. 2015; Pagès et al. 2017). To score splice site consensus sequence, I took the sequences extracted above and utilized the maximum entropy modeling software MaxEntScan (Yeo and Burge 2004).

Gene Neighborhood Distances

A particularly interesting and novel result from my U1 AMO treated mouse cell RNA-sequencing experiments was that many PCPAed genes were located in close genomic proximity to small, up-regulated genes. I decided to globally analyze this potential phenomenon to see if, and how, transcription expression correlates with expression of other nearby protein coding genes. I first had to remove most non-coding RNAs (e.g. micro RNAs, uncharacterized LOC RNAs, predicted and pseudo genes, etc.) from a reference file as many of these non-coding RNAs would show up as hits when checking a gene list for neighboring transcripts, but have little to no experimental data to confirm their function or expression patterns. It should be noted that some non-coding transcripts are of interest to us, for example MALAT1, because of known function in cancer or other disease states or as heavy binders of RBPs (Ji et al. 2003). This required the removal of most non-coding RNAs to be done partially by hand, identifying either groups to remove in batch or singular annotations to remove individually.

Once cleaned, I used BEDtools (Quinlan and Hall 2010) to extract the closest nearby gene from the target list and to compute the intergenic distance between them. Due to the large size of some genes, and importantly PCPAed genes, I needed to test this distance using the TSS, the transcription end site (TES), or both sites together as the reference point for finding the nearest genes. Ultimately, there was minimal difference between the results from utilizing different reference points of the gene, and I decided on using both the TSS

and TES as the reference points so as not to restrict the data any more than necessary. The only remaining variables were the specific gene groupings to compare, as well as the set to use as the reference point. To reiterate, the purpose of this study is to determine if the expression of small, essential genes increased in transcription by being in proximity to larger genes that would be PCPAed in the event of cell stimulation, stress, or activation. As such, I ended up using the small (<12kb) genes that were up-regulated with U1 AMO treatment as the reference set due to the fact that these would best mimic the genes that benefit from a loss of U1 telescripting. For the test group, I use two sets of genes: the down-regulated but not PCPAed, as well as the PCPAed and down-regulated. In order to have a control intergenic distance, I checked both the up-regulated genes to themselves, as well as compared two random sampling of genes to one another. For this sampling, I randomly extracted 20% of the genes per chromosome and calculated their intergenic distance. I calculated this for each chromosome separately in order to not overestimate intergenic distances, and then calculated the median from these data.

Upstream Antisense Transcription

Recently it has been shown that promoter directionality and upstream antisense transcription is known to be highly regulated by U1 levels (Almada et al. 2013). Following with the gene neighborhoods analysis explained above, it made sense to pivot the already written pipeline to target directly the upstream antisense region of protein coding genes. I again started by removing the same non-coding RNA transcripts. At first, I worked to examine whether upstream antisense (UA) transcription is affected by the presence of previously annotated genes in the UA region. To do this, I created reference files that either used a 10kb window UA to all genes or used a window ≤10kb that included the entire region UA to the target gene up until it encountered another annotated transcript. I

then checked for read coverage and read coverage change in these regions in the control or U1 AMO treated samples, and compared the two methods against one another. Using a set 10kb window, regardless of the presence of annotated transcripts, produced less read number variation in comparison to only testing regions without any gene annotations. This is most likely because these known transcripts contain their own promoters and are less dependent on the promoter from the target gene. Because the interest here was in detecting UA transcription changes from PCPA and not testing differences in promoters, I decided to filter out all genes with overlapping or nearby (≤10kb) UA transcripts.

Once this set of reference genes was collected, I grouped those that experienced transcriptional changes after U1 AMO treatment into four groups: either not PCPAed but down-regulated or up-regulated, PCPAed, or unchanged. I then computed both the read coverage and read coverage change within the UA region in four different size ranges: 0.3kb, 1kb, 5kb, and 10kb. This was done in order to examine how far reaching the effect of PCPA or U1 AMO induced transcriptional changes was on UA transcription. A distance of 1kb showed the most acute response to U1 AMO.

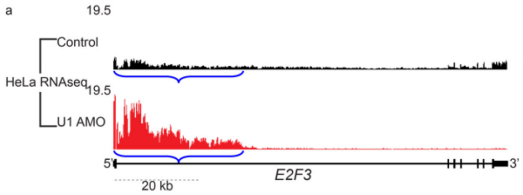
RBP XLIPseq Analysis

As we had many RNA cross-linking immunoprecipitations with high-throughput sequencing (XLIPseq) datasets, I tried to develop a method that would both verify the accuracy of the pull downs and identify potential snRNP interactors. With the large number of gene and pseudogene copies for each snRNA, which often only vary by one nucleotide, it was not feasible to use data aligned to the human genome (hg19 or hg38). During normal alignment and processing, reads from snRNAs would be extremely similar in sequence to gene copies and, thus, multimap to these multiple locations and not be included in the final alignment. As a consequence, I downloaded all annotated sequences for

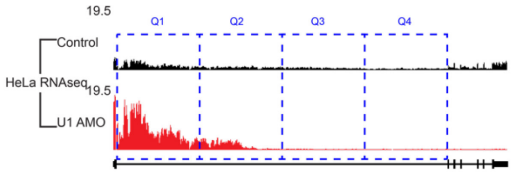
spliceosomal snRNAs (U1, U2, U4, U5, U6, U7, U11, U12, U4atac, U6atac) and two non-spliceosomal snRNAs (7SK and 7SL) from Ensembl's Biomart. I then collated these into fasta files and annotation gene transfer format (GTF) files in order to create Bowtie 2 (Langmead and Salzberg 2012) alignment indexed files. These were then used to align the XLIPseq or RNAseq reads. I also used more stringent cleaning of aligned reads by removing reads less than 20 base pairs that did not align completely to only one sequence. Once processed, I was then able to use R to plot the read density along each snRNA sequence to identify binding locations of the XLIP'd proteins or to identify pre-snRNA sequences in the sample.

In order to determine the RNA binding compared to mRNA output, I simply compared the mapped reads per million (RPM) of each gene in the XLIPseq to the exon RPM value in the 5 minute 5-ethynyl uridine (EU) pulse-labeled RNA. I used the latter sample as a control in this case because the shorter labeling time is more comparable to the 10-minute formaldehyde cross-linking time used in the XLIP data. For calculating the binding loss to RNA after PCPA, I divided gene RPM for each XLIPseq in the U1 AMO treatment by the same value in control.

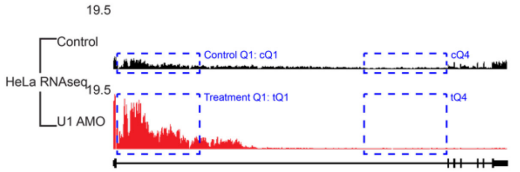
a



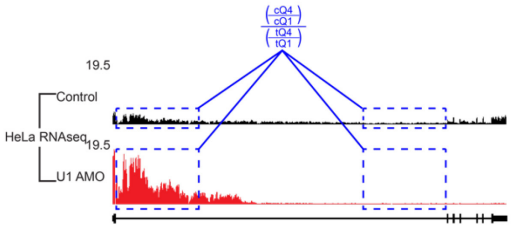
Purpose: Identify the 5' aggregation of reads indicative of PCPA



Divide the intron or gene body into quarters



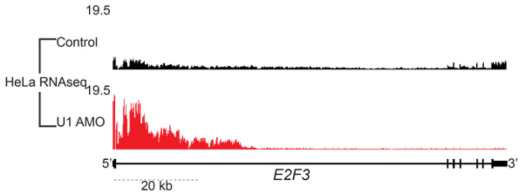
Sum the reads in Q1 and Q4 for each sample



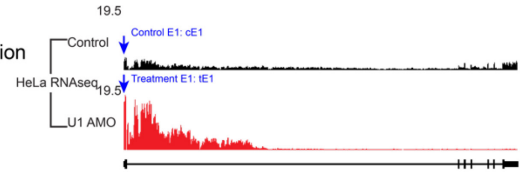
Compare the change in reads after treatment: control Q1/Q4 vs. treatment Q1/Q4

b

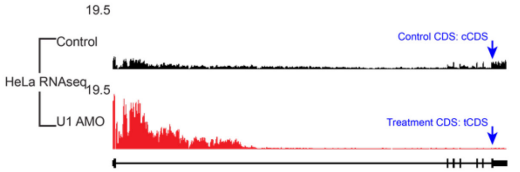
Purpose: Identify the decrease in full-length mRNA production



Sum the first exon reads as a measure of transcription initiation



Sum first terminal coding sequence reads as a measure of full-length transcription



Compare the change in reads after treatment: control CDS/E1 vs. treatment CDS/E1

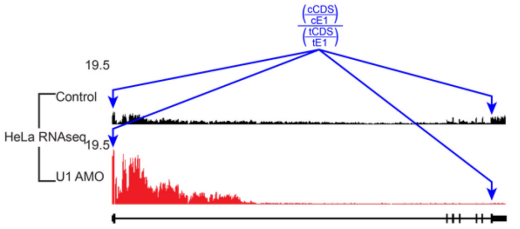


Figure 2.1 PCPA identification workflow for RNAseq data

(a) Description of the method used to identify the 5' side read aggregation in PCPAed genes. Blue brackets indicate the region of accumulated PCPA products as identified by eye. Blue boxes indicate the quartile regions used for the calculation, also shown in blue. (b) Description of the method used to identify the reduction in mRNA output from U1 AMO treatment. Select exons used for this calculation are identified by blue arrows, the calculation itself is also shown in blue. 30 minute labeled 4sU RNA from U1 AMO transfected HeLa cells (red) is compared to control (black) for the indicated regions. For gene structure lines depict introns, boxes depict exons, and thinner boxes depict UTRs. Genomic distances are shown as dashed black lines.

CHAPTER 3: CO-TRANSCRIPTIONAL U1 TELESCRIPTING AND IDENTIFICATION OF PCPAED GENES

Chapter Details

Published in “U1 snRNP telescripting regulates a size-function-stratified human genome” (Oh et al. 2017). Contributing members: Jung-Min Oh and Chie Arai (RNAseq, 3’RACE); Chao Di (help with statistical significance testing and metagene profiling).

Detection of PCPA Events

While U1 telescripting was identified by the Dreyfuss lab in 2010 and expanded upon in 2012, the discovery was limited by the techniques available at the time (Kaida et al. 2010; Berg et al. 2012). Advancements into RNAseq, already described in the Introduction, allowed for concomitant advancements in analysis of new, large datasets. The critical first step in understanding PCPA and its effect on the transcriptome was to identify which genes were sensitive to U1 base-pairing inhibition. The process in its entirety is described in more detail in the Methods section herein, but it is worthwhile to note again that the analytical methodology underwent several revisions before it was finalized.

Briefly, the workflow targets two important changes in transcription related to PCPA: (1) that full-length transcription is decreased by measuring the terminal CDS vs. the first CDS, and (2) that there is a 5’ aggregation of reads in either the whole gene body or in one of the introns (Figure 3.1). These calculations also work to exclude false positives due to decreases in transcription initiation, which would show similar decreases in both the first and terminal exon, as well as intron retention, which are exemplified by near flat levels of intronic reads.

We applied the above criteria to recent 5nmol U1 AMO datasets from Jung-Min Oh and Chie Arai, two postdocs in the lab (Oh et al. 2017). These experiments were carried out in

HeLa cells with a 4 or 8-hour incubation time after U1 AMO transfection, and nascent RNA was labeled for either 30 minutes or 2 hours with 4sU. The RNA was either selected using oligo-dT beads to enrich mRNA with poly(A) tails, or simply depleted of rRNA to capture the total RNA pools. I performed adapter trimming, quality control, and alignment of the reads to the hg19 and hg38 reference genomes using TopHat and STAR. This resulted in 53-167 million mapped reads used for downstream analysis. I then applied the PCPA detection algorithm and identified 3,590 PCPAed genes out of a total 9,744 expressed genes (RPKM ≥ 1) from the 8-hour transfection time point, with a 90% overlap with the 4-hour time point suggesting rapid PCPA onset after U1 AMO treatment (Figure 3.1; full list of genes found in Supplementary Table 1). The identification of these PCPAed transcripts was validated both by visualization of a large subsample of genes on the UCSC Genome Browser, as well as by 3'RACE of select genes performed by Jung-Min Oh. Moreover, the presence and position of these PCPAed transcripts were verified from data external to our lab. For example, tissue-specific 3' polyadenylated transcripts and binding locations of cleavage and polyadenylation factors coincided with the PCPA sites in our samples (Derti et al. 2012; Yao et al. 2012). As with our earlier studies, the polyadenylated nature of these shortened transcripts further confirmed the PCPA mechanism to be due to PAS activation by the CPA machinery, although we still lacked direct evidence that pol II disengaged from the gene as it does at the canonical 3' end. As a consequence, I decided to directly test the effect of U1 AMO on pol II.

Co-transcriptional PCPA Detection with Pol II ChIPseq

I worked to target pol II molecules that were engaged in transcription through chromatin immunoprecipitation followed by high-throughput sequencing (ChIPseq) after treatment of either a control AMO or U1 AMO. This technique utilizes formaldehyde's zero-distance cross-linking followed by antibody pull downs to select for DNA sequences bound by a

target protein. Using this technique, I pulled down pol II bound DNA using three antibodies: one that was general for all pol II through its N-terminal domain, and two that specifically target promoter proximal paused or actively elongating polymerases through CTD Ser2p or Ser5p markers, respectively. An immunoglobulin G (IgG) antibody was used as a non-specific control. As there was no lab protocol for this, I adapted already published methodologies (Lee, Johnstone, and Young 2006; Sun et al. 2011; Weber, Ramachandran, and Henikoff 2014). Testing of the protocol, antibodies, and AMO treatment was done using qPCR on the promoter and exon 1 of the housekeeping gene GAPDH. The AFM promoter was used as a negative control, as it does not express in HeLa cells (Figure 3.2).

I processed and aligned the sequencing data as described above, resulting in 18-135 million mapped reads (Table 1). Background reads from non-specific antibody interactions are to be expected in most IP experiments; as a result, I used the reads from IgG IP, in both control and U1 AMO treatment, to remove unwanted signal from their corresponding pull-downs and to detect true binding peaks using a model-based analysis of ChIPseq (MACS) (Y. Zhang et al. 2008). Visualization of the cleaned data on the UCSC Genome Browser showed positive results for the N-terminal antibody, however the Ser2p and Ser5p pull downs did not have good coverage across genes (Figure 3.3). The lack of adequate coverage was not surprising; previous studies have identified issues with these types of pol II antibody studies, including the fact that phosphoepitope specific antibodies can be affected by the phosphorylation state of neighboring peptides and that CTD specific antibodies may be hindered by crosslinking of the CTD itself to CTD-binding proteins (Mayer et al. 2010; Bowman and Kelly 2014). However, this resulted in an inability to elucidate any mid-gene pausing related to PCPA. Despite this, the N-terminal antibody

proved to contain enough information to demonstrate the dynamics of pol II transcription after U1 AMO.

As I compared the ChIP signal to our RNAseq, it was apparent that inhibition of U1 base pairing induces premature pol II termination. The level of pol II binding decreased to background level usually within a short distance (~1-10 kb) downstream of the PCPA site. This is similar to the already described torpedo model of 3' termination, where the polymerase continues past the PAS and the resulting unnecessary RNA is degraded by the Xrn2 exonuclease (Connelly and Manley 1988; Fong et al. 2015; Nick J. Proudfoot 2016). U1 AMO treatment results in a peak of pol II signal in what is now considered a "termination zone" and may work to facilitate 3' end processing or polymerase release. The similarity in my pol II signal within the premature termination zone further suggests that these PCPAed transcripts are processed in a manner almost identical to those at the canonical 3' end (Figure 3.4). Between the TSS and RNAseq-derived PCPA site, the pol II signal was unchanged or higher than that found in the control. This indicated that transcription initiation was either unaffected or possibly increased by PCPA. The latter result is likely explained by a reduced recycling time in pol II molecules due to the higher turnover caused by PCPA. As more polymerases disengage within a few kb of the TSS, their proximity to the promoter could increase the likelihood of re-engaging with the same gene. Metagene plots of pol II signal across the PCPAed genes from the RNAseq data show all these phenomena in detail (Figure 3.5).

The Relationship Between Gene Size and Function

The identification of PCPAed genes quickly led to the discovery that large genes were disproportionately more affected by U1 base pairing inhibition in comparison to smaller genes (Supplementary Table 1). While the median gene size of all the expressed genes

in the HeLa RNAseq was 23 kb, PCPAed genes had a median size of 39 kb. In contrast, non-PCPAed genes were much shorter, with a median size of 14kb (Figure 3.6a). This was not entirely unexpected, as larger genes stochastically contain more PASs due to their size, largely through an increase in intron length. Within our samples, I verified this intron expansion and PCPA correlation from the expressed genes and found it to be very high (Spearman correlation = 0.9994). It is also worth noting that intron number increases with gene size, although this does not correlate as strongly as intron size to PCPA, where over one third of all PCPA events occurred in the first or second intron.

A more striking result was discovered upon further examination of the smaller, PCPA resistant gene group. There were many genes that not only escaped PCPA, but exhibited increased expression levels without any apparent splicing deficit. These genes appeared to be the counter to those that were PCPAed; as a result, we explored the functional differences between the two sets of genes.

The large numbers for these groups made the use of GO software problematic, as its results were not reliable and resulted in a wide range of functional groups. As a consequence, we further stratified the groups by selecting the most up- and down-regulated genes from both groups. Specifically, I took the top 50% of genes as ranked by fold-change from the up-regulated or PCPAed groups, resulting in 493 and 3,134 genes, respectively. The select groups were more manageable, and the software XGR produced GO terms that were then easily summarized using a package called REVIGO (Fang et al. 2016; Supek et al. 2011). The non-PCPAed and up-regulated genes were highly enriched in cell proliferation, stimuli response, and transcription factors, indicating their use in cell stress response and general housekeeping. In contrast, PCPAed and down-regulated

genes were more diverse, showing enrichment for neuronal development, cell division, and DNA replication and repair (Figure 3.6b and c).

To get a sense of whether this size-function stratification was limited to PCPAed genes and their unaffected counterparts, I expanded the analysis to the entire human genome. I took the enriched GO term groups for the PCPAed and the up-regulated, non-PCPAed genes and extracted all the genes that are categorized by each functional term. The results indicated that the size-function relationship persisted regardless of the sensitivity to U1 AMO in our experiments. For example, tissue differentiation or tissue specific genes are highly enriched in large genes, while cell stimuli response and cell survival are found more often in smaller genes. This implies that PCPA avoidance, by keeping introns small so as to not stochastically generate many PASs, has played a role in the determination of gene size. By maintaining a small size, primary-response genes could be transcribed quickly and not worry about transcription loss due to PCPA. This theory will be discussed in more detail in the next chapter.

These results shown here elucidate the process and important nature of U1 telescripting. My pol II ChIPseq experiment demonstrated that transcript shortening is due to PCPA and is co-transcriptional. Additionally, my analytical work allowed us to globally identify PCPAed transcripts has opened the door for further scientific study into the specifics of U1 telescripting. The establishment of U1 as a transcriptional regulator based on gene size and function is, to my knowledge, the first of its kind. Through transcript length dependent PCPA suppression, U1 can regulate not only transcription, but also gene structure. I sought to examine this theory in more detail through a series of RNAseq experiments of my own, as well as the creation of additional analytical programs.

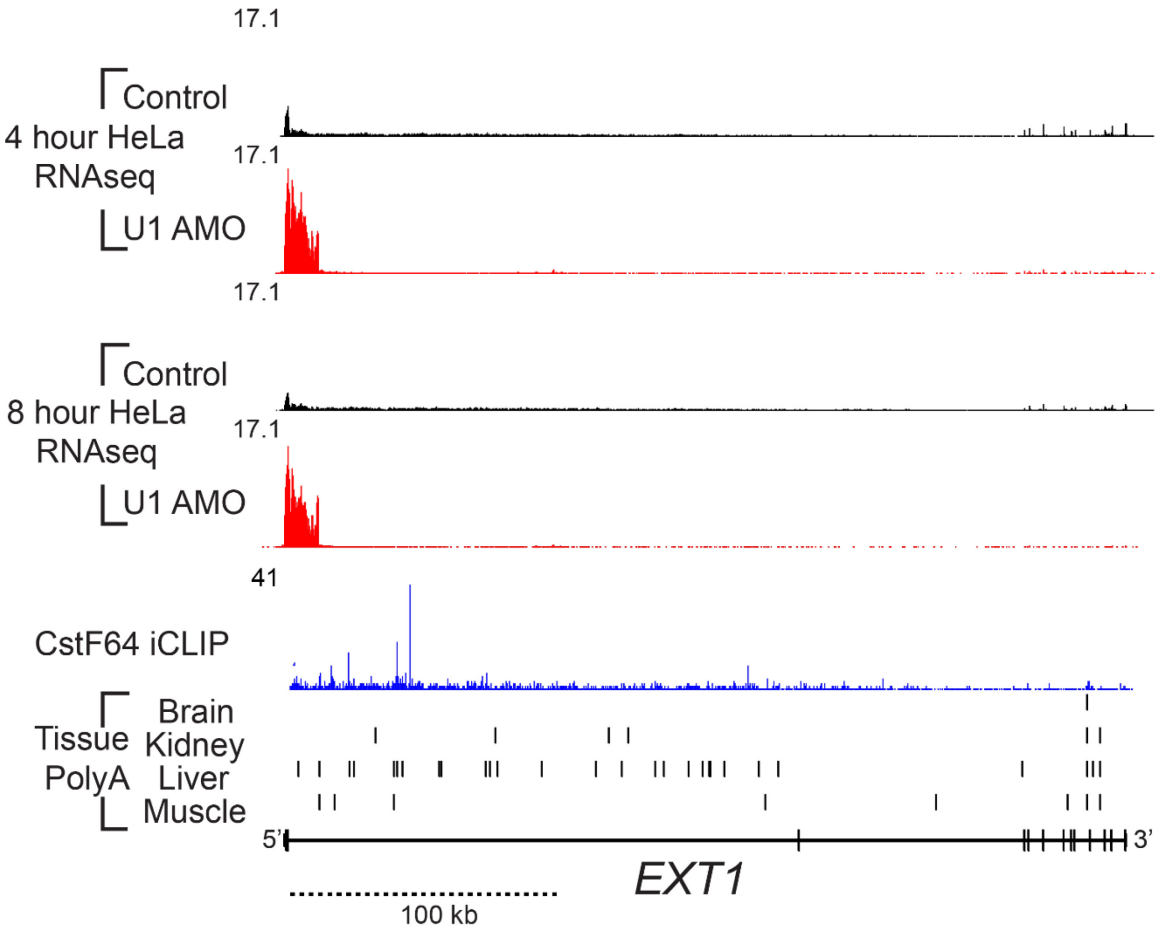

17.1
Control
4 hour HeLa
RNAseq
17.1
U1 AMO
17.1
Control
8 hour HeLa
RNAseq
17.1
U1 AMO
41
CstF64 iCLIP
Tissue
PolyA
Brain
Kidney
Liver
Muscle
5'
3'
EXT1
100 kb

Figure 3.1 U1 base pairing inhibition rapidly induces PCPA in 5' introns

Genome browser view of the gene *EXT1* with 4sU labeled RNA from HeLa cells transfected with either U1 (red) or control (black) AMO is shown as reads aligned to the human genome (hg19). PCPA is demonstrated by the abrupt termination of reads in the U1 AMO transfected tracks in the large first intron. The height of each RNAseq track is scaled to the same value in this figure to demonstrate the large accumulation of reads in the 5' region of the gene due to PCPA. CstF64 iCLIP binding site peaks are shown in blue, and poly(A) sites detected in various human tissues are shown as vertical black bars (Yao et al. 2012; Derti et al. 2012). Numbers to the left of the read distributions show the highest peak height value in the field as normalized to the total mapped reads. For gene structure, lines depict introns, boxes depict exons, and thinner boxes depict UTRs. Genomic distances are shown as dashed black lines.

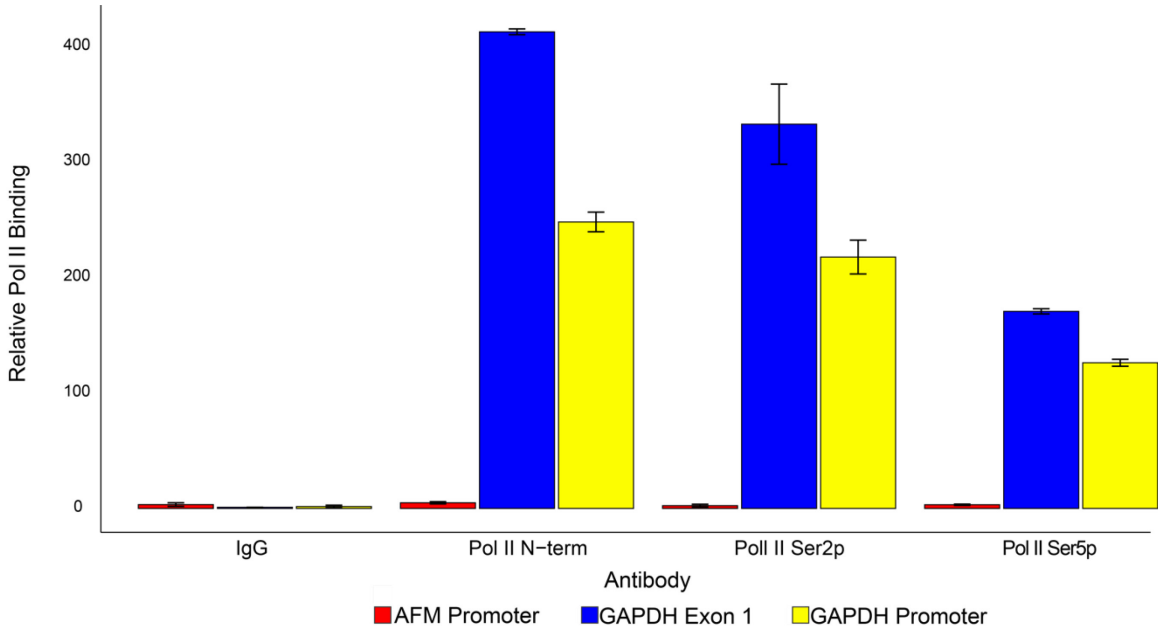

Relative Pol II Binding
400
300
200
100
0
IgG
Pol II N-term
Poll II Ser2p
Pol II Ser5p
Antibody
AFM Promoter
GAPDH Exon 1
GAPDH Promoter

Figure 3.2 Pol II ChIP RT-qPCR

Pol II binding counts to DNA from HeLa cells in GAPDH promoter (yellow) or exon 1 (blue) after formaldehyde cross-linking are shown normalized to the signal from the AFM promoter (red) which does not express in HeLa. Three pol II antibodies (N-terminus, Ser2p, and Ser5p) are shown, as well as IgG as a non-binding control. Error bars depict the standard error from the triplicate experiment.

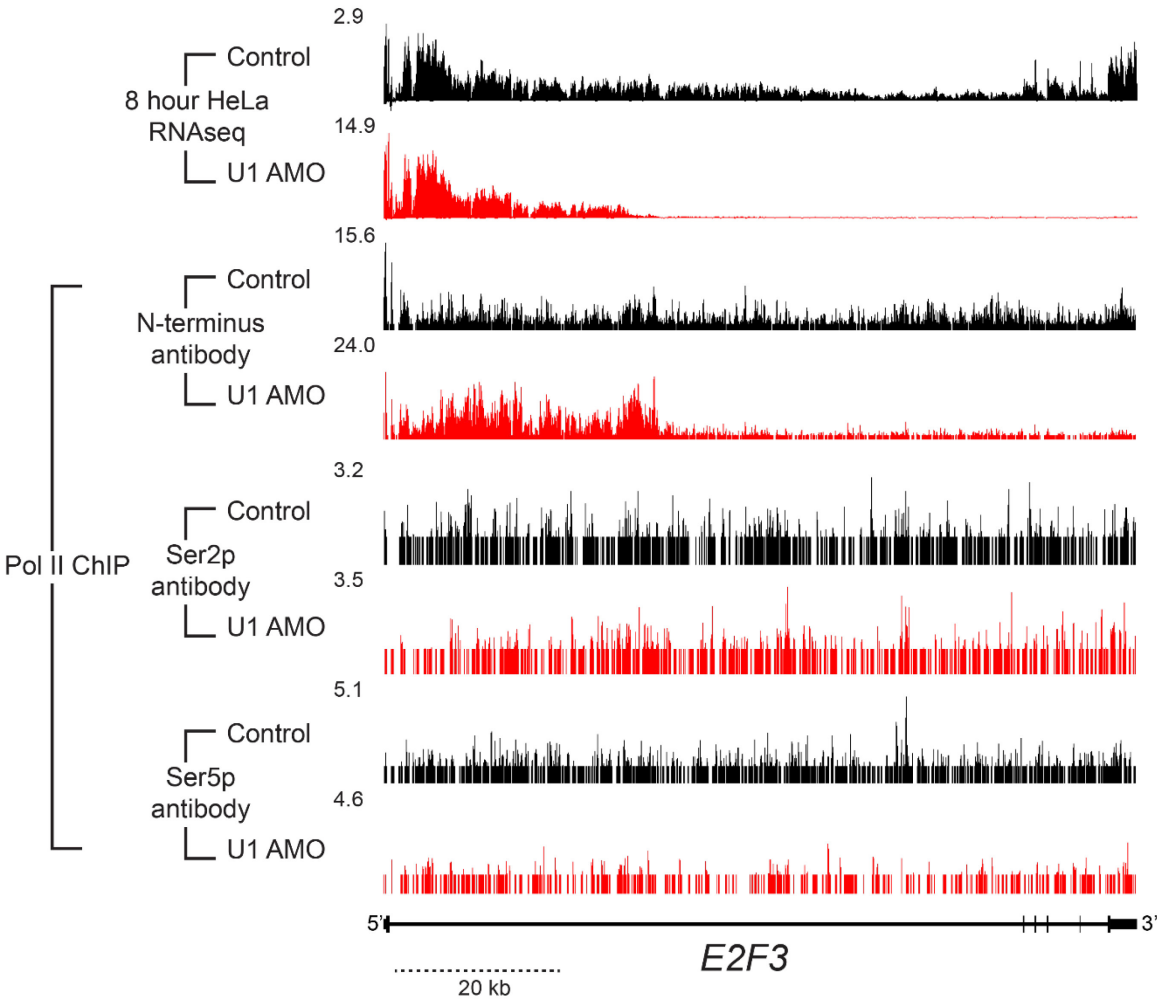

2.9
Control
8 hour HeLa RNAseq
14.9
U1 AMO
15.6
Control
N-terminus antibody
24.0
U1 AMO
Pol II ChIP
3.2
Control
Ser2p antibody
3.5
U1 AMO
5.1
Control
Ser5p antibody
4.6
U1 AMO
5'
3'
E2F3
20 kb

Figure 3.3 Pol II ChIPseq

Genome browser view of the gene *E2F3* with pol II ChIPseq data from HeLa cells transfected with either U1 (red) or control (black) AMO is shown as reads aligned to the human genome (hg19). Numbers to the left of the read distributions show the highest peak height value in the field as normalized to the total mapped reads. For gene structure, lines depict introns, boxes depict exons, and thinner boxes depict UTRs. Genomic distances are shown as dashed black lines. Pull-down data was normalized to the relevant IgG control (U1 or control AMO transfection) using MACS v1.4.4 (Y. Zhang et al. 2008).

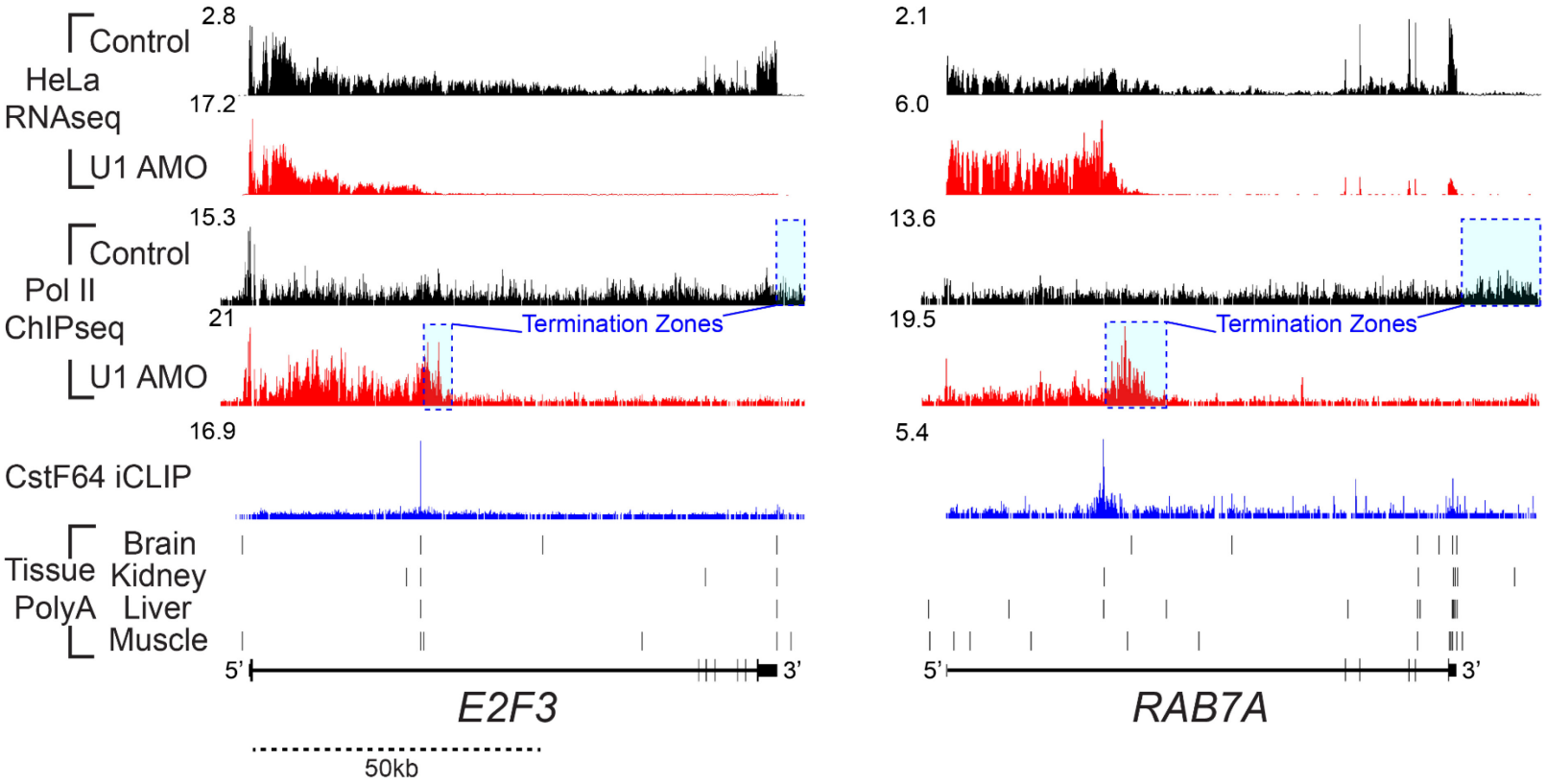

Control
HeLa
RNAseq
U1 AMO
Control
Pol II
ChIPseq
U1 AMO
CstF64 iCLIP
Brain
Tissue
Kidney
PolyA
Liver
Muscle
2.8
17.2
15.3
21
16.9
2.1
6.0
13.6
19.5
5.4
Termination Zones
Termination Zones
5'
3'
5'
3'
E2F3
RAB7A
50kb

Figure 3.4 U1 base pairing inhibition prematurely terminates pol II in gene bodies

Genome browser view of the genes *E2F3* and *RAB7A* with pol II ChIPseq and 30 minute, 4sU labeled RNAseq data from HeLa cells transfected with either U1 (red) or control (black) AMO is shown as reads aligned to the human genome (hg19). Blue boxes highlight the termination zone, and show the similar decline of pol II signal past the TES and PCPA point in control or U1 AMO transfected samples, respectively. CstF64 iCLIP binding site peaks are shown in blue, and poly(A) sites detected in various human tissues are shown as vertical black bars (Yao et al. 2012; Derti et al. 2012). Numbers to the left of the read distributions show the highest peak height value in the field as normalized to the total mapped reads. For gene structure, lines depict introns, boxes depict exons, and thinner boxes depict UTRs. Genomic distances are shown as dashed black lines. Pull-down data was normalized to the relevant IgG control (U1 or control AMO transfection) using MACS v1.4.4 (Y. Zhang et al. 2008). Figure is adapted from Oh *et al.* (Oh et al. 2017).

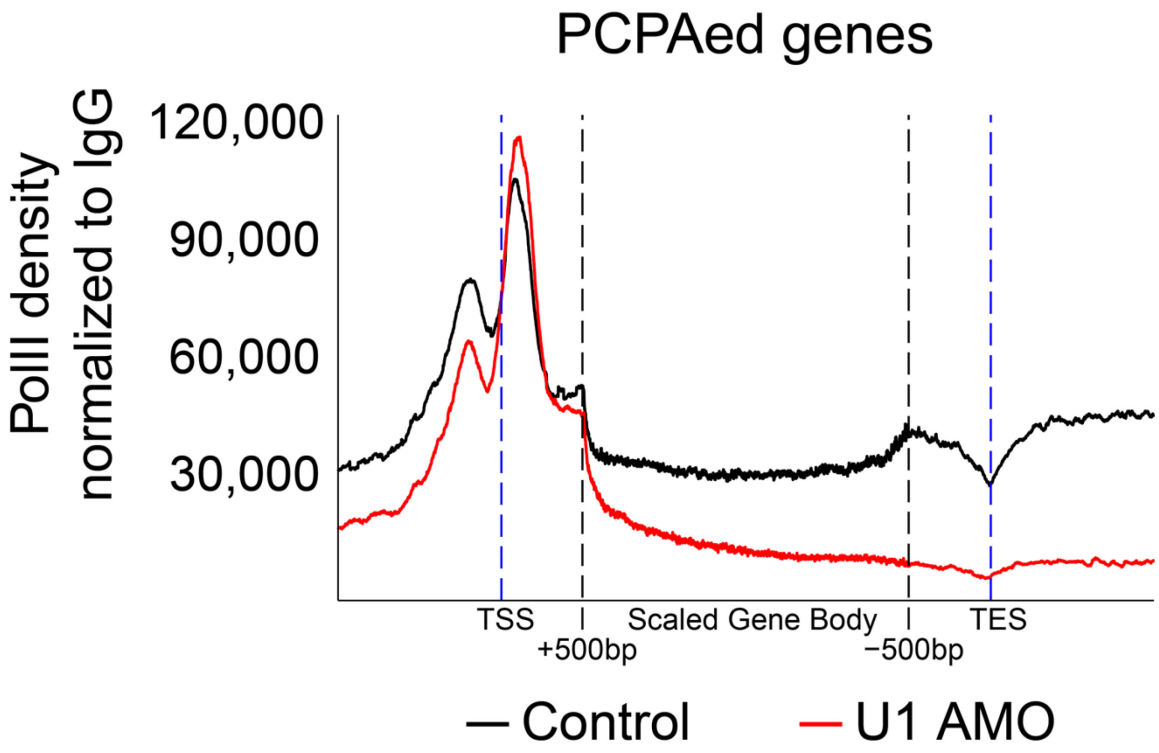

PCPAed genes
PolII density
normalized to IgG
120,000
90,000
60,000
30,000
TSS
+500bp
Scaled Gene Body
−500bp
TES
Control
U1 AMO

Figure 3.5 PCPA is co-transcriptional

Metagene plot of pol II ChIPseq data from PCPAed identified genes from either U1 (red) or control (black) AMO transfected HeLa cells. Plot is shown with absolute distances around the TSS (1,000bp upstream, 500bp downstream) and TES (500bp upstream, 1,000bp downstream), which are depicted as blue dashed lines. The remaining region of the gene body, between the two black dashed lines, is scaled to 2,000bp for all genes in the analysis. Pull-down data was normalized to the relevant IgG control (U1 or control AMO transfection) using MACS v1.4.4 and the metagene profile was generated using CGAT v0.2.5 (Y. Zhang et al. 2008; Sims et al. 2014). Figure is adapted from Oh *et al.* (Oh et al. 2017).

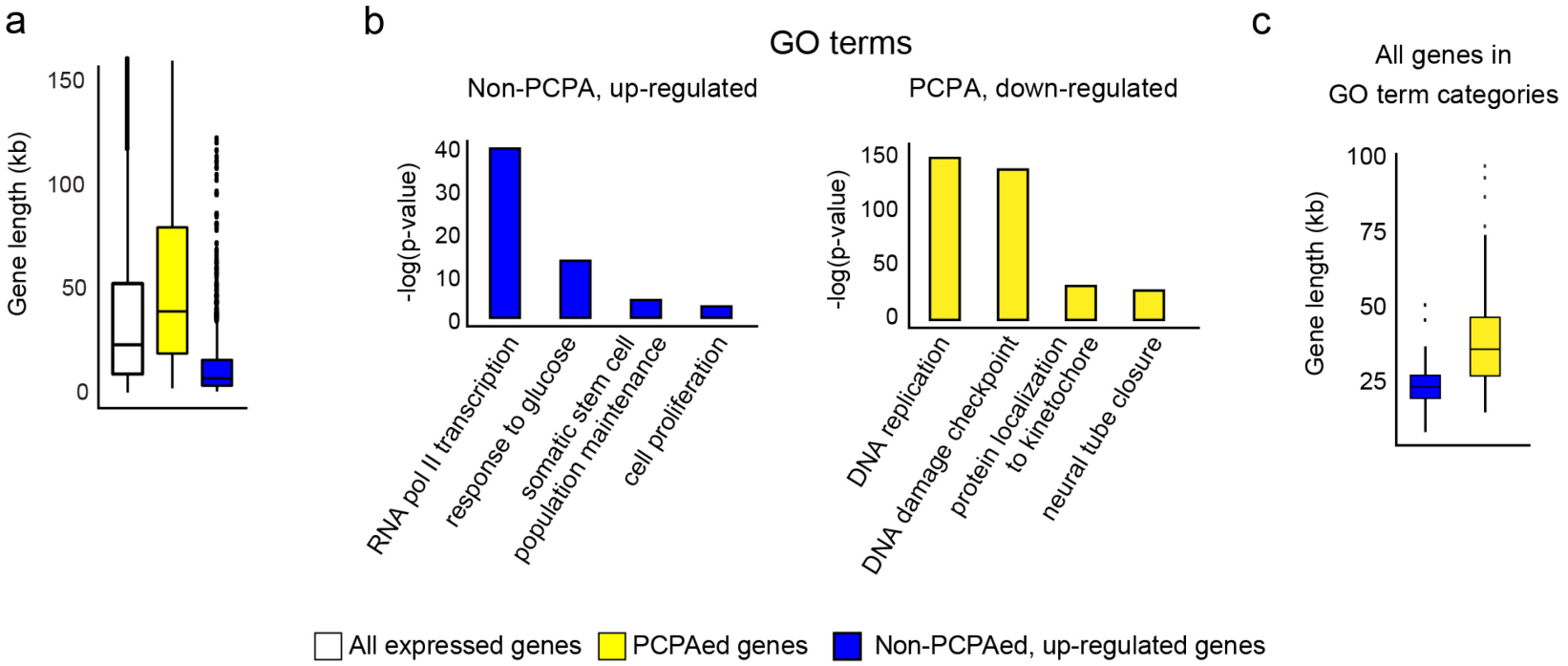

a
Gene length (kb)
150
100
50
0
b
GO terms
Non-PCPA, up-regulated
-log(p-value)
40
30
20
10
0
RNA pol II transcription
response to glucose
somatic stem cell population maintenance
cell proliferation
PCPA, down-regulated
-log(p-value)
150
100
50
0
DNA replication
DNA damage checkpoint
protein localization to kinetochore
neural tube closure
c
All genes in GO term categories
Gene length (kb)
100
75
50
25
All expressed genes
PCPAed genes
Non-PCPAed, up-regulated genes

Figure 3.6 Gene size and function are stratified by sensitivity to U1 base pairing inhibition

(a) Boxplots showing the gene size distribution in all expressed genes (white; RPKM ≥ 1; n = 9,744, median size = 22.8kb), PCPAed genes (yellow; n = 3,590, median size = 39.0kb), or non-PCPAed and up-regulated genes (blue; n = 988, median size = 6.8kb) from 30 minute, 4sU labeled and U1 AMO transfected HeLa cells compared to control. (b) Histogram showing the GO term enrichment ($p < 0.05$) for non-PCPAed and up-regulated or PCPAed and down-regulated genes. Enrichment was calculated using the XGR tool and further functional classification was performed using REVIGO (Fang et al. 2016; Supek et al. 2011). The $-\log_2$(p-value) for top 50% of genes as ranked by fold change for each group is shown. (c) Boxplots showing the gene-size distribution for all human genes classified in the functional groups depicted in b. Median size for functional groups enriched in non-PCPAed and up-regulated genes is 23kb, while median in PCPAed and down-regulated genes is 38kb. Figure is adapted from Oh *et al*. (Oh et al. 2017).

Antibody	Treatment	Input Reads	Mapped Reads	Mapping Percent	Multimapped Reads	Multimapping Percentage
IgG	Control	14,536,565	13,425,928	92.36%	1,148,578	8.55%
	5nmol U1 AMO	107,408,555	97,206,719	90.50%	8,850,480	9.10%
Pol II N-terminus	Control	54,576,053	52,673,238	96.51%	4,442,971	8.43%
	5nmol U1 AMO	55,640,783	52,291,207	93.98%	4,508,621	8.62%
Pol II Ser2p	Control	63,898,661	60,633,978	94.89%	5,400,916	8.91%
	5nmol U1 AMO	50,081,748	47,564,298	94.97%	4,150,087	8.73%
Pol II Ser5p	Control	46,864,571	44,132,302	94.17%	4,130,930	9.36%
	5nmol U1 AMO	45,081,305	42,746,731	94.82%	3,805,272	8.90%

Table 1 Pol II ChIPseq mapping statistics

Numbers for the hg19 read alignment of pol II ChIPseq using TopHat v2.1.1 (Trapnell, Pachter, and Salzberg 2009). Multi-mapped reads are defined as those which align to 20 or more locations in the genome.

CHAPTER 4: TELESCRIPTING IS EVOLUTIONARILY CONSERVED AND STRATIFIES GENOME BY GENE SIZE-FUNCTION

Chapter Details

Published in part in “U1 snRNP telescripting regulates a size-function-stratified human genome” (Oh et al. 2017). Contributing members: Jung-Min Oh and Chie Arai (RNAseq, 3’RACE); Byung-Ran So and Jiannan Guo (XLIPseq); Chao Di (help with metagene profiling and XLIPseq binding calculations).

Gene Size Conservation

The segregation of gene function by size was of interest, at first, due to our interest in a common gene structure found in higher eukaryotes: the majority of intronic sequence in a gene is from extremely large 5’ side introns, typically the first (Figure 4.1a) (Gelfman et al. 2012). We reasoned that this particular layout in many genes was due to the function of U1 telescripting, as the cryptic PASs in the large 5’ introns would be actionable for PCPA early in transcription. This would allow for PCPA to terminate pol II elongation within a matter of minutes, rather than other mechanisms of transcription regulation that could take hours to show cellular affects due to transcribing polymerases stuck in large genes, if the actionable PASs were instead clustered in large 3’ introns one or two hundred kb from the TSS. More specifically, we hypothesized that genes that would be immediately necessary for cell survival or activation in the event of specific stimuli or stress would maintain their small size due to evolutionary pressure to avoid PCPA from an actionable PAS. Genes that are not required for short-term cell survival, however, were selectively allowed to increase in size through intron expansion, which has been very significant to the gene structure in complex vertebrates (Catania and Lynch 2008; Rogozin et al. 2012). I decided to test this hypothesis using analytical methods from the RNAseq data that we currently

had in HeLa, discussed in the previous chapter, and additionally perform similar U1 AMO experiments on three different cell lines: S2 (*Drosophila melanogaster*, fruit fly), ZF4 (*Danio rerio*, zebrafish), and NIH-3T3 (*Mus musculus*, mouse) cells.

Gene Size Expansion

I first decided to try and understand how both the PCPAed and up-regulated genes had changed in size evolutionarily. To do this, I obtained gene homology data from Ensembl and compiled lists of gene sizes across multiple organisms referencing their orthologs that were found in humans. From these data, I extracted all of the genes from different groups as identified in our HeLa U1 AMO: PCPAed, PCPAed and down-regulated (PCPA-down), non-PCPAed (unchanged), non-PCPAed and up-regulated (UP), and all expressed genes (i.e. RPKM ≥ 1). I reasoned that the most impactful of these for my analysis would be the groups most affected from U1 base-pairing inhibition; as a consequence, I focused on the PCPA-down as well as UP gene groups, with a baseline comparison of all expressed genes (Supplementary Table 1). Plotting the distributions of gene size versus phylogenetic distance from human displayed a striking difference between these two groups. The UP genes remained smaller in size than the PCPA-down genes across all the organisms I tested. Moreover, there was a very minor change in their small size throughout the organisms, with a standard deviation of 3.9kb (1.5-11kb). In comparison, orthologous genes from the PCPA-down group showed a strong expansion of gene size across evolution, from 6kb in fly to 43.6kb in mouse with a standard deviation of 12.3kb (Figure 4.1b).

Taken together with the previous results of the functional differences between small and large genes, the marked distinction between the intron expansion in PCPA-down genes and the lack of such expansion in UP genes demonstrates selective evolutionary pressure

for gene size based on function. This provides an alternative solution to the current theories regarding intron size and the purpose of intron expansion, particularly in early introns. The most common of these theories are: the selection hypothesis, which holds that evolution has selected for compact genes to reduce transcription time and energy expenditure; and the genomic design hypothesis, which postulates that gene size is regulated by the need for time-dependent expression control and the inclusion of regulatory elements in introns (Rogozin et al. 2012). Selective intron expansion, to benefit small genes, seems more likely when one takes into consideration our earlier experiments with cell activation, which demonstrated mild PCPA and 3'UTR shortening in some select, large genes after transcription up-regulation (Berg et al. 2012). This physiologically relevant experiment suggested that large genes could be sacrificed in the event of cell activation or the stress response, in order to allow, or even facilitate, small genes to be expressed when needed.

Large Genes Sequester RNA Binding Proteins

Concurrent to my analytical work on the expansion of intron size, Byung-Ran So and Jiannan Guo, two post-doctoral colleagues in the lab, performed a series of high (5nmol) U1 AMO experiments with low concentration formaldehyde crosslinking and immunoprecipitation (XLIP) followed by both high-throughput sequencing (XLIPseq) and mass spectrometry (MS). This ribo-proteomic strategy takes advantage of formaldehyde's ability to induce protein-protein and protein-RNA crosslinking to capture both the protein composition and stoichiometry of complexes associated with the target protein, as well as to determine their binding location on RNA (Yong et al. 2010). In order to verify the specificity of each antibody and the stringency of the wash steps, I compiled the sequences of all known spliceosomal snRNA isoforms and used these as a reference to align the XLIPseq data (Figure 4.2a). There are a large number of snRNA gene copies

and pseudogenes found in the genome with only minor differences between them, often changes of only 1-2 base pairs. As such, during normal alignment and processing, many reads associated with these snRNAs are discarded as multi-mapping reads due to the alignment algorithms. By aligning directly and specifically to the snRNA sequences, I was able to more accurately represent the amount of snRNPs included in each of the pull-downs. This alignment method worked well to demonstrate the validity of the experiment, as spliceosomal proteins demonstrated strong enrichment in snRNA reads with their known interactors. Conversely, non-spliceosomal proteins, such as hnRNPs or the SP2/0 control antibody, did not show high snRNA read enrichment as expected.

While there were many different antibodies used to target a range of RBPs, I examined the general differences between two groups: the hnRNPs (hnRNPA1 and hnRNPC) and the splicing factors (SFs) (SF3B1, U2B", U1A, and U1C). Because of their separate, general binding locations on genes seen in the XLIPseq. hnRNPA1 and hnRNPC interact with pre-mRNA and bind preferentially to introns, whereas SFs are seen near splice sites and across exons (Figure 4.2b). I also included in my calculations the input RNA, a non-specific SP2/0 antibody pull-down, and 5 minute 5-ethynyl uridine (EU) pulse-labeled RNA, both rRNA depleted (total) or poly(A) selected (mRNA) as control samples. The latter is helpful for comparison because the shorter labeling time is more similar to the 10-minute formaldehyde cross-linking time as opposed to the longer (30 minute or 2 hour) labeling times used in earlier experiments. Visualization on the UCSC Genome Browser showed the specific binding locations of the protein binding, as hnRNPC was present across nearly the entire pre-mRNA strand, in comparison to the exon-centric SFs (Figure 4.3a). Metagene analysis across splice sites provided further granularity, demonstrating that hnRNPC was not only enriched in introns, but excluded from exons (Figure 4.3b).

This result supports the theory that the exon definition complex, i.e. SFs and snRNPs, evicts hnRNP proteins prior to splicing (Wongpalee et al. 2016).

In order to determine whether hnRNP and SF binding contributed to mRNA output, I normalized the binding signal for these proteins based on the mRNA output for each gene as measured by exon reads from the EU RNA. When broken down by gene size into 10 bins of equal gene number, it was clear that larger genes bound a significantly large pool of hnRNPs. Specifically, the largest 20% of genes bound >50% of the total hnRNPC on pre-mRNA, but these genes did not exhibit a proportional increase in mRNA level (Figure 4.4a). In smaller genes, hnRNPs were under-represented in comparison to SFs when normalized to the mRNA. A potential (and tempting) explanation for this result is that the large introns sequester several processing factors during transcription due to their larger size and higher number of exons. However, the fact that this increased binding does not stimulate mRNA production suggests that these excess proteins may not be necessary for transcription, but instead may serve another function. For instance, during cell stimulation or activation, both of which cause minimal PCPA and 3'UTR shortening, as explained earlier, the 3' transcription loss from PCPA would then free up a portion of these processing factors that would normally be bound to the RNA of large genes. Not only would this affect hnRNPs, due to the loss of large introns, but SFs as well, due to the higher density of exons in the 3' ends of genes (Bradnam and Korf 2008; Gelfman et al. 2012). Thus, large genes serve, additionally, as sponges, or as a reserve of various RNA binding proteins. Under stress or cell activation conditions, these proteins would be released from PCPAed transcripts and could be utilized by small, acute response genes to boost their transcriptional output.

In order to investigate how much protein binding is lost after PCPA, I calculated the change in sequencing depth normalized reads between control and U1 AMO samples for each XLIP. For this calculation, I used only genes that experienced substantial or near complete PCPA, as determined by a >90% decrease in last exon versus first exon read signal from the EU labeled RNA. Breaking the data down into the gene size bins as used above showed that there was significant loss of hnRNPC binding (>10 times) in the largest gene group when compared with those at median size (24kb) or lower (Figure 4.4b). PCPA induced U1C binding loss, by comparison, was not affected by gene size. Taken together, these observations suggest that large genes hoard certain pre-mRNA processing proteins, specifically hnRNPC.

Recent research has demonstrated not only that RNA processing factors are limited in quantity within a cell, but also that their levels can be insufficient for splicing and processing in certain instances such as meiosis, transcriptional up-regulation, or sequestration due to micro-satellite repeats (Miller et al. 2000; Munding et al. 2013). In comparison to these instances, transcription of large genes can bind many more RBPs given their size alone and yet this resource hoarding is constant in many cells. The sensitivity of large genes to U1 levels relative to transcription, for example in neuronal activation, that leads to even a moderate amount of PCPA could free up many RNA processing resources for use by other, smaller genes (Berg et al. 2012; Oh et al. 2017). Given the size-function stratification mentioned previously, this further supports a mechanism by which non-vital large gene transcription is sacrificed through PCPA during cell stimulation to provide a rapid boost to acute response genes. The common asymmetrical architecture of genes with higher 5’ intron density and 3’ exon density would maximize this mechanism when large genes PCPA early and thereby shorten the lag time in resource turnover.

Organismal RNAseq

Armed with the indirect analytical results that demonstrated intron expansion in PCPAed genes, it was important that I directly examine whether global U1 AMO induced PCPA was isolated to HeLa cells. While we had previously shown that this intron expansion occurs in fruit fly and mouse, these data were not directly comparable to our current RNAseq data and were limited in breadth due to the technology (Berg et al. 2012). To expand on our earlier research, I chose to revisit the mouse and fruit fly systems using the primary NIH-3T3 fibroblast and S2 embryonic cell lines, respectively. I also decided to use a zebrafish cell line, ZF4 embryonic fibroblasts, as its genome was well researched, there was already well-established work with morpholinos, and a few recent spliceosome studies had used zebrafish (Nasevicius and Ekker 2000; Trede et al. 2007; Rösel et al. 2011). In order to be consistent with previous experiments in HeLa cells, I included two U1 AMO time points, at 6 and 8 hours. In addition, I examined if there were any early U1 level sensitive genes in these organisms. I also tested three labeling time points (30, 60, and 120 minutes) so as to be able to isolate both a more accurate snapshot of nascent transcription with the shorter label and also test the data against our older, 120 minute labeled samples.

Before I began my experiments, lab protocol for nascent RNA isolation called for the removal of rRNA prior to 4sU labeled RNA purification. This works well in the highly active HeLa cell lines that were our primary experimental foundation *in vitro*. As these are a highly aggressive and metabolically active cancer cell line, there has never been an issue in the lab with acquiring usable amounts of RNA for experiments using this approach. During the course of my tests with alternative, less active cell lines, such as fibroblasts, I discovered that there was less labeled RNA over the same time period in comparison to HeLa cells. Combined with a maximum input limit for the rRNA removal kits, primarily due

to the binding limit of anti-rRNA sequences on the beads, I would have required extraordinary amounts (10-20x) of harvested cells in order to acquire enough labeled RNA for research purposes. Instead, I reversed the purification procedure by first selecting for the labeled RNA. This was achievable due to the stringent and multi-step washing procedure used in the 4sU selection. In doing so, I was able to achieve greater amounts of purified RNA for both PCR and RNAseq from my quiescent cell lines. RNA quality and purity was verified using an Agilent Bioanlyzer, which confirmed the efficacy of this new method.

I verified the efficacy of the U1 AMO treatments in mouse and fly using RT-qPCR on genes identified as PCPAed from our earlier HIDE-seq experiments. For these validations, I used a 5nmol U1 and control AMO dose to induce strong PCPA and to remain consistent with the previous experiments done in HeLa cells. I targeted specific gene regions by using primers to the first intron upstream of the PCPA site as well as either the junction between exon 1 and exon 2 or both exons separately (Figure 4.5; full data in Table 2). Most time points in the mouse system showed increased intronic signal compared to exon signal after U1 AMO treatment, indicating PCPA within the target genes. Notably, there were a few outliers that exhibited minimal change or even increased exonic signal over the intron. This is most likely due to incomplete purification of 4sU labeled RNA, which would result in contaminant mRNA that had not yet been turned over within the cell. Two genes used in the fly sample showed clear PCPA, however the gene *Ten-m* had increased exonic signal. The latter utilized separate upstream and downstream exon primers for the PCR, as an exon-exon junction spanning primer set was suboptimal in GC content and melting temperature. As a result, it is unclear whether this increase in exonic signal is due to a failed U1 AMO treatment, signal aggregation due to PCPA within the first and second exon regions, or if the gene even PCPAs at all.

Given the overall positive RT-qPCR results, I decided to proceed onto library preparation and RNA-sequencing with the 6 hour U1 AMO treatment time points using 30 minute, 4sU labeled RNA. Along with the high, 5nmol dose described earlier, I included a low, 1nmol U1 AMO treatment in order to check for 3'UTR shortening as had been described previously (Berg et al. 2012). After sequencing, the reads were cleaned of any adapter sequence and aligned to the most up-to-date reference genomes for fly, zebrafish, and mouse (dm6, danRer10, and mm10, respectively) as described previously (Oh et al. 2017). Multi-mapped reads were removed, resulting in between 50 and 175 million mapped reads per sample (Figure 4.6; Table 3). For downstream analysis, read signal for each sample was normalized to its sequencing depth as mapped reads per million (RPM).

Initial analysis of the data revealed that high and low U1 AMO concentrations induced extensive down-regulation in mouse and zebrafish, consistent with the results from our previous experiments in HeLa cells (Table 4). Both fly and mouse presented moderate numbers of up-regulated genes as well, although zebrafish had many fewer than I expected. A potential explanation for this would be a loss of RNA material during the purification and selection process, leading to increased sequencing of background, intergenic RNAs. However, this was not reflected in the alignment distribution of the reads. Instead, the distribution showed only a slight increase in intronic and intergenic reads after U1 AMO treatments. Further study into the mRNA output from zebrafish showed that both U1 AMO doses decreased the overall splicing significantly more than either fly or mouse (Figure 4.7). While I understood that U1 base-pairing inhibition would reduce splicing, the level to which it was seen in zebrafish, and only in zebrafish, was surprising. Depicting the read distribution in a metagene analysis revealed global loss of transcription in the 3' end of exons and introns in zebrafish when compared to mouse and fly (Figure 4.8). As our PCPA detection algorithm is quite stringent, it is highly likely that a large number of genes

experience low levels of background PCPA that escape our statistical and transcription decrease cutoffs. This would reduce the overall splicing amount and number of genes detected as up-regulated.

In the fly system, it was apparent that the response to the U1 AMO dose was much more attenuated than in the other organisms. When viewing these data in comparison to zebrafish, mouse, and human models, moderate PCPA and down-regulation could be explained either by incomplete U1 AMO treatments or that telescripting and transcription in general is not as dependent on U1 levels. Past experiments in the lab, confirming the AMO dose response for U1 5' sequence occupancy, did not include a fly cell It is also possible that the transfection I performed was not of very high efficiency. As the mouse and zebrafish AMO treatments worked well, I proceeded with downstream analysis in all samples.

Visualization of the data using the UCSC Genome Browser confirmed the expected PCPA with high, 5nmol U1 AMO in zebrafish and mouse at similar levels as those observed in HeLa cells, but the fruit fly sample showed very minimal PCPA. Global analysis using the calculations described previously confirmed these results (Table 4). Moreover, analyzing the gene sizes for the up group compared to those that are PCPA-down showed the same size stratification as we had published previously (Figure 4.9; numerical data in Table 5) (Oh et al. 2017). It is important to note that the size of PCPA-down genes increases from fly to zebrafish to mouse, from 16kb to 23kb and 41kb, respectively. In comparison, up genes increased only 1.5kb to 11kb and 7kb across the same organisms, while expressed genes increased from 2kb to 12kb and 17kb. This was expected based on my previous analytical work examining orthologs from the human PCPA-down genes. This size increase may be the primary explanation behind the lack of PCPA in fly, which have much

smaller genes overall. The median size of PCPA-down genes in the other organisms, 20-43kb, is significantly larger than what is seen in fly, 6-16kb.

To examine the conservation of U1 level sensitivity, I compared overlap of the high U1 AMO induced PCPA-down genes between these three organisms against those in human using ortholog data from Ensembl (Table 4). These results showed a significant level of overlap between the PCPA-down genes (30-44%, hypergeometric test p-value < 0.05) for mouse and zebrafish; fly, however, showed non-significant minimal overlap to human (17-20%, hypergeometric test p-value 0.099-0.394). Moreover, when I visually examined some of these overlapping genes on the genome browser, the PCPA site occurred at a similar location within the gene across organisms (Figure 4.10). The ortholog overlap in up genes was significant for all samples (3-34%, hypergeometric test p-value < 0.05), when compared to human. When taken together, these data demonstrate the evolutionarily conserved function of telescripting. The conservation in both PCPA-down and up also strengthens the theory that the boost in small, acute response genes is possible, at least in part, by the transcription decrease from telescripting loss.

Gene Neighborhoods

When examining the PCPA-down genes in mouse on the UCSC Genome Browser, I noted the prevalence of small, up-regulated genes nearby (that is, ~5-50kb away) (Figure 4.11). Given the above conclusions from the hnRNP XLIP and gene size-function analysis, I postulated that the two, contrasting gene groups were positioned in close proximity globally within the genome. To address this, I searched for the closest nearby transcript to each up-regulated gene. I decided to use these as the reference point rather than the PCPA-down genes, as I reasoned this would better demonstrate whether the up-regulation in small genes could be explained from the loss of transcription in PCPA-down

genes. As comparisons, I also did the same analysis for up-regulated genes to other up-regulated genes (UP-UP) and up-regulated to down-regulated (UP-DOWN). For a control, I randomly sampled genes on each chromosome, with the number weighted based on the number of genes each chromosome contains. Calculating the median intergenic distance between these groups demonstrated that up-regulated compared to PCPA (UP-PCPA) and UP-DOWN were significantly closer than both the control and UP-UP distances (Figure 4.12; numerical data in Table 6).

The shorter intergenic distances between UP-DOWN genes was surprising and consistent among all species. This provides more support for the resource sequestration idea suggested earlier; that a transcription decrease of any kind may contribute to up-regulation in nearby genes by freeing up RNA processing factors. Genes unaffected by the mechanism of decrease, in our case by being insensitive to U1 base pairing inhibition, are able to take advantage of resources freed up by neighboring genes. Of course, this does not imply that genes would be required to be in close proximity in order to benefit from these resources, as there are examples of both PCPA-down and up genes that are not near to one another. The generally smaller intergenic distances, however, support the idea that adjacency is beneficial to the small genes, most likely by reducing the distances over which the freed resources must disperse before being available for use. As these calculations only take into account linear distance, and not the three-dimensional space that results from chromatin looping, they act only as a rough measurement for proximity in the nucleus.

Upstream Antisense Transcription

Recently it has been shown that promoter directionality is controlled by U1, as sense direction DNA (i.e. downstream of the TSS in the gene body) contain a higher ratio of

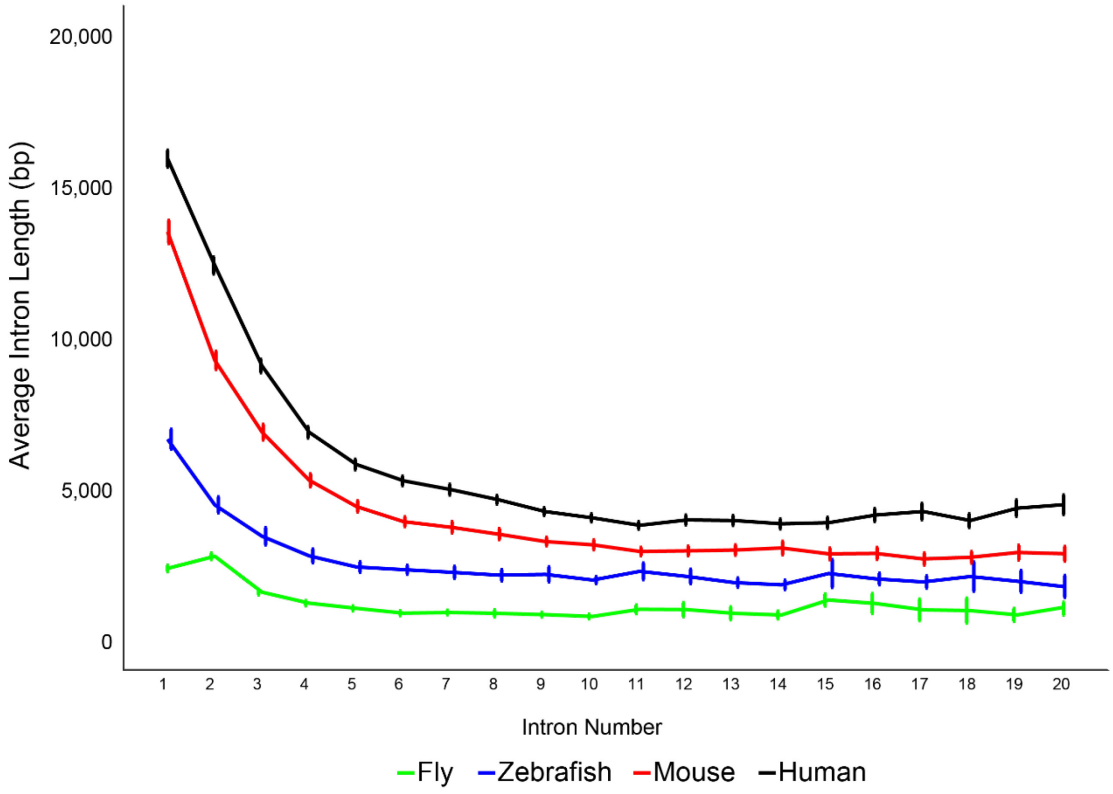
strong U1 binding sites to PASs (Almada et al. 2013). Conversely, the region upstream antisense (UA) of the TSS contains a higher number of strong PASs and fewer U1 binding sites. This allows for early transcription termination in the UA direction, while preserving telescripting within the gene body. The method by which I compiled the analytical work for gene neighborhoods made it very simple to adapt the protocol to examine UA transcription in my organismal RNAseq samples or any other data from the lab. To do this, I extracted a 1kb UA region from all protein coding genes that did not contain another overlapping, annotated transcript. I filtered out genes with previously annotated upstream transcripts because I worried that these would be under the control of their own, independent promoter. Roughly half of expressed genes across all organisms tested contained a UA protein coding gene or non-coding RNA (ncRNA).

When taking into consideration the pol II ChIPseq, I anticipated that U1 AMO induced PCPA would increase the UA transcription since it caused a slight increase in pol II initiation and promoter proximal pausing. To verify this, I analyzed all UA regions broken down by the status of the sense gene transcript after U1 AMO into either: unchanged, up-regulated, down-regulated but not PCPAed, or PCPAed. The results verified my hypothesis and also demonstrated that sense direction transcription throughput is indicative of UA transcription as down-regulated genes showed a strong decrease in UA signal and vice-versa for up-regulated genes (Figure 4.13). This supports the recent studies done on divergent transcription, which claim that most promoters allow polymerases to engage bi-directionally and that the U1 or nucleosome structure then dictate which direction transcription is allowed to continue (Seila et al. 2009).

My experiments with U1 AMO in multiple organisms have further confirmed the presence of U1 telescripting outside of the human system. More importantly, they have

demonstrated the evolution of gene size-function stratification caused by U1. My analysis of disproportionate hnRNP binding, specifically hnRNPA1 and hnRNPC, and the shift of these RNA processing resources from large genes to small genes with U1 AMO, suggest a new explanation as to how the small genes increase mRNA output with U1 base pairing inhibition. This explanation is further supported by the proximity of these small genes to large, RBP rich and U1 AMO sensitive large genes in linear genomic space. The work I have done here helps to cement U1 as a global regulator of transcription, as well as a potential regulator of gene and genome structure, through its function in telescripting.

a


20,000
15,000
10,000
5,000
0
Average Intron Length (bp)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Intron Number
Fly
Zebrafish
Mouse
Human

b

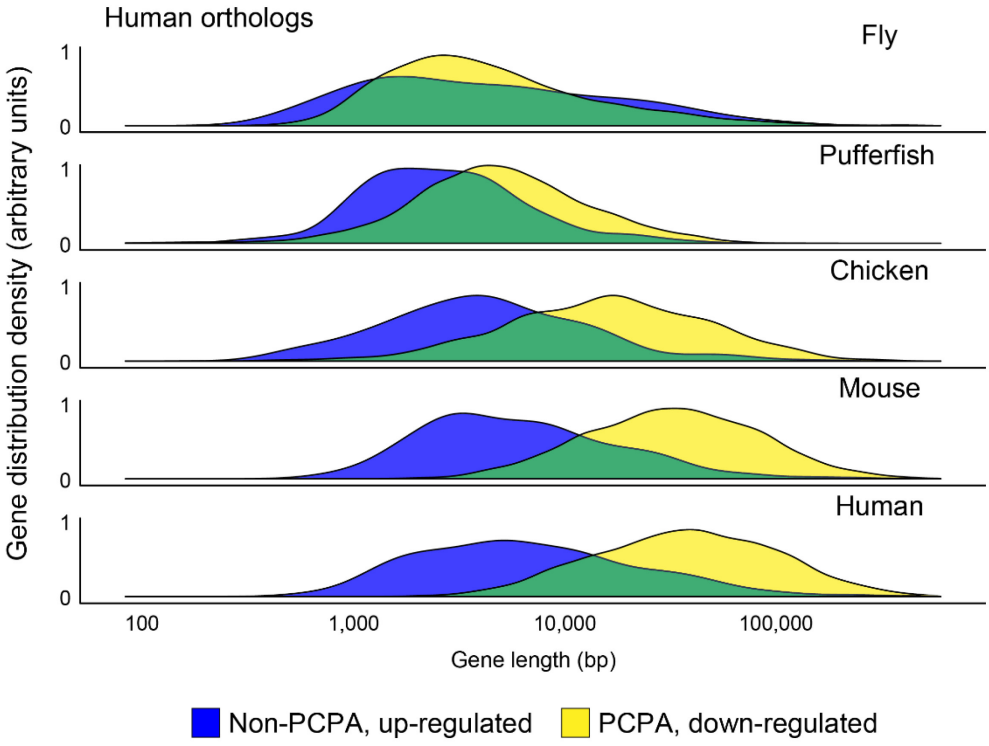

Human orthologs
Fly
Pufferfish
Chicken
Mouse
Human
1
0
Gene distribution density (arbitrary units)
100
1,000
10,000
100,000
Gene length (bp)
Non-PCPA, up-regulated
PCPA, down-regulated

Figure 4.1 U1 telescripting facilitated first intron and gene size expansion

(a) Line graphs showing the average intron size in base pairs for all genes in *Drosophila melanogaster* (Fly; green; dm6), *Danio rerio* (Zebrafish; blue; danRer10), *Mus musculus* (Mouse; red; mm10), and *Homo sapiens* (Human; black; hg19) genomes. Only the first 20 introns of each gene are shown. Vertical lines at each position represent the confidence interval. (b) Density plots showing the gene size distributions for human gene orthologs for genes that were either not-PCPAed and up-regulated (blue) or PCPAed and down-regulated (yellow) after 5nmol U1 AMO transfection in HeLa cells. Distributions are shown for *Drosophila melanogaster* (Fly), *Takifugu rubripes* (Pufferfish), *Gallus gallus* (Chicken), *Mus musculus* (Mouse), and *Homo sapiens* (Human). Figure panel adapted from Venters *et al.* (Venters et al. 2019).

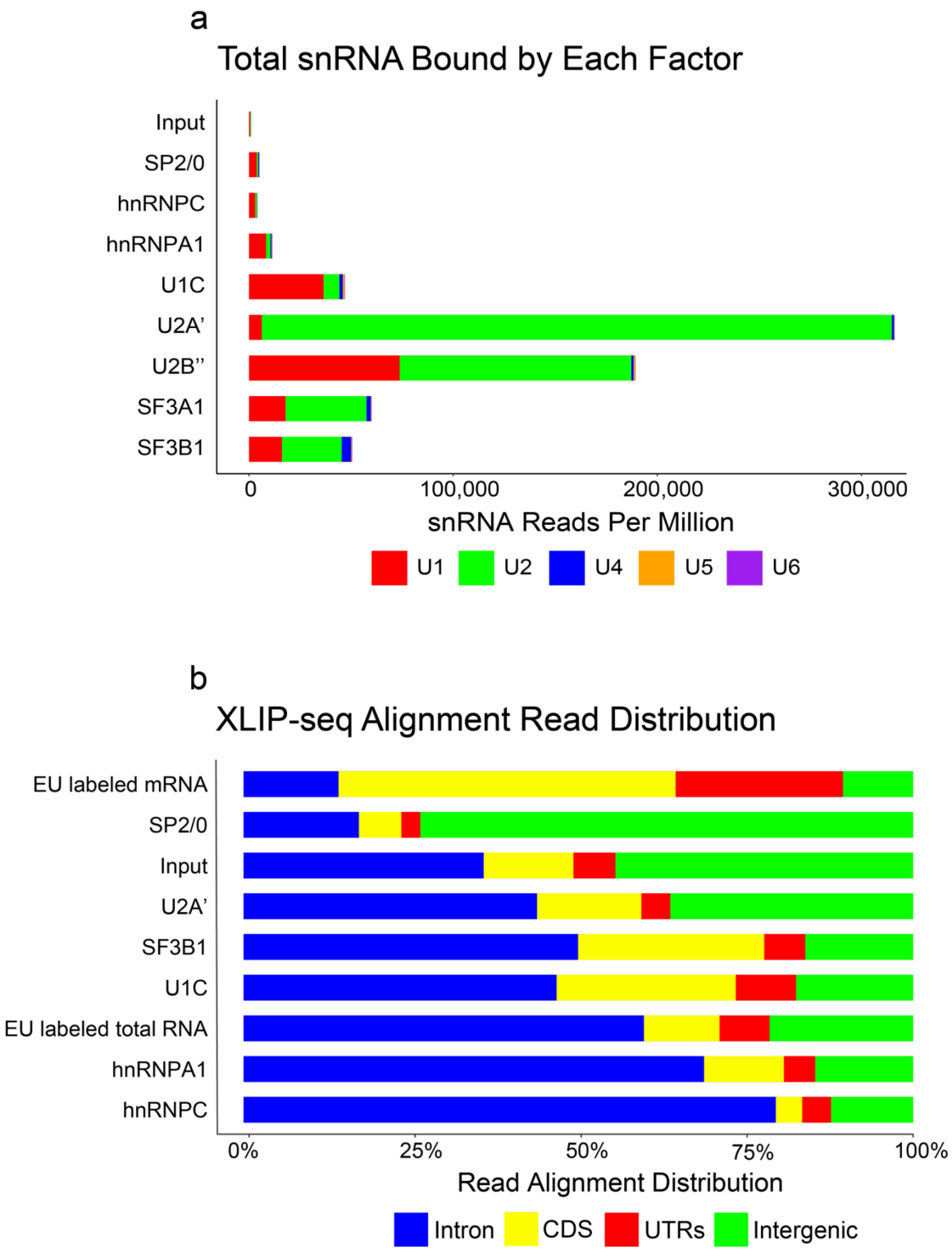

a
Total snRNA Bound by Each Factor
Input
SP2/0
hnRNPC
hnRNPA1
U1C
U2A'
U2B''
SF3A1
SF3B1
0
100,000
200,000
300,000
snRNA Reads Per Million
U1
U2
U4
U5
U6
b
XLIP-seq Alignment Read Distribution
EU labeled mRNA
SP2/0
Input
U2A'
SF3B1
U1C
EU labeled total RNA
hnRNPA1
hnRNPC
0%
25%
50%
75%
100%
Read Alignment Distribution
Intron
CDS
UTRs
Intergenic

Figure 4.2 XLIPseq alignment validates cross-linking procedure

(a) Stacked bar graphs showing the number of reads aligned to each collection of spliceosomal snRNA sequences (U1, red; U2, green; U4, blue; U5, orange; U6, purple) normalized to the total mapped reads (hg19 and snRNA reads combined). (b) Stacked bar graphs showing the distribution of XLIPseq and RNAseq read alignment locations (intron, blue; CDS, yellow; UTR, red; intergenic, green) as a percentage of the total aligned reads.

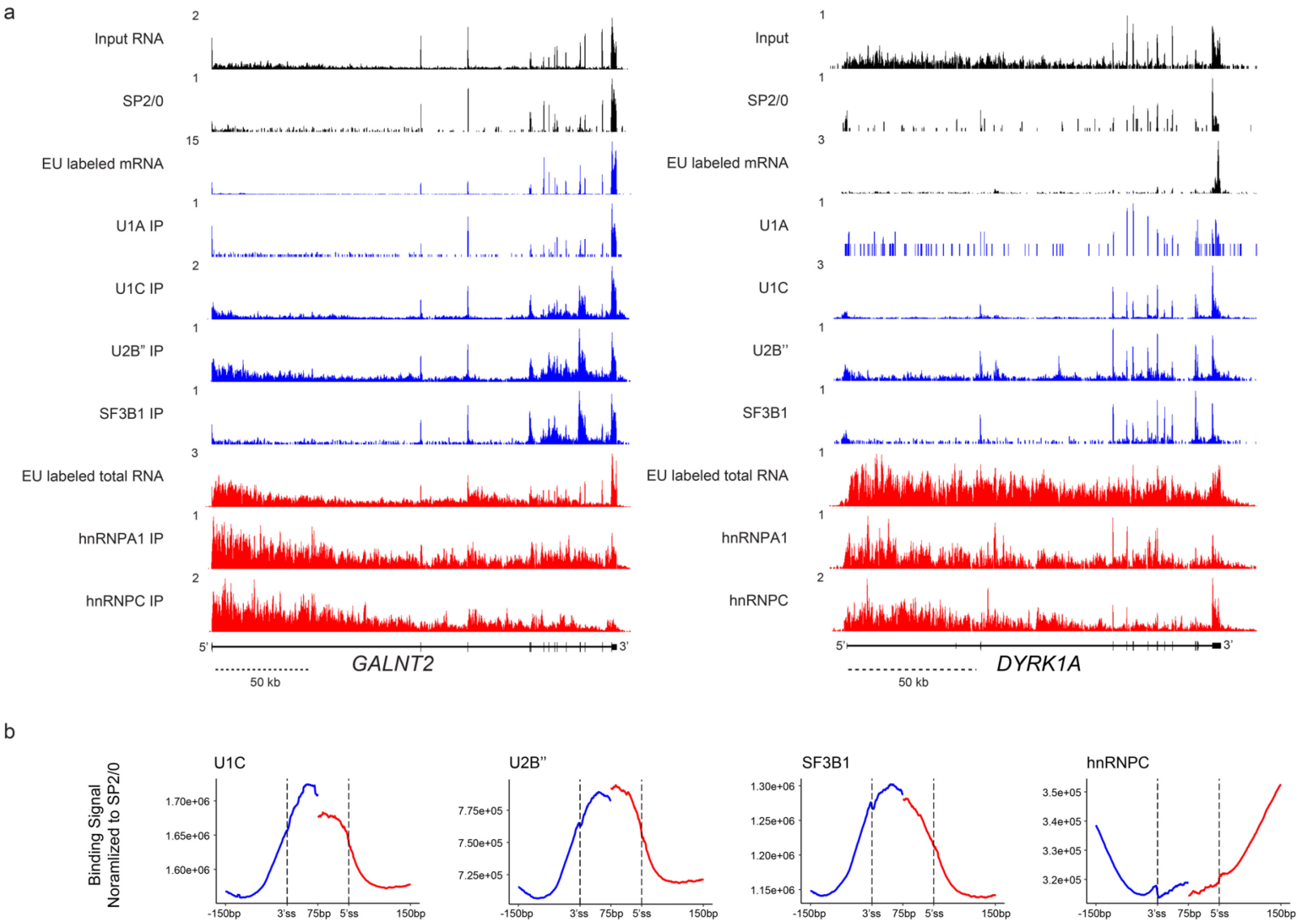

a
Input RNA
SP2/0
EU labeled mRNA
U1A IP
U1C IP
U2B" IP
SF3B1 IP
EU labeled total RNA
hnRNPA1 IP
hnRNPC IP
5'
3'
GALNT2
50 kb
Input
SP2/0
EU labeled mRNA
U1A
U1C
U2B''
SF3B1
EU labeled total RNA
hnRNPA1
hnRNPC
5'
3'
DYRK1A
50 kb
b
Binding Signal
Noramlized to SP2/0
U1C
U2B''
SF3B1
hnRNPC
-150bp
3'ss
75bp
5'ss
150bp

Figure 4.3 hnRNP proteins are excluded from exons

(a) Genome browser views for the genes *GALNT2* (left) and *DYRK1A* (right) with XLIPseq (IP) or RNAseq (RNA) read distributions for control samples (black), mRNA/exon centric binding factors and poly(A) selected mRNAseq (blue), or pre-mRNA binding factors and total RNAseq (red) from HeLa cells aligned to the human genome (hg19). Numbers to the left of the read distributions show the highest peak height value in the field as normalized to the total mapped reads. For gene structure, lines depict introns, boxes depict exons, and thinner boxes depict UTRs. Genomic distances are shown as dashed black lines. (b) Metagene plots for XLIPseq read distributions across 3' splice sites (blue) or 5' splice sites (red). Read signal was normalized to SP2/0 and metagene profile was generated using CGAT v02.5 (Sims et al. 2014). Plotted region includes the 150bp intronic region and 75bp exonic region around each splice site.

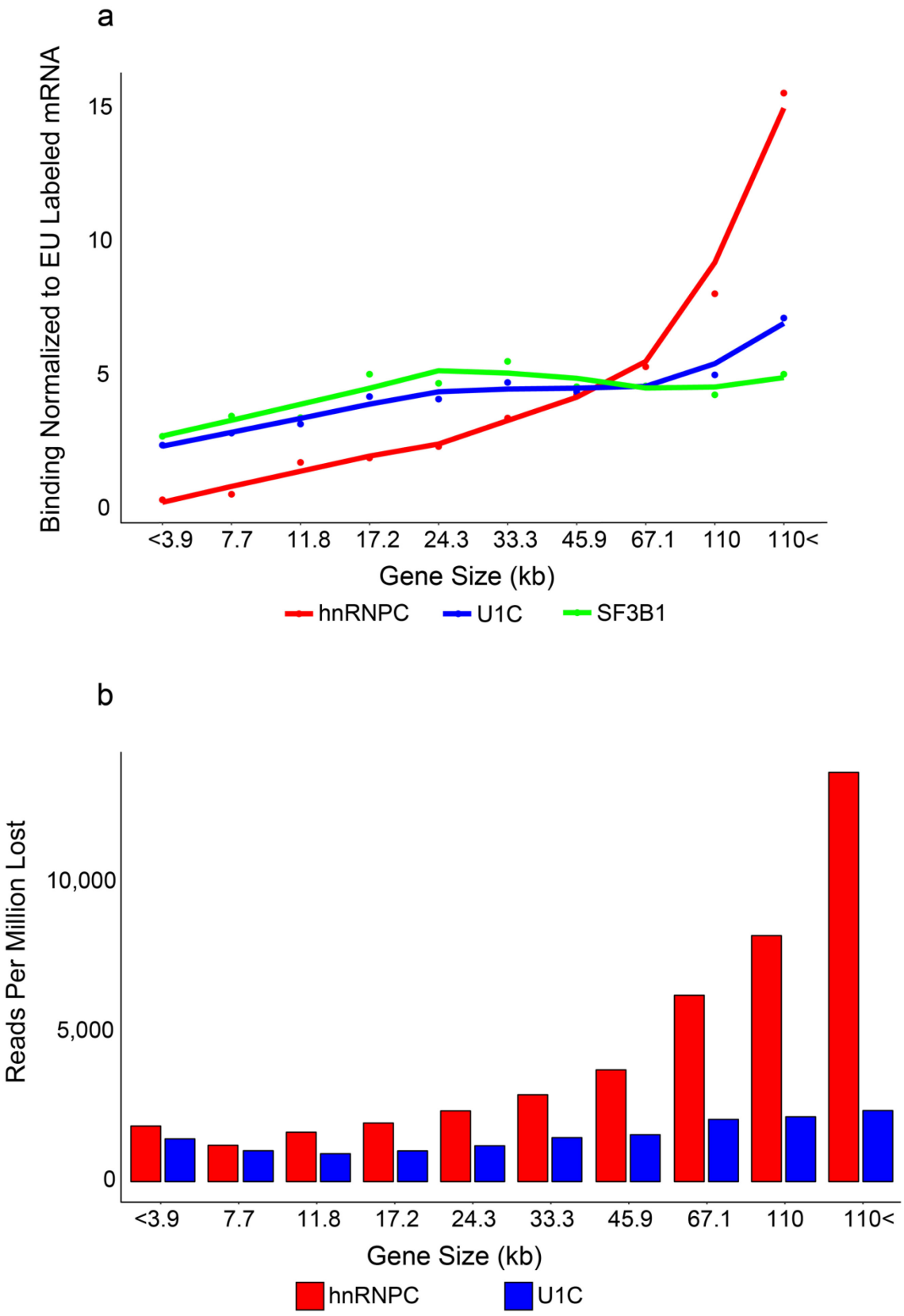

a
Binding Normalized to EU Labeled mRNA
15
10
5
0
<3.9
7.7
11.8
17.2
24.3
33.3
45.9
67.1
110
110<
Gene Size (kb)
hnRNPC
U1C
SF3B1
b
Reads Per Million Lost
10,000
5,000
0
<3.9
7.7
11.8
17.2
24.3
33.3
45.9
67.1
110
110<
Gene Size (kb)
hnRNPC
U1C

Figure 4.4 Large genes sequester disproportionate amounts of hnRNPs

(a) Point and line graph showing the XLIPseq binding signal of hnRNPC (red), U1C (blue), and SF3B1 (green) normalized to the mRNA output as measured by exon signal from poly(A) selected and 5-minute, EU labeled RNAseq data. This binding is calculated for ten gene size bins of equal number. Points represent the mean binding signal, and the line represents the best fit line for these points. (b) Bar graph showing the loss of binding for hnRNPC (red) and U1C (blue) over the same gene size bins as in panel a. Binding loss was calculated as the difference in XLIPseq gene reads from U1 versus control AMO transfected HeLa cells.

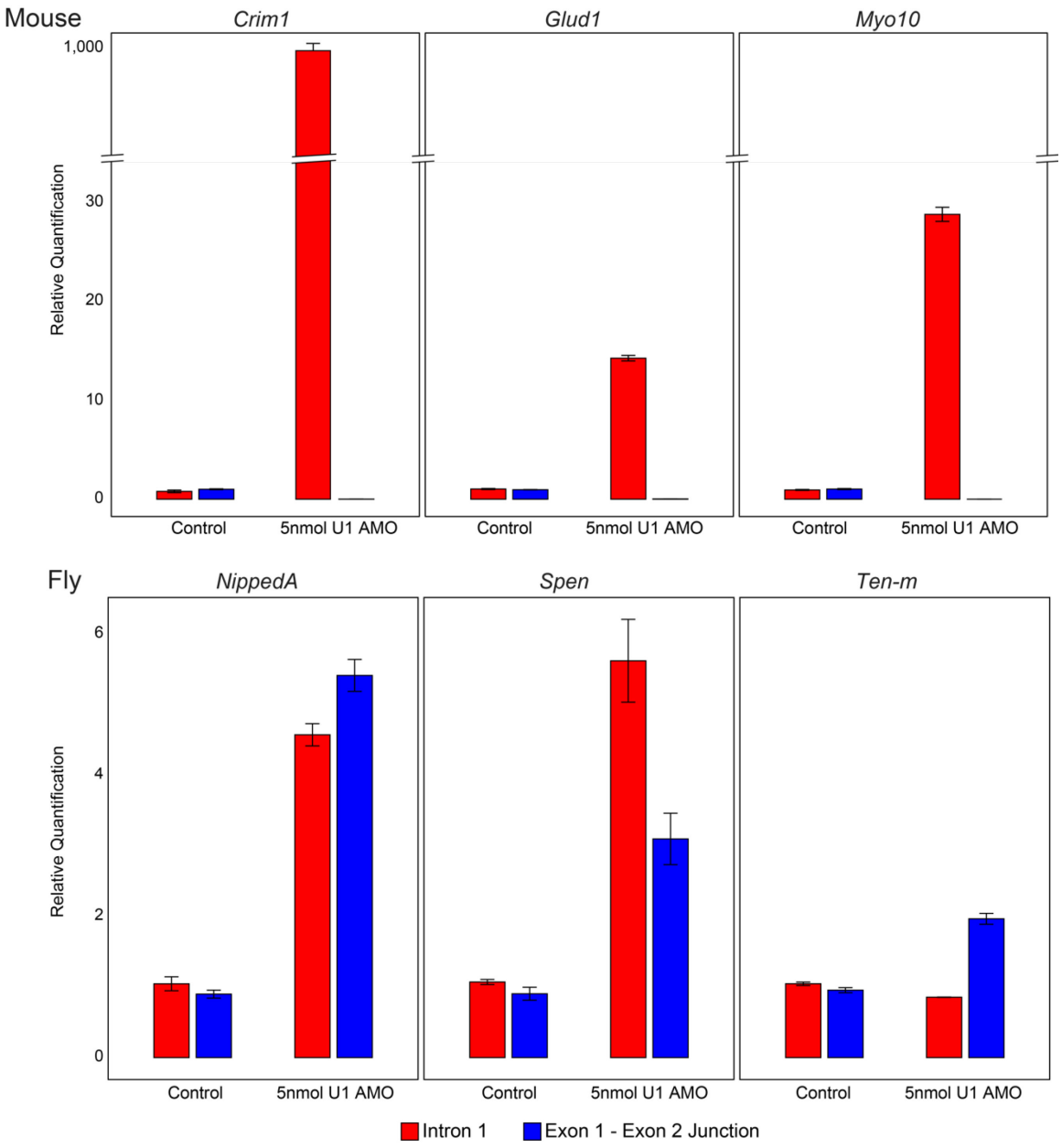

Mouse
Crim1
Glud1
Myo10
1,000
30
20
10
0
Relative Quantification
Control
5nmol U1 AMO
Control
5nmol U1 AMO
Control
5nmol U1 AMO
Fly
NippedA
Spen
Ten-m
6
4
2
0
Relative Quantification
Control
5nmol U1 AMO
Control
5nmol U1 AMO
Control
5nmol U1 AMO
Intron 1
Exon 1 - Exon 2 Junction

Figure 4.5 Mouse and fly RT-qPCR

RNA relative quantification data from U1 or control AMO transfected NIH-3T3 (Mouse, top) or S2 (Fly, bottom) cells in three genes for each organism. Primers for either the first intron (red) or the exon 1 to exon 2 junction is shown to quantitate the PCPA to spliced mRNA ratio, respectively. Quantification is normalized to the control AMO transfected intron 1 signal. Error bars depict the standard error from the triplicate experiment.

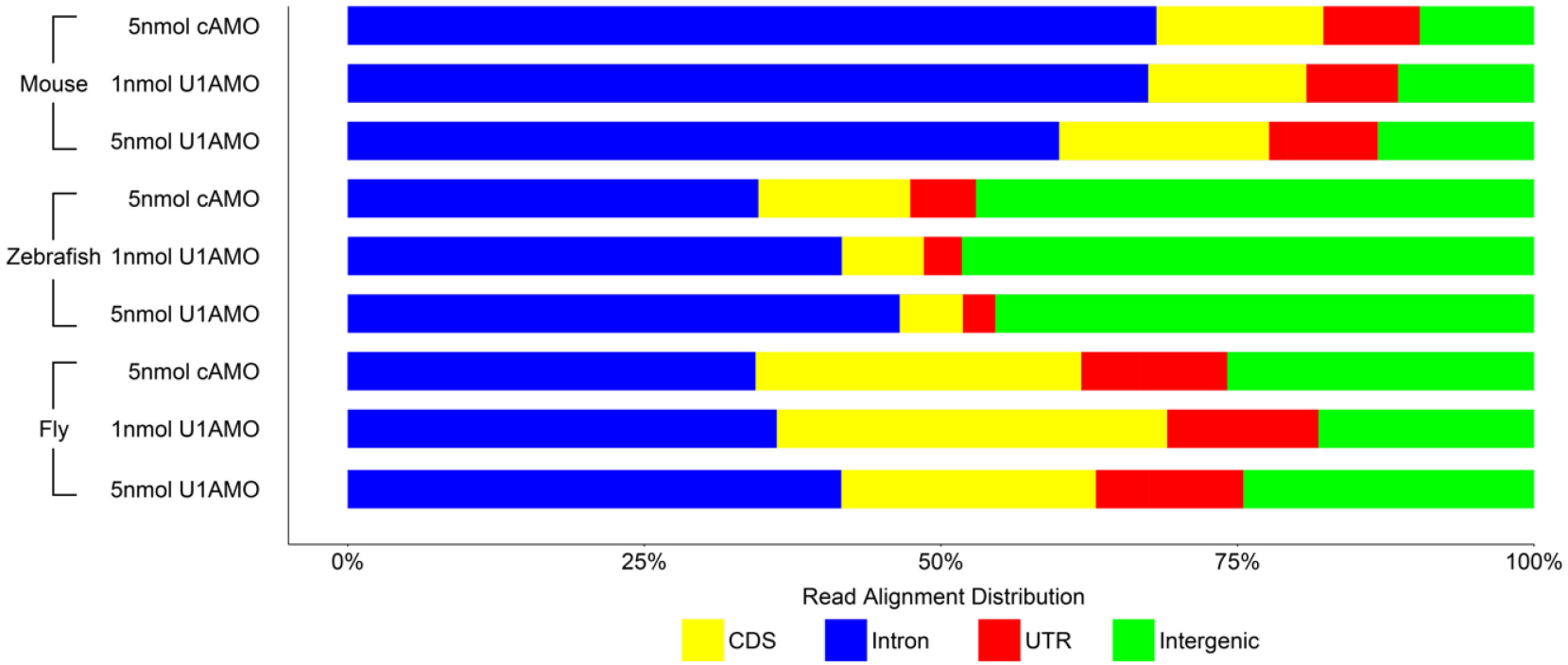

5nmol cAMO
Mouse
1nmol U1AMO
5nmol U1AMO
5nmol cAMO
Zebrafish
1nmol U1AMO
5nmol U1AMO
5nmol cAMO
Fly
1nmol U1AMO
5nmol U1AMO
0%
25%
50%
75%
100%
Read Alignment Distribution
CDS
Intron
UTR
Intergenic

Figure 4.6 Organism RNAseq read alignment distribution

Stacked bar graph showing the distribution of RNAseq read alignment locations (intron, blue; CDS, yellow; UTR, red; intergenic, green) as a percentage of the total aligned reads. Reads were aligned to either mm10 for mouse, danRer10 for zebrafish, or dm6 for fly.

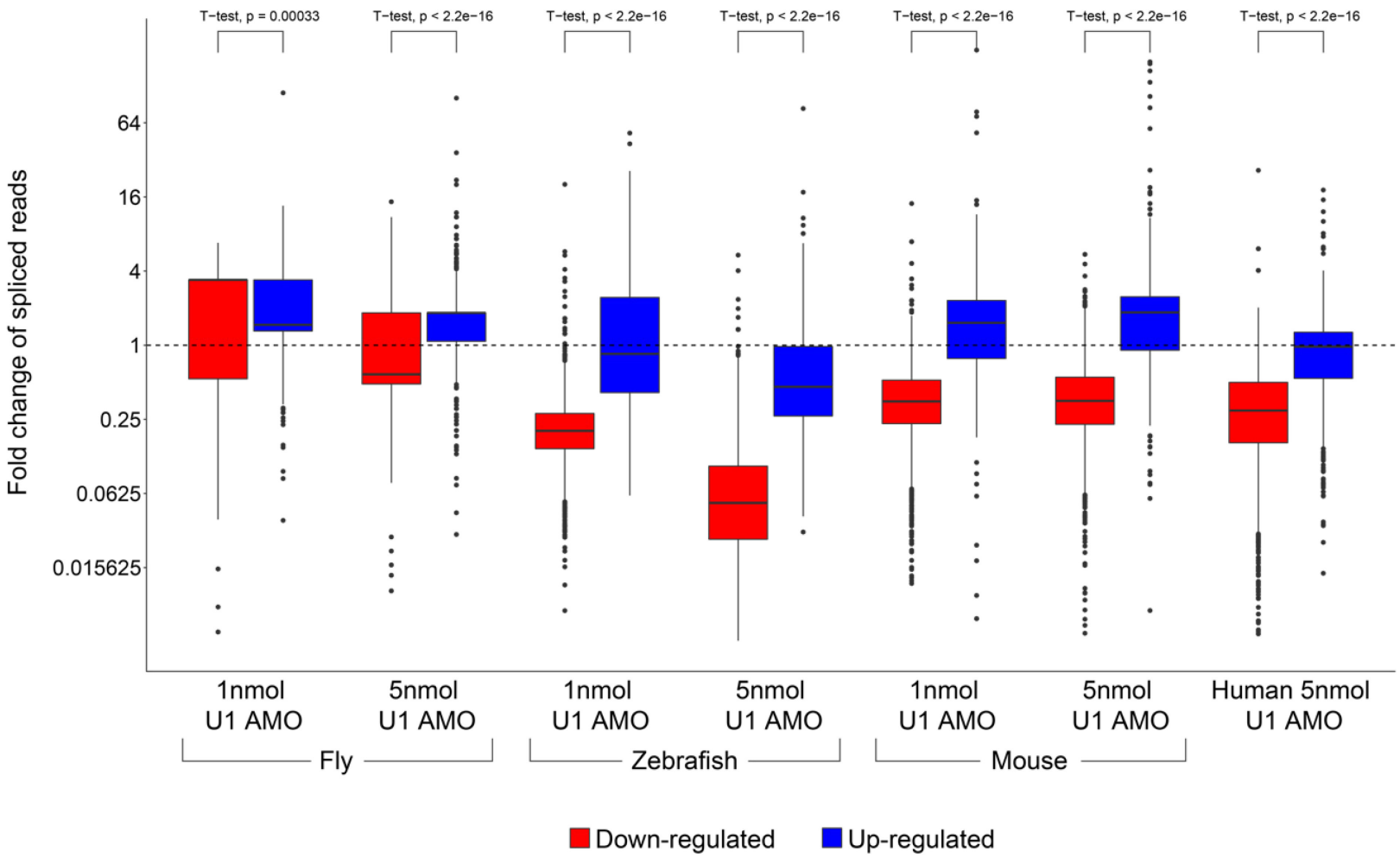

T−test, p = 0.00033
T−test, p < 2.2e−16
T−test, p < 2.2e−16
T−test, p < 2.2e−16
T−test, p < 2.2e−16
T−test, p < 2.2e−16
T−test, p < 2.2e−16
Fold change of spliced reads
64
16
4
1
0.25
0.0625
0.015625
1nmol U1 AMO
5nmol U1 AMO
1nmol U1 AMO
5nmol U1 AMO
1nmol U1 AMO
5nmol U1 AMO
Human 5nmol U1 AMO
Fly
Zebrafish
Mouse
Down-regulated
Up-regulated

Figure 4.7 U1 AMO causes significant, global splicing inhibition in zebrafish

Boxplots showing the $\log_2$ fold change in 30 minute, 4sU labeled RNAseq exon-exon junction reads from either down-regulated (red) or up-regulated (blue) genes in U1 versus control AMO transfected cells for S2 (Fly), ZF4 (Zebrafish), or NIH-3T3 (Mouse) cells. Gene expression changes were calculated using exon RPKM values tested for significance using a Poisson test with a p-value < 0.01 threshold. Distribution differences between the up- and down-regulated gene groups were tested for significance using a Student's t-test. A horizontal dashed line represents the 1-fold change line.

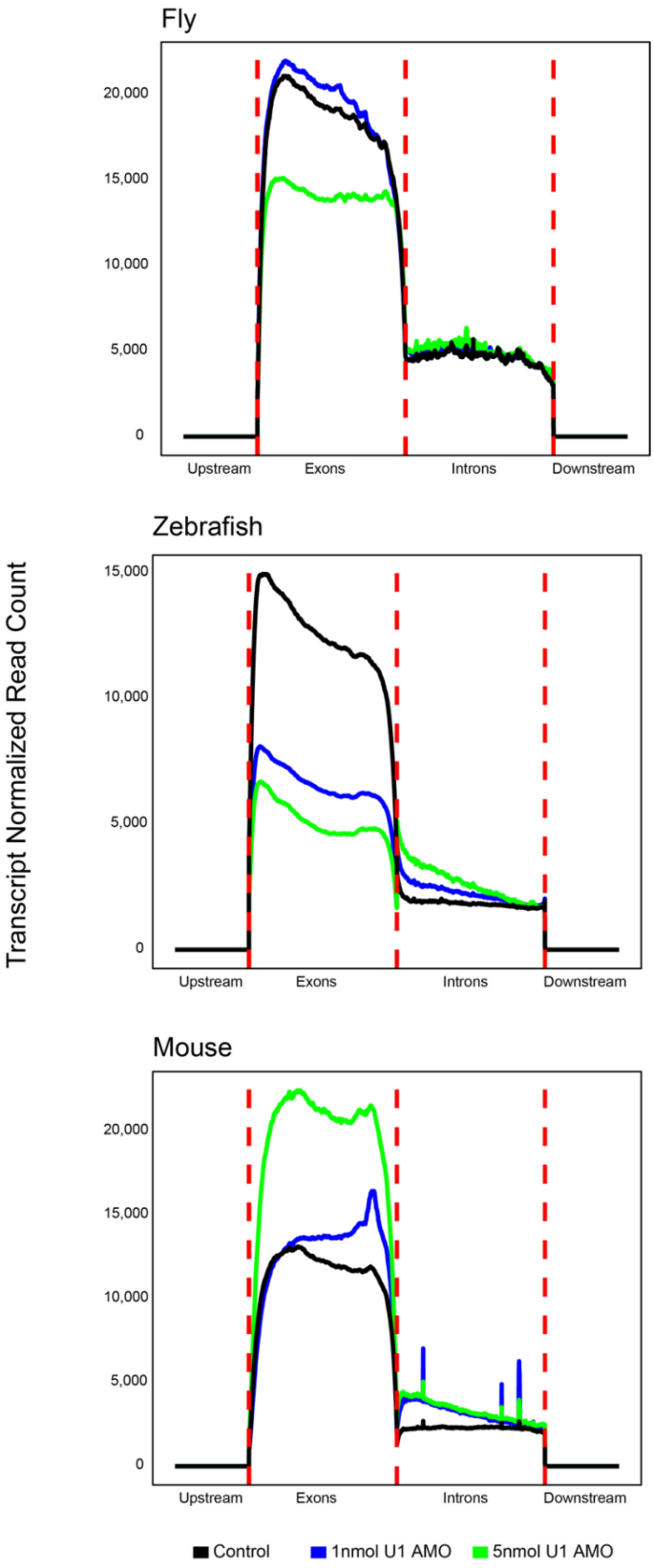

Fly
20,000
15,000
10,000
5,000
0
Upstream
Exons
Introns
Downstream
Zebrafish
15,000
10,000
5,000
0
Upstream
Exons
Introns
Downstream
Transcript Normalized Read Count
Mouse
20,000
15,000
10,000
5,000
0
Upstream
Exons
Introns
Downstream
Control
1nmol U1 AMO
5nmol U1 AMO

Figure 4.8 U1 AMO causes global 3' transcription loss

Metagene plots showing the exon and intron read distribution in control (black), 1nmol U1 (blue), or 5nmol U1 (green) AMO transfected S2 (Fly), ZF4 (Zebrafish), or NIH-3T3 (Mouse) cells from 30 minute, 4sU labeled RNAseq data. Exons and introns for all genes were joined in consecutive order and these regions were then scaled to 2,000bp. RNAseq reads aligned to each reference genome were normalized to the total mapped reads from each sample. Metagene profiles were generated using CGAT v02.5 (Sims et al. 2014).

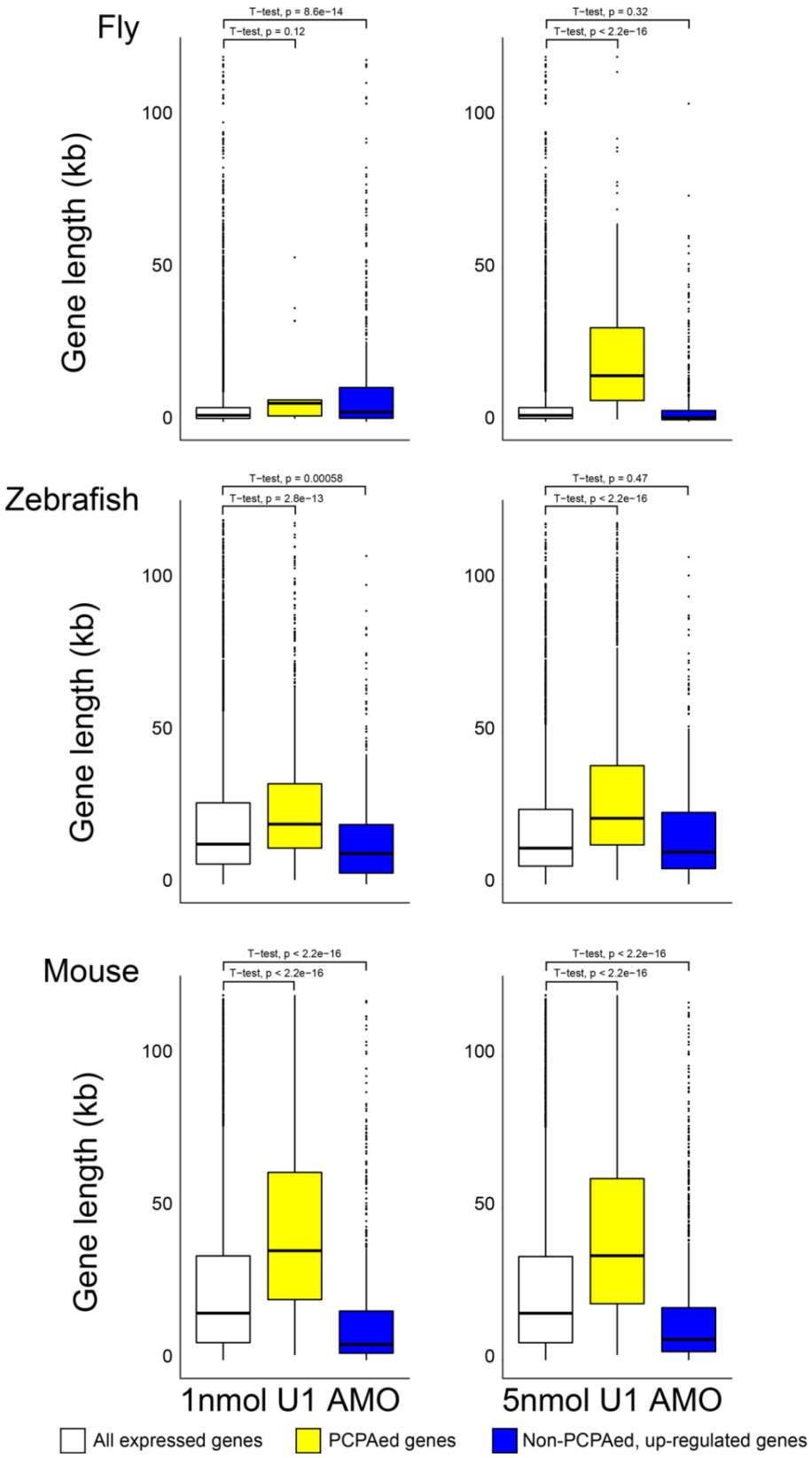

Fly
Zebrafish
Mouse
Gene length (kb)
T−test, p = 8.6e−14
T−test, p = 0.12
T−test, p = 0.32
T−test, p < 2.2e−16
T−test, p = 0.00058
T−test, p = 2.8e−13
T−test, p = 0.47
T−test, p < 2.2e−16
T−test, p < 2.2e−16
T−test, p < 2.2e−16
T−test, p < 2.2e−16
T−test, p < 2.2e−16
0
50
100
1nmol U1 AMO
5nmol U1 AMO
All expressed genes
PCPAed genes
Non-PCPAed, up-regulated genes

Figure 4.9 U1 inhibition selectively PCPAs large genes across evolution

Boxplots showing the gene size distribution in all expressed genes (white; RPKM ≥ 1), PCPAed genes (yellow), or non-PCPAed and up-regulated genes (blue) from U1 versus control AMO transfected S2 (Fly, top), ZF4 (Zebrafish, middle), or NIH-3T3 (Mouse, bottom) cells from 30 minute, 4sU labeled RNAseq data. Distribution differences between the all expressed and two other gene groups were tested for significance using a Student's t-test.

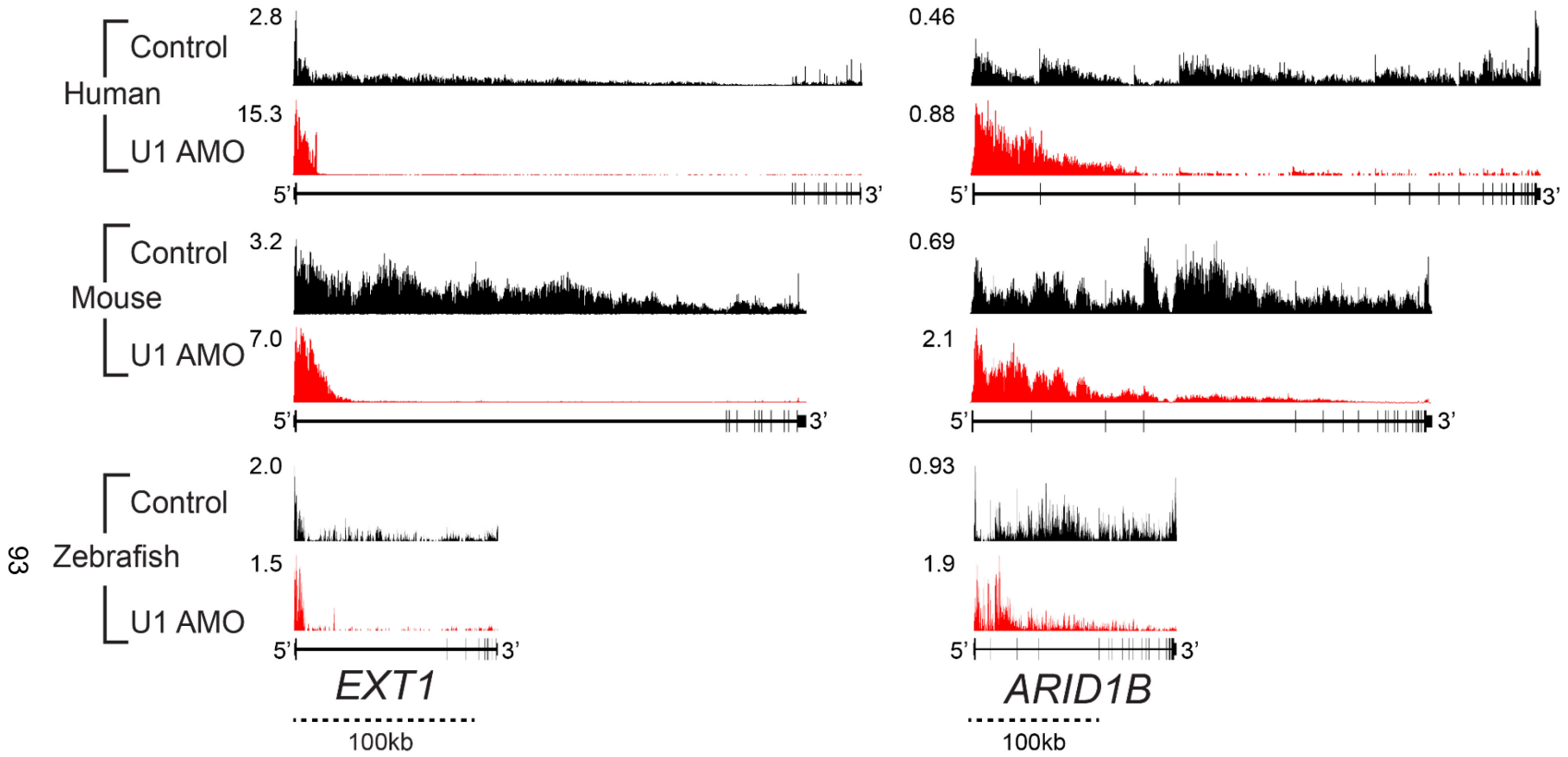

Control
Human
U1 AMO
2.8
15.3
5'
3'
0.46
0.88
5'
3'
Control
Mouse
U1 AMO
3.2
7.0
5'
3'
0.69
2.1
5'
3'
Control
Zebrafish
U1 AMO
2.0
1.5
5'
3'
0.93
1.9
5'
3'
EXT1
100kb
ARID1B
100kb

Figure 4.10 PCPA susceptibility is evolutionarily conserved

Genome browser view for the genes *EXT1* and *ARID1B* with 30 minute, 4sU labeled RNAseq read distributions for control (black) or U1 (red) AMO transfected cells from three organisms aligned to their respective reference genomes. Numbers to the left of the read distributions show the highest peak height value in the field as normalized to the total mapped reads. For gene structure, lines depict introns, boxes depict exons, and thinner boxes depict UTRs. Genomic distances are shown as dashed black lines.

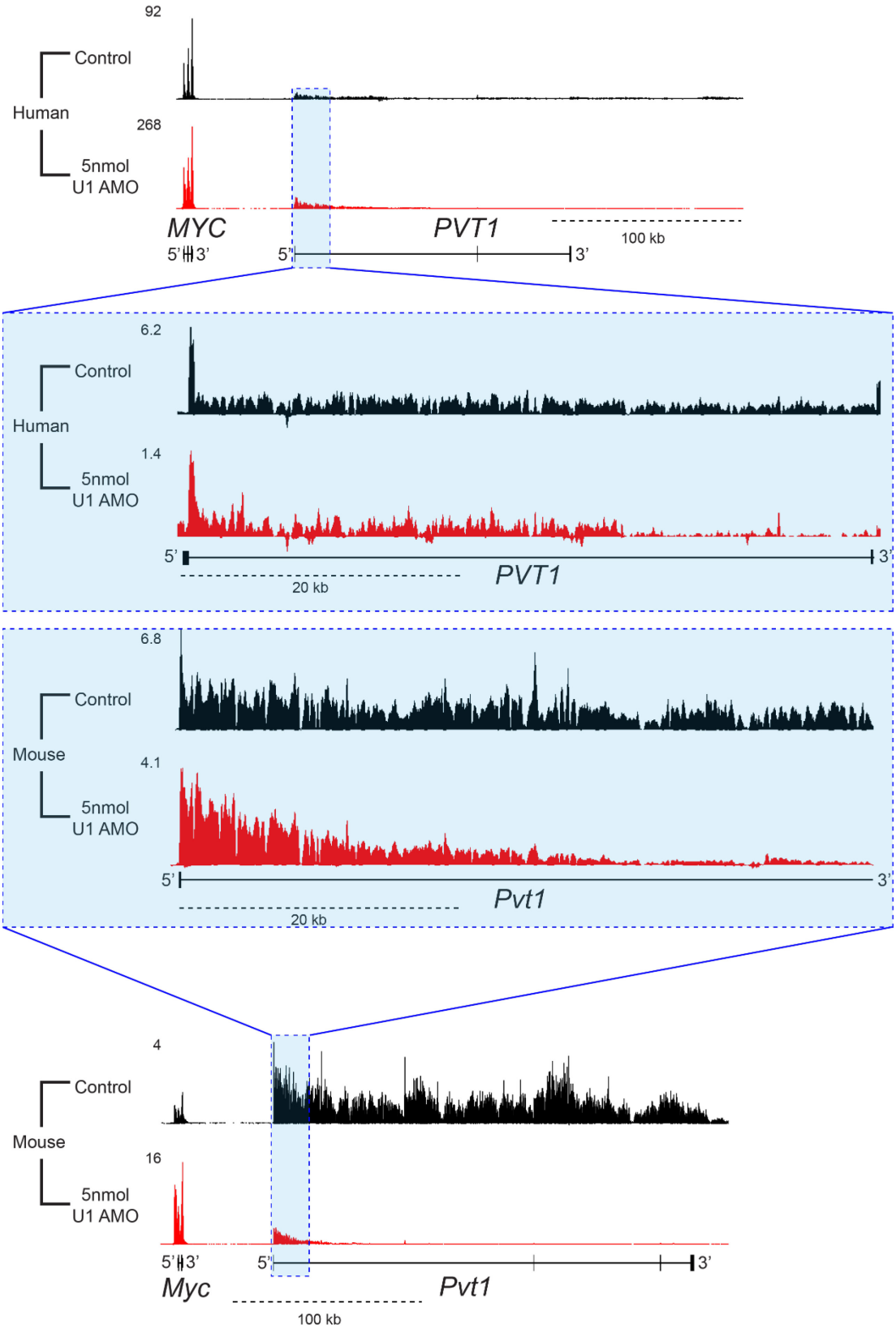

92
Control
Human
268
5nmol
U1 AMO
MYC
PVT1
100 kb
5'
3'
6.2
Control
Human
1.4
5nmol
U1 AMO
5'
3'
20 kb
PVT1
6.8
Control
Mouse
4.1
5nmol
U1 AMO
5'
3'
Pvt1
20 kb
4
Control
Mouse
16
5nmol
U1 AMO
5'
3'
Myc
Pvt1
100 kb

Figure 4.11 PCPAed genes and up-regulated genes are proximally grouped in gene neighborhoods

Genome browser view for the genes *PVT1* and *MYC* genes with 30 minute, 4sU labeled RNAseq read distributions for control (black) or U1 (red) AMO transfected cells from human (top) or mouse (bottom) aligned to their respective reference genomes. *PVT1* was identified as PCPAed, and *MYC* was identified as up-regulated after U1 AMO transfection. The intergenic distance between then (35kb) is much shorter than that found by randomly sampling genes (median = 74kb from hg19). Blue box inset shows the zoomed region of PCPA within the *PVT1* gene for both organisms. Numbers to the left of the read distributions show the highest peak height value in the field as normalized to the total mapped reads. For gene structure, lines depict introns, boxes depict exons, and thinner boxes depict UTRs. Genomic distances are shown as dashed black lines.

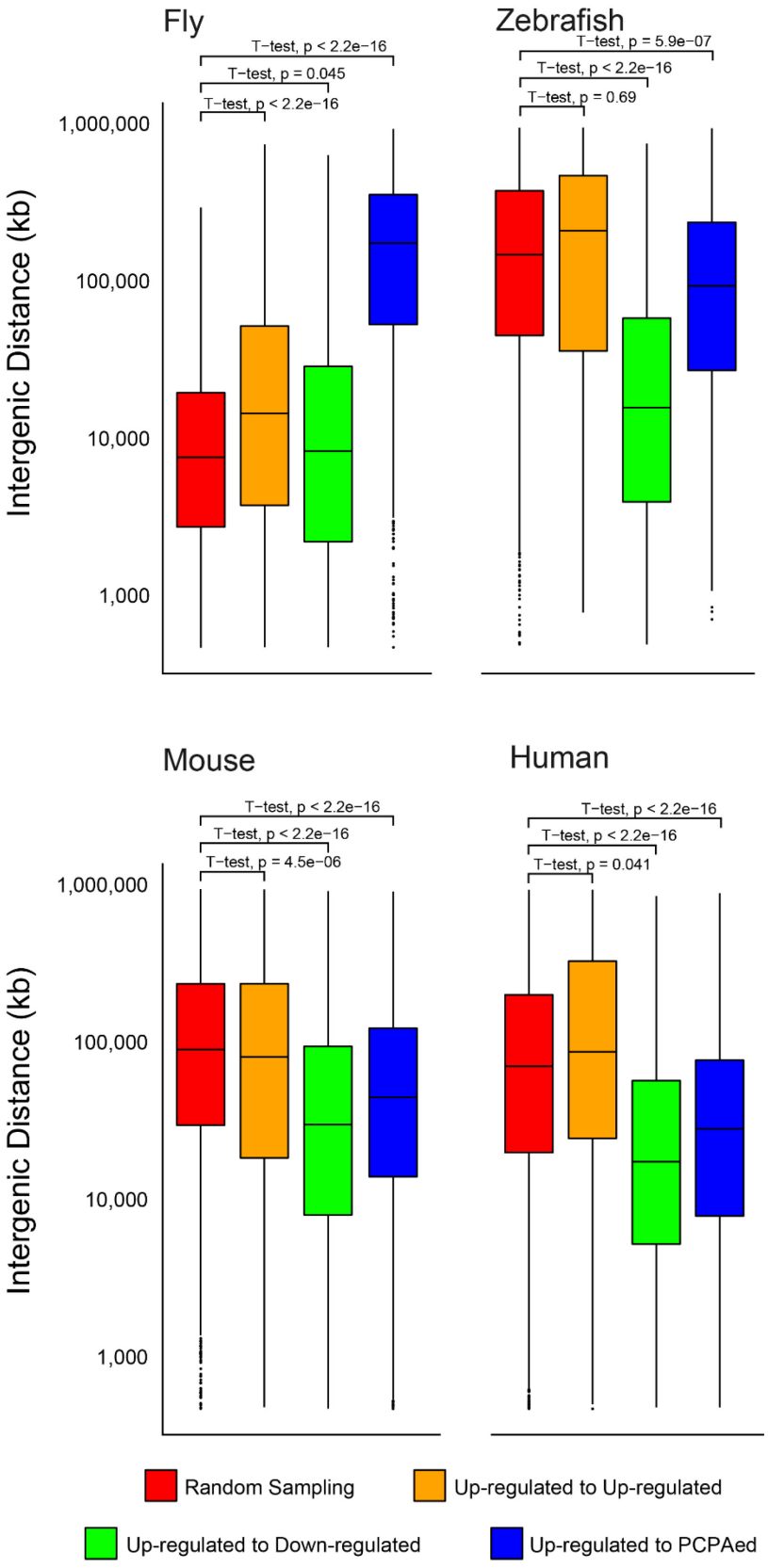

Fly
T−test, p < 2.2e−16
T−test, p = 0.045
T−test, p < 2.2e−16
Zebrafish
T−test, p = 5.9e−07
T−test, p < 2.2e−16
T−test, p = 0.69
Mouse
T−test, p < 2.2e−16
T−test, p < 2.2e−16
T−test, p = 4.5e−06
Human
T−test, p < 2.2e−16
T−test, p < 2.2e−16
T−test, p = 0.041
Intergenic Distance (kb)
1,000,000
100,000
10,000
1,000
Random Sampling
Up-regulated to Up-regulated
Up-regulated to Down-regulated
Up-regulated to PCPAed

Figure 4.12 U1 AMO induced PCPA and down-regulation contributes to proximal gene up-regulation

Boxplots showing intergenic distance distributions between randomly sampled (red) UP-UP (orange), UP-DOWN (green), or UP-PCPA (blue) gene groupings for fly (top-left), zebrafish (top-right), mouse (bottom-left), or human (bottom-right) genes. Genes were identified and intergenic distances were calculated as described in the text. Distribution differences between the all randomly sampled and other three gene groups were tested for significance using a Student's t-test.

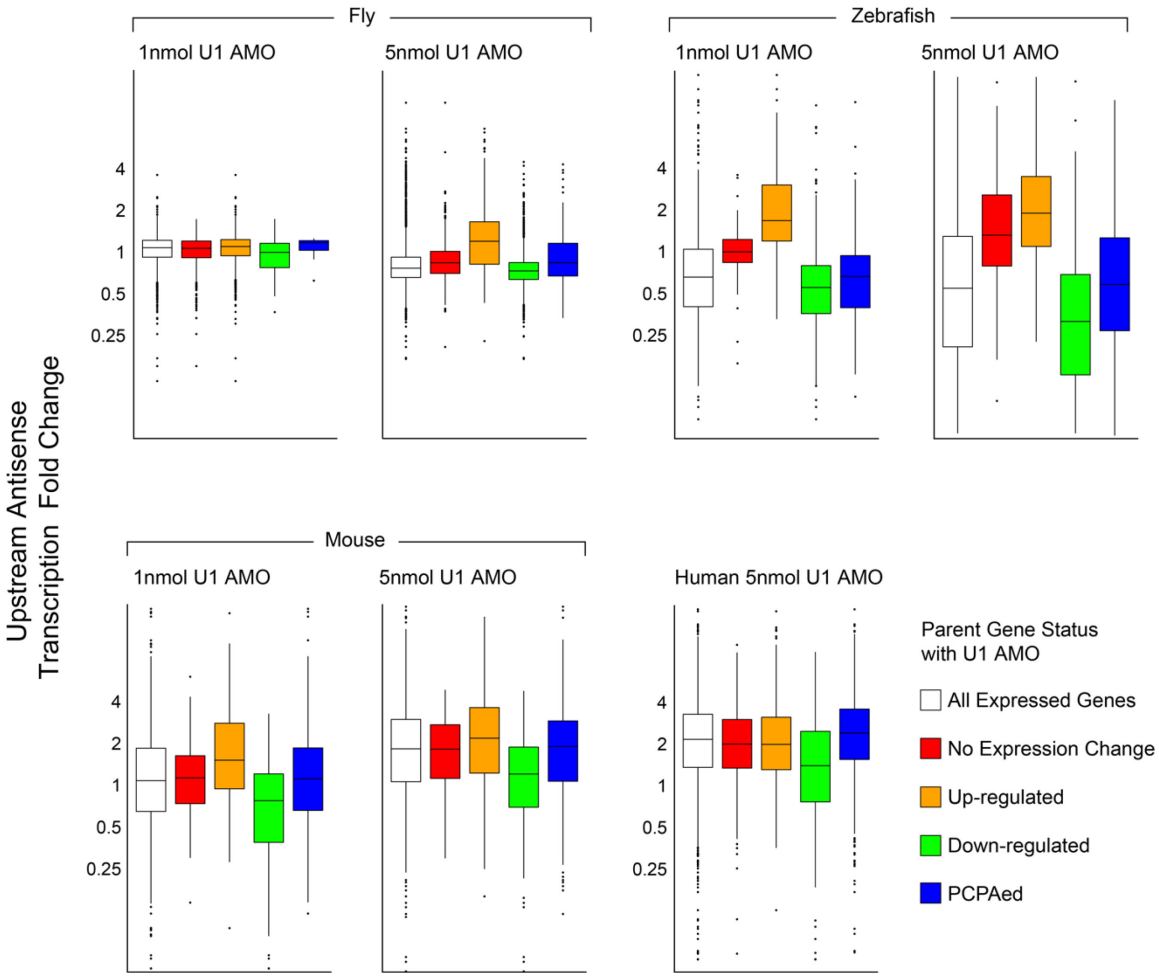

Fly
1nmol U1 AMO
5nmol U1 AMO
Zebrafish
1nmol U1 AMO
5nmol U1 AMO
Upstream Antisense Transcription Fold Change
4
2
1
0.5
0.25
Mouse
1nmol U1 AMO
5nmol U1 AMO
Human 5nmol U1 AMO
Parent Gene Status with U1 AMO
All Expressed Genes
No Expression Change
Up-regulated
Down-regulated
PCPAed

Figure 4.13 Upstream antisense transcription is driven sense gene transcription levels

Boxplots showing the $\log_2$ transcription fold change for U1 versus control AMO transfected cells from fly (top-left), zebrafish (top-right), mouse (bottom-left), or human (bottom-right). Sense genes were identified as either all expressed (RPKM ≥ 1; white), unchanged (red), up-regulated (orange), down-regulated (green), or PCPAed (blue) as described in the text.

Mouse		*Crim1*		*Glud1*		*Myo10*	
Transfection	Labeling	Mean Intron 1 RQ	Mean Exon-Exon RQ	Mean Intron 1 RQ	Mean Exon-Exon RQ	Mean Intron 1 RQ	Mean Exon-Exon RQ
Control, 6 hour	30 minute 4sU	0.795	1.011	1.035	0.966	0.939	1.028
5nmol U1 AMO, 6 hour	30 minute 4sU	1,000.372	0.007	14.283	0.026	28.844	0.000
Control, 6 hour	60 minute 4sU	2.406	2.740	3.187	0.521	0.412	0.443
5nmol U1 AMO, 6 hour	60 minute 4sU	3,650.835	0.001	3.251	0.057	10.366	0.308
Control, 6 hour	120 minute 4sU	1.000	1.000	1.000	0.664	0.000	0.129
5nmol U1 AMO, 6 hour	120 minute 4sU	1,055.067	0.400	0.059	0.585	1.502	0.009
Control, 8 hour	30 minute 4sU	1.000	1.000	1.000	0.002	0.956	1.000
5nmol U1 AMO, 8 hour	30 minute 4sU	49.112	0.191	4.007	0.001	16,443.616	0.714
Control, 8 hour	60 minute 4sU	1.575	0.940	0.972	1.022	404,110.164	37,857.245
5nmol U1 AMO, 8 hour	60 minute 4sU	1,718.194	0.138	2.956	4.459	676,348.774	4,647.712
Control, 8 hour	120 minute 4sU	1.054	0.816	0.052	3.711	607,716.546	16,450.001
5nmol U1 AMO, 8 hour	120 minute 4sU	598.687	0.008	0.614	0.207	244,708.310	1,108.373

Fly		*Nipped-A*		*spen*		*Ten-m*	
Transfection	Labeling	Mean Intron1 RQ	Mean Exon-Exon RQ	Mean Intron 1 RQ	Mean Exon-Exon RQ	Mean Intron 1 RQ	Mean Exon-Exon RQ
Control, 6 hour	30 minute 4sU	1.028	0.900	1.072	0.905	1.047	0.956
5nmol U1 AMO, 6 hour	30 minute 4sU	0.000	5.416	5.622	3.099	0.856	1.967
Control, 6 hour	60 minute 4sU	0.443	0.296	3.168	1.624	0.341	0.705
5nmol U1 AMO, 6 hour	60 minute 4sU	0.308	0.474	1.485	0.574	0.382	0.869
Control, 6 hour	120 minute 4sU	0.129	20.236	4.878	5.158	0.191	1.900
5nmol U1 AMO, 6 hour	120 minute 4sU	0.009	3.435	2.159	0.642	0.546	0.907
Control, 8 hour	30 minute 4sU	1.000	0.967	1.159	1.024	1.075	0.966
5nmol U1 AMO, 8 hour	30 minute 4sU	0.714	0.755	1.196	0.304	0.703	1.247
Control, 8 hour	60 minute 4sU	37,857.245	0.185	0.941	0.460	0.158	1.023
5nmol U1 AMO, 8 hour	60 minute 4sU	4,647.712	0.285	0.136	0.329	0.121	0.908
Control, 8 hour	120 minute 4sU	16,450.001	0.243	0.719	0.335	0.277	0.464
5nmol U1 AMO, 8 hour	120 minute 4sU	1,108.373	0.339	0.310	0.146	0.294	0.383

Table 2 Relative quantification data for all RT-qPCR samples

Mean relative quantification values from 30 minute, 4sU labeled RNA for three genes comparing U1 and control AMO transfected NIH-3T3 (Mouse) and S2 (Fly) cells. Primers for either the first intron or exon 1 to exon 2 were used to quantitate the PCPA to spliced mRNA ratio, respectively. Quantification is normalized to the control AMO transfected intron 1 signal.

Species	Treatment	Input	Mapped	Mapping Percent	Reads Passing Cleaning	Read Percent Passing Filter	Multimapped Reads	Multimapping Percentage
Fly	Control	339,161,138	272,306,848	80.29%	194,896,964	57.46%	88,945,296	32.66%
	1nmol U1 AMO	94,786,010	75,845,989	80.02%	57,415,060	60.57%	19,524,093	25.74%
	5nmol U1 AMO	210,107,152	155,347,350	73.94%	106,504,398	50.69%	55,126,781	35.49%
Zebrafish	Control	159,920,352	110,611,290	69.17%	68,151,246	42.62%	6,075,202	5.49%
	1nmol U1 AMO	369,564,308	252,637,772	68.36%	165,295,256	44.73%	13,961,048	5.53%
	5nmol U1 AMO	116,268,124	76,976,142	66.21%	50,655,878	43.57%	4,963,868	6.45%
Mouse	Control	252,119,472	222,623,222	88.30%	159,565,788	63.29%	4,537,363	2.04%
	1nmol U1 AMO	125,975,914	101,742,818	80.76%	69,120,072	54.87%	2,861,026	2.81%
	5nmol U1 AMO	294,062,990	249,468,934	84.84%	175,115,664	59.55%	6,767,081	2.71%

Table 3 Organism RNAseq mapping statistics

Numbers for the fly (dm6), zebrafish (danRer10), and mouse (mm10) read alignment of RNAseq data using TopHat v2.1.1 (Trapnell, Pachter, and Salzberg 2009). Multi-mapped reads are defined as those which align to 20 or more locations in the genome; read were filtered to include only the primary alignment of properly mapped pairs using Samtools v1.9 (Wysoker et al. 2009).

	Fly		Zebrafish		Mouse		Human
	1nmol U1 AMO	5nmol U1 AMO	1nmol U1 AMO	5nmol U1 AMO	1nmol U1 AMO	5nmol U1 AMO	5nmol U1 AMO
Expressed	9,688	9,635	6,992	4,820	6,991	7,233	9,744
Down-Regulated	453	1,817	6,568	7,266	5,765	5,190	6,409
PCPAed	24	214	1,009	2,520	3,490	3,736	3,590
Overlap with Human	5	37	318	774	1,545	1,620	-
Percent Overlap	20.83%	17.29%	31.52%	30.71%	44.27%	43.36%	-
p-value	0.394	0.099	2.06×10^{-09}	2.26×10^{-19}	0.00	0.00	-
Up-Regulated	653	624	326	344	1,022	2,048	988
Overlap with Human	58	18	38	58	341	341	-
Percent Overlap	17.30%	2.89%	11.66%	17.26%	33.66%	22.22%	-
p-value	1.27×10^{-26}	4.21×10^{-04}	3.42×10^{-04}	6.20×10^{-12}	1.36×10^{-234}	8.75×10^{-234}	-

Table 4 U1 AMO induced PCPA and up-regulated genes is evolutionarily conserved

Numbers of genes expressed (RPKM ≥ 1) or affected by U1 (PCPAed or up-regulated) compared to control AMO transfected fly, zebrafish, mouse, and human cells from 30 minute, 4sU labeled RNAseq. The overlap of the fly, zebrafish, and mouse affected genes were compared for ortholog overlap to human for PCPAed as well as up-regulated groupings, and statistical significance in the overlap was calculated using a hypergeometric distribution.

	Fly		Zebrafish		Mouse		Human	Comparison Across Species	
Gene status	1nmol U1 AMO	5nmol U1 AMO	1nmol U1 AMO	5nmol U1 AMO	1nmol U1 AMO	5nmol U1 AMO	5nmol U1 AMO	Variance	Standard Deviation
All expressed	2,152	2,138	13,861	12,407	17,022	16,962	22,800	76,247,192	8,732
PCPAed	6,185	16,084	20,646	23,225	43,695	41,594	39,000	151,658,055	12,315
Non-PCPAed, Up-regulated	3,283	1,426	10,336	11,095	5,247	7,002	6,800	15,725,141	3,965

Table 5 Median gene sizes for expressed, PCPAed, and up-regulated genes

Numbers showing the median gene size for all expressed (RPKM ≥ 1), PCPAed, or not-PCPAed and up-regulated gene groups in fly, zebrafish, mouse, and human cells for U1 compared to control AMO transfections from 30 minute, 4sU labeled RNAseq. The variance and standard deviation across these organisms are also calculated.

	Median Intergenic Distance Between Gene Groups (bp)			
Species	Random Gene Sampling	UP-UP	UP-DOWN	UP-PCPA
Fly	5,127	7,617	4,332	188,564
Zebrafish	178,317	607,286	15,437	99,264
Mouse	104,576	91,262	28,199	43,720
Human	74,743	198,294	14,306	22,857

Table 6 Median intergenic distances for various gene groups

Numbers showing the median intergenic distances between randomly sampled, UP-UP, UP-DOWN, or UP-PCPA gene groupings for fly, zebrafish, mouse, or human genes from 30 minute, 4sU labeled RNAseq after transfection with either control or U1 AMO. Genes were identified and intergenic distances were calculated as described in the text.

CHAPTER 5: NOVEL EXON SKIPPING AND READ-THROUGH SPLICING IS CAUSED BY LOW U1 BASE-PAIRING INHIBITION

Chapter Details

Unpublished. Contributing members: Chie Arai, Eric Babiash, Maura Jones (data interpretation and discussion).

Multi-Exon Skipping

To this point, much of the discussion of my research has revolved around U1 telescripting and gene structure as it relates to PCPA. As U1's importance was first shown by the discovery of its role in splicing through binding of the 5'ss, it would have been remiss of me not to take a closer look at how splicing was affected by U1 AMO treatments. There were many software packages already available with which to study alternative splicing, though each have some drawbacks. For example, MISO required pre-processed annotation files of known isoforms and cannot discover novel splice formats, while JUM and MAJIQ were computationally intensive and only analyze specific splicing event types (Katz et al. 2010; Q. Wang and Rio 2018; Vaquero-Garcia et al. 2016). I decided to analyze splicing changes observed in our samples from the simpler starting point of using the *de novo* splice junctions from the alignment and knock exon-intron boundaries. From this analysis, I could then easily explore additional avenues of research such as a deeper examination of specific mis-splicing events.

To begin this analysis, I calculated the overall splicing, as well as splice junctions that map to all known, annotated splice sites and those that do not. Overall splicing levels decreased with U1 AMO in almost every human, mouse, zebrafish, and fly RNAseq sample, with the exception of 5nmol U1 AMO in mouse and 1nmol U1 AMO in fly (Figure 5.1a). As discussed previously in terms of PCPA, the fly cells were less affected overall

by U1 base pairing inhibition. With this in mind, the minimal changes in splicing are not out of place when considering the fewer PCPA events and expression changes in the fly sample. However, the increased splicing in mouse with 5nmol U1 AMO was surprising; especially given that an expected decrease was seen with the 1nmol U1 AMO treatment. Closer examination of these junctions in the mouse sample showed that they resulted from the extremely high up-regulation found in small (median size 6.9kb) genes, for example *Myc*, *Fos*, or *Gadd45a*.

Despite U1's requirements for exon definition, U1 AMO only produced minor increases in non-annotated compared to annotated splice junctions when analyzed globally in mouse and human samples (Figure 5.1b). In the fly and zebrafish, however, U1 AMO did reduce overall splicing fidelity considerably. I should note that both of these organisms had high baseline levels of non-annotated splice junctions, which appear to result from two different contributing sources. Since I calculated splicing fidelity from previously known and annotated exon-exon junctions, which are in turn annotated using combinations of experimental data and sequence-based inferences, the decreased fidelity in zebrafish is most likely due to the paucity of scientific knowledge about their genes and gene structures relative to mouse or human. In fly, which is one of the most well-studied organisms genetically, I found that many of the non-annotated splice junctions came from high copy number genes, ribosomal RNAs that were not fully depleted from the sample during cleaning, or non-coding RNAs that are in close proximity in the genome. These were most likely the result of sequencing and/or alignment error where similar sequences were called upon to align to nearby genes rather than splicing within a single gene. When added together, these type of splice junctions accounted for 23-37% of the total non-annotated splicing found in fly, compared to <10% for mouse and 13-40% for zebrafish.

Thus, U1 base pairing inhibition does reduce overall splicing fidelity in all the organisms studied; however, the effect was less dramatic than I had initially anticipated.

When visualizing the junctions in all samples using the UCSC Genome Browser, I noticed the loss of splicing in many genes, especially from PCPA, down-regulation, and decreased splicing. As reported here, there were also a specific subset of genes where splicing was unaffected or even increased with U1 AMO. Visualization revealed that inhibiting U1 base pairing created many instances (between 11-4,341 genes across all organisms) of splicing that occurred across multiple exons which did not associate to known splice isoforms (Figure 5.2; full list of genes found in Supplementary Table 2). Additionally, there were rare, aberrant splicing events observed, where splice junctions occurred within introns, between introns and exons, or in intergenic space. While it was nearly impossible to distinguish if these were the result of a technical error in sequencing or in alignment, most of these were <5 reads in depth, usually only a single read, in comparison to the ≥100 read average for all junctions. These low numbers could perhaps be ascribed to a background error of some type; however, additional factors suggested caution in this interpretation. Specifically, these multi-exon skipping events utilized annotated 5' or 3' splice sites and were present at lower levels and frequency in control conditions, where they skipped the same exons. Moreover, these non-canonical splicing events incorporated higher read numbers than could be expected for technical error alone. As such, I decided to compile these splicing events together as an indicator of mis-splicing due to U1 base pairing inhibition.

Having seen several promising splicing events, I now wanted to systematically identify these novel junctions. Given the high sequencing depth in our samples and the fact that there were already non-annotated splice junctions at a very low read number, I decided to

begin by filtering out all splicing events that contained less than five reads. This was done to remove any junction that could be simply a technical error. I then split the splicing events into those that were from annotated and consecutive exons (annotated), and those that were not, but which were still spliced from within a single gene. Within these, I defined the multi-exon skipping junctions as those that spliced directly from known 5' and 3' splice sites. The remaining events then were put into a group consisting of junctions utilizing novel splice sites (aberrant or other); these were often offset from annotated splice sites by only a few bases, suggesting that they resulted either from base calling error during sequencing or were produced from poorly defined 5' or 3' splice sites (although a few spliced within introns or from an exon into an intron). I decided to count both the number of events for each splicing category, for analyzing breadth, and the number of reads grouped into each category, for analyzing depth. As U1 AMO affected the overall splicing levels, it was important that I compare the multi-exon skipping and aberrant splicing levels to the number of reads in annotated exon-exon junctions. This ensured that any increase or decrease in the former two would be taken in the context of what splicing was occurring within the sample as a whole.

I began with the fly, zebrafish, and mouse RNAseq data. With regard to depth, the total amount of multi-exon skipping increased by 1-2 fold after U1 AMO treatment in all organisms (Table 7). The breadth of skipping, however, was not as consistent across organisms as it decreased for the 1nmol doses in mouse and fly then increased again at 5nmol, while the opposite was true for zebrafish. Higher U1 AMO doses were expected to induce extensive PCPA, which would reduce the number of exons transcribed from the 3' ends of genes. In contrast, more attenuated PCPA from a lower U1 AMO dose should, in theory, have allowed for more exon skipping. As a consequence, I examined the splicing events in mouse and fly for explanations for their reduced skipping and aberrant splicing

events at 1nmol U1 AMO. Consistent with the results discussed in the previous chapter, the fly exhibited much more general up-regulation rather than down-regulation at 1nmol U1 AMO. This increased mRNA output could explain the decreased levels of exon skipping in comparison to annotated exon-exon junctions. As mentioned above, the 1nmol U1 AMO mouse data showed an overall decrease in splicing, which may have been due to lower sequencing levels when compared to the control and 5nmol U1 AMO (69 versus 159 and 175 million mapped reads, respectively). While the lower depth did not heavily impact general analyses and transcriptome changes, the already lower levels of expression seen in the mis-splicing events could have meant that a previously low level splicing change in a deeply sequenced dataset (>100 million reads) would not possess enough reads to pass a statistical filter if the total sequencing depth was lowered by half.

I decided to also include a HeLa low U1 AMO titration (0.01, 0.05, 0.25, 0.5 and 1nmol) RNAseq dataset, carried out by a post-doctoral colleague Jung-Min Oh, into this examination to help determine whether incomplete U1 base pairing inhibition can stimulate higher levels of mis-splicing in the human system. It is important to note that these low levels of U1 AMO do not cause severe PCPA, and so allow for the investigation of other transcriptome changes that would otherwise be masked by widespread transcription termination. These samples demonstrated a dose dependent increase in the depth of multi-exon skipping, as well as a sizable increase in the breadth at 1nmol. Surprisingly, there were a significant number of aberrant splicing events in all HeLa samples as well, but not a correlative increase in depth at each splicing location. This observation could have been explained by the fact that these samples were more deeply sequenced than the fly, zebrafish, and mouse RNAseq, and, thus, contained more statistical power to identify mis-splicing events. Alternatively, the higher transcriptional output from highly

active and cancerous HeLa cells could have already contained more alternative splicing errors at baseline.

At first, it seemed likely that the breadth of multi-exon skipping would depend on the number of exons in a gene; I reasoned that the more splice sites there were, the more likely it was for mis-splicing to occur. However, this proved somewhat untrue, as multi-exon skipping was only mildly correlated with exon number (Spearman correlation, 0.1-0.64 over all four organisms tested) and with aberrant splicing (Spearman correlation, 0.13-0.43 over all four organisms tested); this was also the case for the depth of mis-splicing (Spearman correlation, 0.09-0.54 and 0.11-0.4 over all four organisms tested) (Table 8). By comparison, the annotated splicing events were highly correlated with exon number (Spearman correlation, 0.68-0.96), which was to be expected as they provided more splicing opportunities. There was also no correlation between skipping breadth or depth, or changes to them after U1 AMO, and gene size (data not shown).

In order to determine a potential rationale for which exons are skipped and which are included, I decided to examine the 5' and 3' splice site sequences. Most splice sites conform to a series of consensus sequences so that they can both be recognized by U1 and U2 and carry out the transesterification reaction steps in splicing (Mount 2000; Wahl, Will, and Lührmann 2009; Will and Lührmann 2011). There are multiple portions of consensus within both the 5' and 3' splice site; the most invariant of these are the dinucleotides both at the 5' side of the intron, GT spanning the exon-intron boundary, and at the 3' side of the intron, AG at the intron terminus. It seemed likely to me that this exon skipping could be due to weaker 5' or 3' splice site sequences, i.e. ones that do not conform as well to consensus sequences. To examine this, I used two tools: the first was sequence logos to look for changes in base pair enrichment at the 5' and 3' splice sites

and the second was maximum entropy modeling (MaxEnt) on these regions to test for conformity to known splice site sequences (Crooks et al. 2004; Yeo and Burge 2004). I was only able to accomplish the latter in mouse and human, as there were readily available trained splice site databases only for these organisms.

For examining the sequence logos over the 5' and 3' splice site, I extracted the target regions in each sample so as to include 9 nucleotides of the intron as well as 3 nucleotides of the flanking exon. As I split each group by splice site, inclusion or exclusion for skipping, as well as by sample, I ended up with 60 separate logos. However, logo sequence did not change significantly when comparing amongst AMO treatments (data not shown). This is most likely due to the presence of some exon skipping events in the control, albeit at lower levels than in U1 AMO. Thus, to consolidate the logos into more manageable pieces, I combined the sequences for all AMO treatments in each species (Figure 5.3). Both the 5' and 3' splice sites had very strong dinucleotide invariance in exons that were not skipped during splicing. By comparison, these same 5' and 3' dinucleotides had a small number of loci (5-15%) with altered sequences in skipped exons in all organisms. While the change was rather small, this highlights the importance of this specific two base pair consensus sequence in maintaining inclusion for an exon during transcription.

For a more quantitative analysis, I scored the splice site sequences used in included or excluded exons against all known splice sites in the genome. Using the MaxEnt scoring system resulted in lower median scores for the skipped 5' and 3' splice sites, as well as a considerable number of negative score outliers, in comparison to those exons that were included (Figure 5.4). It is important to note that these type of scoring systems have no recognized cutoff for determining usage. In this case, the lower values in the skipped exons indicated less overall conformity to consensus splice site sequences, which could

reduce overall exon definition by U1 and U2. Additionally, or perhaps alternatively, the extreme negative scores seen in some outliers could be the underlying reason for exclusion as they indicate a splice site that may not be identifiable by one of the snRNPs. Here, a potential explanation is that the associated exon relies heavily on either U1 or U2, but not both, to initiate its definition before the other is recruited and the process of splicing begins.

As the difference between splice site sequences for included and excluded exons within samples was greater than the sequences across samples, this suggests that multi-exon skipping is present at low levels in normal cells. The baseline level of exon skipping is likely from pol II not slowing or pausing long enough at exons, a process known to play a role in cassette exon inclusion or exclusion (Carrillo Oesterreich, Bieberstein, and Neugebauer 2011; Oesterreich et al. 2016). In this scenario, splice site sequences that conform less to the overall consensus may require more time for definition when recruiting U1 or U2. Another possible explanation to the presence of baseline multi-exon skipping lies in the branch point sequence, where U2 binds. Any disruption here, similar to a loss of U1 binding, would reduce exon definition and promote exon exclusion. The position of this site upstream of the 3'ss, however, can vary by as much as a few hundred nucleotides, making it difficult to map precisely. I am in the process of scanning for this sequence in the included or skipped exons detected from my analysis. Despite the difficulty, this is a promising opportunity, and should be explored further.

Read Through Transcription and Splicing

During visualization of the data on the UCSC Genome Browser, I noticed between a few instances of read-through transcription between neighboring, same strand genes (Figure 5.5; between 7-154 genes across all organisms, full list of genes found in Supplementary

Table 3). Importantly, many of these cases also included splicing from one gene into another, where the intergenic space becomes part of a novel intron between exon skipping as described above. Transcription between these conjoined genes would then, most likely, end at the canonical 3' end of the downstream gene, although U1 AMO treatments could induce transcription termination within the downstream gene rather than at the canonical terminus. Read through transcription has been described previously, and as many as 800 conjoined genes have been identified, mainly through computational methods and then validated experimentally (Akiva et al. 2006; Prakash et al. 2010; Kim et al. 2012). The novel conjoined transcripts could then produce chimera proteins, potentially increasing protein complexity within the genome (Thomson et al. 2000). Open reading frame prediction for these chimeras, however, suggested that the majority of them produce non-functional proteins, as only 16% utilized conserved reading frames (Prakash et al. 2010).

I extracted all the read through splicing events from the same samples that I used in my multi-exon skipping examinations: my U1 AMO treatments in fruit fly, zebrafish, mouse together with the low U1 AMO dose response from Jung-Min Oh. Also, similar to the multi-exon skipping analysis, the changes in levels of read through splicing must take into consideration the changes in annotated exon-exon junction splicing. In order to make this comparison, I averaged the splicing from the flanking genes across which the read through was occurring. In comparison to the control, U1 AMO increased the overall levels of read through splicing in almost all samples, with the exception of the lose dose (1nmol) in mouse and the very low dose (0.01nmol) in human (Table 9). When comparing the distribution of the read through events, however, not all samples were significantly increased in comparison to the control (Figure 5.6). This is most likely because the overall read numbers in human are skewed by a few outliers that increase in depth after U1 AMO.

As I was interested in what could cause the read through splicing, I looked for correlation between the depth at each site compared to the intergenic distance between the flanking genes, as well as their splicing depth and expression. None of these, however, demonstrated any significant correlation (data not shown). Instead, I examined the events and flanking genes visually to both verify the read through and manually discover potential cases of interest. There was at least one instance where the flanking genes were noteworthy due to correlations to disease; in particular, read through between *UBE2D3* and *MANBA* that can be seen with 1nmol U1 AMO in HeLa (Figure 5.5).

UBE2D3 encodes a protein that is a part of an enzymatic family which acts to conjugate ubiquitin to specific target proteins, which in turn causes the degradation of the target by the proteasome (Jensen et al. 1995). The E2 family of ubiquitin ligases, in particular, prevents the accumulation of the tumor suppressor p53 in unstressed cells (Saville et al. 2004). Regulation of p53 has been heavily studied and is believed to serve many roles in cells; one of the most well-known of these is that upregulation of the protein in response to stress induces apoptosis in order to prevent the survival of potentially malignant cells (Ryan, Phillips, and Vousden 2001). p53's proper function is critical for protecting against cancer development. The second gene in this read through transcription event, *MANBA*, encodes the β-mannosidase enzyme that functions in the lysosome to degrade disaccharides. Mutations in this gene cause β-mannosidosis, a lysosomal storage disease that affects neurological development but is rare in humans (Alkhayat, Kraemer, and Leipprandt 1998; Blomqvist et al. 2019). In addition, polymorphic CA repeats in *MANBA* have also recently been linked to colorectal cancer in Swedish patients (Gao et al. 2008).

Due to the very recent discovery of these read through splicing events, this research, even in its infancy, provides an exciting new area of study for transcription regulation by U1.

The lower U1 AMO dose where these events occur in the human system are more physiologically relevant than complete base pairing inhibition that leads to extensive PCPA. As mentioned previously, low U1 AMO mimics transcriptional changes that are seen with neuronal activation (Berg et al. 2012). As such, cell stress or stimulation could also induce read through transcription and splicing between genes such as *UBE2D3* and *MANBA*. There are several possible explanations for the unresolved elements of *MANBA*'s role in neurological development and p53's many roles including cancer regulation, if the chimera protein of UBE2D3:β-mannosidase turns out to be stable *in vivo*. For example, it could act outside of normal regulatory pathways to repress p53 in the early stages of malignancy, or it could potentially form toxic aggregates. Further studies could determine whether this fusion is produced in certain cancer cell or patient populations, as well as whether the chimera protein is stable and functional.

Overall, this is an interesting result which has potentially important implications; current research into read through splicing is ongoing by several of my laboratory colleagues. While we have not done much analysis into these events, it is curious as to why U1 inhibition increases these numbers as a reduction in telescripting, from less free U1, should induce shorter transcripts. A potential theory for this is that mild U1 inhibition, which is also critical in defining the terminal exon, causes polymerases to transcribe past the last 3'ss without pausing for splicing. Much more can be done experimentally and analytically for these events. My work above highlighted only one of 80-120 read through splicing events found in human; as a consequence, it is possible that there are additional clinically relevant or scientifically compelling examples to study. I hope to help continue this work going forward.

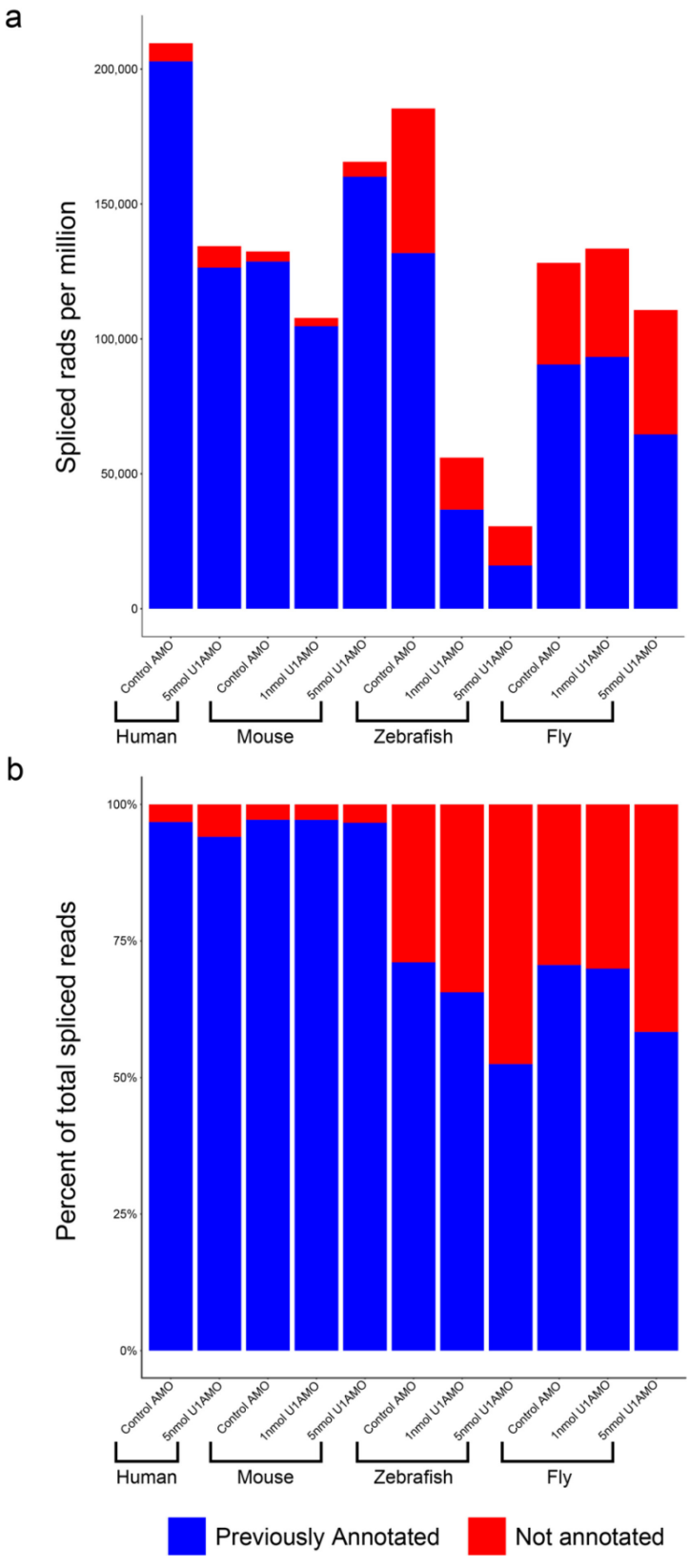

a
Spliced rads per million
200,000
150,000
100,000
50,000
0
Control AMO
5nmol U1AMO
Control AMO
1nmol U1AMO
5nmol U1AMO
Control AMO
1nmol U1AMO
5nmol U1AMO
Control AMO
1nmol U1AMO
5nmol U1AMO
Human
Mouse
Zebrafish
Fly
b
Percent of total spliced reads
100%
75%
50%
25%
0%
Control AMO
5nmol U1AMO
Control AMO
1nmol U1AMO
5nmol U1AMO
Control AMO
1nmol U1AMO
5nmol U1AMO
Control AMO
1nmol U1AMO
5nmol U1AMO
Human
Mouse
Zebrafish
Fly
Previously Annotated
Not annotated

Figure 5.1 U1 AMO reduces overall splicing and splicing fidelity

(a) Stacked bar graphs showing total *de novo* spliced read numbers, identified by TopHat v2.1.1 (Trapnell, Pachter, and Salzberg 2009), normalized to the total mapped read number for control or U1 AMO transfected fly, zebrafish, mouse, or human cells. Spliced reads are grouped by those that align to consecutive, annotated 5' and 3' splice sites (Previously Annotated, blue) and those that do not (Not Annotated, red). (b) Stacked bar graphs showing the same data as in a, but scaled to be a representation of splice junction type by total percent.

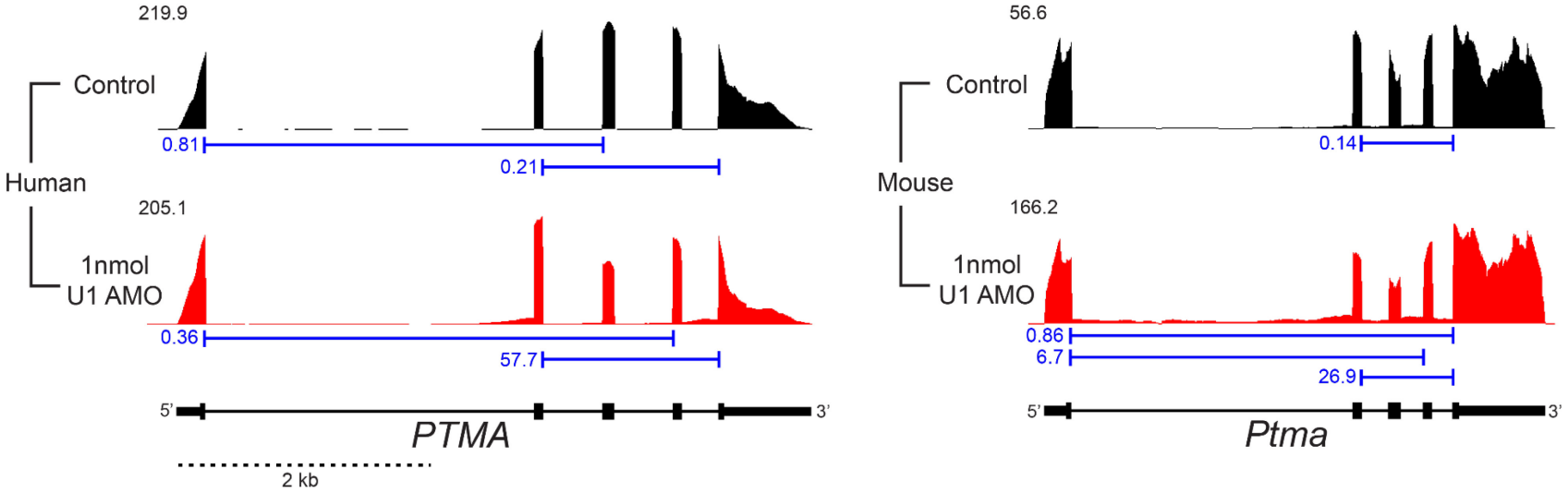

219.9
Control
0.81
0.21
Human
205.1
1nmol
U1 AMO
0.36
57.7
5'
3'
PTMA
2 kb
56.6
Control
0.14
Mouse
166.2
1nmol
U1 AMO
0.86
6.7
26.9
5'
3'
Ptma

Figure 5.2 U1 AMO increases multi-exon skipping in human and mouse

Genome browser view of the *PTMA* gene with 30 minute, 4sU labeled RNAseq reads and spliced junctions spanning multiple exons from control (black) and U1 (red) AMO treated NIH-3T3 (Mouse) and HeLa (Human) cells. Blue lines with vertical end-brackets show the splicing location of multi-exon skipping junctions. Numbers to the left of the read distributions and splice junctions show the highest peak height value in the field and spliced reads, respectively, as normalized to the total mapped reads. For gene structure, lines depict introns, boxes depict exons, and thinner boxes depict UTRs. Genomic distances are shown as dashed black lines.

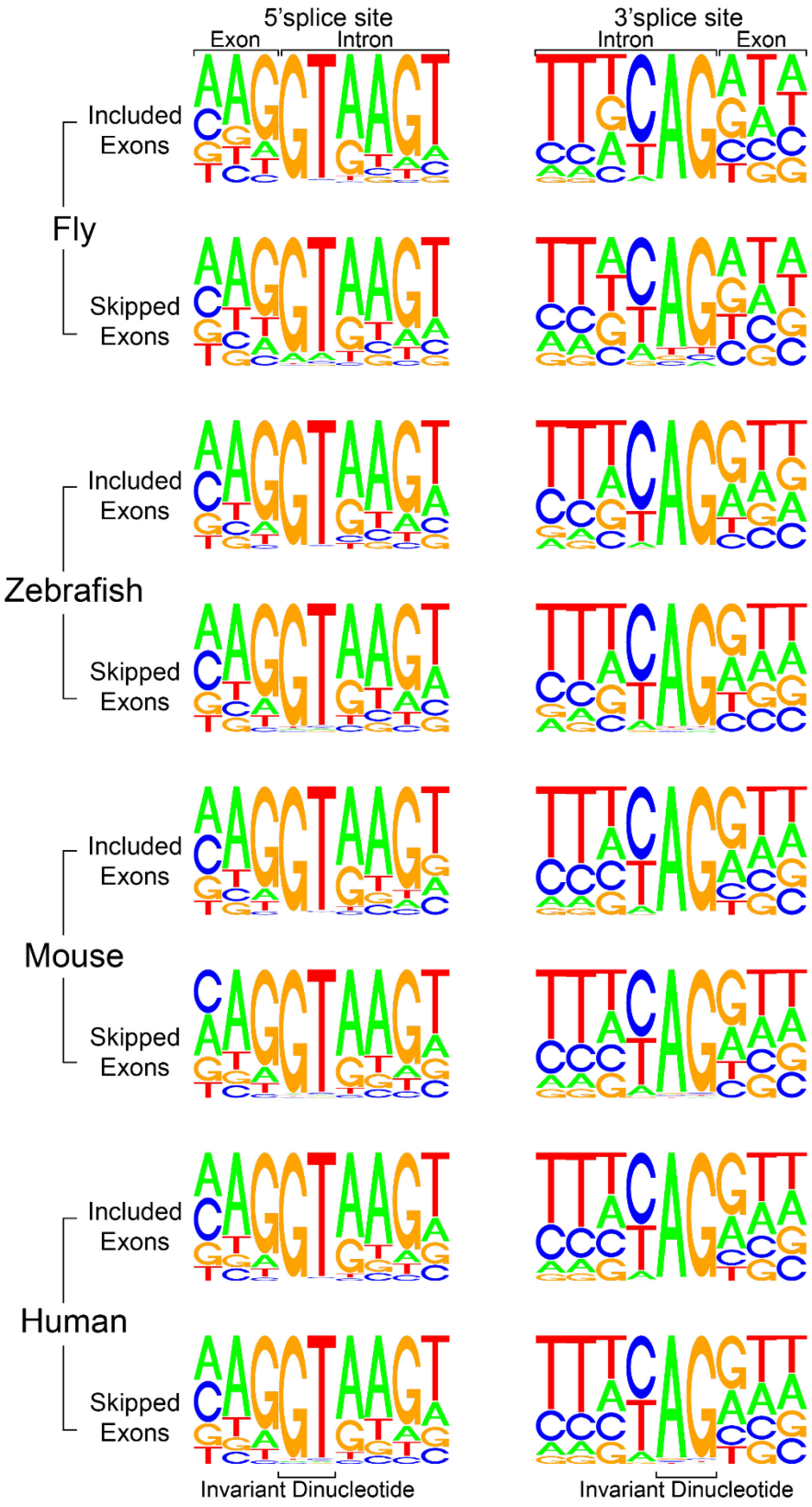

5'splice site
Exon
Intron
3'splice site
Intron
Exon
Fly
Included Exons
Skipped Exons
Zebrafish
Included Exons
Skipped Exons
Mouse
Included Exons
Skipped Exons
Human
Included Exons
Skipped Exons
Invariant Dinucleotide
Invariant Dinucleotide

Figure 5.3 Skipped exons contain less invariance at the critical dinucleotides at the 5’ and 3’ splice sites

Nine nucleotide splice site logos for the 5’ and 3’ splice sites for fly, zebrafish, mouse, and human organisms. Taller nucleotide letters indicate higher prevalence of that base at the position. Splice sites are grouped by those that flank exons which are always included in previously annotated splice isoforms (Included), and those that are excluded from a novel multi-exon skipping event (Skipped) as identified in the text from 30 minute, 4sU labeled and U1 AMO transfected RNAseq datasets. Sequences contain three nucleotides of the exon and six nucleotides of the intron as bracketed above the logos. The invariant dinucleotides are bracketed below the logos.

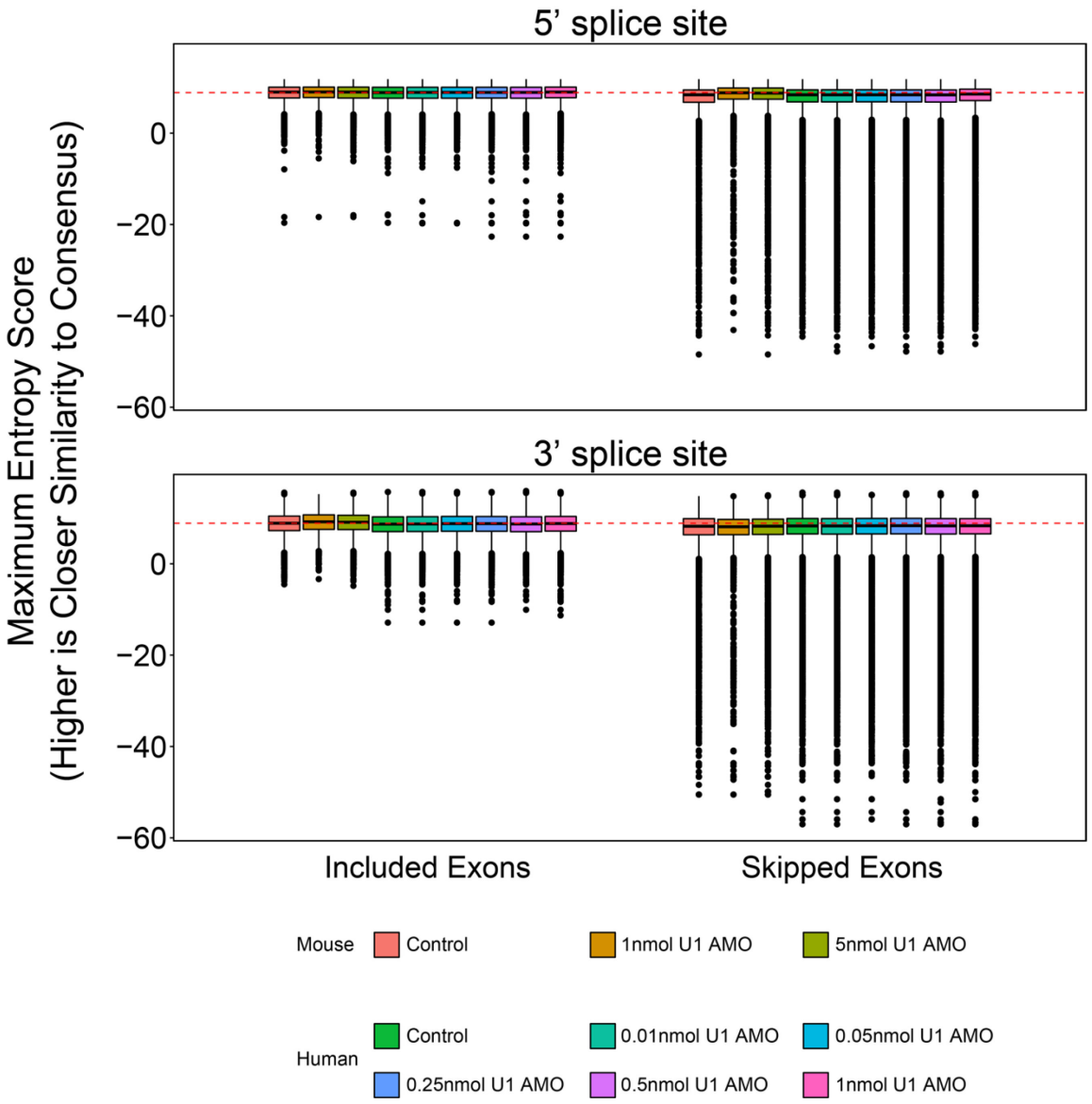

5’ splice site
3’ splice site
Maximum Entropy Score
(Higher is Closer Similarity to Consensus)
0
−20
−40
−60
Included Exons
Skipped Exons
Mouse
Control
1nmol U1 AMO
5nmol U1 AMO
Human
Control
0.01nmol U1 AMO
0.05nmol U1 AMO
0.25nmol U1 AMO
0.5nmol U1 AMO
1nmol U1 AMO

Figure 5.4 Skipped exons have lower splice site consensus sequence

Boxplots showing the maximum entropy score for 5' and 3' splice site sequences in mouse and human for exons that are either skipped or included in multi-exon skipping events as identified in the text from 30 minute, 4sU labeled and U1 AMO transfected RNAseq datasets. Splice site consensus strength was measured using MaxEntScan (Yeo and Burge 2004). Red dashed line indicates the mean maximum entropy score (8.8) of the included 5' and 3' splice site scores. Values above zero represent scores closer to the consensus splice site sequences, negative values represent less consensus sequences.

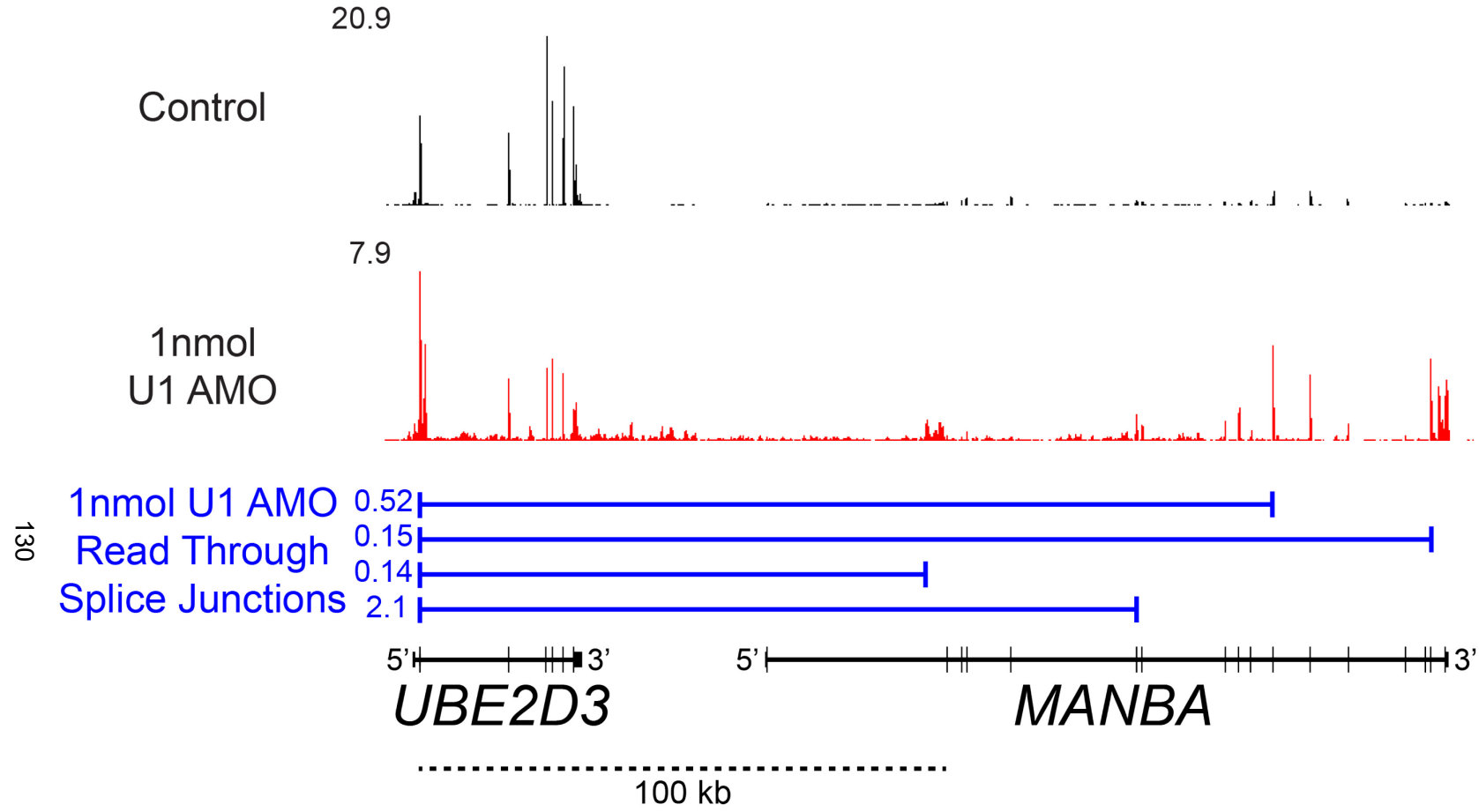

20.9
Control
7.9
1nmol
U1 AMO
1nmol U1 AMO
Read Through
Splice Junctions
0.52
0.15
0.14
2.1
5'
3'
5'
3'
UBE2D3
MANBA
100 kb

Figure 5.5 Low U1 AMO read through transcription and splicing

Genome browser view of the genes *UBE2D3* and *MANBA* with 30 minute, 4sU labeled RNAseq reads and read through spliced junctions from control (black) and U1 (red) AMO treated and HeLa cells. Blue lines with vertical end-brackets show the splicing location of read through splice junctions from the low U1 AMO condition. Numbers to the left of the read distributions and splice junctions show the highest peak height value in the field and spliced reads, respectively, as normalized to the total mapped reads. For gene structure, lines depict introns, boxes depict exons, and thinner boxes depict UTRs. Genomic distances are shown as dashed black lines.

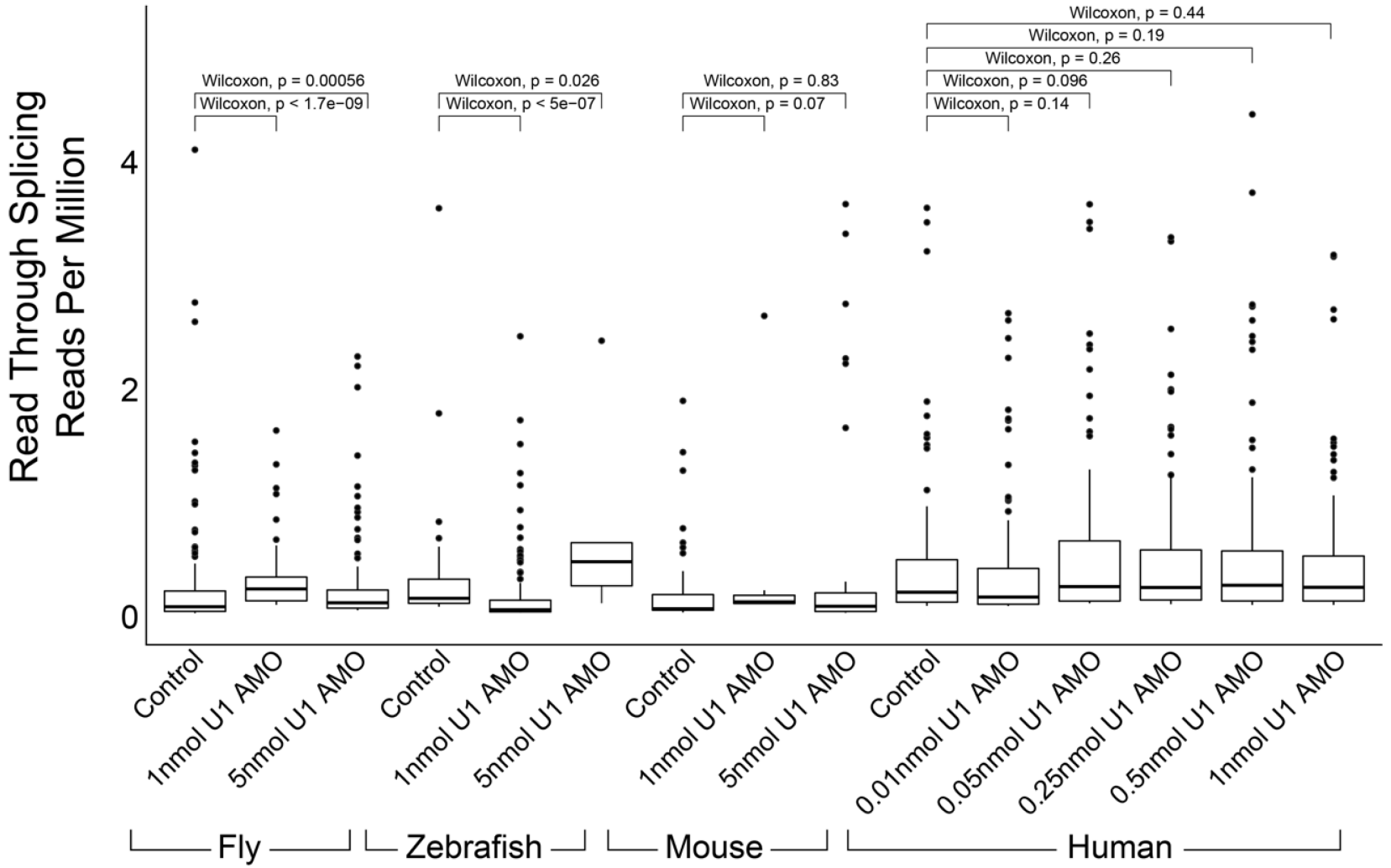

Read Through Splicing
Reads Per Million
0
2
4
Wilcoxon, p = 0.00056
Wilcoxon, p < 1.7e−09
Wilcoxon, p = 0.026
Wilcoxon, p < 5e−07
Wilcoxon, p = 0.83
Wilcoxon, p = 0.07
Wilcoxon, p = 0.44
Wilcoxon, p = 0.19
Wilcoxon, p = 0.26
Wilcoxon, p = 0.096
Wilcoxon, p = 0.14
Control
1nmol U1 AMO
5nmol U1 AMO
Control
1nmol U1 AMO
5nmol U1 AMO
Control
1nmol U1 AMO
5nmol U1 AMO
Control
0.01nmol U1 AMO
0.05nmol U1 AMO
0.25nmol U1 AMO
0.5nmol U1 AMO
1nmol U1 AMO
Fly
Zebrafish
Mouse
Human

Figure 5.6 Low U1 AMO increases read through transcription

Boxplots showing the distribution of reads that splice between two adjacent, same strand transcripts for 30 minute, 4sU labeled RNAseq data of fly, zebrafish, mouse, and human cells transfected with either control or U1 AMO. Read values were normalized to the total mapped reads in each sample. Distribution differences between the control and U1 AMO transfected samples were tested for significance using a Wilcoxon signed-rank test.

a

		Splicing Event Type Reads Per Million				Percent of Total		
Organism	Treatment	Annotated	Multi-Exon	Other	Total	Annotated	Multi-Exon	Other
Fly	Control	194,137	127	23,446	217,709.93	89.17%	0.06%	10.77%
	1nmol U1 AMO	200,030	148	17,251	217,429.47	92.00%	0.07%	7.93%
	5nmol U1 AMO	138,488	266	35,243	173,997.31	79.59%	0.15%	20.26%
Zebrafish	Control	264,204	807	8,495	273,505.41	96.60%	0.29%	3.11%
	1nmol U1 AMO	71,442	2,985	3,299	77,726.69	91.91%	3.84%	4.24%
	5nmol U1 AMO	29,905	552	1,714	32,170.74	92.96%	1.72%	5.33%
Mouse	Control	242,047	582	4,023	246,652.08	98.13%	0.24%	1.63%
	1nmol U1 AMO	192,669	1,170	2,035	195,873.12	98.36%	0.60%	1.04%
	5nmol U1 AMO	299,868	2,049	4,099	306,015.60	97.99%	0.67%	1.34%
Human	Control	2,428,084	16,494	54,879	2,499,457	97.14%	0.66%	2.20%
	0.01nmol U1 AMO	467,243	3,830	12,159	483,232	96.69%	0.79%	2.52%
	0.05nmol U1 AMO	493,516	4,091	13,405	511,013	96.58%	0.80%	2.62%
	0.25nmol U1 AMO	444,571	4,201	14,105	462,877	96.05%	0.91%	3.05%
	0.5nmol U1 AMO	464,880	4,530	13,063	482,473	96.35%	0.94%	2.71%
	1nmol U1 AMO	245,106	9,293	9,406	263,804	92.91%	3.52%	3.57%

b

		Number of Splicing Events By Type				Percent of Total		
Organism	Treatment	Annotated	Multi-Exon	Other	Total	Annotated	Multi-Exon	Other
Fly	Control	77,746	37,460	1,296	116,502	66.73%	32.15%	1.11%
	1nmol U1 AMO	72,348	15,938	678	88,964	81.32%	17.92%	0.76%
	5nmol U1 AMO	74,140	24,984	1,648	100,772	73.57%	24.79%	1.64%
Zebrafish	Control	154,398	17,350	4,532	176,280	87.59%	9.84%	2.57%
	1nmol U1 AMO	120,466	17,994	33,926	172,386	69.88%	10.44%	19.68%
	5nmol U1 AMO	48,826	3,408	3,026	55,260	88.36%	6.17%	5.48%
Mouse	Control	192,074	35,000	9,912	236,986	81.05%	14.77%	4.18%
	1nmol U1 AMO	124,088	6,724	5,292	136,104	91.17%	4.94%	3.89%
	5nmol U1 AMO	130,710	21,764	13,668	166,142	78.67%	13.10%	8.23%
Human	Control	582,370	79,605	247,365	909,340	64.04%	8.75%	27.20%
	0.01nmol U1 AMO	117,255	17,504	55,295	190,054	61.70%	9.21%	29.09%
	0.05nmol U1 AMO	116,850	15,641	50,516	183,007	63.85%	8.55%	27.60%
	0.25nmol U1 AMO	117,845	17,035	54,345	189,225	62.28%	9.00%	28.72%
	0.5nmol U1 AMO	117,525	19,464	56,702	193,691	60.68%	10.05%	29.27%
	1nmol U1 AMO	101,290	28,368	40,365	170,023	59.57%	16.68%	23.74%

Table 7 U1 AMO increases both multi-exon and other splicing events

(a) Numbers showing the splice junction reads from 30 minute, 4sU labeled RNAseq of fly, zebrafish, mouse, and human cells transfected with either control (black) or U1 (red) AMO. Also shown is the percentage of reads in each group as a fraction of the total spliced reads. (b) Same as in a, except showing the total splicing events (i.e. breadth).

		Spearman Correlation of Exon Number to Splicing Event Type			Spearman Correlation of Exon Number to Splicing Depth By Event Type		
Organism	Treatment	Annotated	Multi-Exon	Other	Annotated	Multi-Exon	Other
Fly	Control	0.877	0.182	0.431	0.588	0.181	0.401
	1nmol U1 AMO	0.848	0.152	0.376	0.598	0.152	0.364
	5nmol U1 AMO	0.866	0.213	0.401	0.597	0.212	0.388
Zebrafish	Control	0.962	0.267	0.228	0.573	0.264	0.218
	1nmol U1 AMO	0.952	0.637	0.258	0.526	0.535	0.241
	5nmol U1 AMO	0.683	0.227	0.118	0.326	0.217	0.115
Mouse	Control	0.961	0.319	0.362	0.589	0.303	0.323
	1nmol U1 AMO	0.957	0.107	0.128	0.479	0.093	0.119
	5nmol U1 AMO	0.962	0.209	0.142	0.403	0.154	0.108
Human	Control	0.943	0.346	0.408	0.576	0.303	0.336
	0.01nmol U1 AMO	0.944	0.352	0.421	0.591	0.306	0.339
	0.05nmol U1 AMO	0.942	0.341	0.387	0.571	0.294	0.313
	0.25nmol U1 AMO	0.942	0.355	0.412	0.589	0.309	0.335
	0.5nmol U1 AMO	0.942	0.361	0.442	0.601	0.318	0.367
	1nmol U1 AMO	0.940	0.389	0.394	0.555	0.322	0.323

Table 8 Mis-splicing does not correlate with exon number

Numbers showing the Spearman correlation of splicing events and splicing reads to exon number as broken down by splice junction type. Shown is the 30 minute, 4sU labeled RNAseq data for fly, zebrafish, mouse, and human cells transfected with either control (black) or U1 (red) AMO.

		Read Through Splicing		Flanking Genes	
Organism	Treatment	Events	Reads Per Million	Average Reads Per Million	Read Through as a Percent of Flanking
Fly	Control	153	38.9	3,330.2	1.17%
	1nmol U1 AMO	65	21.7	1,660.6	1.31%
	5nmol U1 AMO	128	32.5	2,333.7	1.39%
Zebrafish	Control	34	12.9	507.1	2.55%
	1nmol U1 AMO	145	25.8	783.6	3.30%
	5nmol U1 AMO	8	5.3	53.9	9.85%
Mouse	Control	48	11.4	439.7	2.59%
	1nmol U1 AMO	7	3.5	517.5	0.68%
	5nmol U1 AMO	36	19.0	875.9	2.17%
Human	Control	88	44.0	10,559.0	0.42%
	0.01nmol U1 AMO	113	46.9	10,695.8	0.44%
	0.05nmol U1 AMO	102	59.8	7,381.4	0.81%
	0.25nmol U1 AMO	97	52.5	9,758.4	0.54%
	0.5nmol U1 AMO	110	62.7	10,618.2	0.59%
	1nmol U1 AMO	121	55.9	8,691.8	0.64%

Table 9 U1 AMO increases the amount of read through splicing

Numbers showing the number of read through splicing events and reads for 30 minute, 4sU labeled RNAseq data fly, zebrafish, mouse, and human cells transfected with either control (black) or U1 (red) AMO. Also shown is the average number of annotated splice junction reads from the flanking (upstream and downstream) genes between the read through events as well as the read through splicing as a percentage of this flanking gene splicing.

CHAPTER 6: IMPACT OF THIS WORK AND FUTURE DIRECTIONS

U1 Telescripting Has Shaped Gene Structure Evolution

When I began work in the Dreyfuss lab, the most recent publications from the group had demonstrated the existence of U1 telescripting, a function separate to its role in splicing (Kaida et al. 2010; Berg et al. 2012). This discovery challenged the commonly-held belief in the RNA processing and transcription fields, that pol II processivity is extremely robust. Before this, the existing theory regarding transcription was that complete elongation through the entire length of a transcript is inexorable once the polymerase was engaged past the promoter. Our lab's research into U1 telescripting has demonstrated that this is not true, and that full-length transcription is not the default state in metazoans. The genomic tiling array and HIDE-seq experiments conducted in our lab leading to these discoveries were soon replaced with larger and more general RNAseq experiments. As I highlighted earlier in this work, the novelty of telescripting meant that there were no available tools or methods to identify or study PCPA or transcription shortening. As a consequence, I created the tools necessary for the analysis of telescripting from RNAseq datasets. I combined this computational work with biological experiments to examine the mechanism of PCPA and the conservation of U1 telescripting in different organisms.

My pol II ChIPseq experiments demonstrated that loss of telescripting, achieved through U1 AMO but also possible due to transcription up-regulation, caused PCPA through transcription termination at actionable PASs within large introns of genes (Oh et al. 2017). This observation was critical to the understanding of telescripting as a co-transcriptional process, rather than from a post-transcriptional degradation of 3' side pre-mRNA. Importantly, this also showed that transcription initiation was not inhibited by U1 AMO, but

was increased slightly in PCPAed genes. We postulate that this is most likely due to increased polymerase recycling from prematurely terminated transcripts in the same gene.

My analytical work, calculating the important parameters in PCPAed genes from RNAseq data, allowed for the identification of U1 AMO induced transcription termination. This also revealed the stratification in gene size between PCPAed genes and up-regulated genes with U1 base pairing inhibition. Moreover, the difference in functional enrichment across these different gene size groups is an important piece of evidence that helps in our understanding of how U1 telescripting acted as one potential cause behind gene size expansion. Using orthology data and U1 AMO RNAseq in fly, mouse, and zebrafish, I have demonstrated that large transcripts, which typically contain more neuronal and developmental genes, experienced the most extreme intron expansion (Bertagnolli et al. 2013; Gabel et al. 2015). These larger introns also increase the likelihood of an actionable PAS arising stochastically, making these transcripts more susceptible to PCPA.

The up-regulation of small genes with U1 AMO was a surprising result, and was consistent across the organisms I studied. The genes, highly enriched for cell stimuli response or growth functions, continued to splice and produce full-length mRNA even with virtually no free U1 present in these cells. Existing possible explanations for this result included, for example, that U1 AMO transfections could produce a stress response, or that specific down-regulated genes could be transcription inhibitors. However, my analysis of hnRNP binding after inducing PCPA suggests an alternative mechanism, one where large genes are sacrificed in order to provide critical pre-mRNA processing resources to smaller, more acute cell response genes. Further evidence for this was found in my organism RNAseq, which showed that these two functionally disparate gene groups are also found in close

proximity across evolution. This shorter intergenic distance would make the transfer of these resources much quicker and more reliable.

While there has been research into the evolution of gene size and the purpose of large introns, our work provides a new and sharply different perspective (Bradnam and Korf 2008; Catania and Lynch 2008; Chorev and Carmel 2012; Gelfman et al. 2012; Rogozin et al. 2012). Centered around U1 telescripting, intron size expansion has been allowed, or even promoted, by genome evolution in non-critical “luxury” genes, rather than occurring randomly and tolerated by metazoan cells. Evolutionary pressure to maintain the size of small genes while allowing intron expansion in large genes is also supported by the data that intron length and splice site strength have been shown to be inversely linked (Gelfman et al. 2012). Introns flanked by weak 5’ss are shorter than average, while longer introns are found in closer proximity to stronger 5’ss. This purifying selection against intron expansion near weak splice sites could be explained by the necessity for stronger U1 interactions in order to prevent PCPA from occurring during the common life cycle of a cell. The enrichment for cell proliferation and glucose response in small genes further supports this idea.

The identification of PCPA and the stratified size-function genome has formed a foundation for a new approach to study transcription and the RNA processing machinery. One of the most pressing questions is the mechanism of PCPA itself. Though our data has shown that the transcripts are cleaved and polyadenylated, which suggest a process at least similar to that found at canonical 3’ gene ends, the actual mechanism of action remains unknown. Understanding which factors are directly involved at PCPA sites, through XLIPseq or other cross-linking experiments (CLIPseq or PAR-CLIP) with U1 AMO, would be a good first step towards elucidating this mechanism. While there has recently

been confirmation that U1 binds throughout introns, a study of the change in locations with U1 AMO or cell stimulation could help determine the change in binding that elicits PCPA (Engreitz et al. 2014). This could also be achieved with XLIP or CLIPseq with antibodies for U1 specific proteins such as U170K, U1A, or U1C.

On the matter of up-regulation with U1 AMO in small genes, it is important to understand how these maintain mRNA output when U1 cannot base pair to the 5'ss. While U1 protein XLIP/CLIP as mentioned above would help for further studies, mass spectrometry and XLIPseq of other spliceosomal proteins would provide an alternative method to test this. If, for example, U1 was not base pairing to the 5'ss in these genes, pull downs of U2 specific proteins (ex. U2A', U2B", SF3A1, or SF3B1) or exon junction complex proteins (ex. PRP8, RNPS1, or Y14) could show the composition of the spliceosome at different stages on these introns.

More narrow identification of the PCPA site, perhaps even down to the level of base pair specificity, has been a long-time goal in the lab. However, current RNAseq data may not be sufficiently informative to identify many PCPA sites, from immediate transcript termination to multiple less active sites across an intron, with acceptable accuracy due to the varied types of PCPA as seen manually on a genome browser. Despite this, other mathematical and statistical approaches could be applied to these data in an attempt for more precise PCPA site determination. For example, there are software packages available which identify 3'UTR shortening that could be adapted to examine introns or the entire gene (W. Wang, Wei, and Li 2014; Xia et al. 2014). Another analytical method that might prove useful would be to use 3'-end sequencing, which is typically accomplished through oligo-dT primers to target poly(A) sequences mRNA ends (de Hoon and Hayashizaki 2008; Beck et al. 2010; Shepard et al. 2011). A potential drawback of this

method is that the oligo-dT can potentially also prime to genomic poly(A) sequences within a gene and cause false positives during the analytical process. These specific types of sequencing experiments also limit the amount of data that is able to be used in downstream work, although this can be offset by doing total RNA sequencing in conjunction.

Deeper Implications for U1's Role in Exon Definition

Substantial research has been done investigating both mis-splicing, especially in regards to disease, and conjoined genes; yet the identification of new read through transcription sites and splicing events is still an important area in need of further research. My analytical work on multi-exon skipping provides a fast alternative to discovering novel splicing events at the gene level. While my workflow is simpler and more specific than other splicing programs available, its simplicity does not detract from its ability to detect novel splicing events. The exon skipping I discovered with U1 AMO transfected RNAseq data, highlights the importance of invariance at the critical dinucleotide splice site sequence. Reporter constructs using these lower consensus splice sites would help verify the presence of multi-exon skipping. From an analytical perspective, globally identifying the U2 binding sites in order to compare this sequence in skipped versus included exons could further shed light on the mis-splicing. Additionally, analyzing sequencing experiments from U2 or other spliceosomal protein knock downs may induce more of these skipping events, and allow for identification of any that were missed in our datasets, with the potential to build on our current understanding.

An unexpected result from my RNAseq work and analysis was the identification of novel read through transcription and splicing. Although this very recent result was not the main focus of my work, this discovery has created many research opportunities for the lab going

forward. Going forward, the laboratory should work to identify events that are currently not in any conjoined gene databases and examine them for significance. For example, one could consider whether the flanking genes are known to be associated with any disease states. From there, the chimera protein product can be assessed from sequence data and examined for function, cellular localization, stability, and any other pertinent criteria. Localization and stability studies can be accomplished through the creation of an antibody targeted to the fusion region of the two proteins. In the case of disease association, a more advanced study could express the chimera protein in an animal model, such as zebrafish or mouse, to search for chimeric protein accumulation.

Final Remarks

This thesis presents a series of experimental and analytical examinations that demonstrate the importance of U1 telescripting. The consequence of a transcriptional regulator that links gene size and function is significant, and this is, to my knowledge, the first evidence that these two gene features are connected. U1's function in splicing has been extensively researched, but this has proven to be only the tip of the iceberg with regards to its function in transcription regulation. There is a certain elegance in what we now know to be nature's utilization of U1: its 5' sequence of 9-nucleotides is necessary for transcription to continue far enough within a gene so that the same sequence can then be used for splicing (Venters et al. 2019). It is not clear, however, which of these functions came first in evolution. Perhaps further work will identify U1's origin and initial function; it may also identify even more mechanisms or processes regulated by this small ribonucleoprotein.

APPENDIX

Code deposition

Code for the projects listed in the Methods section can be found on GitHub (https://github.com/ChrisVenters/DreyfussLab) with the following sub-directories:

1. PCPA detection software: https://github.com/ChrisVenters/DreyfussLab/tree/master/PCPA_detection
2. Ortholog extraction and overlap: https://github.com/ChrisVenters/DreyfussLab/tree/master/Ortholog_data
3. Multi-exon skipping and mis-splicing identification: https://github.com/ChrisVenters/DreyfussLab/tree/master/MultiExon_skipping
4. Read through splicing identification: https://github.com/ChrisVenters/DreyfussLab/tree/master/Read_through_splicing
5. Gene neighborhoods/intergenic distance calculations: https://github.com/ChrisVenters/DreyfussLab/tree/master/Gene_neighborhoods
6. Upstream antisense region creation: https://github.com/ChrisVenters/DreyfussLab/tree/master/Upstream_antisense_regions
7. snRNA alignments for XLIPseq: https://github.com/ChrisVenters/DreyfussLab/tree/master/snRNA_alignment

PCR primers

Mouse CRIM1 Exon 1 – Exon 2 Junction, Forward:
CTCCATCACCGAGTACGAAGTG

Mouse CRIM1 Exon 1 – Exon 2 Junction, Reverse:
GGTTTTCATTGCAGGGTTCAA

Mouse CRIM1 Intron 1, Forward:
CGGGCGCGGAACAAA

Mouse CRIM1 Intron 1, Reverse:
CCCCGGGAACCCTCTCT

Mouse GLUD1 Exon 1 – Exon 2 Junction, Forward:
GCAAGGGAGGTATCCGTTACAG

Mouse GLUD1 Exon 1 – Exon 2 Junction, Reverse:
TCATTAAGGAAGCCAGTGCTTTT

Mouse GLUD1 Intron 1, Forward:
GCCTCGGTGCCTGTCATG

Mouse GLUD1 Intron 1, Reverse

GTCGGCTGCTGGCGTTA

Mouse MYO10 Exon 1, Forward:
GGAGCACTTCGCCCAGAA

Mouse MYO10 Exon 1, Reverse:
CCCAACCACAGCCTTTGTCT

Mouse MYO10 Exon 2, Forward:
CGGGTCTGGCTAAGAGAAAATG

Mouse MYO10 Exon 2, Reverse:
GCCTTCTGCACAGGAATTTACA

Mouse MYO10 Intron 1, Forward:
GCCGGCCCCGAGTCT

Mouse MYO10 Intron 1, Reverse:
TGCGAGCGCAGGACAAA

Fly NIPPEDA Exon 1 – Exon 2 Junction, Forward:
AAATACAATGCAACACATACGTAAACTG

Fly NIPPEDA Exon 1 – Exon 2 Junction, Reverse:
TTGTCTTTACATGTTGCCTTAAGCTT

Fly NIPPEDA Intron 1, Forward:
TTCGGTGCCCTAGTAACAATCTTT

Fly NIPPEDA Intron 1, Reverse:
CACAATGCACGGTCGTTTTAA

Fly SPEN Exon 1, Forward:
CGCGCGTCGTTTGCA

Fly SPEN Exon 1, Reverse:
GCTCGCCGCCGTTGT

Fly SPEN Exon 2, Forward:
GGAGGGCGCCCATTTAAA

Fly SPEN Exon 2, Reverse:
GGAGCATTGCGTCGTATGC

Fly SPEN Intron 1, Forward:
CGCAAATGGTAAGAATGGATCA

Fly SPEN Intron 1, Reverse:

TTTCCGTGAATGGATTATTTGCT

Fly TENM Exon 1, Forward:
CGCACGGTCGGAATTGTC

Fly TENM Exon 1, Reverse:
CGCTGTTTTCGCTACTTTTCG

Fly TENM Exon 2, Forward:
CGGCACCACCGGATGTT

Fly TENM Exon 1, Reverse:
CCATTTTGCATGCGACTCAT

Fly TENM Intron 1, Forward:
GCAGTGTGGTCAGTGGAATCA

Fly TENM Intron 1, Reverse:
CGCTCTATTTGGGCCACTAAA

Human AFM Promoter, Forward:
TGTGCATACTTAGCCTGTGGACTT

Human AFM Promoter, Reverse:
TTACCTTTGTGTTTCGCTGGAA

Human GAPDH Promoter, Forward:
GGTGCGTGCCCAGTTGA

Human GAPDH Promoter, Reverse:
CTACTTTCTCCCCGCTTTTTTTT

Human GAPDH Exon 1, Forward:
CCTCCCGCTTCGCTCTCT

Human GAPDH Exon 1, Reverse:
GGCGACGCAAAAGAAGATG

BIBLIOGRAPHY

Akiva, Pinchas, Amir Toporik, Sarit Edelheit, Yifat Peretz, Alex Diber, Ronen Shemesh, Amit Novik, and Rotem Sorek. 2006. “Transcription-Mediated Gene Fusion in the Human Genome.” *Genome Research* 16 (1): 30–36.

Alkhayat, A. H., S. A. Kraemer, and J. R. Leipprandt. 1998. “Human β-Mannosidase cDNA Characterization and First Identification of a Mutation Associated with Human β-Mannosidosis.” *Human Molecular Genetics*. https://academic.oup.com/hmg/article-abstract/7/1/75/640017.

Almada, Albert E., Xuebing Wu, Andrea J. Kriz, Christopher B. Burge, and Phillip A. Sharp. 2013. “Promoter Directionality Is Controlled by U1 snRNP and Polyadenylation Signals.” *Nature* 499 (7458): 360–63.

Baralle, Francisco E., and Jimena Giudice. 2017. “Alternative Splicing as a Regulator of Development and Tissue Identity.” *Nature Reviews. Molecular Cell Biology* 18 (7): 437–51.

Baserga, Susan J., and Joan A. Steitz. 1993. “The Diverse World of Small Ribonucleoproteins.” *Cold Spring Harbor Monograph Series* 24: 359–359.

Battle, Daniel J., Chi-Kong Lau, Lili Wan, Hongying Deng, Francesco Lotti, and Gideon Dreyfuss. 2006. “The Gemin5 Protein of the SMN Complex Identifies snRNAs.” *Molecular Cell* 23 (2): 273–79.

Beck, Andrew H., Ziming Weng, Daniela M. Witten, Shirley Zhu, Joseph W. Foley, Phil Lacroute, Cheryl L. Smith, et al. 2010. “3’-End Sequencing for Expression Quantification (3SEQ) from Archival Tumor Samples.” *PloS One* 5 (1): e8768.

Bembom, Oliver. 2014. “Sequence Logos for DNA Sequence Alignments.” http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.431.3748.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57 (1): 289–300.

Berg, Michael G., Larry N. Singh, Ihab Younis, Qiang Liu, Anna Maria Pinto, Daisuke Kaida, Zhenxi Zhang, et al. 2012. “U1 snRNP Determines mRNA Length and Regulates Isoform Expression.” *Cell* 150 (1): 53–64.

Bertagnolli, Nicolas M., Justin A. Drake, Jason M. Tennessen, and Orly Alter. 2013. “SVD Identifies Transcript Length Distribution Functions from DNA Microarray Data and Reveals Evolutionary Forces Globally Affecting GBM Metabolism.” *PloS One* 8 (11): e78913.

Blomqvist, Maria, Marie Falkenberg Smeland, Julia Lindgren, Per Sikora, Hilde Monica Frostad Riise Stensland, and Jorge Asin-Cayuela. 2019. “Beta-Mannosidosis Caused by a Novel Homozygous Intragenic Inverted Duplication in MANBA.” *Cold*

Spring Harbor Molecular Case Studies, March. https://doi.org/10.1101/mcs.a003954.

Bowman, Elizabeth A., and William G. Kelly. 2014. "RNA Polymerase II Transcription Elongation and Pol II CTD Ser2 Phosphorylation: A Tail of Two Kinases." *Nucleus* 5 (3): 224–36.

Bradnam, Keith R., and Ian Korf. 2008. "Longer First Introns Are a General Property of Eukaryotic Gene Structure." *PloS One* 3 (8): e3093.

Carrillo Oesterreich, Fernando, Nicole Bieberstein, and Karla M. Neugebauer. 2011. "Pause Locally, Splice Globally." *Trends in Cell Biology* 21 (6): 328–35.

Catania, Francesco, and Michael Lynch. 2008. "Where Do Introns Come From?" *PLoS Biology* 6 (11): e283.

Cauchi, Ruben J. 2010. "SMN and Gemins: 'We Are Family' … or Are We?: Insights into the Partnership between Gemins and the Spinal Muscular Atrophy Disease Protein SMN." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 32 (12): 1077–89.

Chorev, Michal, and Liran Carmel. 2012. "The Function of Introns." *Frontiers in Genetics* 3 (April): 55.

Connelly, S., and J. L. Manley. 1988. "A Functional mRNA Polyadenylation Signal Is Required for Transcription Termination by RNA Polymerase II." *Genes & Development* 2 (4): 440–52.

Cramer, P., K-J Armache, S. Baumli, S. Benkert, F. Brueckner, C. Buchen, G. E. Damsma, et al. 2008. "Structure of Eukaryotic RNA Polymerases." *Annual Review of Biophysics* 37: 337–52.

Cramer, P., D. A. Bushnell, and R. D. Kornberg. 2001. "Structural Basis of Transcription: RNA Polymerase II at 2.8 Angstrom Resolution." *Science* 292 (5523): 1863–76.

Crooks, Gavin E., Gary Hon, John-Marc Chandonia, and Steven E. Brenner. 2004. "WebLogo: A Sequence Logo Generator." *Genome Research* 14 (6): 1188–90.

Darzacq, Xavier, Yaron Shav-Tal, Valeria de Turris, Yehuda Brody, Shailesh M. Shenoy, Robert D. Phair, and Robert H. Singer. 2007. "In Vivo Dynamics of RNA Polymerase II Transcription." *Nature Structural & Molecular Biology* 14 (9): 796–806.

David, Charles J., Alex R. Boyne, Scott R. Millhouse, and James L. Manley. 2011. "The RNA Polymerase II C-Terminal Domain Promotes Splicing Activation through Recruitment of a U2AF65-Prp19 Complex." *Genes & Development* 25 (9): 972–83.

Derti, Adnan, Philip Garrett-Engele, Kenzie D. Macisaac, Richard C. Stevens, Shreedharan Sriram, Ronghua Chen, Carol A. Rohl, Jason M. Johnson, and Tomas

Babak. 2012. “A Quantitative Atlas of Polyadenylation in Five Mammals.” *Genome Research* 22 (6): 1173–83.

Diatchenko, L., Y. F. Lau, A. P. Campbell, A. Chenchik, F. Moqadam, B. Huang, S. Lukyanov, et al. 1996. “Suppression Subtractive Hybridization: A Method for Generating Differentially Regulated or Tissue-Specific cDNA Probes and Libraries.” *Proceedings of the National Academy of Sciences of the United States of America* 93 (12): 6025–30.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21.

Du, Hansen, and Michael Rosbash. 2002. “The U1 snRNP Protein U1C Recognizes the 5’ Splice Site in the Absence of Base Pairing.” *Nature* 419 (6902): 86–90.

Egloff, Sylvain, and Shona Murphy. 2008. “Cracking the RNA Polymerase II CTD Code.” *Trends in Genetics: TIG* 24 (6): 280–88.

Engreitz, Jesse M., Klara Sirokman, Patrick McDonel, Alexander A. Shishkin, Christine Surka, Pamela Russell, Sharon R. Grossman, Amy Y. Chow, Mitchell Guttman, and Eric S. Lander. 2014. “RNA-RNA Interactions Enable Specific Targeting of Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites.” *Cell* 159 (1): 188–99.

Fang, Hai, Bogdan Knezevic, Katie L. Burnham, and Julian C. Knight. 2016. “XGR Software for Enhanced Interpretation of Genomic Summary Data, Illustrated by Application to Immunological Traits.” *Genome Medicine* 8 (1): 129.

Fischer, U., Q. Liu, and G. Dreyfuss. 1997. “The SMN-SIP1 Complex Has an Essential Role in Spliceosomal snRNP Biogenesis.” *Cell* 90 (6): 1023–29.

Fischer, Utz, Clemens Englbrecht, and Ashwin Chari. 2011. “Biogenesis of Spliceosomal Small Nuclear Ribonucleoproteins.” *Wiley Interdisciplinary Reviews. RNA* 2 (5): 718–31.

Flavell, Steven W., Tae-Kyung Kim, Jesse M. Gray, David A. Harmin, Martin Hemberg, Elizabeth J. Hong, Eirene Markenscoff-Papadimitriou, Daniel M. Bear, and Michael E. Greenberg. 2008. “Genome-Wide Analysis of MEF2 Transcriptional Program Reveals Synaptic Target Genes and Neuronal Activity-Dependent Polyadenylation Site Selection.” *Neuron* 60 (6): 1022–38.

Fong, Nova, Kristopher Brannan, Benjamin Erickson, Hyunmin Kim, Michael A. Cortazar, Ryan M. Sheridan, Tram Nguyen, Shai Karp, and David L. Bentley. 2015. “Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition.” *Molecular Cell* 60 (2): 256–67.

Gabanella, Francesca, Matthew E. R. Butchbach, Luciano Saieva, Claudia Carissimi, Arthur H. M. Burghes, and Livio Pellizzoni. 2007. “Ribonucleoprotein Assembly

Defects Correlate with Spinal Muscular Atrophy Severity and Preferentially Affect a Subset of Spliceosomal snRNPs." *PloS One* 2 (9): e921.

Gabel, Harrison W., Benyam Kinde, Hume Stroud, Caitlin S. Gilbert, David A. Harmin, Nathaniel R. Kastan, Martin Hemberg, Daniel H. Ebert, and Michael E. Greenberg. 2015. "Disruption of DNA-Methylation-Dependent Long Gene Repression in Rett Syndrome." *Nature* 522 (7554): 89–93.

Gao, Jingfang, Gunnar Arbman, Lujun He, Fang Qiao, Zhiyong Zhang, Zengren Zhao, Johan Rosell, and Xiao-Feng Sun. 2008. "MANBA Polymorphism Was Related to Increased Risk of Colorectal Cancer in Swedish but Not in Chinese Populations." *Acta Oncologica* 47 (3): 372–78.

Gelfman, Sahar, David Burstein, Osnat Penn, Anna Savchenko, Maayan Amit, Schraga Schwartz, Tal Pupko, and Gil Ast. 2012. "Changes in Exon-Intron Structure during Vertebrate Evolution Affect the Splicing Pattern of Exons." *Genome Research* 22 (1): 35–50.

Gurskaya, N. G., L. Diatchenko, A. Chenchik, P. D. Siebert, G. L. Khaspekov, K. A. Lukyanov, L. L. Vagner, O. D. Ermolaeva, S. A. Lukyanov, and E. D. Sverdlov. 1996. "Equalizing cDNA Subtraction Based on Selective Suppression of Polymerase Chain Reaction: Cloning of Jurkat Cell Transcripts Induced by Phytohemaglutinin and Phorbol 12-Myristate 13-Acetate." *Analytical Biochemistry* 240 (1): 90–97.

Guthrie, C., and B. Patterson. 1988. "Spliceosomal snRNAs." *Annual Review of Genetics* 22: 387–419.

Hall, S. L., and R. A. Padgett. 1994. "Conserved Sequences in a Class of Rare Eukaryotic Nuclear Introns with Non-Consensus Splice Sites." *Journal of Molecular Biology* 239 (3): 357–65.

Heidemann, Martin, Corinna Hintermair, Kirsten Voß, and Dirk Eick. 2013. "Dynamic Phosphorylation Patterns of RNA Polymerase II CTD during Transcription." *Biochimica et Biophysica Acta* 1829 (1): 55–62.

Hodnett, J. L., and H. Busch. 1968. "Isolation and Characterization of Uridylic Acid-Rich 7 S Ribonucleic Acid of Rat Liver Nuclei." *The Journal of Biological Chemistry* 243 (24): 6334–42.

Holstege, F. C., U. Fiedler, and H. T. Timmers. 1997. "Three Transitions in the RNA Polymerase II Transcription Complex during Initiation." *The EMBO Journal* 16 (24): 7468–80.

Hoon, Michiel de, and Yoshihide Hayashizaki. 2008. "Deep Cap Analysis Gene Expression (CAGE): Genome-Wide Identification of Promoters, Quantification of Their Expression, and Network Inference." *BioTechniques* 44 (5): 627–28, 630, 632.

Huber, Wolfgang, Vincent J. Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S. Carvalho, Hector Corrada Bravo, et al. 2015. “Orchestrating High-Throughput Genomic Analysis with Bioconductor.” *Nature Methods* 12 (2): 115–21.

Iñiguez, Luis P., and Georgina Hernández. 2017. “The Evolutionary Relationship between Alternative Splicing and Gene Duplication.” *Frontiers in Genetics* 8 (February): 14.

Jensen, J. P., P. W. Bates, M. Yang, R. D. Vierstra, and A. M. Weissman. 1995. “Identification of a Family of Closely Related Human Ubiquitin Conjugating Enzymes.” *The Journal of Biological Chemistry* 270 (51): 30408–14.

Ji, Ping, Sven Diederichs, Wenbing Wang, Sebastian Böing, Ralf Metzger, Paul M. Schneider, Nicola Tidow, et al. 2003. “MALAT-1, a Novel Noncoding RNA, and Thymosin beta4 Predict Metastasis and Survival in Early-Stage Non-Small Cell Lung Cancer.” *Oncogene* 22 (39): 8031–41.

Jonkers, Iris, Hojoong Kwak, and John T. Lis. 2014. “Genome-Wide Dynamics of Pol II Elongation and Its Interplay with Promoter Proximal Pausing, Chromatin, and Exons.” *eLife* 3 (April): e02407.

Kaida, Daisuke, Michael G. Berg, Ihab Younis, Mumtaz Kasim, Larry N. Singh, Lili Wan, and Gideon Dreyfuss. 2010. “U1 snRNP Protects Pre-mRNAs from Premature Cleavage and Polyadenylation.” *Nature* 468 (7324): 664–68.

Kaida, Daisuke, Hajime Motoyoshi, Etsu Tashiro, Takayuki Nojima, Masatoshi Hagiwara, Ken Ishigami, Hidenori Watanabe, et al. 2007. “Spliceostatin A Targets SF3b and Inhibits Both Splicing and Nuclear Retention of Pre-mRNA.” *Nature Chemical Biology* 3 (9): 576–83.

Katz, Yarden, Eric T. Wang, Edoardo M. Airoldi, and Christopher B. Burge. 2010. “Analysis and Design of RNA Sequencing Experiments for Identifying Isoform Regulation.” *Nature Methods* 7 (12): 1009–15.

Kim, Dae-Soo, Dong-Wook Kim, Min-Young Kim, Seong-Hyeuk Nam, Sang-Haeng Choi, Ryong Nam Kim, Aram Kang, Aeri Kim, and Hong-Seog Park. 2012. “CACG: A Database for Comparative Analysis of Conjoined Genes.” *Genomics* 100 (1): 14–17.

Komarnitsky, P., E. J. Cho, and S. Buratowski. 2000. “Different Phosphorylated Forms of RNA Polymerase II and Associated mRNA Processing Factors during Transcription.” *Genes & Development* 14 (19): 2452–60.

Kotake, Yoshihiko, Koji Sagane, Takashi Owa, Yuko Mimori-Kiyosue, Hajime Shimizu, Mai Uesugi, Yasushi Ishihama, Masao Iwata, and Yoshiharu Mizui. 2007. “Splicing Factor SF3b as a Target of the Antitumor Natural Product Pladienolide.” *Nature Chemical Biology* 3 (9): 570–75.

Krämer, A., F. Mulhauser, C. Wersig, K. Gröning, and G. Bilbe. 1995. “Mammalian Splicing Factor SF3a120 Represents a New Member of the SURP Family of

Proteins and Is Homologous to the Essential Splicing Factor PRP21p of Saccharomyces Cerevisiae." *RNA* 1 (3): 260–72.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.

Lau, Chi-Kong, Jennifer L. Bachorik, and Gideon Dreyfuss. 2009. "Gemin5-snRNA Interaction Reveals an RNA Binding Function for WD Repeat Domains." *Nature Structural & Molecular Biology* 16 (5): 486–91.

Lee, Tong Ihn, Sarah E. Johnstone, and Richard A. Young. 2006. "Chromatin Immunoprecipitation and Microarray-Based Analysis of Protein Location." *Nature Protocols* 1 (2): 729–48.

Lerner, M. R., J. A. Boyle, S. M. Mount, S. L. Wolin, and J. A. Steitz. 1980. "Are snRNPs Involved in Splicing?" *Nature* 283 (5743): 220–24.

Lerner, M. R., and J. A. Steitz. 1979. "Antibodies to Small Nuclear RNAs Complexed with Proteins Are Produced by Patients with Systemic Lupus Erythematosus." *Proceedings of the National Academy of Sciences of the United States of America* 76 (11): 5495–99.

Lianoglou, Steve, Vidur Garg, Julie L. Yang, Christina S. Leslie, and Christine Mayr. 2013. "Ubiquitously Transcribed Genes Use Alternative Polyadenylation to Achieve Tissue-Specific Expression." *Genes & Development* 27 (21): 2380–96.

Liu, Q., U. Fischer, F. Wang, and G. Dreyfuss. 1997. "The Spinal Muscular Atrophy Disease Gene Product, SMN, and Its Associated Protein SIP1 Are in a Complex with Spliceosomal snRNP Proteins." *Cell* 90 (6): 1013–21.

Magana-Mora, Arturo, Manal Kalkatawi, and Vladimir B. Bajic. 2017. "Omni-PolyA: A Method and Tool for Accurate Recognition of Poly(A) Signals in Human Genomic DNA." *BMC Genomics* 18 (1): 620.

Matter, Nathalie, and Harald König. 2005. "Targeted 'Knockdown' of Spliceosome Function in Mammalian Cells." *Nucleic Acids Research* 33 (4): e41.

Mayer, Andreas, Michael Lidschreiber, Matthias Siebert, Kristin Leike, Johannes Söding, and Patrick Cramer. 2010. "Uniform Transitions of the General RNA Polymerase II Transcription Complex." *Nature Structural & Molecular Biology* 17 (10): 1272–78.

Mayr, Christine. 2017. "Regulation by 3'-Untranslated Regions." *Annual Review of Genetics* 51 (November): 171–94.

Mayr, Christine, and David P. Bartel. 2009. "Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells." *Cell* 138 (4): 673–84.

Meister, G., D. Bühler, R. Pillai, F. Lottspeich, and U. Fischer. 2001. “A Multiprotein Complex Mediates the ATP-Dependent Assembly of Spliceosomal U snRNPs.” *Nature Cell Biology* 3 (11): 945–49.

Miller, J. W., C. R. Urbinati, P. Teng-Umnuay, M. G. Stenberg, B. J. Byrne, C. A. Thornton, and M. S. Swanson. 2000. “Recruitment of Human Muscleblind Proteins to (CUG)(n) Expansions Associated with Myotonic Dystrophy.” *The EMBO Journal* 19 (17): 4439–48.

Mount, S. M. 2000. “Genomic Sequence, Splicing, and Gene Annotation.” *American Journal of Human Genetics* 67 (4): 788–92.

Mount, S. M., I. Pettersson, M. Hinterberger, A. Karmas, and J. A. Steitz. 1983. “The U1 Small Nuclear RNA-Protein Complex Selectively Binds a 5’ Splice Site in Vitro.” *Cell* 33 (2): 509–18.

Munding, Elizabeth M., Lily Shiue, Sol Katzman, John Paul Donohue, and Manuel Ares Jr. 2013. “Competition between Pre-mRNAs for the Splicing Machinery Drives Global Regulation of Splicing.” *Molecular Cell* 51 (3): 338–48.

Nasevicius, A., and S. C. Ekker. 2000. “Effective Targeted Gene ‘Knockdown’ in Zebrafish.” *Nature Genetics* 26 (2): 216–20.

Niibori, Yosuke, Fumihiko Hayashi, Keiko Hirai, Minoru Matsui, and Kaoru Inokuchi. 2007. “Alternative poly(A) Site-Selection Regulates the Production of Alternatively Spliced Vesl-1/homer1 Isoforms That Encode Postsynaptic Scaffolding Proteins.” *Neuroscience Research* 57 (3): 399–410.

Odawara, Jun, Akihito Harada, Tomohiko Yoshimi, Kazumitsu Maehara, Taro Tachibana, Seiji Okada, Koichi Akashi, and Yasuyuki Ohkawa. 2011. “The Classification of mRNA Expression Levels by the Phosphorylation State of RNAPII CTD Based on a Combined Genome-Wide Approach.” *BMC Genomics* 12 (October): 516.

Oesterreich, Fernando Carrillo, Lydia Herzel, Korinna Straube, Katja Hujer, Jonathon Howard, and Karla M. Neugebauer. 2016. “Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II.” *Cell* 165 (2): 372–81.

Oh, Jung-Min, Chao Di, Christopher C. Venters, Jiannan Guo, Chie Arai, Byung Ran So, Anna Maria Pinto, et al. 2017. “U1 snRNP Telescripting Regulates a Size-Function-Stratified Human Genome.” *Nature Structural & Molecular Biology* 24 (11): 993–99.

Padgett, R. A., S. M. Mount, J. A. Steitz, and P. A. Sharp. 1983. “Splicing of Messenger RNA Precursors Is Inhibited by Antisera to Small Nuclear Ribonucleoprotein.” *Cell* 35 (1): 101–7.

Pagès, H., P. Aboyoun, R. Gentleman, and S. DebRoy. 2017. “Biostrings: Efficient Manipulation of Biological Strings.” *R Package Version* 2 (0).

Patel, Abhijit A., and Joan A. Steitz. 2003. “Splicing Double: Insights from the Second Spliceosome.” *Nature Reviews. Molecular Cell Biology* 4 (12): 960–70.

Pellizzoni, Livio, Jennifer Baccon, Juri Rappsilber, Matthias Mann, and Gideon Dreyfuss. 2002. “Purification of Native Survival of Motor Neurons Complexes and Identification of Gemin6 as a Novel Component.” *The Journal of Biological Chemistry* 277 (9): 7540–45.

Pomeranz Krummel, Daniel A., Chris Oubridge, Adelaine K. W. Leung, Jade Li, and Kiyoshi Nagai. 2009. “Crystal Structure of Human Spliceosomal U1 snRNP at 5.5 A Resolution.” *Nature* 458 (7237): 475–80.

Prakash, Tulika, Vineet K. Sharma, Naoki Adati, Ritsuko Ozawa, Naveen Kumar, Yuichiro Nishida, Takayoshi Fujikake, Tadayuki Takeda, and Todd D. Taylor. 2010. “Expression of Conjoined Genes: Another Mechanism for Gene Regulation in Eukaryotes.” *PloS One* 5 (10): e13284.

Proudfoot, Nick J. 2016. “Transcriptional Termination in Mammals: Stopping the RNA Polymerase II Juggernaut.” *Science* 352 (6291): aad9926.

Proudfoot, N. J., and G. G. Brownlee. 1976. “3’ Non-Coding Region Sequences in Eukaryotic Messenger RNA.” *Nature* 263 (5574): 211–14.

Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features.” *Bioinformatics* 26 (6): 841–42.

Rogozin, Igor B., Liran Carmel, Miklos Csuros, and Eugene V. Koonin. 2012. “Origin and Evolution of Spliceosomal Introns.” *Biology Direct* 7 (1): 11.

Rösel, Tanja Dorothe, Lee-Hsueh Hung, Jan Medenbach, Katrin Donde, Stefan Starke, Vladimir Benes, Gunnar Rätsch, and Albrecht Bindereif. 2011. “RNA-Seq Analysis in Mutant Zebrafish Reveals Role of U1C Protein in Alternative Splicing Regulation.” *The EMBO Journal* 30 (10): 1965–76.

Ryan, K. M., A. C. Phillips, and K. H. Vousden. 2001. “Regulation and Function of the p53 Tumor Suppressor Protein.” *Current Opinion in Cell Biology* 13 (3): 332–37.

Sandberg, Rickard, Joel R. Neilson, Arup Sarma, Phillip A. Sharp, and Christopher B. Burge. 2008. “Proliferating Cells Express mRNAs with Shortened 3’ Untranslated Regions and Fewer microRNA Target Sites.” *Science* 320 (5883): 1643–47.

Saville, Mark K., Alison Sparks, Dimitris P. Xirodimas, Julie Wardrop, Lauren F. Stevenson, Jean-Christophe Bourdon, Yvonne L. Woods, and David P. Lane. 2004. “Regulation of p53 by the Ubiquitin-Conjugating Enzymes UbcH5B/C in Vivo.” *The Journal of Biological Chemistry* 279 (40): 42169–81.

Scotto-Lavino, Elizabeth, Guangwei Du, and Michael A. Frohman. 2006. “3’ End cDNA Amplification Using Classic RACE.” *Nature Protocols* 1 (6): 2742–45.

Seila, Amy C., Leighton J. Core, John T. Lis, and Phillip A. Sharp. 2009. “Divergent Transcription: A New Feature of Active Promoters.” *Cell Cycle* 8 (16): 2557–64.

Shepard, Peter J., Eun-A Choi, Jente Lu, Lisa A. Flanagan, Klemens J. Hertel, and Yongsheng Shi. 2011. “Complex and Dynamic Landscape of RNA Polyadenylation Revealed by PAS-Seq.” *RNA* 17 (4): 761–72.

Shilatifard, Ali, Ronald C. Conaway, and Joan Weliky Conaway. 2003. “The RNA Polymerase II Elongation Complex.” *Annual Review of Biochemistry* 72 (March): 693–715.

Shi, Yongsheng, and James L. Manley. 2015. “The End of the Message: Multiple Protein-RNA Interactions Define the mRNA Polyadenylation Site.” *Genes & Development* 29 (9): 889–97.

Sims, David, Nicholas E. Ilott, Stephen N. Sansom, Ian M. Sudbery, Jethro S. Johnson, Katherine A. Fawcett, Antonio J. Berlanga-Taylor, Sebastian Luna-Valero, Chris P. Ponting, and Andreas Heger. 2014. “CGAT: Computational Genomics Analysis Toolkit.” *Bioinformatics* 30 (9): 1290–91.

So, Byung Ran, Lili Wan, Zhenxi Zhang, Pilong Li, Eric Babiash, Jingqi Duan, Ihab Younis, and Gideon Dreyfuss. 2016. “A U1 snRNP-Specific Assembly Pathway Reveals the SMN Complex as a Versatile Hub for RNP Exchange.” *Nature Structural & Molecular Biology* 23 (3): 225–30.

Sun, Hao, Jiejun Wu, Priyankara Wickramasinghe, Sharmistha Pal, Ravi Gupta, Anirban Bhattacharyya, Francisco J. Agosto-Perez, Louise C. Showe, Tim H-M Huang, and Ramana V. Davuluri. 2011. “Genome-Wide Mapping of RNA Pol-II Promoter Usage in Mouse Tissues by ChIP-Seq.” *Nucleic Acids Research* 39 (1): 190–201.

Supek, Fran, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. 2011. “REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms.” *PloS One* 6 (7): e21800.

Thomson, T. M., J. J. Lozano, N. Loukili, R. Carrió, F. Serras, B. Cormand, M. Valeri, et al. 2000. “Fusion of the Human Gene for the Polyubiquitination Coeffector UEV1 with Kua, a Newly Identified Gene.” *Genome Research* 10 (11): 1743–56.

Tian, Bin, and James L. Manley. 2017. “Alternative Polyadenylation of mRNA Precursors.” *Nature Reviews. Molecular Cell Biology* 18 (1): 18–30.

Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. 2009. “TopHat: Discovering Splice Junctions with RNA-Seq.” *Bioinformatics* 25 (9): 1105–11.

Trede, Nikolaus S., Jan Medenbach, Andrey Damianov, Lee-Hsueh Hung, Gerhard J. Weber, Barry H. Paw, Yi Zhou, et al. 2007. “Network of Coregulated Spliceosome Components Revealed by Zebrafish Mutant in Recycling Factor p110.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (16): 6608–13.

Vaquero-Garcia, Jorge, Alejandro Barrera, Matthew R. Gazzara, Juan González-Vallinas, Nicholas F. Lahens, John B. Hogenesch, Kristen W. Lynch, and Yoseph Barash. 2016. “A New View of Transcriptome Complexity and Regulation through the Lens of Local Splicing Variations.” *eLife* 5 (February): e11752.

Venters, Christopher C., Jung-Min Oh, Chao Di, Byung Ran So, and Gideon Dreyfuss. 2019. “U1 snRNP Telescripting: Suppression of Premature Transcription Termination in Introns as a New Layer of Gene Regulation.” *Cold Spring Harbor Perspectives in Biology* 11 (2). https://doi.org/10.1101/cshperspect.a032235.

Wada, Youichiro, Yoshihiro Ohta, Meng Xu, Shuichi Tsutsumi, Takashi Minami, Kenji Inoue, Daisuke Komura, et al. 2009. “A Wave of Nascent Transcription on Activated Human Genes.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (43): 18357–61.

Wahl, Markus C., Cindy L. Will, and Reinhard Lührmann. 2009. “The Spliceosome: Design Principles of a Dynamic RNP Machine.” *Cell* 136 (4): 701–18.

Wang, C., K. Chua, W. Seghezzi, E. Lees, O. Gozani, and R. Reed. 1998. “Phosphorylation of Spliceosomal Protein SAP 155 Coupled with Splicing Catalysis.” *Genes & Development* 12 (10): 1409–14.

Wang, Eric T., Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. 2008. “Alternative Isoform Regulation in Human Tissue Transcriptomes.” *Nature* 456 (7221): 470–76.

Wang, Qingqing, and Donald C. Rio. 2018. “JUM Is a Computational Method for Comprehensive Annotation-Free Analysis of Alternative Pre-mRNA Splicing Patterns.” *Proceedings of the National Academy of Sciences of the United States of America* 115 (35): E8181–90.

Wang, Wei, Zhi Wei, and Hongzhe Li. 2014. “A Change-Point Model for Identifying 3’UTR Switching by next-Generation RNA Sequencing.” *Bioinformatics* 30 (15): 2162–70.

Weber, Christopher M., Srinivas Ramachandran, and Steven Henikoff. 2014. “Nucleosomes Are Context-Specific, H2A.Z-Modulated Barriers to RNA Polymerase.” *Molecular Cell* 53 (5): 819–30.

Weinberg, R. A., and S. Penman. 1968. “Small Molecular Weight Monodisperse Nuclear RNA.” *Journal of Molecular Biology* 38 (3): 289–304.

Will, Cindy L., and Reinhard Lührmann. 2011. “Spliceosome Structure and Function.” *Cold Spring Harbor Perspectives in Biology* 3 (7). https://doi.org/10.1101/cshperspect.a003707.

Wongpalee, Somsakul Pop, Ajay Vashisht, Shalini Sharma, Darryl Chui, James A. Wohlschlegel, and Douglas L. Black. 2016. “Large-Scale Remodeling of a

Repressed Exon Ribonucleoprotein to an Exon Definition Complex Active for Splicing." *eLife* 5 (November): e19743.

Workman, Eileen, Luciano Saieva, Tessa L. Carrel, Thomas O. Crawford, Don Liu, Cathleen Lutz, Christine E. Beattie, Livio Pellizzoni, and Arthur H. M. Burghes. 2009. "A SMN Missense Mutation Complements SMN2 Restoring snRNPs and Rescuing SMA Mice." *Human Molecular Genetics* 18 (12): 2215–29.

Wysoker, A., T. Fennell, J. Ruan, N. Homer, and G. Marth. 2009. "The Sequence Alignment/map (SAM) Format and SAMtools." *Bioinformatics* .

Xia, Zheng, Lawrence A. Donehower, Thomas A. Cooper, Joel R. Neilson, David A. Wheeler, Eric J. Wagner, and Wei Li. 2014. "Dynamic Analyses of Alternative Polyadenylation from RNA-Seq Reveal a 3′-UTR Landscape across Seven Tumour Types." *Nature Communications* 5 (November): 5274.

Yang, Yan, Wencheng Li, Mainul Hoque, Liming Hou, Steven Shen, Bin Tian, and Brian D. Dynlacht. 2016. "PAF Complex Plays Novel Subunit-Specific Roles in Alternative Cleavage and Polyadenylation." *PLoS Genetics* 12 (1): e1005794.

Yao, C., J. Biesinger, J. Wan, L. Weng, Y. Xing, X. Xie, and Y. Shi. 2012. "Transcriptome-Wide Analyses of CstF64-RNA Interactions in Global Regulation of mRNA Alternative Polyadenylation." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1211101109.

Yeo, Gene, and Christopher B. Burge. 2004. "Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals." *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 11 (2-3): 377–94.

Yong, Jeongsik, Tracey J. Golembe, Daniel J. Battle, Livio Pellizzoni, and Gideon Dreyfuss. 2004. "snRNAs Contain Specific SMN-Binding Domains That Are Essential for snRNP Assembly." *Molecular and Cellular Biology* 24 (7): 2747–56.

Yong, Jeongsik, Mumtaz Kasim, Jennifer L. Bachorik, Lili Wan, and Gideon Dreyfuss. 2010. "Gemin5 Delivers snRNA Precursors to the SMN Complex for snRNP Biogenesis." *Molecular Cell* 38 (4): 551–62.

Yong, Jeongsik, Lili Wan, and Gideon Dreyfuss. 2004. "Why Do Cells Need an Assembly Machine for RNA-Protein Complexes?" *Trends in Cell Biology* 14 (5): 226–32.

Zhang, Yong, Tao Liu, Clifford A. Meyer, Jérôme Eeckhoute, David S. Johnson, Bradley E. Bernstein, Chad Nusbaum, et al. 2008. "Model-Based Analysis of ChIP-Seq (MACS)." *Genome Biology* 9 (9): R137.

Zhang, Zhenxi, Francesco Lotti, Kimberly Dittmar, Ihab Younis, Lili Wan, Mumtaz Kasim, and Gideon Dreyfuss. 2008. "SMN Deficiency Causes Tissue-Specific Perturbations in the Repertoire of snRNAs and Widespread Defects in Splicing." *Cell* 133 (4): 585–600.