The Dissertation Committee for Nathan David Elrod Certifies that this is the approved version of the following dissertation:

THE ROLE OF INTEGRATOR SUBUNIT 11 IN PROMOTOR-PROXIMAL ATTENUATION OF PROTEIN CODING GENES

Committee:

Eric J. Wagner, Ph.D., Mentor

Mariano A. Garcia-Blanco, M.D., Ph.D., Chair

Andrew Routh, Ph.D.

Marc C, Morais, Ph.D.

Karen Adelman, Ph.D.

THE ROLE OF INTEGRATOR SUBUNIT 11 IN PROMOTOR-PROXIMAL ATTENUATION OF PROTEIN CODING GENES

by

Nathan David Elrod, B.S., M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas Medical Branch

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas Medical Branch

August 2021

Dedication

This work is dedicated to my family, friends, mentors, and pets that have endeavored to keep me sane throughout this process.

Acknowledgements

I would Like to acknowledge the contributions of my many collaborators both in the works represented here and the other work that I have performed. Without such excellent colleges, science would cease. I would also like to thank Sandra Perdue and Ruth Rea for keeping me on track near the end of this writing process and being such excellent cheerleaders throughout.

THE ROLE OF INTEGRATOR SUBUNIT 11 IN PROMOTOR-PROXIMAL ATTENUATION OF PROTEIN CODING GENES

Publication No._____

Nathan David Elrod. Ph.D.

The University of Texas Medical Branch, 2021

Supervisor: Eric J. Wagner

The Integrator complex is a 14-subunit protein complex that interacts with RNA polymerase II (RNAPII) in metazoans during transcription of multiple forms of coding and non-coding RNA. Originally described as being responsible for the 3' cleavage and termination of uridine-rich small nuclear RNA, further roles have been discovered in the cleavage of long non-coding RNA, enhancer RNA, telomerase RNA, and coding messenger RNAs. While many studies have identified a presence at the transcriptional start site of mRNA, Integrator's function at these mRNA has yet to be identified.

In this study, we used multiple biochemical and sequencing techniques to elucidate a role for Integrator at these mRNA start sites. We found that the endonuclease activity of Integrator subunit 11 (IntS11) was responsible for the attenuation of mRNA through promoter-proximal termination of a large set of paused mRNAs. We also identified a novel interaction between IntS11 and CG7044 that could inhibit IntS11's endonuclease activity.

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

UTMB	University of Texas Medical Branch
GSBS	Graduate School of Biomedical Science
TDC	Thesis and Dissertation Coordinator
RNAPII	RNA Polymerase II
IntS11	Integrator Subunit 11
mRNA	Messenger RNA
CTD	Carboxyl Terminal Domain of RNAPII
snRNA	Uridine Rich small nuclear RNA
eRNA	Enhancer RNA
lncRNA	Long non-coding RNA
TSS	Transcription Start Site
PIC	Pre-Initiation Complex
SNP	Single Nucleotide Polymorphisms
eQTL	Expression Quantitative Trait Loci
Ser5P	Serine 5 Phosphorylation
Ser7P	Serine 7 Phosphorylation
Ser2P	Serine 2 Phosphorylation
CPSF	Cleavage and Poly-adenylation Specificity Factor
Hsp	Heat Shock Protein
scRNA	Short Capped RNA
LTR	Long Terminal Repeat

NELF	Negative Elongation Factor
DSIF	DRB Sensitivity Inducing Factor
P-TEFb	Positive Transcription Elongation Factor b
PP2A	Protein Phosphatase 2A
CPSF73	Cleavage and Polyadenylation Specificity Factor of 73 kDa
CPSF100	Cleavage and Polyadenylation Specificity Factor of 100 kDa
IntS9	Integrator Subunit 9
IntS1	Integrator Subunit 1
IntS12	Integrator Subunit 12
IntS4	Integrator Subunit 4
CM	Integrator Cleavage Module
IntS5	Integrator Subunit 5
IntS8	Integrator Subunit 8
IntS3	Integrator Subunit 3
IntS6	Integrator Subunit 6
IntS2	Integrator Subunit 2
IntS7	Integrator Subunit 7
IntS10	Integrator Subunit 10
IntS14	Integrator Subunit 14
SOSS	Sensor of Single Stranded DNA
hSSB	Human Single Stranded DNA Binding Protein
ssDNA	Single Stranded DNA
TR	Telomerase TNA

DRSC	*Drosophila* RNAi Screening Center
RNAi	RNA interference
dsRNA	Double Stranded RNA
$CuSO_4$	Copper Sulphate
RT-qPCR	Reverse Transcriptase PCR
ChIP-qPCR	Chromatin Immuno-Precipitated PCR
MtnA	Transcription of Drosophila Metallothionein A
UTR	Untranslated Region
WT	Wild-type
NSAF	Normalized Spectral Abundance Factor
TF	Transcription Factors
Trp	Triptolide
NNS	Nrd1-Nab3-Sen1
EGF	Epidermal Growth Factor

Chapter 1. Introduction

INTRODUCTION AND BACKGROUND

RNA Polymerase-II transcription cycle

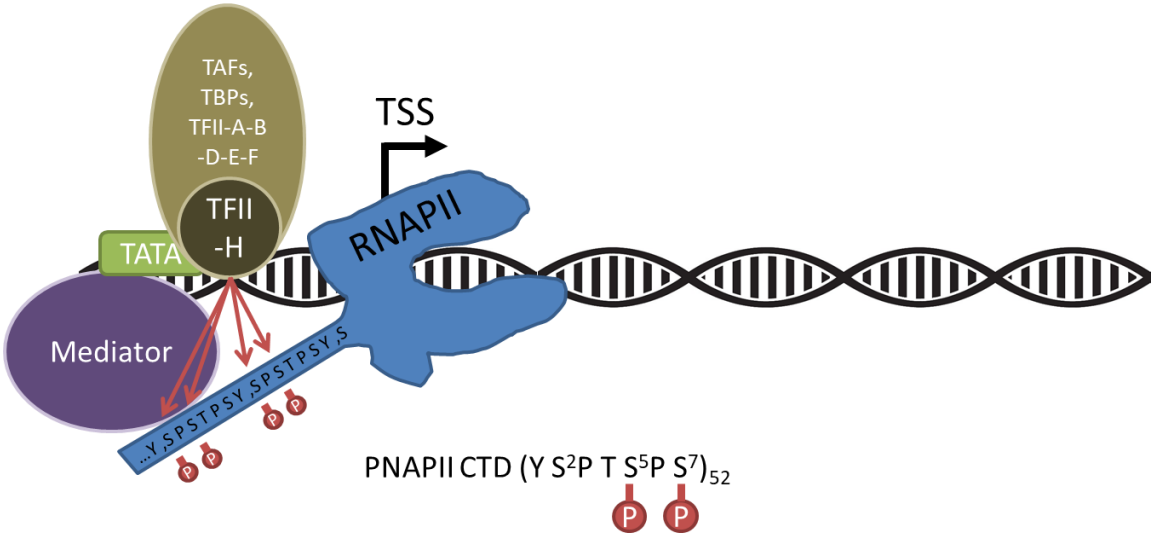
One of key defining features of cellular life is the involved process of converting the genetic information stored as DNA into the various RNA and protein products that sustain function at cellular and organismal levels. A central player in this process is the DNA-dependent RNA Polymerase-II (RNAPII), which was first described as a separate enzyme in eukaryotes in 1969 (Roeder and Rutter 1969). The core enzyme of RNAPII is composed of 12 subunits in humans and yeast, RPB1-12, of which RPB1 and RPB2 form the catalytic core (Roeder and Rutter 1969). The largest of these, RPB1, has an extended carboxyl terminal domain (CTD) that consists of 52 heptad repeats of the consensus sequence YSPTSPS in humans and 26 repeats in yeast and is highly conserved between species (Chapman et al. 2008). The key function of RNAPII is to catalyze the formation of a poly-nucleic acid RNA sequence complementary to a DNA sequence. The main RNAPII core also is assisted by a host of components that function as helicases, RNA binding proteins, and other accessory functions in order to form the total holoenzyme.

The main outputs of RNAPII include protein coding messenger RNA (mRNA), uridine rich small RNA (snRNA), enhancer RNA (eRNA), and long non-coding RNA (lncRNA). Of main importance to this discussion is its primary role at protein coding genes in transcribing messenger RNA (mRNA). During the production of mRNA, RNAPII goes through three main stages; initiation, elongation, and termination, each of which provide vital points in the regulation and control of final output. Since that initial discovery, a number of other factors have been identified that can influence RNAPII's transcription of DNA into RNA at those stages.

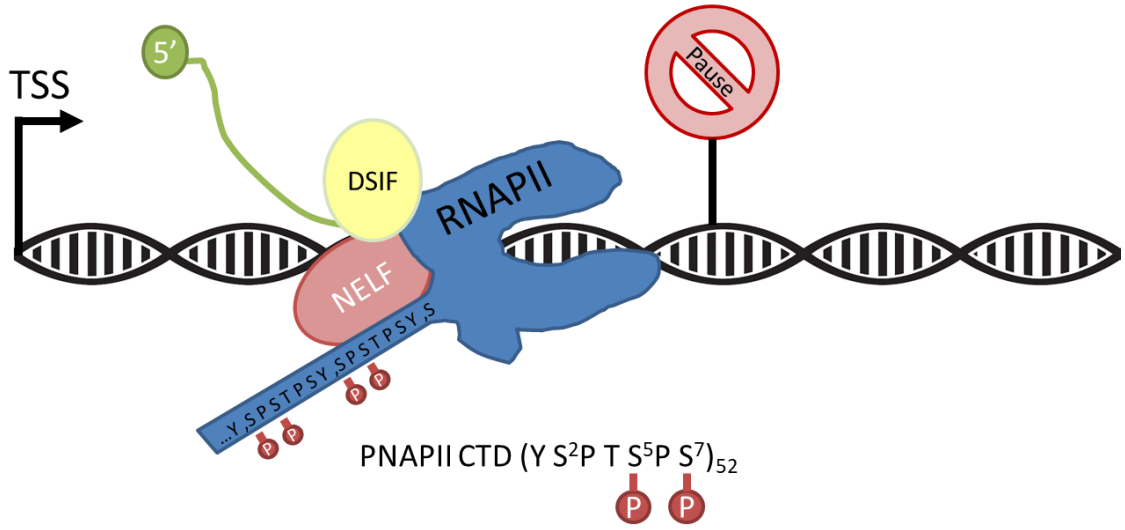
The initiation phase of RNAPII involves the recruitment of RNAPII to the transcription start site (TSS) of DNA and the initial formation of the RNA strand of around 20 nucleotides before entering elongation (Conaway and Conaway 1993). The initiation phase provides some of the greatest amount of overall control on mRNA production through the ability to increase or decrease the amount of RNAPII recruited to the TSS and the number of modifications to the DNA structure and RNAPII itself. The control of mRNA levels is critical to the maintenance of protein levels in the cell.

The beginning of the initiation phase involves a number of general transcription factors (TFII-A, -B, -D, -E, -F, -H), TATA-binding proteins (TBPs), and gene specific transcription associated factors (TAFs) are recruited to the promoters of mRNA to form the Pre-Initiation Complex (PIC) (Conaway and Conaway 1993; Zawel and Reinberg 1993; Zawel and Reinberg 1995; Burley and Roeder 1996). These factors are responsible for unwinding the DNA, recruiting RNAPII, and initiation of transcription through post-translational modification of RNAPII's CTD. Initially, the transcriptional initiation of RNAPII was thought to be regulated mainly through modulation of DNA accessibility (Kadonaga et al. 1988; Nacheva et al. 1989), however, studies have since shown a wide number of factors that influence RNAPII initiation and recruitment. These factors can act in either *cis* or *trans* and include such modifiers as the Mediator complex (Kelleher et al. 1990; Flanagan et al. 1991), activated transcription factors (MAPK pathways, NRFII, HSPs, etc.), and chromatin histone modification (Nacheva et al. 1989; Green et al. 1995; Zhang and Liu 2002; Nguyen et al. 2009) (Figure 1.1a).

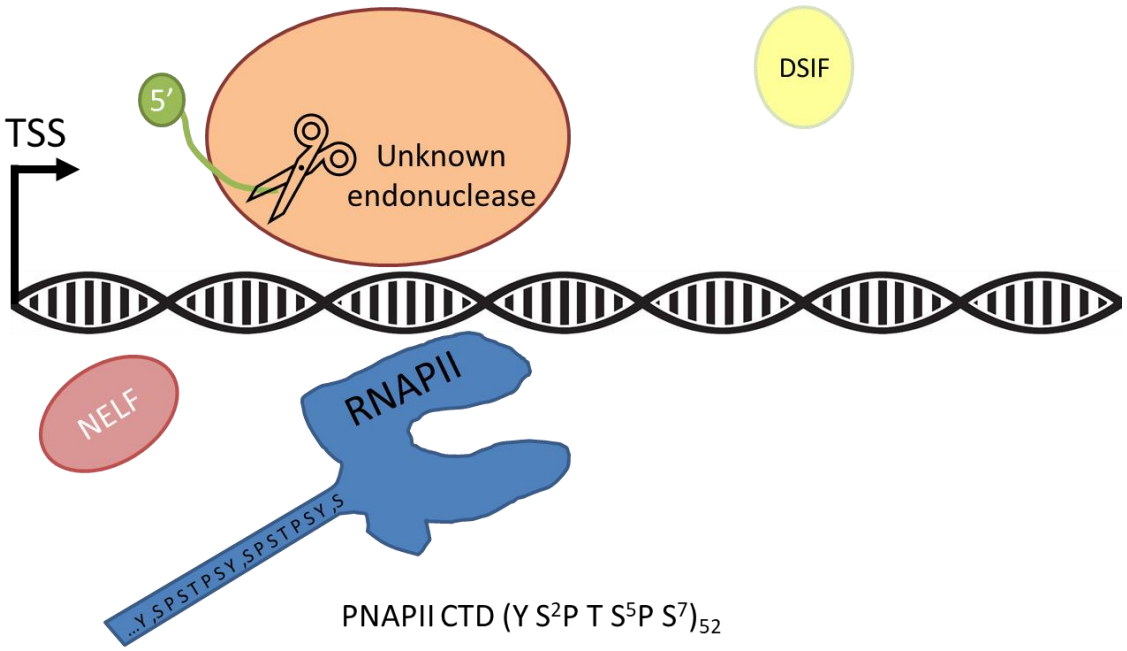
a


TAFs,
TBPs,
TFII-A-B
-D-E-F
TFII
-H
TATA
Mediator
TSS
RNAPII
...Y,S P S T P S Y ,S P S T P S Y ,S
PNAPII CTD (Y S2P T S5P S7)52

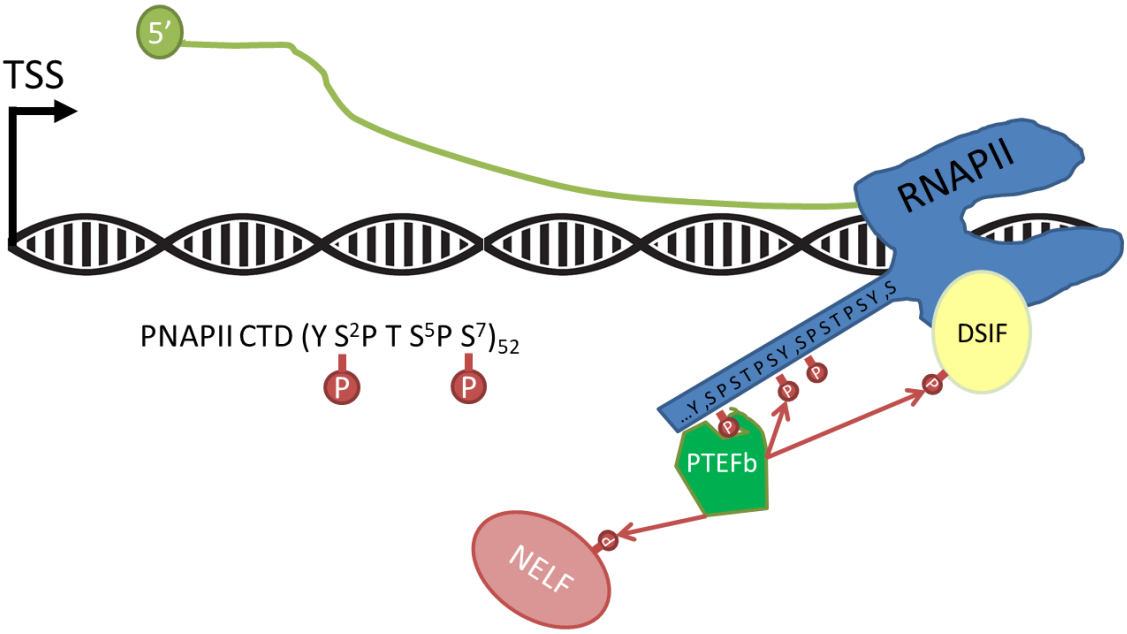
b


5′
TSS
DSIF
NELF
RNAPII
Pause
...Y,S P S T P S Y ,S P S T P S Y ,S
PNAPII CTD (Y S2P T S5P S7)52

c


5′
TSS
Unknown
endonuclease
DSIF
NELF
RNAPII
…Y,S P S T P S Y, S P S T P S Y, S
PNAPII CTD (Y S2P T S5P S7)52

d


5′
TSS
RNAPII
PNAPII CTD (Y S2P T S5P S7)52
P
P
DSIF
…Y, S P S T P S Y, S P S T P S Y, S
PTEFb
NELF

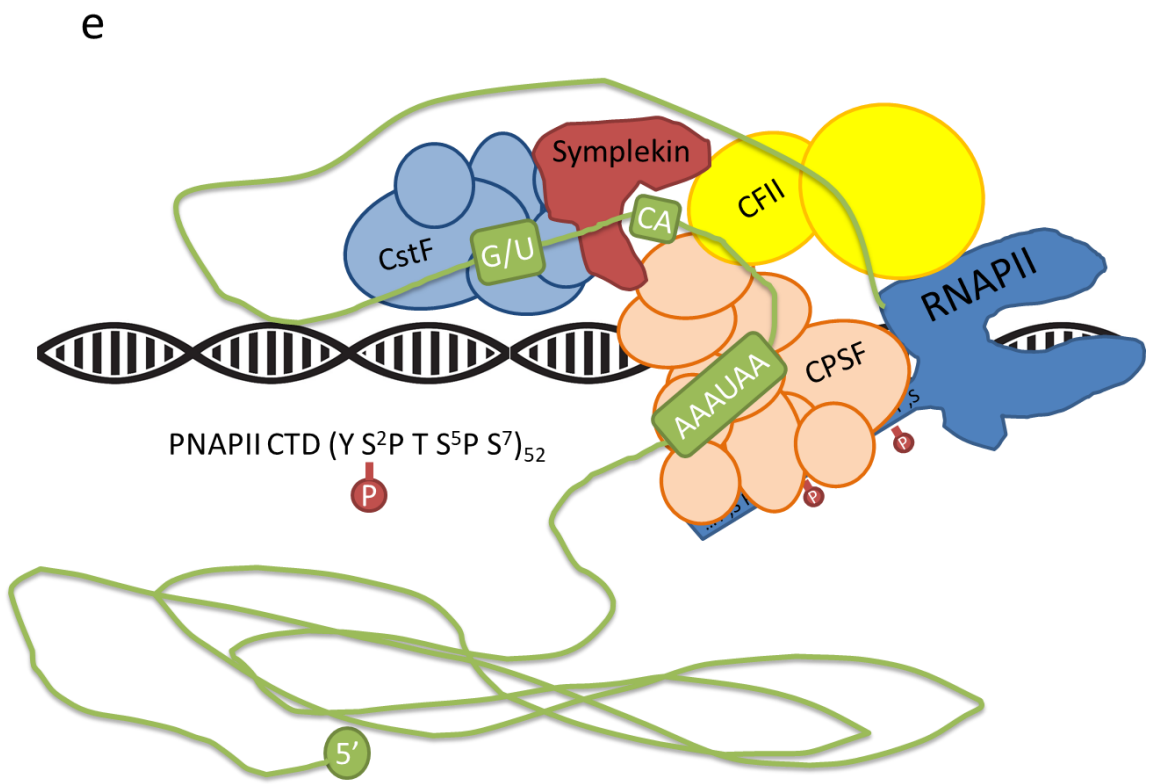


Figure 1.1: Schematic of the stages of RNA Polymerase II mRNA transcription of an intronless gene

(a) RNAPII Initiation Complex
(b) Paused RNAPII complex
(c) Promoter Proximal termination of RNAPII
(d) Actively elongating RNAPII
(e) Fully elongated mRNA with Cleavage and Poly-adenylation complex

Recently, single nucleotide polymorphisms (SNPs) have been identified as expression quantitative trait loci (eQTLs) that affect expression levels (Consortium et al. 2017). Furthermore, many of these regulatory features are subject to their own regulation via post-translational modification, cellular state changes, or changes in cofactor association (Darnell et al. 1994; Ho et al. 1996; Kline and Morimoto 1997). In all, there is an astounding amount of complexity in just transcriptional initiation that increases in intricacy as one travels down the evolutionary timeline.

Immediately following initiation, RNAPII enters a phase of active elongation with a number of genes entering a pausing state (Figure 1.1b) before either undergoing promoter proximal termination (Figure 1.1c) or continuing elongation (Figure 1.1d). This pausing step and the regulatory factors around it will be discussed in depth in the next section, but for the genes that enter productive elongation a number of changes to both the mRNA and RNAPII occur. Once RNAPII enters active elongation, there is a shift to rapid processivity of adding nucleotides to the growing RNA chain.

One of the first processes that occurs following elongation is the addition of a 7-methylguanososine cap by a by a series of enzymes that are recruited via a phosphorylated serine 5 (Ser5P) of RNAPII's CTD repeat (Furuichi 2015; Shuman 2015). This phorporlyation of Ser5 (and Ser7) is performed by TFIIH which had been activated by the mediator complex bringing in unphosphorylated RNAPII during initiation and additionally leads to dissociation from mediator and the PIC (Sogaard and Svejstrup 2007). Additionally, mRNA undergoes splicing to remove non-coding introns co-transcriptionally. These splicing events are driven by snRNA binding with additional modification by RBPs to influence isoform selection.

During elongation, RNAPII's CTD contiues to undergo further modification with a rapid decrease in Ser5P signal followed by a gradual decrease in Ser7P sites and an increase in Ser2P modification. This change is coordinated by a number of enzymes (transferases, kinases, phosphotases, etc.). This change in CTD signal coincides with

changes in recruited cofactors associated with RNAPII leading to changes in processivity and ultimately to termination.

The next highly coordinated step in mRNA production is the process of termination and poly-adenylation through the concerted efforts of a number of co-factors (Manley et al. 2021) (Figures 1.1e and 1.2; Table 1.1). The proper termination event is driven both by site consensus depending on the RNA type being produced and by various RBP co-factors. In the case of mRNA, this is determined in the main by the cleavage and poly-adenylation specificity factor (CPSF) family of proteins and their associated factors.

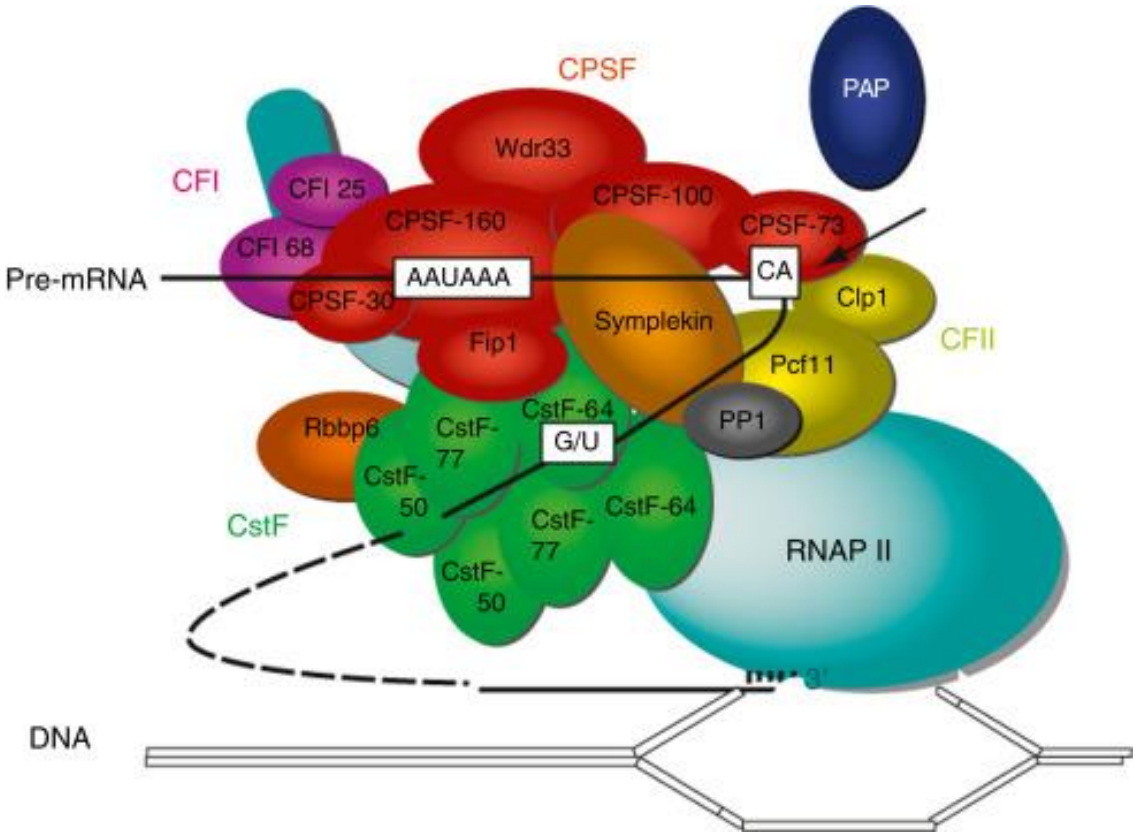


Figure 1.2: A schematic model of the mammalian pre-mRNA 3'-end processing machinery.

(Figure reprinted with permission from (Manley et al. 2021))

Table 1.1 Characteristics of the protein factors of the mammalian pre-mRNA 3′-end processing complex

Protein factor (processing step)	***Subunits***	***Yeast homologue (sub-complex)***	***Sequence characteristics***	***Protein function***	***Interacting proteins***
CPSF (cleavage and polyadenylation)	CPSF-160	Cft 1p/Yhh1p (CPF)	Three possible β-propellers	Binds the AAUAAA sequence	CstF-77, Pol II CTD, PAP, Fip1, TFIID
	CPSF-100	Cft2p/Ydh1p (CPF)	Non-metal binding β-lactamase domain		CPSF-73 CstF-64 symplekin
	CPSF-73	Brr5p/Ysh1p (CPF)	Metallo β-lactamase domain	Endo-nuclease	CPSF-100, CstF64, symplekin
	CPSF-30	Yth1p (CPF)	Five zinc fingers and one zinc knuckle	Binds U-rich RNA sequences	Fip1
	Fip 1	Fip1p (CPF)	Pro-rich sequence. RD-rich sequence. Arg-rich sequence		PAP, CPSF-160, CPSF-70, CstF-77
	WDR33	Pfs2 (PFI)	WD repeats		
CstF (cleavage)	CstF-77	Rna14p (CFIA)	HAT domain, Proline rich sequence	Scaffolding protein, links CstF and CPSF	CPSF-160, CstF-64, CstF-50 Fip1
	CstF-64	Rna15p (CFIA)	RRM, pro/gly-rich sequence. MEARA/G pentapeptide motif	Binds to G/U rich sequences	CstF-77, Symplekin
	CstF-50		Seven WD40 repeats	Regulatory role during DNA damage	CstF-77, Pol II CTD, BARD1
CFIm (cleavage)	CFI-25		NUDIX domain, PAP interaction domain	Helps binding AAUAAA	PAP, CFI-68, PABPII
	CFI-68		RBD, SR protein homology in C-terminus	Helps binding AAUAAA	CFI-25

CFIIm (cleavage)	hClp1	Clp1p (CFIA)	Walker A and B motifs for ATP binding	Tethers CPSF with CF Im	CFlm, CPSF-100, CPSF-73, CstF-64, symplekin
	Pcf11	Pcf11p (CFIA)	polll CTD interacting motif, two zinc fingers		Pol II CTD
Symplekin (cleavage)		Pta1p (CPF)	HEAT fold	Mediates interaction CPSF and CstF	CstF, CPSF. Ssu72, pol II CTD
PAP (cleavage and polyadenylation)		Pap1p	Catalytic core at N-terminus, C-terminus contains RBS, bipartite NLS, ST-rich region	Catalyzes the addition of the poly(A) tail to cleaved mRNA, non specific activity by itself	CPSF-160, Fip1, CFlm
PABPII (polyadenylation)		Pab1p	Two RRM domains	Responsible for processive elongation and control of poly(A) tail length, stabilizes the tail by binding	CPSF-30
Pol II CTD (cleavage)		Pol II CTD	YSPTSPS repeats (52 in humans)	Essential for co-transcriptional recruitment of CPSF and CstF and for cleavage	CPSF-160, CstF-77, CstF-50, Pcf11
PP1 (polyadenylation)		Glc7p		Type 1 protein phosphatase	
RBBP6		Mpe1p (CPF)	RS domain, RING finger, zinc knuckle, DWNN, pro-rich		p53, Rb

(Table reprinted with permission from (Manley et al. 2021))

Specifically, the endonuclease activity of the CPSF73 and CPSF100 heterodimer cleaves the mRNA at the consensus site before a poly-a tail is added. CPSF73 and CPSF100 are members of the metallo-β-lactamase and β-CASP domain containing family of nucleases. These endonucleases have evolved to coordinate a trapped zinc metal ion in a catalytic core of two closely related proteins acting as a heterodimer that both contain metallo-β-lactamase and β-CASP domains. In the case of the CPSF73/100 heterodimer, this trapped zinc ion catalyzes the reaction to cleave the phospho-diester backbone of RNA along the strand. One of the proteins in the pair has an intact metallo-β-lactamase domain while the other is mutated (CPSF73 and CPSF100 respectively in this case) (Mandel et al. 2006) (Figures 1.3)

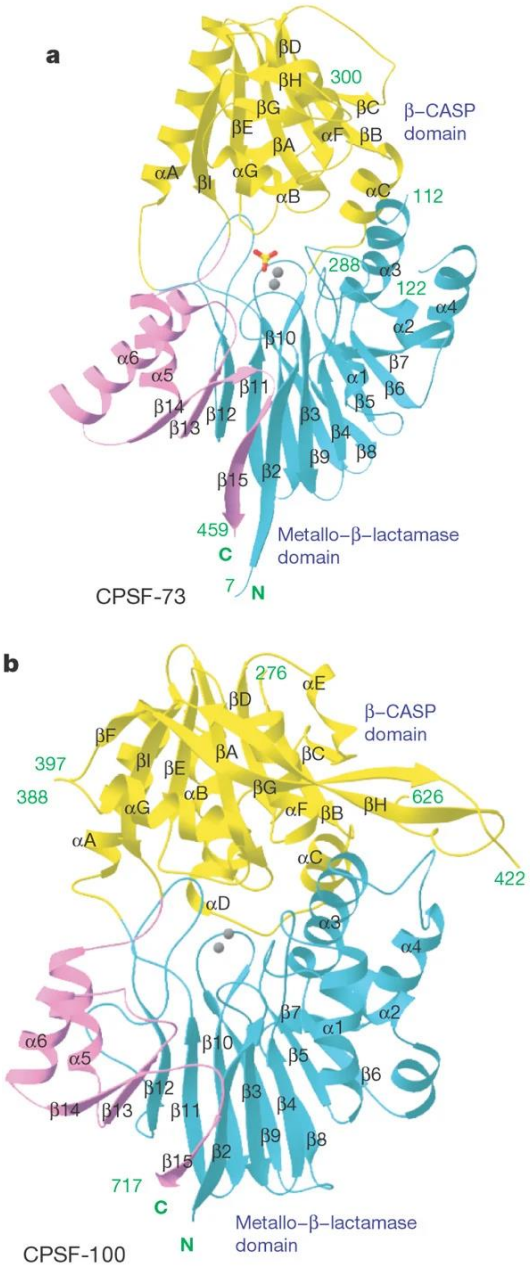


Figure 1.3: Structures of human CPSF-73 and yeast CPSF-100 (Ydh1).

(a) Schematic representation of the structure of human CPSF-73. The b-strands and a-helices are labelled, and the two zinc atoms in the active site are shown as grey spheres. The sulphate ion is shown as a stick model.
(b) Schematic representation of the structure of yeast CPSF-100. The zinc atoms in the CPSF-73 structure are shown for reference.
(Figure and legend text reprinted with permission from (Mandel et al. 2006))

Following the RNA cleavage event, the free 3-prime end undergoes poly-adenylation by other members of the CPSF complex. The free 5-prime end, no longer protected by a methyl cap, undergoes attack by the XRN exonuclease. XRN continues to progressively cleave nucleotides off the 5-prime end until it reaches RNAPII and pulls the RNA strand from RNAPII's active site causing it to cease transcription and fall off the DNA coding strand in the so called "torpedo model" of termination (Eaton et al. 2018; Eaton et al. 2020). This is followed by a complete de-phosphorylation of the CTD which resets RNAPII to be readied for a new round of initiation.

RNA Polymerase-II pausing

In metazoan species at a large number of coding genes, mRNA transcription undergoes an additional regulatory step in the form of promoter proximal pausing. The discovery of this event followed work in the 1970s and early 1980s wherein it was discovered that not all transcription initiation events led to full length mRNA (Fraser et al. 1978; Gariglio et al. 1981). A number of experiments by the Lis lab on *Drosophila melanogaster* heat shock protein (Hsp) led to the determination that a 20-60 nucleotide nascent RNA (scRNA) was produced downstream form Hsp promoters (Rougvie and Lis 1988; Rasmussen and Lis 1993). Similar in nature to the paused RNA polymerase found in the lambda late gene of *Escherichia coli* the term was given to RNAPII in these promoter-proximal Hsp genes as "paused" (Grayhack et al. 1985; Rougvie and Lis 1990). Further work into the 1990s revealed similar sites in a number of human genes and in the HIV long terminal repeat (LTR) (Kao et al. 1987; Krumm et al. 1992; Strobl and Eick 1992; Plet et al. 1995). Importantly, the HIV LTR produces a 59-nucleotide-long RNA as a product of premature promoter-proximal termination and while there had not been any proven levels of promoter-proximal termination in mRNA genes it was suggested to occur (Kao et al. 1987). Later work confirmed the presence of promoter-proximal termination but the exact enzymatic system for this termination remained unclear (Brannan et al. 2012).

The number of coding genes that display paused RNAPII that is capable of then proceeding into active elongation is ~30% (Core and Lis 2008; Larschan et al. 2011; Min et al. 2011) (Figure 1.4).

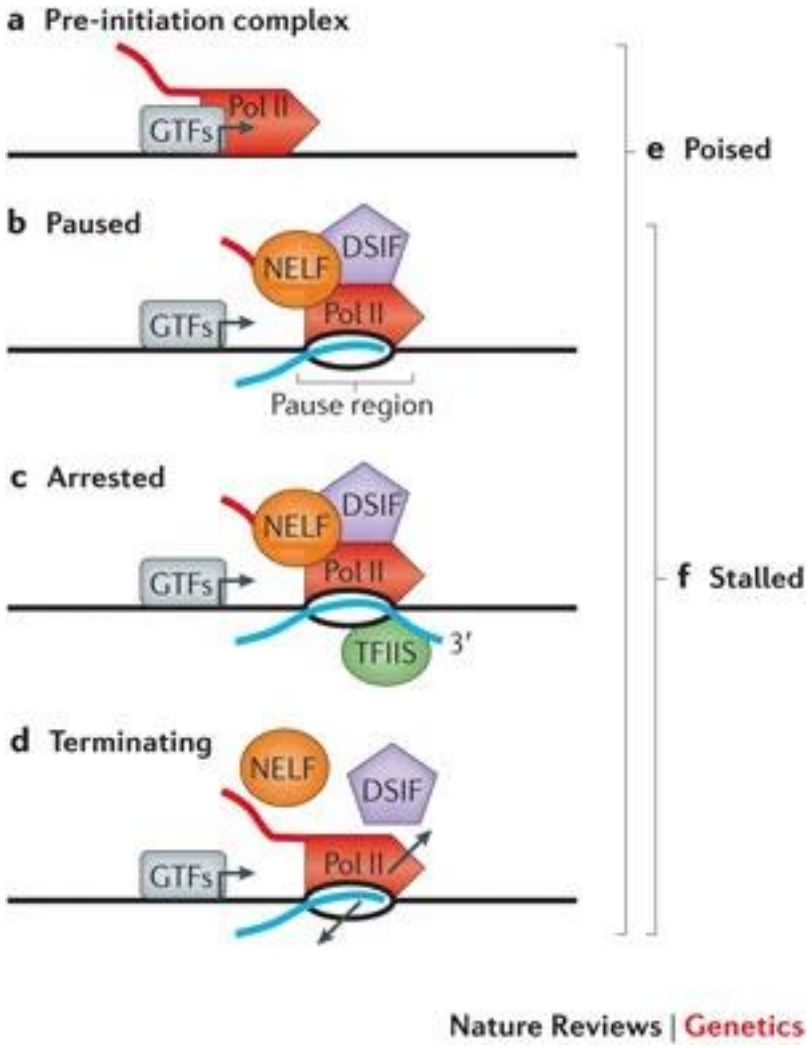
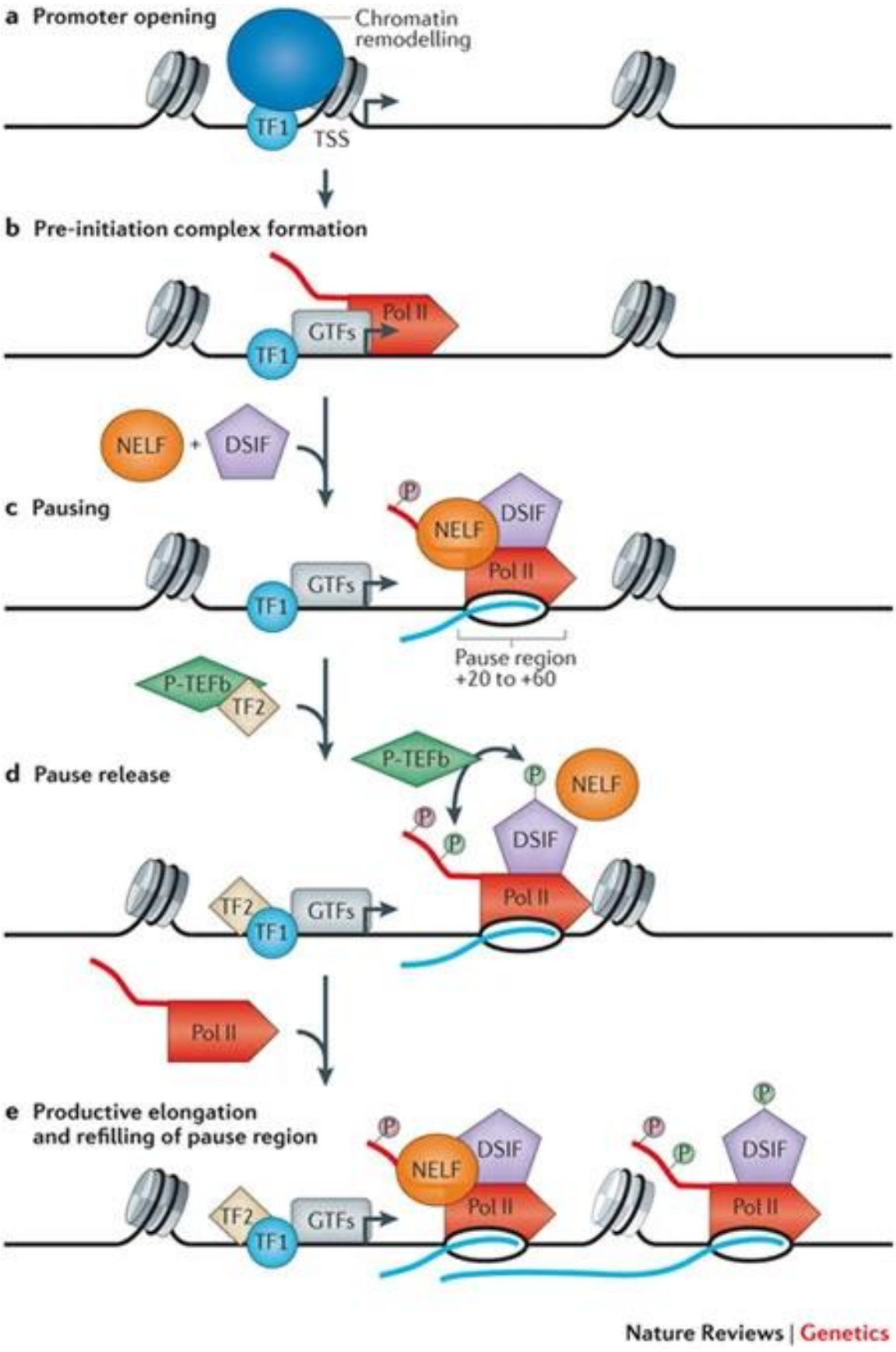


Figure 1.4: Defining the terms used to describe promoter-associated Pol II complexes.

The promoter region is depicted with the transcription start site (TSS) labelled with an arrow. RNA polymerase II (Pol II) is illustrated as a red rocket. The general transcription factors (GTFs; grey oval) are shown centered at the TSS (arrow). The pause-inducing factors negative elongation factor (NELF; orange oval), DRB-sensitivity-inducing factor (DSIF; purple pentagon) and transcript cleavage factor TFIIS (green circle) are shown. The nascent RNA transcript is shown in blue, and a bracket indicates the pausing region, usually 20–60 nucleotides downstream from the TSS.
(a) Pre-initiation complex: an entry form of Pol II in a complex with general transcription factors in which the polymerase is bound to the promoter DNA but has not yet initiated RNA synthesis.
(b) Paused: an early elongation complex that has transiently halted RNA synthesis. Paused polymerase is fully competent to resume elongation, remaining stably engaged and associated with the nascent RNA. The 3′ end of the RNA may have 'frayed' slightly from the Pol II active site in a manner that would slow further RNA synthesis, but the RNA is properly aligned with the active site. Two protein complexes, DSIF and NELF, reduce the rate of elongation and facilitate the establishment of the stably paused state.

(c) Arrested: a stably engaged elongation complex wherein the polymerase has backtracked along the DNA template, such that the RNA 3′ end is displaced from the active site. Restart of an arrested complex usually requires TFIIS, which induces Pol II to cleave the nascent RNA at the active site, creating a new 3′ end that is properly aligned with the Pol II active site and releasing a short (2–9-nucleotide) 3′ RNA.
(d) Terminating: an unstable elongation complex that is in the process of dissociating from the DNA template and releasing the nascent RNA. The released Pol II could have the potential rapidly to reinitiate transcription and to 'recycle' at the promoter.
(e) Poised: a generic term that simply indicates that Pol II is located near the TSS but does not specify anything about its transcriptional status. It can include any of the above complexes (a–d).
(f) Stalled: a term that indicates Pol II is engaged in transcription but that makes no assumptions about its ability to resume synthesis. This term includes paused, arrested and terminating complexes (b–d, above).
(Figure and legend text reprinted with permission from (Adelman and Lis 2012))

The two main factors resulting in the pause of RNAPII are the DRB-sensitivity-inducting-factor (DSIF; made of SPT4 and SPT5)(Wada et al. 1998) and the negative-elongation-factor (NELF) (Yamaguchi et al. 1999). NELF is associated with a Ser5P marker on the CTD while DSIF is recruited in complex with NELF (Adelman and Lis 2012). Release into active elongation is accomplished through the kinase positive-elongation-factor-b (P-TEFb) recruitment to activated genes through a variety of recognition co-factors (Adelman and Lis 2012). P-TEFb phosphorylates the CTD to disassociate NELF and phosphorylates DSIF changing its state to an active promoter of elongation (Wada et al. 1998) (Figure 1.5).



Nature Reviews | Genetics

Figure 1.5: Establishment and release of paused Pol II.

The promoter region is shown with the transcription start site (TSS) labelled with an arrow. Nucleosomes are depicted in grey, and RNA polymerase II (Pol II) is illustrated as a red rocket. The nascent RNA transcript is shown in blue. Factors that are involved in the establishment or release of paused Pol II, such as DRB sensitivity-inducing factor (DSIF; purple pentagon), negative elongation factor (NELF; orange oval) and positive transcription elongation factor b (P-TEFb; green diamond) are indicated.
(a) Promoter opening often involves binding a sequence-specific transcription factor (shown here as TF1, light blue circle) that brings in chromatin remodelers (blue oval) to remove nucleosomes from around the TSS and to render the promoter accessible for recruitment of the transcription machinery.
(b) Pre-initiation complex formation involves the recruitment of a set of general transcription factors (GTFs; grey oval) and Pol II, which is also facilitated by binding specific transcription factors (also shown as TF1 for simplicity). This step precedes the initiation of RNA synthesis.
(c) Pol II pausing occurs shortly after transcription initiation and involves the association of pausing factors DSIF and NELF. The paused Pol II is phosphorylated on its carboxy-terminal heptapeptide repeat domain (CTD; shown in pink). The region in which pausing takes place is indicated on the figure.
(d) Pause release is triggered by the recruitment of the P-TEFb kinase (green diamond), either directly or indirectly by a transcription factor (shown here as TF2; beige diamond). P-TEFb kinase phosphorylates the DSIF–NELF complex to release paused Pol II and also targets the CTD (shown in green). Phosphorylation of DSIF–NELF dissociates NELF from the elongation complex and transforms DSIF into a positive elongation factor that associates with Pol II throughout the gene.
(e) In the presence of both TF1 and TF2, escape of the paused Pol II into productive elongation is rapidly followed by entry of another Pol II into the pause site, allowing for efficient RNA production. When the gene is activated, some nucleosome disruption is likely, as depicted by the lighter colouring of the downstream nucleosome.
(Figure and legend text reprinted with permission from (Adelman and Lis 2012))

The Integrator complex

The Integrator complex is a metazoan protein complex composed of 14 confirmed subunits with a number of possible additional subunits in the process of identification. A number of biochemical assays have shown an association between Integrator with both the C-terminal tail (CTD) of the large subunit of RNA polymerase II (RPBI) and protein phosphatase 2A (PP2A) (Baillat et al. 2005; Egloff et al. 2010; Chen et al. 2012; Huang et al. 2020). Integrator included subunits 1-12 in its initial discovery in 2005 (Baillat et al. 2005). Genome-wide screens later identified additional subunits 13 and 14 (Chen et al. 2012). Key to Integrator's function are subunits 9 and 11 which form an active endonuclease similar to CPSF73/100.

Integrator complex structure. The complete physical structure of the complete Integrator complex remains unknown. There has been work in the field involving sub-sets heterodimers or heterotrimers of subunits that were shown to interact through various biochemical methods. These have been used to demonstrate a number of interaction surfaces and binding partners. However, the structure of a complete complex containing all the subunits and how they interact with each other remains elusive.

The most complete structure of the complex was published in 2020 and constituted nine partial subunits and PP2A (Zheng et al. 2020). While this gave new insights into the structure of the Integrator complex, a majority of information is incomplete to the point where previously identified interactions where lost and resolution was not able to capture key previously identified interacting components (Figure 1.6). Specifically, it is bereft of some of Integrator subunits 3, 10, 12, 13, and 14 and large portions of other subunits. The largest subunit, IntS1, was shown to interact with IntS2 and IntS7 to form a scaffolding/backbone module for the complex. With the exception of IntS11 and the subunits that were not modeled, every subunit interacts with a backbone component (Zhang et al. 2020).

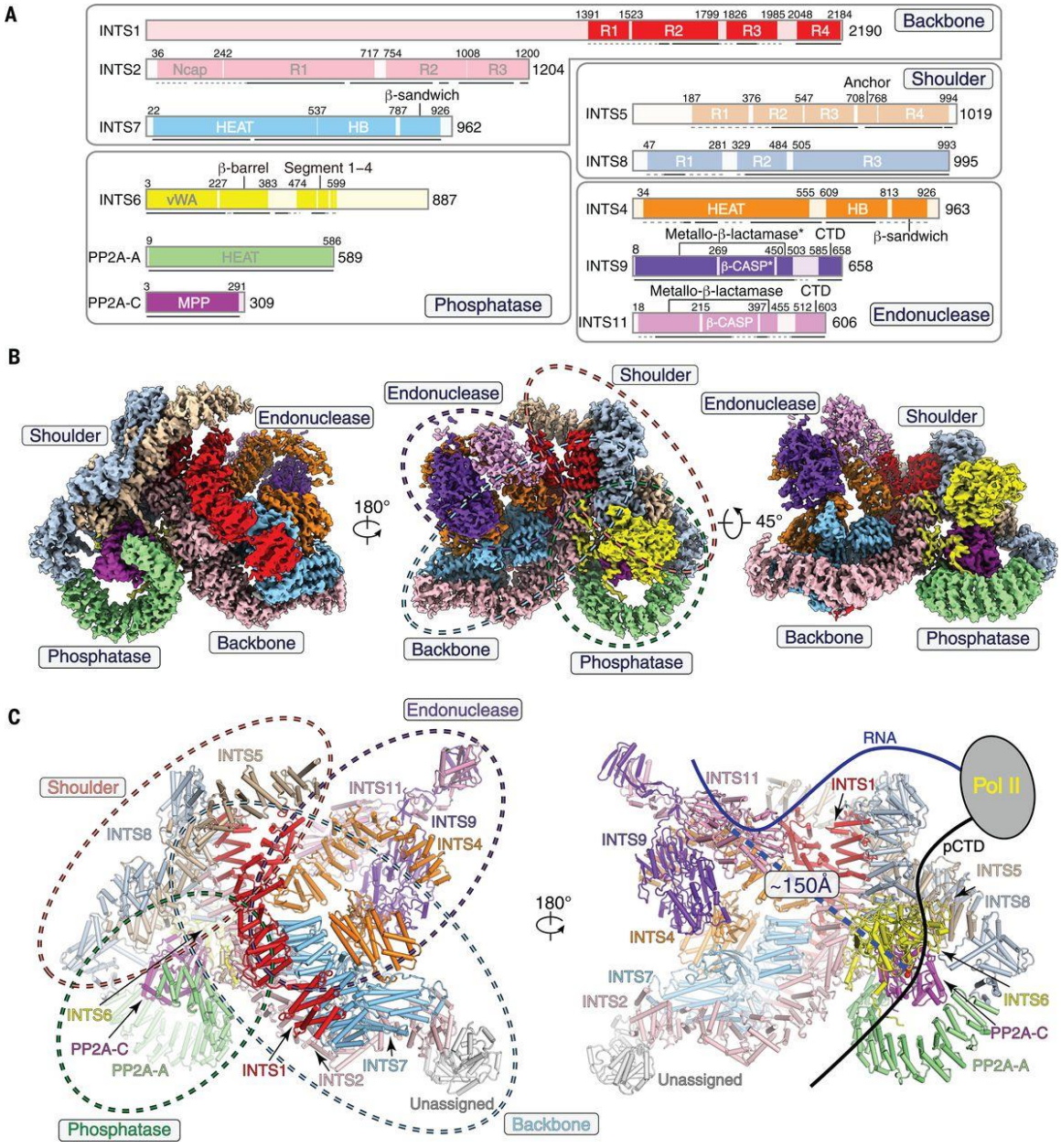


Figure 1.6: Overall structure of the INTAC complex.

(A) Schematic modular organization and domain structures. Residues at domain boundaries are indicated. Solid and dashed lines indicate regions that were modeled with residues and poly-alanine, respectively. The color scheme is indicated and used throughout all figures. R1, R2, R3, and R4, repetitive helix hairpin domains; HEAT, HEAT repeat domain; HB, helix bundle domain; MPP, metallophosphatase domain. The * represents an inactive catalytic domain of INTS9.

(B) The 3.5-Å-resolution cryo-EM map of the INTAC complex in three different views, with subunit surfaces colored as in (A). The four modules are indicated with dashed circles.
(C) Cartoon model of the INTAC structure shown in two different views. The putative binding sites of RNA transcript on INTS11 and phosphorylated CTD (pCTD) of Pol II on PP2A-C are indicated. The zinc and magnesium cations at the catalytic centers of INTS11 and PP2A-C are shown as gray and red spheres, respectively, and their distance is indicated by the blue dashed line.
(Figure and legend text reprinted with permission from (Zheng et al. 2020))

Apart from Cryo-EM structures, a large number of interactions between small sets of subunits have been identified through other biochemical methods and more targeted Cryo-EM studies. Yeast two hybrid was employed to detect the interaction between IntS1 and a small domain of IntS12 (Chen et al. 2013). IntS4 interacts with the IntS9 and IntS11 heterodimer similar to Symplekin in the CPSF73/100 complex (Dominski et al. 2005; Wu et al. 2017; Albrecht et al. 2018; Pfleiderer and Galej 2021). The interaction of IntS4 is critical to Integrator's endonuclease function of the IntS9 and IntS11 heterodimer and loss of IntS4 shows a similar phenotype to loss of IntS9/11 (Albrecht et al. 2018). Cryo-EM structures exist of the critical regions of the Carboxyl-termini of IntS9 and IntS11 interaction surfaces, and of the IntS4/9/11 heterotrimer, which has been named the Integrator Cleavage Module (CM) (Wu et al. 2017; Pfleiderer and Galej 2021). The interaction between IntS5, IntS8, and the pr65 scaffold subunit of the PP2A phosphatase was demonstrated by yeast two hybrid and later followed by affinity purification and cryo-EM structure (Huang et al. 2020; Zheng et al. 2020; Pfleiderer and Galej 2021). A dimer of IntS3 interacts with IntS6 via the IntS3 carboxy-terminus playing a critical role in functions separate from of the remainder of the Integrator complex (Zhang et al. 2013; Li et al. 2021).

In 2020 there was a partial structure of the IntS13 and IntS14 interacting domains in which they dimerize to form a shared domain due to their shared sequence similarity. This study also showed the formation of a heterotrimer of IntS10/13/14, and an interaction between IntS13 and the CM (Sabath et al. 2020). Concurrent with this study it had been determined that IntS13 interacts with both the CM and IntS10 and IntS14 through yeast two hybrid analysis while additionally identifying regions of IntS13's Carboxy-terminus important to interaction with the CM without disrupting interaction with IntS10 and IntS14 (Mascibroda et al. 2020). It was also demonstrated through Co-IP experiments that IntS10/13/14 could stably co-purify each other (Mascibroda et al. 2020; Pfleiderer and Galej 2021)

The association between the CTD and Integrator has been firmly established from the discovery of the complex (Baillat et al. 2005). However, even with all the structural and biochemical studies performed to date, the precise set of subunits involved in direct interaction between Integrator and RNAPII remains elusive.

While most Integrator subunits complex together exclusively, there have been other complexes discovered with certain subunits outside of the canonical Integrator complex. As described above, IntS3 dimerizes with itself to interact with IntS6 (Li et al. 2021). These two subunits also form components of the sensor of single-stranded DNA (SOSS) and the human single-stranded DNA binding protein (hSSB) complex (Huang et al. 2009; Zhang et al. 2013). IntS3, also termed SOSS-A, interacts with hSSB1/2 (SOSS-B1/2) and SOSS-C to bind ssDNA (single-stranded DNA) via its N-terminus (Huang et al. 2009; Ren et al. 2014). These four proteins recognize DNA damage as part of the homologous recombination DNA damage repair process (Zhang et al. 2013).

Integrator function. One of the functional keys to the complex is the enzymatic subunit Integrator subunit 11 (IntS11), which is an RNA endonuclease with both a β-CASP and metallo-β-lactamase domain. IntS11 has been described as a functional paralog of CPSF73 via sequence and structural analysis. As described above, CPSF73 is an RNA endonuclease which cuts mRNA during RNAPII transcriptional termination and facilitates the addition of the poly-A tail as part of the CPSF machinery (Dominski et al. 2005; Wu et al. 2017). Much like CPSF73 has a homodimer binding partner in CPSF100, IntS11 has a homodimer binding partner in IntS9 (Integrator Subunit 9), which is a paralog of CPSF100. The main difference between the two homodimer pairs (IntS9/11 vs CPSF 73/100) is in sequence divergence in their CTDs (Albrecht and Wagner 2012; Wu et al. 2017) (Figure 1.7). IntS11's endonuclease function is to cleave the 3' end of a variety of RNPII transcriptional RNA products to facilitate termination of RNAPII in a CPSF independent manner.

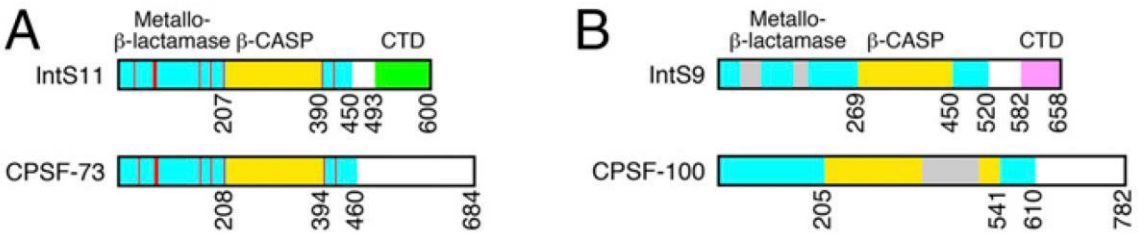


Figure 1.7: Domain comparisons between Integrator and CPSF endonucleases

(A) Domain organizations of human IntS11 and CPSF-73. The metallo-β-lactamase and β-CASP domains are shown in cyan and yellow, respectively. The conserved residues in the active site are indicated by red lines. The CTD of IntS11 is shown in green. CPSF-73 also has a CTD, but its sequence is highly divergent from that of IntS11, and its exact boundary is not known.
(B) Domain organizations of human IntS9 and CPSF-100. The CTD of IntS9 is shown in pink. An insert in the β-CASP domain of CPSF-100 and two inserts in the metallo-β-lactamase domain of IntS9 are shown in gray. (Figure and legend text reprinted with permission from (Wu et al. 2017))

The Integrator complex has been identified as playing a key role in RNAPII transcription of multiple forms of coding and noncoding RNA. These include uridine-rich small nuclear RNAs (snRNA), enhancer RNA (eRNA), telomerase RNA (TR), and some messenger RNA (mRNA) (Baillat and Wagner 2015; Beckedorff et al. 2020; Mendoza-Figueroa et al. 2020; Thomas et al. 2020; Beltran et al. 2021; Kirstein et al. 2021; Rosa-Mercado et al. 2021). Ribonucleoproteins and snRNAs combine to form the spliceosome which carries out splicing of mRNA. During RNAPII transcription of snRNA, the Integrator complex recognizes a sequence known as the downstream 3' box shortly after a terminal stem-loop structure and cleaves the nascent RNA downstream of this location and causes RNAPII termination. This allows the snRNA to be further processed and incorporated into the spliceosome. Loss of Integrator function leads to read-through and misprocessing of the snRNA.

Long noncoding RNAs (lncRNA) transcribed from an enhancer region are referred to as eRNAs. These enhancers regions are composed of a distal cis-regulatory regions to which transcription factors and RNAPII bind. Chromatin looping by cohesion of these eRNAs at the target mRNA promoter regions promotes gene expression (Kagey et al. 2010; Ong and Corces 2011). Integrator function has been found to be key to eRNA transcription as the Integrator complex cleaves the nascent RNA to complete RNAPII transcription. Loss of Integrator function leads to 3' end misprocessing and loss of function of eRNA structures (Lai et al. 2015).

SIGNIFICANCE OF THE STUDY

While a number of ChIP-Seq experiments have shown the association of Integrator with the TSS of a number of paused mRNA genes, the actual role of Integrator at these paused gene promoter sites has yet to have been elucidated. The main purpose of our study was to determine if the endonuclease function IntS11 may play a role at these TSSs and how that endonuclease function may be regulated.

Chapter 2. Methods

DROSOPHILA CELL LINES

Drosophila DL1 cells were cultured at 25°C in Schneider's *Drosophila* medium (Thermo Fisher Scientific 21720024), supplemented with 10% (v/v) fetal bovine serum (HyClone SH30910.03), 1% (v/v) penicillin-streptomycin (Thermo Fisher Scientific 15140122), and 1% (v/v) L-glutamine (Thermo Fisher Scientific 35050061). Drosophila S2 cells from the DGRC were grown in SFX-Insect serum free media (Thermo Fisher Scientific SH3027802).

RNAI

Double-stranded RNAs from the DRSC (*Drosophila* RNAi Screening Center) were generated by *in vitro* transcription (MEGAscript kit, Thermo Fisher Scientific AMB13345) of PCR templates containing the T7 promoter sequence on both ends. Primer sequences are provided in Table 2.1. Knockdown experiments in 6-well dishes were then performed by bathing $1.5x10^6$ cells with 2 μg of dsRNA, followed by incubation for 60 hours of standard cell culture conditions. For RNAi + rescue experiments cells were incubated for 60 hours in the presence of dsRNA and media was supplemented with a final concentration of 100 μM $CuSO_4$ to induce expression of the RNAi-resistant transgenes.

RT-QPCR

Total RNA was isolated using Trizol and cDNA was reverse transcribed using M-MLV Reverse Transcriptase (Thermo Fisher Scientific 28025) according to the manufacturer's instructions. Random hexamers were used for cDNA synthesis and RT-qPCR was then carried out in triplicate using Bio-Rad iTaq Universal SYBR Green Supermix (Bio-Rad 1725120). Primers are in Table 2.1.

Table 2.1: Primer List

Northern Blot Probes	**Sequence**	
MtnA probe	AAGATGCAGCGCCTCTACTC	
RpS6 probe	GTGATCAGGCGCTGAATTTTGGGG	
MtnA +1 to +50 probe	GTTGCACTGAGATGATTCACTTGATTTTGCTGCTGACCACAACTGATGCA	
tRNA:val4:70 BCb probe	GTTTCCGCCCGGGATCGAAC	
eGFP probe	GTCACGAACTCCAGCAGGAC	
RpL32 probe	GACGCACTCTGTTGTCGATACC	
MtnA -1 to -20 probe	CATTGGCCAGATGCTCTCGG	
MtnA +1 to +20 probe	TGCTGACCACAACTGATGCA	
MtnA +21 to +40	GATGATTCACTTGATTTTGC	
MtnA +41 to +60	TAGGCCTTTAGTTGCACTGA	
MtnA +61 to +80	AAAGGTAGGTATGGGCTATT	
MtnA +75 to +105	CTCGAACTTGTTCACTTGTTTACAAAAAAGG	
MtnA +101 to +120	TTGAGTTGTATTTCCTCGAA	
MtnA +121 to +140	GCATGGGCAAGGCATCTTGA	
bantam miRNA probe	AATCAGCTTTCAAAATGATCTCA	
CG6770 probe	CTTCTGCTTTTTTTTGTTCACTTTTCGCTGAGCTGCGAACGGAATTGAAT	
pst probe	TTCATGGTTTCCTAATCTATTCGCCATCCCTAAGTTGTCACCTTTCCGTT	
CG2247 probe	ACACTGAACACGACAACTTCTTTCGGAAATATAAAACGAACCTACAAGCG	
Sirup probe	TCGGCTGGGAGTCAATGTCACTTGCTTGTTGCTTCGCGTCTGAATTGCAA	

CG8620 probe	ATCTGTTGCTTATGTTTCGACTGCTTCACTTGAACACTTGAATTTGACTG	
U1 probe	CTTCGTGATCACGGTTAACCTCT	
U2 probe	ACAGATACTACACTTTGATCTTAGCCA	
U4:39B probe	ACCTCAGGAGGACTTCATTGG	
U5 probe	GGCGAAAGATTTATTCGACAATTGAAGAGAAAC	
U6 probe	TTCCAATTTTAGTATATGTTCTGCCG	
MtnA 45-55/MCS antisense	CGACCTCGAGCTTTAGTTGC	
MtnA 30-40/MCS antisense	CGACCTCGAGGATGATTCAC	
MtnA 10-20/MCS antisense	CGACCTCGAGTGCTGACCAC	
RT-qPCR Primers	**Forward Primer**	**Reverse Primer**
MtnA	ACTGCGGATCTGACTGCAAG	AAGATGCAGCGCCTCTACTC
RpL32	TACAGGCCCAAGATCGTGAAG	GACGCACTCTGTTGTCGATACC
DCP2	CGCAAGGAGAAGCAGCAACAACTT	TGACTGGCTGCTGTGGATTGTACT
MtnA pre-mRNA	GCCTTGCCCATGCGGAAGC	CTGGAAAGAAGTAGAATTTAAAAATTAGTACATGCTGGTACATC
RpL32 pre-mRNA	TGCTAAGCTGTCGGTGAGTG	CATTTGTGCTGCAAGGAGAC
DCP2 pre-mRNA	CATGGTTTATGGACTTTGAAAACA	GCATCGATTAGGTCCGTGAT
Pepck1	GAACTGACGGACTCTGCTTAC	GGTGCGTTCGGGATCACAA
Hml	TGGTTATGGCGGGATAAAGACG	GTTGCCCTGACTTCCCTGG
CG8620	GCAATCCGATAACGTGGCAC	CAAGGCCATGTCCTCGACTT
Sirup	TGGGCAAGCTGGATGAAT	CGTATACGGATTGGTCTGATTG

CG2247	AACGTGGATCTTTCGACTCAC	CGGCGTTCAAATTGACTCTTG
pst	AAACTGCAACGGAAACTGAAAAA	ACGGAATCGAAAATGATCTGACG
CG6770	ACACCAACCACTTCGATCCC	GCTTGGTCAGAATCTTGCGG
Pepck1 pre-mRNA	CACAAACAAAATGCCTGAGC	TTGAATGCGTTTCGAGTGAC
Hml pre-mRNA	TAACCGATGATGACGACGAG	GACGTATTTATTCCGCTTTACGA
Sirup pre-mRNA	CCCAAATGGGCAAACAAG	AGAACGTTAGCATCGCCACT
CG2247 pre-mRNA	CCAGCATCTGTAGCATAATAACACA	CGGCGTTCAAATTGACTCTT
pst pre-mRNA	TCAAAGTATTTGGGGTAATGACG	GACAGAGTTTGCCCCTTAGC
MTF-1	CAGGAGCGGCCCTACAAATG	TGTGTCCTTCGGTGGGTCTT
MED9	CGTTAACATGTTTAAGAACAACGTG	TCCTTCTTCACTCCCTCCA
MED15	GGTCTGGTAGCCAAGCTCTTT	GGCAAGATTCTGAAGTGCGTT
Rrp40	AGAGCAAGCGGGTGATACTC	GGGTTCCTTGTGTCGAAGTGG
Mtr4	GAAGAGCTGTTCGACTGTTTTGA	GCTTGGCGCTTATTTCCTTTC
Mpp6	GGCAGCATGAGCAAGAAGTT	TGGCATGCGGACTATCACTA
Dis3	AACGAGGTGAAGCACAGGAG	GGCGGTCTCATCTGGTTCG
Rrp6	CCCGCGCCCTTTACCTAAG	ACTCCTGCACCATCTCAAACT
IntS1	TGCACATTCTGTTCGCCAATC	CGTCTATGTAAATGCGGAGCAG
IntS2	TGAGGATGTACGATGTATCGCC	ACGGCCAGGATCTCTTTGC
IntS3	CCTCCTACAAAACGTAGCTCG	CCGTATATGCAGTAGTTGCCAA
IntS4	CCGCGAATAGCAGGGAACTG	CCATTTCCACGTCCTCATAGGAT
IntS5	CGCCAGAACCTGTTGGATCAG	CAGTGCGAGCTTGATGAGATT
IntS6	GCCAGAAGGCGTATGTGAATG	GTAGCGATCTCCCAGGCAAT
IntS7	TAAGCTGGCCGATTACTTTGTC	CTTGTCCAGATGATTCTCGCTC

IntS8	TTGCAGAATCTAAACCAGACGC	GGGCCTTTGTGGGCTATGG
IntS9	ATGCGATTGTATTGTCTCAGCG	GAAATTCAGGACTGTCTGCTCC
IntS10	TACGAGGCTTATCTACTGGAACG	CAGCTCCGTATGTTGATTCTGG
IntS11	TGGAGGACATGCGGAAGGT	TCTGGTGAAGTGTGACGGGA
IntS12	TGGAGGATGAAGCGAACTTTTC	CGAATCGGAGGGCTCCTTG
IntS13	CATGTGCGGTTTATTGTCTCGG	GCGCCATATTTTGCGTGCT
IntS14	CCTGCACGGCGATTACTACTT	GCCCTTCAGTGAGACCTTGTC
eve	CATGCACGGATACCGAACCT	AGGCATTCATTTGGCGAGGG
Kah	GACTTCCCATCGGCAACTGA	GAGCTGGTGAAATGCCCCAC
l(1)G0469	ATTCGCAATCAGGTGTCGGT	CTCGTCGTTACCTGTCGCTT
Kal1	GCCACACAAGGAGAAGGTCT	TTCTGGAGTCGTCCAACACG
GstS1	TGGCTGGAGAGAGTGAGAGT	GCTTTCTCAATGCAATTCCCTCAA
IRSp53	CGTTCAGAGCTCGAAGAACCAT	CTCGGCACCGTGTTGGTAT
CG9896	GTACGCTTCTGTCACTCGGT	GGTTTCGTTTGGTCTGGAGC
dmGlut	TTTCGTGATACCCCAGCGAG	ATCGCCACCAACACTGGTT
CG31431	CGTCGTGTCAGTTGTCCATC	GTTTTCGCTGCATTGCTTAAAGT
E(spl)m2-BFM	ATGACACCAAGTCAACGCCA	AAGAAGGTGCCATTGTCCGT
Fen1	CTGGCTGAACAACTTTGCCC	CACATAGAGGCCACCTGCAA
CG6006	ATGAACTTGAACGCCGCAAC	GGCCCTACGGTCTTATTTGCT
wdp	GGTCGTGTGTGATCTCCGAT	CCTCTCTTGGCGTCTTATCTTCT
SP1029	AGAGTGCAGCTTGAACGGAA	TTCATTACCAAAATTGTGTCACTCT
CG42240	AGTTAGTGACGGTGCCGATG	GCGTATCCCCAAACGACTGA
ChIP-qPCR Target	**Forward Primer**	**Reverse Primer**

Intergenic	GGCCGACGAGATGGGTCTG	GTAGGACGTGATACACACAT
MtnA 5' end	TGCATCAGTTGTGGTCAGCAGCAAAATC	AGGTATGGGCTATTTAGGCC
MtnA 3' end	GACGCAGTTAGGCATCAATTACT	CGCTGCATCTTGTCTCTCTACA
Pepck1	GTTTGACCCTCTCACTCGGC	ATCCGGGGCCTTTTATACCC
Hml	GGCGTTGTGGCTGCTTTTTA	CCTCGTCGTCATCATCGGTT
CG8620	AAAGCTCATCGACCGAATGC	GCTGCGGAATCTGTTGCTTA
Sirup	GACATTGACTCCCAGCCGAA	CCGTTTGTCTGGTCACGGAT
CG2247	GAGAGGAAACCTGACGCACA	AAGCGAAACGGTCCGAAAGA
pst	ATTATCTCTTCCGCGTAGCCG	ACCATGAAGTTACCCGCACC
CG6770	TATAAAAGCCGCTGCTCGAC	GGCTGCTTTCCAAATTTCTC
DCP2	TGCTCAAAACTCGCCTTCCT	GCTTCCCTGGTCGCTAAAGA
eve	TTACCATTTGGCGAGGGAGG	GCGCAGCGGTATAAAAGGG
Kah	CCCAGGAAGAACTGAGCGTT	AGAGGATGGGAGACCGAGAC
l(1)G0469	CACTTTTCCGCGCGTTTTCC	ATAAAGCGACGTGCCCGAAA
Kal1	ATCAACAGTTCCCAGGAGCG	CGGGCAGTGCGATACATTTTT
GstS1	GGTGCTACGAAATGAGGTGGT	GCGGCAAGCGTATAAAAGCA
IRSp53	TTCGCTATTCCCGATACGGC	GCACGCCCACCCAGTTATTA
CG9896	TTACTTTGGTCTGGAGCCGC	TTTGGATACTTGGCGCTCGG
dmGlut	CGGCGCTTATCTGCTCTCA	GCAGCGGAGAAGGAGAATGT
CG31431	ATTGTTCGGGTGGGAAAGCA	AGCAAACACACTCACTGTGGCTA
E(spl)m2-BFM	TCAGCGGTTTCCACACGTTA	TGCTTGGGTTATAGCCGCTC
Fen1	CACATTGGTAGTTGCGCTCAC	AATACAACACTCGGCTGCGG
CG6006	GAGCATCGAATCTGCCAACG	AATTTCGGACCTCGTACCGC

wdp	CGCTATTGGAACCCCCGATT	GGCTTGGCACTCTCCTTCTC
SP1029	TTCCGTTCAAGCTGCACTCT	TCGACGTTTCACTGGCTCTC
CG42240	GTGCAACGATGCGTATGAGT	GCTGCTTGCCTTTTTCCCTC
U1	GCTGAGTTGACCTCTGCGATTA	CTTTTAAAATTTATTGCAGATGTCGG
U2	CCCGGTATTGCAGTACCGCCGGGA	ATCCTACCATTCGAATTTGCATGTAAA
DCP2	TGCTCAAAACTCGCCTTCCT	GCTTCCCTGGTCGCTAAAGA
Cyp12e1	TGCTGATATTCCGGGACCCA	TTTCCTTCACTTACCGCCCG
MAGE	CGAGATATGCGGTCACACCA	TCACAACAGTCTACCGGTGC
Scamp	AGCTCTTCTATCGCCTCACC	GTGCGAGTCAGTGCGTTTTT
SA	CCCTTGTTTCCGATTGCGTC	TGACCAAAGCGTCCGATTGA
eTSS1	TCGAGATATGCGGTCACACC	GGGGCTCACAACAGTCTACC
eTSS2	TACAACATCTACGGCTGCGA	GGGGAAATGTCAAACGCTCG
eTSS3	GCTCTCAAGACCGTTCGGAAT	TCTGTACGTTTGCTTGTGTGTT
eTSS4	TGACTCTAAGCCAGGGACCA	AGCCGTGTCCACATCTCATC
eTSS5	GGAGCGTAGTCGGCAATCAT	CACGATTTGTTCAACCGCGA
eTSS6	TGGCTGGCCACACTAATACA	TATAGGCCCCGACTGGGATT
eTSS7	GCGTTTCAACTCTCATCGCC	GTCGCACAGACACACCTACA
eTSS8	GAGCGCTGTTGCCGATTTTC	CCCAACGCCCCTTTTTCAGT
eTSS9	ATCTTGCCTCGCAAAACCCT	CGCTTTGGCGTGCTAATGAA
eTSS10	TTCGGCACGAAACAAATGCC	CTGGTCACACATCCCATCCC
eTSS11	TAACCAGGGTCGGCACAAAG	ACAGTCTAAAAGATGACAGCATTG
3' RACE	**Sequence**	
3' RNA adapter	[Phos]GAUCGUCGGACUGUAGAACUCUGAAC[dT5F]	

RT primer and PCR reverse	AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA	
MtnA First 20 cycles PCR forward	CAAAATCAAGTGAATCATCTC (+22 from TSS)	
MtnA Nested PCR forward	ATCATCTCAGTGCAACTAAA (+35 from TSS)	
dsRNA Target	**Forward Primer**	**Reverse Primer**
βgal (Control)	TAATACGACTCACTATAGGGCTGGCGTAATAGCGAAGAGG	TAATACGACTCACTATAGGGCATTAAAGCGAGTGGCAACA
MTF-1	TAATACGACTCACTATAGGGCCGCTGACGGATGCCT	TAATACGACTCACTATAGGG GGGTGCGCCAGTCCTG
MED9	TAATACGACTCACTATAGGGTGGATTTGTCGCCAAACAAT	TAATACGACTCACTATAGGGCACAATGTCGTAGATTATCGG
MED15	TAATACGACTCACTATAGGGCCGGAACTGCCTCTAACTTG	TAATACGACTCACTATAGGGTGTTGCATGGCATTTACGTT
Rrp40	TAATACGACTCACTATAGGGCAGCCTCCATATCGTATCTC	TAATACGACTCACTATAGGGCGAGTTGACGCAGACCA
Mtr4	TAATACGACTCACTATAGGGGTGCTCACCGAGGAGGAT	TAATACGACTCACTATAGGGCAGTGCAGCTTGATTTTGG
Mpp6	TAATACGACTCACTATAGGGAATGCCATCCAAATCAAAGC	TAATACGACTCACTATAGGGTGTCTTGGTCGGATACCTCC
Dis3	TAATACGACTCACTATAGGGATCATCGTAACGATTGACACA	TAATACGACTCACTATAGGGCTTCATTGTCCACTTCCCAC
Rrp6	TAATACGACTCACTATAGGGTTATCGTTGTATATCGTCAACAT	TAATACGACTCACTATAGGGGCATCTCCCTTGGAAGACT
IntS1	TAATACGACTCACTATAGGGATTAAGGGCATGTCGTCGTC	TAATACGACTCACTATAGGGGAATGTGCAGGTTGGTGTTG
IntS2	TAATACGACTCACTATAGGGTTCTTCGAGGGACAGCAA	TAATACGACTCACTATAGGGTGCTTGAGCGTGAGCTTA
IntS3	TAATACGACTCACTATAGGGTCATGAAACTGGGCTCATAAA	TAATACGACTCACTATAGGGCTGATAATGGTAGGTCACGT
IntS4	TAATACGACTCACTATAGGGCCTGTGGCGCCCTTATAC	TAATACGACTCACTATAGGGTTCGGGCGTCTCGAAAA

IntS5	TAATACGACTCACTATAGGG TCATGCTCAATGCCTTTCAC	TAATACGACTCACTATAGG GTCGTGTCCGAGTAGTTGG TG
IntS6	TAATACGACTCACTATAGGG GCTTGTTTTCGCTTGTCCTC	TAATACGACTCACTATAGG GTTTTCTGCGTGATGTGCTT C
IntS7	TAATACGACTCACTATAGGG ATCCCATGCTAGCTCGTTT	TAATACGACTCACTATAGG GAAAGGTGCACGGATGCT G
IntS8	TAATACGACTCACTATAGGG AATCGCTACTTAACAACTACA C	TAATACGACTCACTATAGG GAGCGAGGCCAACGAGT
IntS9	TAATACGACTCACTATAGGG GGTCTTTTGTGGCCATCCTA	TAATACGACTCACTATAGG GTAAATTCGATCCAGCTTC CG
IntS10	TAATACGACTCACTATAGGG GCTGGGAGCCCTTCCTG	TAATACGACTCACTATAGG GAGGCTTTGCACCAGACTG
IntS11	TAATACGACTCACTATAGGG CTGTGAGCCAAAGAACGTCA	TAATACGACTCACTATAGG GCACCTTCACCTCAACGGA TT
IntS12	TAATACGACTCACTATAGGG AATGCGGACGAGATCATCA	TAATACGACTCACTATAGG GGCTGGCATTCGTGGAACT
IntS13	TAATACGACTCACTATAGGG ACGCACCTCATACTGAACC	TAATACGACTCACTATAGG G TTCGGCGGTGCGATAG
IntS14	TAATACGACTCACTATAGGG GAACTGGGCCGCACTAC	TAATACGACTCACTATAGG GGTTTTCTTTTTGTTGCTCT GTC
IntS11 UTR	TAATACGACTCACTATAGGG GCAACGAAGAAGCATCCCAT TG	TAATACGACTCACTATAGG GGGAACGTTTATGTATATT TTGATACTC
Cloning Primers	**Forward Primer**	**Reverse Primer**
IntS1 pUbi-p63E-FLAG	GGCCGCGGCCGCTGATCGCG GGAAAGGAAGCGGCTCC	GGCCTCTAGATTAATAGCC ATGCTGGATCACTAGAG
IntS5 pUbi-p63E-FLAG	GGCCGAATTCTGCTGCGCCA GAACCTGTTGGATCAG	GGCCTCTAGATTAATCTAT TTCAACGATCTGCAGCCGG G
IntS8 pUbi-p63E-FLAG	GGCCACTAGTCCCGGACATC AAGATAACGCCCTTG	GGCCACTAGTCTACAGCAA ATACTGCTTGGCC
IntS11 pUbi-p63E-FLAG	GGCCACTAGTCCCGGACATC AAGATAACGCCCTTG	GGCCCTCGAGCTAGCACAT ATTCTGCAGCACATTC
IntS11 E203Q site directed mut.	CCGGACTTGCTGATCTCCCAG AGCACCTACGCCACTACC	GGTAGTGGCGTAGGTGCTC TGGGAGATCAGCAAGTCCG G

IntS11 pMT-FLAG-puro	GGCCACTAGTCCCGGACATCAAGATAACGCCCTTG	GGCCCTCGAGCTAGCACATATTCTGCAGCACATTC
eGFP pMT-FLAG-puro	GGCCGGATCCAGTGAGCAAGGGCGAGGAG	GGCCACGCGTTTACTTGTACAGCTCGTCCATGCCGAG
CG7044 pMT-FLAG-puro	GGCCGGATCCACAAAAGCAAGAAGAGACGC	GGCCGCGGCCGCTTAGTAACAGTCCGAAATCATGTT
CG7044 DEL 966-974 pMT-FLAG-puro	GGCCGGATCCACAAAAGCAAGAAGAGACGC	GGCCGCGGCCGCTTAGTCGCGCTTCAGCGCACCG
CG7044 DEL 947-974 pMT-FLAG-puro	GGCCGGATCCACAAAAGCAAGAAGAGACGC	GGCCGCGGCCGCTTAGCACTCGAACCACTCCTGGTG

CHARACTERIZATION OF 3' ENDS OF MTNA SMALL RNAS USING 3' LIGATION-MEDIATED RACE

DL1 cells were treated with Mtr4 dsRNAs for 3 days and $CuSO_4$ was added for the last 14 h. Total RNA was isolated using Trizol and 2 μg was ligated to 10 pmol of the 3' RNA adapter oligo (Sigma) using T4 RNA Ligase I (NEB M0204S) following the manufacturer's protocol at 20°C for 6 h. The reaction was acid-phenol:chloroform extracted (Thermo Fisher Scientific AM9720) and ethanol precipitated. Reverse transcription (RT) was performed using M-MLV (Thermo Fisher Scientific 28025013) using the 3' RACE RT Primer. 2 μL of cDNA was used as a template for 20 PCR cycles using PFU and a gene-specific MtnA 5' forward primer and the RT Primer (95°C melting 15 s, 60°C annealing 15 s, 72°C extension 30 s). 2 μL of this PCR was then added to a new reaction for an additional 20 PCR cycles using the nested forward and RT primers. The resultant PCR products were cloned using the Zero Blunt TOPO PCR Cloning Kit (Thermo Fisher Scientific) and sequenced via sanger sequencing (Genewiz). Primers are in Table 2.1.

ANALYSIS OF PROTEIN EXPRESSION BY WESTERN BLOTTING AND IMMUNOFLUORESCENCE

For Western blotting, cells were gently washed in PBS and then resuspended in RIPA buffer (150 mM NaCl, 1% Triton X-100, 50 mM Tris pH 7.5, 0.1% SDS, 0.5% sodium-deoxycholate, and protease inhibitors [Roche 11836170001]). Lysates were passed 10 times through a 28.5 gauge needle and clarified by centrifugation at 20,000x*g* for 20 min at 4°C. Lysates were then resolved on a NuPAGE 4-12% Bis-Tris gel (Thermo Fisher Scientific NP0323) and transferred to a PVDF membrane (Bio-Rad 1620177). Primary antibody incubations (IntS9 [guinea pig], IntS11 [rabbit](Ezzeddine et al. 2011) or alpha-tubulin (rabbit, abcam ab15246) were all done at room temperature for 2 hours with a 1:1000 dilution in 5% milk in TBS-0.1% Tween. Conjugated secondary antibodies against

rabbit (GE Healthcare NA934) or guinea pig (Sigma AP108P) were incubated at room temperature for 90 minutes with 1:10000 dilution in TBS-0.1% Tween. Membranes were processed using SuperSignal West Pico Chemiluminescent Substrate (Thermo Fisher Scientific PI34080).

NORTHERN BLOTTING

Total RNA was isolated using Trizol (Thermo Fisher Scientific 15596018) as per the manufacturer's instructions. Small RNAs were separated by 8% denaturing polyacrylamide gel electrophoresis (National Diagnostics EC-833) and electroblotted/UV crosslinked to Hybond N+ membrane (GE Healthcare RPN303B). ULTRAhyb-oligo hybridization Buffer (Thermo Fisher Scientific AM8663) was used as per the manufacturer's instructions. All oligonucleotide probe sequences are provided in Table 2.1. Blots were viewed and quantified with the Typhoon 9500 scanner (GE Healthcare) and quantified using ImageQuant (GE Healthcare).

CHROMATIN IMMUNOPRECIPITATION (CHIP)-QPCR

A 10-cm dish of 5 x 10^7 DL1 cells was harvested into a 15 mL tube and centrifuged at 1,500x *g* for 2 min. Cells were then washed with 10 mL PBS and centrifuged at 1,500x *g* for 2 min. The cell pellet was resuspended in 10 mL of Fixing Buffer (50 mM Hepes pH 7.5, 100 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0 with 1% formaldehyde) and incubated at room temperature for 30 min. 0.5 mL of 2.5 M glycine was then added (final concentration of 0.125 M) and incubated at room temperature with rotation for 5 min, centrifuged at 1,500 *g* for 2 min, and washed two times with 10 mL PBS. Cells were lysed using lysis buffer (50 mM HEPES pH 7.9, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100) for 10 min on ice and centrifuged at 1,500 *g* for 2 min. The pellet was then washed 2x in Wash Buffer (10 mM Tris-HCl pH 8.1, 200 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0) and resuspended in 1 mL Shearing Buffer

(0.1% SDS, 1 mM EDTA, 10 mM Tris-HCl pH 8.1). The suspension was sonicated at 4°C using a Covaris S220 machine to obtain 500 bp DNA fragments in TC12x12 tubes with AFA fiber (Settings: Time- 15 min, Duty Cycle- 5%, Intensity- 4, Cycles per Burst- 200, Power mode Frequency- Sweeping, Degassing mode- Continuous, AFA Intensifier- none, Water level- 8). To the 1 mL of sheared chromatin, 115 µL of 10% Triton X-100 and 34 µL 5 M NaCl was added per ml of sheared chromatin, so that the final concentration of the sample is 1% Triton X-100 and 150 mM NaCl. Sheared chromatin was pre-cleared with protein A/G beads and 10 µL was reserved as input control. For each IP sample, 100 µL of sheared chromatin was diluted to 1 mL using IP Buffer (0.1% SDS, 1 mM EDTA, 10 mM Tris-HCl pH 8.1, 1% Triton X-100, 150 mM NaCl) and incubated overnight at 4°C with 10 µL of serum. The next day, lysates were immunoprecipitated with protein A/G beads for 2 h at 4°C and washed once with low salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Hepes pH 7.9, 150 mM NaCl), twice with high salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Hepes pH 7.9, 500 mM NaCl), once with LiCl buffer (100 mM Tris-HCl pH 7.5. 0.5 M LiCl, 1% NP-40, 1% Sodium Deoxycholate), and once with TE. Immunocomplexes were eluted and de-crosslinked at 65°C overnight with Proteinase K and RNase A. DNA was extracted by phenol-chloroform and ethanol precipitated. DNA was resuspended in 100 µL, and 2 µL was used for each qPCR reaction. Primers are in Table 2.1.

QUANTIFICATION AND STATISTICAL ANALYSIS

For RT-qPCRs statistical significance for comparisons of means was assessed by Student's t-test. Unless otherwise indicated, the comparison was to the control RNAi treated samples. Statistical details and error bars are defined in each figure legend.

GENERATION OF TRANSCRIPT ANNOTATIONS

All transcript annotations for *D. melanogaster* r5.57 were downloaded from flybase.org in GTF format and filtered such that only "exon" entries for the feature types considered for re-annotation remained. Annotations from chrY, chrM, and random chromosomes were also excluded. Unique "gene_id" values were assigned to each transcript, such that those grouped and represented by a single member in TSS-based analyses were identical. Precise TSS locations employed were based on high-resolution Start-seq data as described previously (Nacheva et al. 1989; Henriques et al. 2013). The start location of each transcript was adjusted to the observed TSS from Start-seq when this resulted in truncation, rather than extension of the model. If the observed TSS fell within an intron, all preceding exons were removed, and the transcript start was set to the beginning of the following downstream exon. Gene annotations for the human genome (hg19, GRCh37 genome build July 2019) were downloaded from gencodegenes.org in GTF format and filtered such that only "gene" entries for the "protein_coding" feature type remained. Annotations from chrM, and random chromosomes were also excluded.

TSS CLUSTERING BASED ON PROMOTER POL II HALF-LIVES UPON TRP TREATMENT

TSS clustering was accomplished as described in (Henriques et al. 2018) using k-medoids clustering based on the Clustering Large Applications (CLARA) object in R.

FEATURES ASSOCIATED WITH GENES WITH SHORT-LIVED PROMOTER POL II OCCUPANCY

A comprehensive repertoire of ChIP-seq datasets from (Lim et al. 2013; Weber et al. 2014; Baumann and Gilmour 2017; Henriques et al. 2018; Kaye et al. 2018) and ChIP-chip from the modENCODE database (mod et al. 2010; Ho et al. 2014) was used representing a total of 111 datasets that include transcription factors, chromatin remodelers and histone modifications.

To find features enriched at protein-coding transcription start sites with short-lived promoter Pol II occupancy a similar approach to the web-based tool ORIO (Lavender et al. 2017) was taken. Analysis of all datasets was anchored on the TSS locations of protein-coding transcripts based on high-resolution Start-seq data (see generation of transcript annotations above). A total of 8389 protein-coding TSSs, in which a decay rate could be calculated, was used. A rank order was given to the TSS feature list based on the decay rate clustering. Read coverage for each dataset used was determined at each TSS using a window that originates 500 nucleotides upstream of the TSS and extends downstream by twenty 50 nt non-overlapping bins, with total window size of 1000 nucleotides. Correlative analysis was then performed considering read coverage values. A total read coverage value was found for each genomic feature by adding the coverage from the datasets across all bins in a genomic window. Clustering methods were then applied to total read coverage values considering both the datasets and individual genomic features. To group datasets, the Pearson and Spearman correlation value for each pair of datasets was determined by comparing feature coverage values. To group the datasets, the correlation value for each pair of datasets is found by comparing feature coverage values. Datasets were then grouped by hierarchical clustering.

ATAC-SEQ LIBRARY GENERATION AND MAPPING

ATAC-seq libraries from 3 independent biological replicates were generated. 50,000 *Drosophila* S2 cells were incubated in CSK buffer (10 mM PIPES pH 6.8, 100 mM NaCl, 300 mM sucrose, 3 mM $MgCl_2$, 0.1% Triton X-100) on ice for 5 min. An aliquot of 2.5 µl of Tn5 Transposase was added to a total 25 µl reaction mixture and genomic DNA was purified using a Qiagen MinElute PCR purification kit (Qiagen) following manufacturer's instructions. After PCR amplification, DNA fragments were purified with AMPure XP (1:3 ratio of sample to beads). Libraries were sequenced using a paired-end 150 bp cycle run on an Illumina NextSeq 500

Paired-end reads were filtered for adapter sequence and low quality 3' ends using cutadapt 1.14, discarding those containing reads shorter than 20 nt (-m 20 -q 10), and removing a single nucleotide from the 3' end of all trimmed reads to allow successful alignment with bowtie 1.2.2 to the dm3 genome assembly. The parameters used in each alignment were: up to 2 mismatches, a maximum fragment length of 1000 nt, and uniquely mappable, and unmappable pairs routed to separate output files (-m1, -v2, -X1000, --un). Non-duplicate reads mapping uniquely to dm3, representative of short fragments (> 20 nt and < 150 nt), were separated, and fragment centers determined in 25 nucleotide windows resolution, genome-wide, and expressed in bedGraph format. Combined bedGraphs for all replicates were generated by summing counts per bin for all replicates.

RNA-SEQ LIBRARY GENERATION AND MAPPING

DL1 cells were treated for 60 h with a control (Beta-galactosidase) dsRNA or a dsRNA to deplete either IntS9 or IntS11 (see RNAi details above) followed by total RNA isolation with Trizol (Thermo Fisher Scientific 15596026) following manufacturer's instructions. RNA quality was confirmed with a BioAnalyzer (Agilent). Using Oligo d(T)25 Magnetic Beads (NEB S1419S), polyA+ RNA from 2.5 μg of total RNA was then enriched and RNA-seq libraries prepared using the ClickSeq library preparation method using a 1:35 azido-nucleotide ratio (Jaworski and Routh, 2018). Libraries were sequenced using a single-end 75 bp cycle run on an Illumina NextSeq 500.

Sequencing reads were filtered (requiring a mean quality score ≥20), trimmed to 50 nt, and then mapped to the dm3 reference genome using STAR 2.5.2b. Default parameters were used except that multimappers were reported randomly (outMultimapperOrder Random), spurious junctions were filtered (outFilterType BySJout), minimum overhang for non-annotated junctions was set to 8 nucleotides (alignSJoverhangMin 8), and non-canonical alignments were removed (outFilterIntronMotifsRemoveNoncanonical Unannotated).

MISO Analysis

Mixture of Isoform analysis (MISO) (Katz et al. 2010) was performed using the latest stable build (ver. 0.5.4) following the directions for an exon-centric analysis on the documents section of the developer's site (http://miso.readthedocs.io/en/fastmiso/). Differential expression was compared between the control (Beta-galactosidase) and IntS9-depleted RNA-seq BAM files for retained introns, skipped exons, alternative 5' splice sites, alternative 3' splice sites, and mutually excluded exons using the *Drosophila* annotations mentioned above. The results were then filtered using the developer suggested default settings to contain only events with: (a) at least 10 inclusion reads, (b) 10 exclusion reads, such that (c) the sum of inclusion and exclusion reads is at least 30, and (d) the $\Delta\Psi$ is at least 0.25 with a (e) Bayes factor of at least 20, and (a)-(e) are true in one of the samples. Using this filter, locations of alternative splicing events were compared to Flybase annotated chromosomal regions using the UCSC genome browser table browser to identify the FBgnIDs of affected genes.

All Flybase genes that included any splicing event that passed filter in MISO were removed from the list of active genes, such that a total of 9,499 active genes were investigated for the effects of IntS9 depletion.

Differentially expressed genes in RNA-seq

Read counts were calculated per gene, in a strand-specific manner, based on annotations described in the modified transcript annotations section above, using featureCounts (Liao et al. 2014). Differentially expressed genes were identified using DESeq2 v1.18.1(Anders and Huber, 2010) under R 3.3.1. For Control versus IntS9-depletion comparisons RNA-seq size factors were determined based on DESeq2 (Control [βgal]: 1.1861939, 1.4205182, 1.2440253; IntS9-dep.: 1.0780809, 0.9979663, 0.8519904), and at an adjusted p-value threshold of <0.0001 and fold-change > 1.5, 886 genes (out of 9499) were identified as differentially expressed upon IntS9 depletion in DL1 cells. For

Control versus IntS11-depletion or rescue samples comparisons RNA-seq size factors were determined based on DESeq2 (Control [βgal]: 1.3346867, 1.8951248, 0.6622473; IntS11-dep.: 0.8673446, 0.9127478, 0.9793937; IntS11-dep. + WT rescue: 1.1305191, 1.0792675, 0.7458915; IntS11-dep. + E203Q rescue: 1.1589313, 1.1588886, 0.7106579) and fold-changes calculated. For Control versus IntS11-depletion chromatin RNA-seq size factors were determined based on DESeq2 (Control: 1.1315534, 1.1665893; IntS11-dep.: 0.8940834, 0.8515502;) and at an adjusted p-value threshold of <0.0001 and fold-change > 1.5, 1283 genes (out of 17262) were identified as differentially expressed upon IntS11 depletion in HeLa cells. UCSC Genome Browser tracks displaying mean read coverage were generated from the combined replicates per condition, normalized as in the differential expression analysis.

SEQUENCING, MAPPING, AND DATA ANALYSIS OF CHIP-SEQ

For IntS1 and IntS12 ChIP-seq, DL1 cells were crosslinked for 30 min with 1% formaldehyde. Material was then sheared using the Covaris S220 system and immunoprecipitations for 3 (IntS1 and IntS12) independent biological replicates were carried out with 10 µl anti-IntS1 or anti-IntS12 antibodies per 3 x 10^7 cells. Additionally, 3 independent biological replicates of input material were carried through this process. Immunoprecipitated and input material was phenol-chloroform purified and ChIP-seq libraries were prepared using the NEBNext Ultra II DNA library kit (NEB) according to the manufacturer's instructions with 35ng of DNA of each sample. IntS1, IntS12 and input ChIP-seq libraries were then sequenced using a paired-end 75 bp cycle run on the Illumina NextSeq system with standard sequencing protocols. Raw sequences were aligned at full length against the dm3 version of the *Drosophila* genome using Bowtie version 1.2.2 (Langmead et al. 2009) with a maximum allowed mismatch of 2 (-m1 –v2). The yield of uniquely mappable reads for each set of biological replicates is listed below.

Datasets were mapped as described above against the dm3 version of the *Drosophila* genome. The genomic location of mapped reads was compiled using custom scripts and visually examined using the UCSC genome browser in bedGraph format. ChIP-seq hit locations were filtered based on fragment length. The 3 biological replicates of each ChIP-seq dataset were combined and binned in 25 bp windows for visualization in bedGraph files. IntS1 and IntS12 were down-sampled by a factor of 1.202985486 and 1.411913925, respectively to match the number of reads in the input dataset. To remove background signal, input signal was subtracted from IntS1 and IntS12 datasets and bedGraphs were generated with 25 bp windows for visualization.

INTS1 AND INTS12 CHIP-SEQ PEAK CALLING AND ANNOTATION

IntS1 and IntS12 ChIP-seq peaks were called with Homer (v4.9) using (-style factor) and input as background (-i). Filtering based on local signal was set to 3 (-L 3) and fold-change signal over input was also set to 3 (-F 3). 490 IntS1 and 553 IntS12 peaks were identified. A peak was assigned to enhancer TSSs (eTSSs) if the peak center would be within ± 500 bp from the eTSS. A total of 691 eTSSs were found to be bound by at least one Integrator subunit.

METAGENE ANALYSIS

Composite metagene distributions were generated by summing sequencing reads at each indicated position with respect to the TSS and dividing by the number of TSSs included within each group. These were plotted across a range of distances. Heatmaps were generated using Partek Genomics Suite version 6.15.0127.

IDENTIFICATION OF START-SEQ READS WITH NON-TEMPLATED 3' END RESIDUES

Start-seq from Rrp40-depleted S2 cells was published previously (Henriques et al. 2013) and is available for download from GEO (GSE49078). Data were analyzed as

described previously (Henriques et al. 2013). Briefly, Start-RNA reads were trimmed to 26 nt and aligned to the *D. melanogaster* reference genome index with Bowtie version 1.2.2, maintaining unique alignments and allowing 2 mismatches (-m1 -v2). To account for the different depths of sequencing across the data sets, all data sets were normalized by uniquely mappable reads. To then identify Start-RNAs with non-templated 3' end residues, reads that initially failed to align with the above Bowtie parameters were specifically trimmed at the 3' end to remove terminal A nucleotides. Reads trimmed of at least 3 A's with at least 18 nt remaining after trimming were aligned to the genome (note that reads with >26 nt remaining after trimming were further trimmed at the 5' end to 26mers) and counted as uniquely-aligned Start-RNAs. The percentage and location of Start-seq reads ending in 3 or more A residues (out of total Start-seq reads mapping to that gene) was calculated for each gene in all the groups.

PRO-SEQ LIBRARY PREPARATION AND DATA ANALYSIS

DL1 cells treated for 60 h with a control (Beta-galactosidase) dsRNA or a dsRNA targeting IntS9 were permeabilized as described below. All temperatures were at 4°C or ice cold unless otherwise specified. Cells were washed once in ice-cold 1x PBS and resuspended in Buffer W (10 mM Tris-HCl pH 8.0, 10% glycerol, 250 mM sucrose, 10 mM KCl, 5 mM $MgCl_2$, 0.5 mM DTT, protease inhibitors cocktail (Roche), and 4 u/mL RNase inhibitor [SUPERaseIN, Ambion]) at the cell density of 2×10^7 cells/mL. 9x volume of Buffer P (10 mM Tris-HCl pH 8.0, 10% glycerol, 250 mM sucrose, 10 mM KCl, 5 mM $MgCl_2$, 0.5 mM DTT, 0.1% Igepal, protease inhibitors cocktail (Roche), 4 u/mL RNase inhibitor [SUPERaseIN, Ambion]) was then immediately added. Cells were gently resuspended and incubated for up to 2 min on ice. Cells were then recovered by centrifugation (800 x *g* for 4 min) and washed in Buffer F (50 mM Tris-HCl pH 8.0, 40% glycerol, 5 mM $MgCl_2$, 0.5 mM DTT, 4 u/mL RNase inhibitor [SUPERaseIN, Ambion]). Washed permeabilized cells were finally resuspended in Buffer F at a density of 1×10^6

cells/30 µL and immediately frozen in liquid nitrogen. Permeabilized cells were stored in -80°C until usage.

PRO-seq run-on reactions were carried out as follows: 1×10^6 permeabilized cells spiked with 5×10^4 permeabilized mouse embryonic stem cells were added to the same volume of 2x Nuclear Run-On reaction mixture (10 mM Tris-HCl pH 8.0, 300 mM KCl, 1% Sarkosyl, 5 mM $MgCl_2$, 1 mM DTT, 200 µM biotin-11-A/C/G/UTP (Perkin-Elmer), 0.8 u/µL SUPERaseIN inhibitor [Ambion]) and incubated for 5 min at 30°C. Nascent RNA was extracted using a Total RNA Purification Kit following the manufacturer's instructions (Norgen Biotek Corp.). Extracted nascent RNA was fragmented by base hydrolysis in 0.25 N NaOH on ice for 10 min and neutralized by adding 1x volume of 1 M Tris-HCl pH 6.8. Fragmented nascent RNA was bound to 30 µL of Streptavidin M-280 magnetic beads (Thermo Fisher Scientific) in Binding Buffer (300 mM NaCl, 10 mM Tris-HCl pH 7.4, 0.1% Triton X-100). The beads were washed twice in High salt buffer (2 M NaCl, 50 mM Tris- HCl pH 7.4, 0.5% Triton X-100), twice in Binding buffer, and twice in Low salt buffer (5 mM Tris-HCl pH 7.4, 0.1% Triton X-100). Bound RNA was extracted from the beads using Trizol (Invitrogen) followed by ethanol precipitation.

For the first ligation reaction, fragmented nascent RNA was dissolved in H_2O and incubated with 10 pmol of reverse 3' RNA adaptor (5'p-rNrNrNrNrNrGrAr-UrCrGrUrCrGrGrArCrUrGrUrArGrArArCrUrCrUrGrArArC-/3'InvdT/) and T4 RNA ligase I (NEB) under manufacturer's conditions for 2 h at 20°C. Ligated RNA was enriched with biotin-labeled products by another round of Streptavidin bead binding and washing (two washes each of High, Binding and Low salt buffers and one wash of 1x Thermo Pol Buffer (NEB)). To decap 5' ends, the RNA products were treated with RNA 5' Pyrophosphohydrolase (RppH, NEB) at 37°C for 30 min followed by one wash of High, Low and T4 PNK Buffer. To repair 5' ends, the RNA products were treated with Polynucleotide Kinase (PNK, NEB) at 37°C for 30 min.

5' repaired RNA was ligated to reverse 5' RNA adaptor (5'-rCrCrUrUrGrG-rCrArCrCrCrGrArGrArArUrUrCrCrA-3') with T4 RNA ligase I (NEB) under manufacturer's conditions for 2 h at 20°C. Adaptor ligated nascent RNA was enriched with biotin-labeled products by another round of Streptavidin bead binding and washing (two washes each of High, Binding and Low salt buffers and one wash of 1x SuperScript IV Buffer [Thermo Fisher Scientific]), and reverse transcribed using 25 pmol RT primer (5'-AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGA-3') for TRU-seq barcodes (RP1 primer, Illumina). A portion of the RT product was removed and used for trial amplifications to determine the optimal number of PCR cycles. For the final amplification, 12.5 pmol of RPI-index primers (for TRU-seq barcodes, Illumina) was added to the RT product with Phusion polymerase (NEB) under standard PCR conditions. Excess RT primer served as one primer of the pair used for the PCR. The product was amplified 12~14 cycles and beads size selected (ProNex Purification System, Promega) before being sequenced in NextSeq 500 machines in a mid-output 150 bp cycle run.

PRO-seq libraries from 3 independent biological replicates (DL1 control (βgal) RNAi or IntS9 RNAi) were generated. Paired-end reads were trimmed to 42 nt, for adapter sequence and low quality 3' ends using cutadapt 1.14, discarding those containing reads shorter than 20 nt (-m 20 -q 10), and removing a single nucleotide from the 3' end of all trimmed reads to allow successful alignment with Bowtie 1.2.2. Remaining pairs were paired-end aligned to the mm10 genome index to determine spike-normalization ratios based on uniquely mapped reads. Mappable pairs were excluded from further analysis, and unmapped pairs were aligned to the dm3 genome assembly. Identical parameters were utilized in each alignment described above: up to 2 mismatches, maximum fragment length of 1000 nt, and uniquely mappable, and unmappable pairs routed to separate output files (-m1, -v2, -X1000, --un). Pairs mapping uniquely to dm3, representing biotin-labeled RNA 3' ends, were separated, and strand-specific counts of the 3' mapping positions determined at single nucleotide resolution, genome-wide, and expressed in bedGraph format with

"plus" and "minus" strand labels swapped for each 3' bedGraph, to correct for the "forward/reverse" nature of Illumina paired-end sequencing. Counts of pairs mapping uniquely to spike-in RNAs (mouse genome) were determined for each sample. Uniquely mappable reads were determined, and a normalization factor calculated. In this case, the samples displayed highly comparable recovery of spike-in reads, thus only normalization based on the DESeq2 size factors was used for each bedGraph. Combined bedGraphs were generated by summing counts per nucleotide of both replicates for each condition.

Read counts were calculated per gene, in a strand-specific manner, based on annotations described in the modified transcript annotations section above, using featureCounts (Liao et al. 2014). This quantification procedure includes signal only in the gene body (+250 from TSS to annotated gene end). Differentially expressed genes were identified using DESeq2 v1.18.1 (Anders and Huber 2010) under R 3.3.1. PRO-seq size factors were determined based on DESeq2 (for Control: 1.0029079, 1.2830936, 0.8962051; IntS9-dep.: 0.9151691, 0.9156818, 1.0672821). At an adjusted p-value threshold of <0.0001 and fold-change >1.5, 1,414 mRNA genes were identified as differentially expressed upon IntS9-depletion in DL1 cells. UCSC Genome Browser tracks displaying mean read coverage were generated from the combined replicates per condition, normalized as in the differential expression analysis.

GENOMIC STATISTICAL TESTS

For RNA-seq, PRO-seq, and ChIP-seq experiments, statistical significance for comparisons was assessed by Mann-Whitney (pairwise tests) test. Statistical details and error bars are defined in each figure legend. To test for the significant overlap between IntS9-upregulated or IntS9-downregulated genes in RNA-seq and PRO-seq, a hypergeometric test was used from a total of 9499 active mRNA genes.

GENE ONTOLOGY ANALYSIS

Gene Ontology analysis was performed using DAVID (v6.8) online tool with standard parameters (https://david.ncifcrf.gov/home.jsp).

PROTEIN EXPRESSION AND PURIFICATION.

Drosophila IntS11 and CG7044 were cloned into the same pFL vector and co-expressed in insect cells using Multibac technology (Geneva Biotech) (Sari et al. 2016). A 6xHis tag was added to the N terminus of IntS11. Bacmids expressing IntS11 and CG7044 were generated in DH10EMBacY competent cells (Geneva Biotech) by transformation. High5 cells were grown in ESF 921 medium (Expression Systems) by shaking at 120 rpm at 27 °C until the density reached 2×10^6 cells·ml^{-1}. Cells were infected with 16 ml IntS11-CG7044 P2 virus and harvested after 48 h.

For purification, the cell pellet was resuspended and lysed by sonication in 100 ml of buffer containing 20 mM Tris (pH 8.0), 250 mM NaCl, 2 mM β-mercaptoethanol (βME), 5% (v/v) glycerol, and one tablet of protease inhibitor mixture (Sigma). The cell lysate was then centrifuged at 13,000 rpm for 40 min at 4 °C. The protein complex was purified from the supernatant via nickel affinity chromatography. The protein complex was further purified by a HiTrap Q column (GE Healthcare) and a Hiload 16/60 Superdex 200 column (GE Healthcare). The IntS11-CG7044 complex was concentrated to 1 mg·ml^{-1} in a buffer containing 20 mM Tris (pH 8.0), 100 mM NaCl, and 2 mM dithiothreitol (DTT), and stored at –80 °C. The protein concentration of the freshly purified samples was measured with a NanoDrop spectrophotometer (Thermo Fisher Scientific).

EM SPECIMEN PREPARATION AND DATA COLLECTION.

All specimens for cryo-EM were frozen with an EM GP2 plunge freezer (Leica) set at 20 °C and 99% humidity. Cryo-EM imaging was performed in the Simons Electron

Microscopy Center at the New York Structural Biology Center using Leginon (Suloway et al. 2005).

For the IntS11-CG7044 complex, a 3.5 μL aliquot at 0.18 $mg \cdot ml^{-1}$ was applied to one side of a Quantifoil 400 mesh 1.2/1.3 gold grid with graphene oxide support film (Quantifoil). After 30 s, the grid was blotted for 1.5 s on the other side and plunged into liquid ethane. 1,603 image stacks were collected on a Titan Krios electron microscope at New York Structural Biology Center, equipped with a K3 direct electron detector (Gatan) at 300 kV with a total dose of 51 e^{-} $Å^{-2}$ subdivided into 40 frames in 2 s exposure using Leginon. The images were recorded at a nominal magnification of 81,000× and a calibrated pixel size of 1.083 Å, with a defocus range from −1 to −2.5 μm.

IMAGE PROCESSING.

For both cryo-EM datasets, image stacks were motion-corrected and dose-weighted using RELION 3.1 (Zivanov et al. 2018). For the IntS11-CG7044 dataset, the CTF parameters were determined with CTFFIND4 (Rohou and Grigorieff 2015) in cryoSPARC (Punjani et al. 2017). 2,028,174 particles were auto-picked and subjected to 2D classification and *ab initio* reconstruction in cryoSPARC to generate eight initial 3D models. These models were then used in heterogeneous refinement against all the particles in cryoSPARC, and the good particles were then used in another round of heterogeneous refinement. 406,222 particles were then selected and imported to RELION for CTF refinement and Bayesian polishing. The polished particles were then imported back to cryoSPARC for homogeneous refinement, yielding a final map at 3.54 Å resolution.

MODEL BUILDING.

Atomic models for CG7044, IntS11, IntS9 and IntS4 were built manually into the cryo-EM density with Coot (Emsley and Cowtan 2004). Homology models for *Drosophila* IntS9 and IntS11 were generated with I-TASSER (Roy et al. 2010), based on the structures

of human CPSF100 and CPSF73 (Zhang et al. 2020). The atomic models were improved by real-space refinement with the program PHENIX (Liebschner et al. 2019).

PLASMID CONSTRUCTION AND STABLE CELL LINES GENERATION.

For mutation analysis of *Drosophila* IntS9, IntS11, and CG7044, truncation primers were used to clone CG7044-Δ947-974 and CG7044-Δ966-974. Wild-types and the mutants of IntS11 and CG7044 were subsequently cloned into the pMT-3xFLAG-puro vector (Elrod et al. 2019) to inducibly express in DL1 cells. All plasmids were sequenced to confirm identity. To generate cells stably expressing the FLAG-IntS11-WT, FLAG-IntS11-K462E, FLAG-CG7044-WT, FLAG-CG7044-Δ947-974, FLAG-CG7044-Δ966-974, and eGFP control transgenes, $2x10^6$ cells were first plated in regular maintenance media in a 6-well dish overnight. 2 μg of expressing plasmids were transfected using Fugene HD (Promega, #E2311). After 24 hours, 2.5 μg/mL puromycin was added to the media to select and maintain the cell population. Primers are Table 2.1.

NUCLEAR EXTRACT PREPARATION.

Five 150 mm dishes of each condition of confluent cells (pretreated with 500 μM $CuSO_4$ for 24 hours) were collected and washed in cold PBS before being resuspended in ten times volumes of the cell pellet of Buffer A (10 mM Tris pH 8, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, and 0.2 mM PMSF). Resuspended cells were allowed to swell during a 15-minute rotation at 4 °C. After pelleting down at 1,000 g for 10 minutes, two times volumes of the original cell pellet of Buffer A were added and cells were homogenized with a dounce pestle B for 40 strokes on ice. Nuclear and cytosolic fractions were then separated by centrifuging at 800 g for 10 minutes. To attain a nuclear fraction, the pellet was washed once with Buffer A before being resuspended in two times volumes of the original cell pellet of Buffer C (20 mM Tris pH 8, 420 mM NaCl, 1.5 mM $MgCl_2$, 25% (v/v) glycerol, 0.2 mM EDTA, 0.5 mM PMSF, and 0.5 mM DTT). The samples were then

homogenized with a dounce pestle B for 20 strokes on ice and rotated for 30 minutes at 4 °C before centrifuging at 15,000 g for 30 minutes at 4 °C. Finally, supernatants were collected and subjected to dialysis in Buffer D (20 mM HEPES, 100 mM KCl, 0.2 mM EDTA, 0.5 mM DTT, and 20% (v/v) glycerol) overnight at 4 °C against 3.5 kDa MWCO membrane (Spectrum Laboratories, #132720). Prior to any downstream applications, nuclear extracts were centrifuged again at 15,000 g for 3 minutes at 4 °C to remove any precipitate.

WESTERN BLOTTING AND ANTI-FLAG AFFINITY PURIFICATION.

To check protein expression, cells were lysed directly in wells in 2X SDS sample buffer (120 mM Tris pH 6.8, 4% SDS, 200 mM DTT, 20% (v/v) glycerol, and 0.02% bromophenol blue). Lysates were incubated at room temperature with periodic swirling prior to a 10-minute boiling at 95 °C and a short sonication. Denatured protein samples were then resolved in a 10% SDS-PAGE and transferred to a PVDF membrane (Bio-Rad, #1620177). Blots were probed by custom-designed *Drosophila* antibodies as previously described (Elrod et al. 2019) diluted in PBS-0.1% Tween supplemented with 5% nonfat milk. To detect proteins from 293T lysate, anti-hIntS11 (Bethyl, #A301-274A), anti-hIMPK (Thermo, #PA5-21629), anti-GFP (Clontech, #632381), anti-alpha Tubulin (abcam, #ab15246), and anti-GAPDH (Thermo, #MA5-15738) were used at the dilution suggested by the manufacturer.

To purify FLAG-tagged Integrator complexes, 1 mg of nuclear extract was mixed with 40 μL anti-Flag M2 affinity agarose slurry (Sigma, #A2220) equilibrated in binding buffer (20 mM HEPES pH 7.4, 100 mM KCl, 10% (v/v) glycerol, 0.1% NP-40) and rotated for 2 hours at 4 °C. Following the two-hour incubation/rotation, five sequential washes were carried out in binding buffer with a 10-minute rotation at 4 °C followed by a 500 g centrifugation at 4 °C. After the final wash, the binding buffer supernatant was removed using a pipette and the protein complexes were eluted from the anti-FLAG resin by adding

40 μL of 2X sample buffer and boiled at 95 °C for five minutes. For input samples, nuclear extracts were mixed with 5X loading buffer and boiled, and 1/10 volume of the immunoprecipitation reaction was loaded on SDS-PAGE.

MASS SPECTROMETRY SAMPLE DIGESTION.

The samples were prepared similar to as described (Andersson et al. 2015). Briefly, the agarose bead-bound proteins were washed several times with 50mM Triethylammonium bicarbonate (TEAB) pH 7.1, before being solubilized with 40 mL of 5% SDS, 50mM TEAB, pH 7.55 followed by a room temperature incubation for 30 minutes. The supernatant containing the proteins of interest was then transferred to a new tube, reduced by making the solution 10mM Tris(2-carboxyethyl)phosphine (TCEP) (Thermo, #77720), and further incubated at 65 °C for 10 minutes. The sample was then cooled to room temperature and 3.75 mL of 1M iodoacetamide acid was added and allowed to react for 20 minutes in the dark after which 0.5 mL of 2M DTT was added to quench the reaction. Then, 5 mL of 12% phosphoric acid was then added to the 50 mL protein solution followed by 350 mL of binding buffer (90% Methanol, 100mM TEAB final; pH 7.1). The resulting solution was administered to an S-Trap spin column (Protifi, Farmingdale NY) and passed through the column using a bench top centrifuge (30 s spin at 4,000 g). The spin column was then washed three times with 400 mLof binding buffer and centrifuged (1200rpm, 1min). Trypsin (Promega, #V5280, Madison, WI) was then added to the protein mixture in a ratio of 1:25 in 50mM TEAB, pH = 8, and incubated at 37 °C for 4 hours. Peptides were eluted with 80uL of 50mM TEAB, followed by 80 mL of 0.2% formic acid, and finally 80 mL of 50% acetonitrile, 0.2% formic acid. The combined peptide solution was then dried in a speed vacuum (room temperature, 1.5 hours) and resuspended in 2% acetonitrile, 0.1% formic acid, 97.9% water and aliquoted into an autosampler vial.

NANOLC MS/MS ANALYSIS.

Peptide mixtures were analyzed by nanoflow liquid chromatography-tandem mass spectrometry (nanoLC-MS/MS) using a nano-LC chromatography system (UltiMate 3000 RSLCnano, Dionex, Thermo Fisher Scientific, San Jose, CA). The nanoLC-MS/MS system was coupled on-line to a Thermo Orbitrap Fusion mass spectrometer (Thermo Fisher Scientific, San Jose, CA) through a nanospray ion source (Thermo Scientific). A trap and elute method was used to desalt and concentrate the sample, while preserving the analytical column. The trap column (Thermo Scientific) was a C18 PepMap100 (300um X 5mm, 5um particle size) while the analytical column was an Acclaim PepMap 100 (75 mm X 25 cm) (Thermo Scientific). After equilibrating the column in 98% solvent A (0.1% formic acid in water) and 2% solvent B (0.1% formic acid in acetonitrile (ACN)), the samples (2 mL in solvent A) were injected onto the trap column and subsequently eluted (400 nL/min) by gradient elution onto the C18 column as follows: isocratic at 2% B, 0-5 min; 2% to 32% B, 5-39 min; 32% to 70% B, 39-49 min; 70% to 90% B, 49-50 min; isocratic at 90% B, 50-54 min; 90% to 2%, 54-55 min; and isocratic at 2% B, until the 65 minute mark.

All LC-MS/MS data were acquired using XCalibur, version 2.1.0 (Thermo Fisher Scientific) in positive ion mode using a top speed data-dependent acquisition (DDA) method with a 3 s cycle time. The survey scans (m/z 350-1500) were acquired in the Orbitrap at 120,000 resolution (at m/z = 400) in profile mode, with a maximum injection time of 100 msec and an AGC target of 400,000 ions. The S-lens RF level was set to 60. Isolation was performed in the quadrupole with a 1.6 Da isolation window, and CID MS/MS acquisition was performed in profile mode using rapid scan rate with detection in the ion-trap using the following settings: parent threshold = 5,000; collision energy = 32%; maximum injection time 56 msec; AGC target 500,000 ions. Monoisotopic precursor selection (MIPS) and charge state filtering were on, with charge states 2-6 included.

Dynamic exclusion was used to remove selected precursor ions, with a ± 10 ppm mass tolerance, for 15 s after acquisition of one MS/MS spectrum.

Database Searching. Tandem mass spectra were extracted and charge state deconvoluted using Proteome Discoverer (Thermo Fisher, version 2.2.0388). Deisotoping was not performed. All MS/MS spectra were searched against a Uniprot Drosophila database (version 04-04-2018) using Sequest. Searches were performed with a parent ion tolerance of 5 ppm and a fragment ion tolerance of 0.60 Da. Trypsin was specified as the enzyme, allowing for two missed cleavages. Fixed modification of carbamidomethyl (C) and variable modifications of oxidation (M) and deamidation were specified in Sequest.

Chapter 3. Integrator's Endonuclease Activity Regulates Transcriptional Activation of a Quick Response Gene

This chapter contains text and figures reprinted with permission from Tatomer DC*, Elrod ND*, Liang D, Jonathan M, Wagner EJ, Cherry S, Wilusz JE: The Integrator complex cleaves nascent mRNAs to attenuate transcriptional attenuation of Methallothionein. Genes and Development Sep;33/21-21/1525 doi:10.1101/gad.330167.119 2019 (* indicates co-first author).

INTRODUCTION

In response to physiological cues, environmental stress, or exposure to pathogens, specific transcriptional programs are induced. These responses are often coordinated, rapid, and robust, in part because many metazoan genes are maintained in a poised state with RNA polymerase II (RNAPII) engaged prior to induction. In addition to promoter-proximal pausing, there are many regulatory steps post transcription initiation that dictate the characteristics and fate of mature transcripts. For example, alternative splicing and/or 3′ end processing events can lead to the production of multiple isoforms from a single locus, and these transcripts can have distinct stabilities, translation potential, or sub-cellular localization.

It is particularly important that genes produce full-length functional mRNAs and mechanisms such as telescripting, involving U1 snRNP, actively suppress premature cleavage and polyadenylation events in eukaryotic cells (Kaida et al. 2010; Berg et al. 2012; Venters et al. 2019). Nevertheless, many promoters are known to generate short unstable RNAs (Kapranov et al. 2007; Xu et al. 2009; Porrua and Libri 2015). This suggests that pre-mature transcription termination may often occur, thereby limiting RNAPII elongation and production of full-length mRNAs. Moreover, this process can be regulated (Brannan et al. 2012; Wagschal et al. 2012; Chalamcharla et al. 2015; Chiu et al. 2018).

For example, it was recently shown that the cleavage and polyadenylation factor PCF11 stimulates premature termination to attenuate the expression of many transcriptional regulators in human cells (Kamieniarz-Gdula et al. 2019). Potentially deleterious truncated transcripts generated by premature termination are often removed from cells by RNA surveillance mechanisms, including those mediated by the RNA exosome. However, the full repertoire of cellular factors and cofactors that control the metabolic fate of nascent RNAs, especially during the early stages of transcription elongation, is still unknown.

We thus performed an unbiased genome-scale RNAi screen in Drosophila cells to reveal factors that control the output of a model inducible eukaryotic promoter. Transcription of Drosophila Metallothionein A (MtnA), which encodes a metal chelator, is rapidly induced when the intracellular concentration of heavy metals (e.g., copper or cadmium) is increased (Figure 3.1A). This increase in transcriptional output is dependent on the MTF-1 transcription factor, which re-localizes to the nucleus upon metal stress and binds to the MtnA promoter (Smirnova et al. 2000). Our RNAi screen identified MTF-1 and other known regulators of MtnA transcription (Marr et al. 2006), but also surprisingly identified the Integrator complex as a potent inhibitor of MtnA during copper stress. Integrator harbors an endonuclease that cleaves snRNAs and enhancer RNAs (Baillat et al. 2005; Lai et al. 2015), and we find that Integrator can likewise cleave nascent MtnA transcripts to limit mRNA production. Using RNA-seq, we find hundreds of additional Drosophila protein-coding genes whose expression increases upon Integrator depletion. Focused studies on a subset of these genes confirmed that Integrator can cleave these nascent RNAs, thereby limiting productive transcription elongation. Altogether, we propose that Integrator-catalyzed premature termination can function as a widespread and potent mechanism to attenuate expression of protein-coding genes.

RESULTS

Genome scale RNAi screening reveals the Integrator complex as a potent inhibitor of the MtnA promoter during copper stress

To identify regulators of an archetype inducible transcription program, the Drosophila MtnA promoter was cloned upstream of an intronless, non-polyadenylated eGFP reporter (Figure 3.1B). Maturation of this mRNA is independent of many canonical mRNA processing events, and we thus reasoned that high throughput RNAi screening using this reporter should primarily identify transcriptional and translational regulators. Drosophila DL1 cells stably maintaining the reporter were treated with double stranded RNAs (dsRNAs) for 3 d and copper was added for the final 6 h to activate the MtnA promoter and eGFP expression (Figure 3.1B). Automated microscopy and image analysis was then used to quantify eGFP fluorescence. A total of 232 factors were required for eGFP expression during copper stress, including ribosomal subunits and well characterized transcriptional regulators such as RNAPII, Mediator subunits (e.g., MED9 and MED15), and the MTF-1 transcription factor (Figures 3.1C,D (Marr et al. 2006; Gunther et al. 2012). Unexpectedly, nearly all 14 subunits of the Integrator (Int) complex (for review, see (Baillat and Wagner 2015)) were among the most potent negative regulators identified (Figures 3.1C,D). Of the >10,000 genes screened, depletion of the IntS8 subunit resulted in the largest increase in eGFP expression.

RT-qPCR and western blotting confirmed that the dsRNAs resulted in depletion of the Integrator subunits and increases in eGFP reporter expression at both the protein and mRNA levels. These increases in expression upon Integrator depletion were observed regardless of the ORF downstream from the MtnA promoter, as we also observed an increase when eGFP was replaced with nano-Luciferase. Likewise, the increases in expression were not dependent on the mRNA 3′ end processing signals downstream from the ORF, as similar increases were observed when a canonical polyadenylation signal was present.

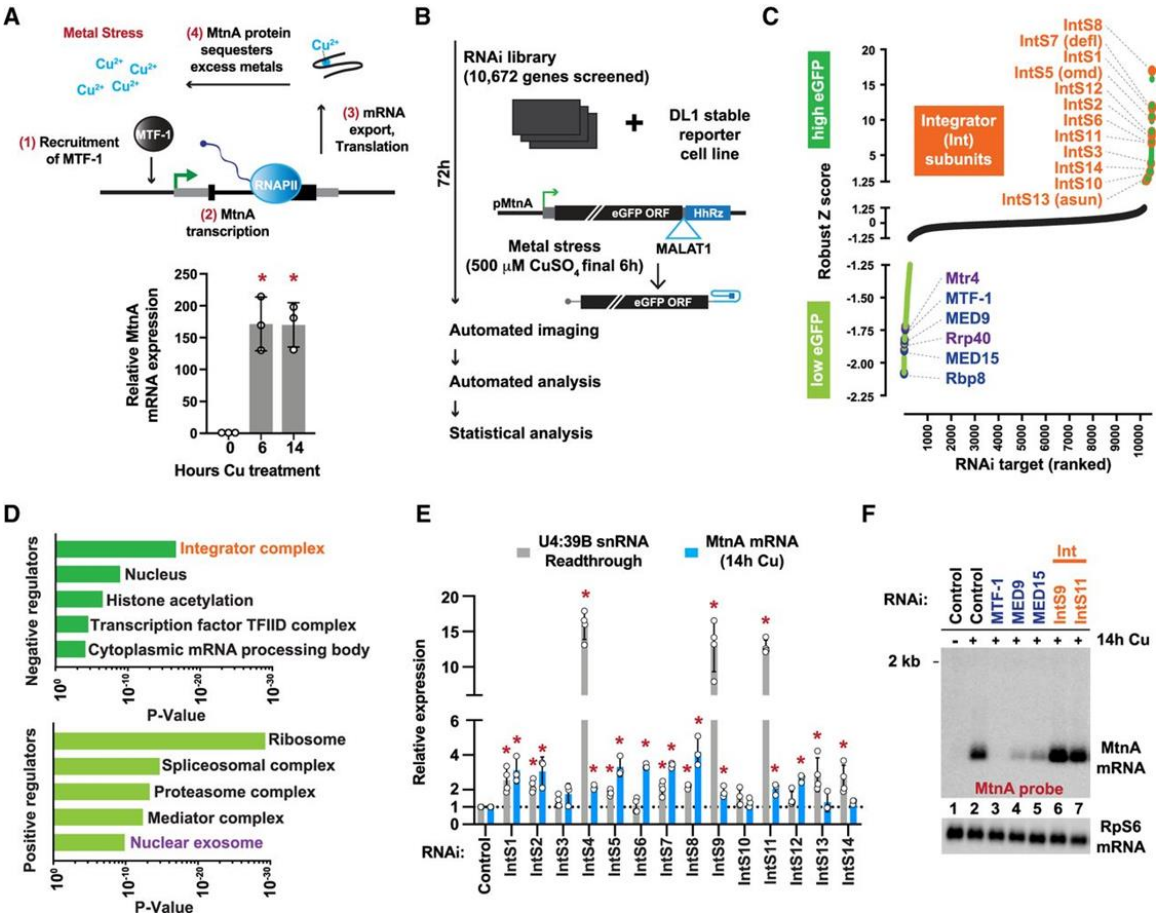


Figure 3.1: The Integrator complex inhibits expression from the MtnA promoter during copper stress

(A, top) Upon metal stress, the tran-scription factor MTF-1 binds and induces transcription from the MtnA promoter, resulting in production of a protein that sequesters the excess metals to alleviate the stress. (Bottom) Drosophila DL1 cells were treated with 500 μM copper sulfate (CuSO4) for the indicated times, and RT-qPCR was used to measure endogenous MtnA mRNA expression. Data from three independent experiments were normal-ized to RpL32 mRNA expression and are shown as mean ± SD, (∗) $P < 0.05$.
(B) RNAi screen pipeline using DL1 cells stably maintaining an eGFP reporter driven by the MtnA promoter. The self-cleaving hammerhead ribozyme (HhRz) (Dower et al. 2004) generates the eGFP mRNA 3′.
end, which is then stabilized by the MALAT1 triple helix structure (Wilusz et al. 2012).
(C) Robust Z-scores of eGFP integrated intensity are shown. RNAi treatments that resulted in increased (Z-score >1.3, dark green) or decreased (Z-score < −1.3, light green) eGFP expression are marked, including Integrator subunits (orange), transcription regulators (blue), and RNA exosome components (purple).
(D) Gene ontology (GO) analysis was performed to identify categories of genes that are enriched among the negative (Z-score >1.3) and positive (Z-score < −1.3) regulators of the eGFP reporter.

(E) DL1 cells were treated with dsRNAs for 3 d to induce RNAi and depletion of the indicated factors. Expression of endogenous MtnA mRNA (after 14 h $CuSO_4$ treatment) was quantified by RT-qPCR, and read-through transcription downstream from the U4:39B snRNA was quantified by northern blotting. Data are shown as mean ± SD, $N \geq 3$.(∗) $P < 0.05$.
(F) Representative northern blot of endogenous MtnA mRNA isolated from DL1 cells treated with dsRNA to induce RNAi of the indicated factor.
(Figure and legend text reprinted with permission from Tatomer DC, Elrod ND, Liang D, Jonathan M, Wagner EJ, Cherry S, Wilusz JE: The Integrator complex cleaves nascent mRNAs to attenuate transcriptional attenuation of Methallothionein. Genes and Development Sep;33/21-21/1525 doi:10.1101/gad.330167.119 2019.)

The Integrator complex is present at the endogenous MtnA locus during copper stress and represses MtnA pre-mRNA levels

The Integrator complex has been implicated in a myriad of diseases, interacts with RNAPII, and contains the IntS11 RNA endonuclease that generates the 3′ ends of spliceosomal snRNAs (Baillat et al. 2005; Baillat and Wagner 2015). We confirmed that depleting subunits of the Integrator cleavage module (IntS4, IntS9, or IntS11) resulted in increased snRNA readthrough transcription (Figure 3.1E (Ezzeddine et al. 2011; Albrecht et al. 2018). Nevertheless, because mature snRNAs have long half-lives (Fury and Zieve 1996), their levels only marginally decreased over the 72 h time course of the experiment.

Integrator has also been implicated in transcription regulation at enhancers and at a subset of EGF-responsive mRNAs (Gardini et al. 2014; Lai et al. 2015), so we hypothesized that the Integrator complex could be directly acting at the MtnA promoter. Indeed, depletion of Integrator subunits resulted in increased expression of the endogenous MtnA mRNA (Figures 3.1E,F) and pre-mRNA (Figure 3.2A) to a similar extent during copper stress. Transcription of the MtnA locus is also induced by cadmium stress (Figure 3.2B), but we found that depletion of Integrator subunits strikingly had no effect on MtnA pre-mRNA levels under these conditions (Figure 3.2C). This indicates that the Integrator complex can regulate the output of a protein-coding gene in a context specific manner.

To explore the underlying basis for this distinct regulation of MtnA expression, we used chromatin immunoprecipitation (ChIP)-qPCR to examine the recruitment of Integrator subunits to the MtnA locus upon copper or cadmium stress. We found that IntS1 and IntS12 were recruited to the endogenous MtnA locus upon copper stress (especially to the 5′ end), but their recruitment was significantly less robust during cadmium stress (Figure 3.2D). These results suggest that Integrator-mediated repression of MtnA during copper stress is due to Integrator recruitment to the chromatin eliciting direct effects on transcription.

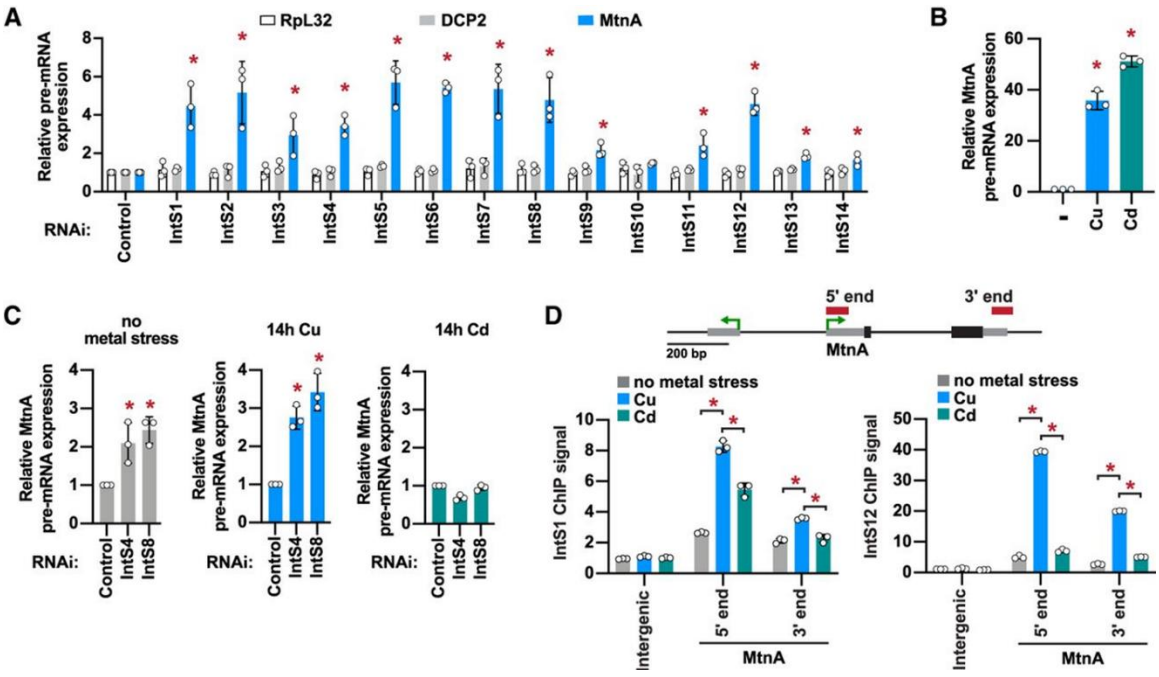


Figure 3.2: The Integrator complex is present at the MtnA locus during copper stress and attenuates MtnA transcription

(A) DL1 cells were treated with dsRNAs for 3 d to induce RNAi and depletion of the indicated factors. A total of 500 µM CuSO4 was added for the last 14 h. RT-qPCR was then used to measure the pre-mRNA levels of endogenous RpL32, DCP2, and MtnA. Primer pairs that span an intron–exon boundary were used to specifically amplify pre-mRNAs. Data were normalized to RpL32 mRNA expression and are shown as mean ± SD, N =3. (∗) P < 0.05.
(B) DL1 cells were unstressed (–) or treated with 500 µM CuSO4 or 50 µM CdCl2 for 14 h, and RT-qPCR was then used to measure MtnA pre-mRNA levels. Data were normalized to RpL32 mRNA expression and are shown as mean ± SD, N =3.(∗) P < 0.05.
(C) DL1 cells were treated with dsRNAs for 3 d and 500 µM CuSO4 or 50 µM CdCl2 was added for the final 14 h, as indicated. RT-qPCR was then used to measure MtnA pre-mRNA levels. Data were normalized to RpL32 mRNA expression and are shown as mean ± SD, N =3. (∗) P < 0.05.
(D) The MtnA locus with the locations of ChIP amplicons. Recruitment of IntS1 and IntS12 in unstressed cells (gray) or after the cells had been treated with CuSO4 (blue) or CdCl2 (green) for 14 h was measured using ChIP-qPCR. Data are shown as fold change relative to the IgG control (mean ± SD, N = 3). (∗) P < 0.05.
(Figure and legend text reprinted with permission from Tatomer DC, Elrod ND, Liang D, Jonathan M, Wagner EJ, Cherry S, Wilusz JE: The Integrator complex cleaves nascent mRNAs to attenuate transcriptional attenuation of Methallothionein. Genes and Development Sep;33/21-21/1525 doi:10.1101/gad.330167.119 2019.)

The IntS11 endonuclease activity is required for Integrator dependent regulation of MtnA expression

To define how Integrator functions at MtnA during copper stress, we first addressed whether the RNA endonuclease activity of IntS11 is required. The endogenous IntS11 protein was depleted using a dsRNA targeting either the IntS11 ORF or 3′ untranslated region (UTR) (Figure 3.3A), and this resulted in increased expression of endogenous MtnA mRNA relative to treatment with a control dsRNA (Figure 3.3B, lanes 1–3). Expression of a wild-type (WT) IntS11 transgene (Figure 3.3A) in cells treated with the IntS11 3′ UTR dsRNA restored MtnA expression to levels similar to control treated cells (Figure 3.3B, lane 4 vs.6), whereas expression of a catalytically dead (E203Q) (Baillat et al. 2005) IntS11 transgene did not (Figure 3.3B, lane 7 vs. 9). As a control, we confirmed that increased MtnA expression was observed when cells were treated with a dsRNA targeting the IntS11 ORF (Figure 3.3B, lanes 5, 8), which depletes both endogenous and exogenously expressed IntS11 (Figure 3.3A). This requirement for IntS11 endonuclease activity was also observed with the eGFP reporter driven by the MtnA promoter. Given that the E203Q mutation does not disrupt Integrator complex integrity (Baillat et al. 2005), these collective results indicate that the IntS11 endonuclease activity is required for Integrator to negatively regulate the transcriptional output of the MtnA promoter.

The Integrator complex cleaves nascent MtnA mRNAs to trigger transcription termination

By immunoprecipitating Integrator subunits followed by immunoblotting (Figure 3.3C) or mass spectrometry (Figure 3.3D), we found that Integrator interacts with the nuclear RNA exosome, which catalyzes 3′–5′ degradation of many RNAs (for review, see (Zinder and Lima 2017)). Interestingly, our genome-scale RNAi screen identified many RNA exosome core components and cofactors (including Rrp40 and Mtr4) as positive regulators of the MtnA promoter during copper stress (Figures 3.1C,D). This suggests an

interplay between the RNA exosome and Integrator complex and that the RNA exosome may also function in controlling MtnA transcriptional output. We confirmed that RNA exosome core components and cofactors were efficiently depleted by RNAi, and that this resulted in the expected ribosomal RNA processing defects as well as decreased output from the MtnA promoter during copper stress (Figure 3.3E). We thus hypothesized that Integrator may cleave nascent MtnA transcripts to prematurely terminate RNAPII transcription (as Integrator is enriched near the 5′ end of the MtnA genomic locus during copper stress) (Figure 3.2D) and that these cleaved transcripts are targeted for rapid degradation by the RNA exosome (Figure 3.3F).

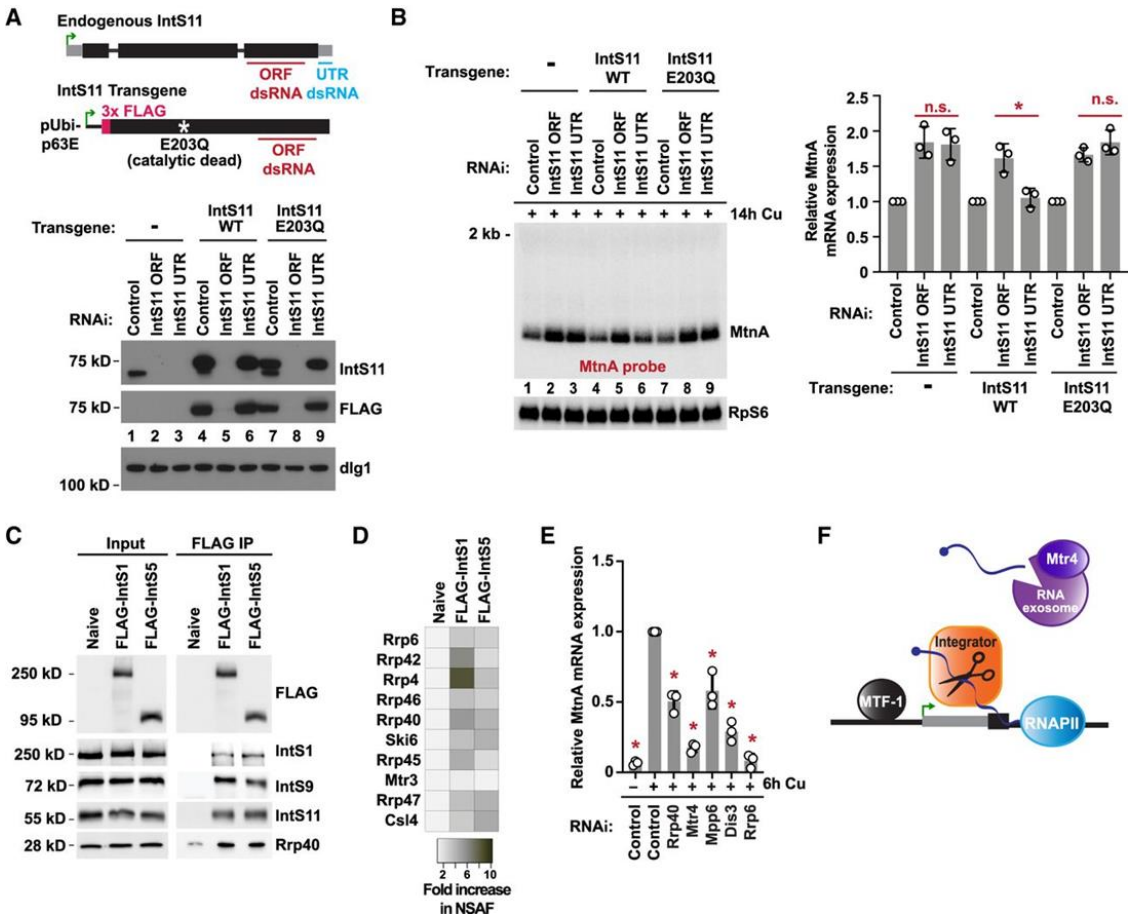


Figure 3.3: The IntS11 endonuclease activity and the RNA exosome regulate MtnA transcript levels

(A, top) Schematic of IntS11 knockdown/plasmid rescue strategy. The ORF dsRNA (red) depletes IntS11 generated from both the endogenous locus and transgenes that are driven by the ubiquitin-63E promoter (pUbi-p63E), while the UTR dsRNA (blue) only depletes endogenous IntS11. (Bottom) DL1 cells (lanes 1–3) or DL1 cells stably maintaining Flag-tagged wild-type (WT; lanes 4–6) or catalytically dead (E203Q; lanes 7–9) IntS11 transgenes were treated with the indicated dsRNAs for 3 d. Western blot analysis was then used to examine IntS11 protein levels. dlg1 was used as a loading control. Representative blots are shown.
(B) Representative northern blot of the endogenous MtnA mRNA in DL1 cells stably maintaining IntS11 transgenes that had been treated with the indicated dsRNAs and CuSO4. Expression was quantified using ImageQuant and data are shown as mean ± SD, N = 3. (*) $P < 0.05$.
(C) Western blot analysis of S2 nuclear extracts that were subjected to purification using Flag affinity resin. (Left panel) Western blots showing input levels of the denoted proteins in nuclear extracts derived from naïve S2 cells, S2 cells stably expressing Flag-IntS1, or S2 cells stably expressing Flag-IntS5. (Right panel) same as left panel except that Flag immunoprecipitates were probed.

(D) Heat map showing the relative enrichment of RNA exosome components within Flag-Integrator subunit purifications as determined by mass spectrometry. Normalized spectral abundance factor (NSAF) values were quantified as previously described (Zybailov et al. 2006), and fold-enrichment of each RNA exosome subunit in Integrator purifications relative to purification from naïve extract is given.
(E) DL1 cells were treated with dsRNAs for 3 d to induce RNAi and depletion of the indicated factors. CuSO4 was added for the last 6 h. Northern blots were then used to quantify expression of endogenous MtnA mRNA and data are shown as mean ± SD, N = 3. (*) $P < 0.05$.
(F) Model for Integrator-dependent premature termination. After cleavage of the nascent MtnA transcript by IntS11, the small RNAs are degraded by the RNA exosome.
(Figure and legend text reprinted with permission from Tatomer DC, Elrod ND, Liang D, Jonathan M, Wagner EJ, Cherry S, Wilusz JE: The Integrator complex cleaves nascent mRNAs to attenuate transcriptional attenuation of Methallothionein. Genes and Development Sep;33/21-21/1525 doi:10.1101/gad.330167.119 2019.)

To test this model, we first treated cells with dsRNAs to deplete the exosome-associated RNA helicase Mtr4 (Lubas et al. 2011). This resulted in a reduction in full length endogenous MtnA mRNA expression (Figure 3.3E) and, concomitantly, a number of small RNAs were detected from the MtnA locus, including prominent transcripts with lengths of ~85 and ~110 nt (RNAs marked in orange) (Figure 3.4A, lane 4). These small RNAs were dependent on the MTF-1 transcription factor (Figure 3.4A, lane 6) and capped at their 5′ ends, as they could be degraded by a 5′ phosphate dependent exonuclease only after treatment with Cap-Clip Acid Pyrophophatase, which hydrolyzes cap structures to generate 5′ monophosphate groups. Northern blots using multiple probes further indicated that the small RNAs have the same TSS as MtnA mRNA and ligation-mediated 3′ RACE revealed that these small RNAs had detectable 3′ oligoadenylation, a mark known to facilitate RNA degradation by the RNA exosome (LaCava et al. 2005)

We next co-depleted Mtr4 and Integrator subunits and observed that these small RNAs were completely eliminated and that full-length MtnA mRNA expression was restored (Figure 3.4A, lanes 8,10,12). This suggests that generation of the small RNAs is dependent on Integrator. In support of this, in cells in which the endogenous IntS11 protein had been depleted by RNAi (Figure 3.4B, lane 4), re-expression of a wild-type (WT) IntS11 transgene restored expression of the MtnA small RNAs (Figure 3.4B, lane 8), whereas expression of a catalytically dead (E203Q) (Baillat et al. 2005) IntS11 transgene did not (Figure 3.4B, lane 12).

Together, the data in Figures 3.1–4 support a model in which the Integrator complex is recruited to the active MtnA locus during copper stress to cleave nascent RNAs, thereby facilitating premature termination (Figure 3.3F). This role for Integrator at MtnA is mechanistically related to its function at snRNA genes, where Integrator both cleaves the nascent snRNA and promotes RNAPII termination/recycling. Rather than Integrator attaining a novel function at protein-coding genes (Arnold et al. 2013; Gardini et al. 2014; Stadelmayer et al. 2014), our data show that the Integrator endonuclease activity has been

“repurposed” at the MtnA gene. Once generated, the prematurely terminated small RNAs are actively targeted for 3′–5′degradation by the nuclear RNA exosome, at least in part due to 3′ oligoadenylation. Degradation of these small RNAs appears to be critical for enabling subsequent rounds of MtnA transcription, as the output of the MtnA promoter is reduced when the RNA exosome is depleted from cells (Figures 3.1C, 3.3E). The requirement of the RNA exosome for MtnA transcription is, however, abrogated when Integrator is depleted or catalytically inactive. These results indicate a potential epistatic relationship in which Integrator cleaves nascent RNAs that must be subsequently degraded to allow production of more full-length mRNA transcripts in copper stressed cells.

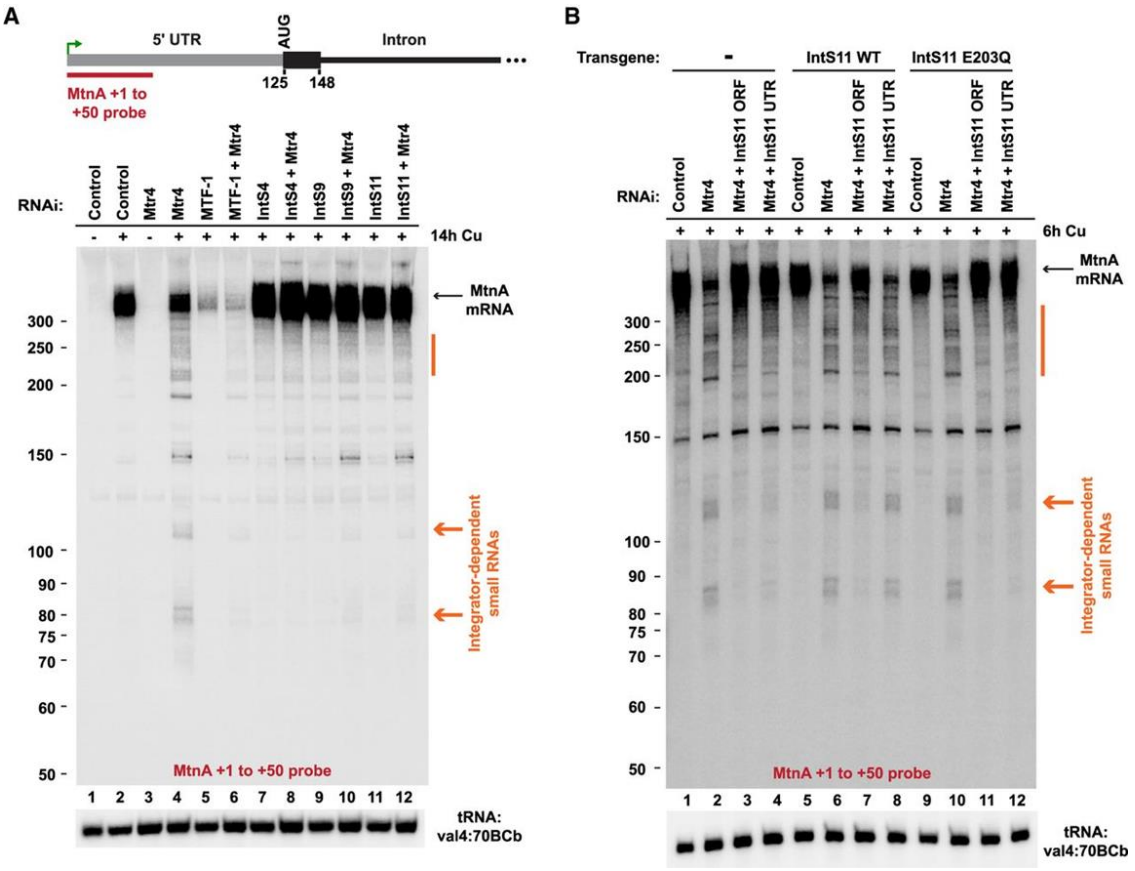


Figure 3.4: The Integrator complex cleaves nascent MtnA RNAs to catalyze premature transcription termination

(A) Northern blotting was used to analyze RNAs generated from the endogenous MtnA locus in DL1 cells treated with the indicated dsRNAs and CuSO4. Full-length MtnA mRNA (black arrow) and Integrator-dependent small RNAs (orange) are indicated.
(B) Parental DL1 cells (lanes 1–4) or DL1 cells stably expressing WT (lanes 5–8) or catalytically inactive (lanes 9–12) IntS11 transgenes were treated with the indicated dsRNAs and CuSO4. Northern blotting was then performed as in A.
(Figure and legend text reprinted with permission from Tatomer DC, Elrod ND, Liang D, Jonathan M, Wagner EJ, Cherry S, Wilusz JE: The Integrator complex cleaves nascent mRNAs to attenuate transcriptional attenuation of Methallothionein. Genes and Development Sep;33/21-21/1525 doi:10.1101/gad.330167.119 2019.)

Many Drosophila protein-coding genes are controlled by the Integrator complex

As Integrator-dependent termination events potently attenuate transcription from the MtnA promoter during copper stress, we next asked whether additional protein-coding genes are similarly regulated. IntS9, which forms a heterodimer with IntS11 and is essential for its endonuclease activity (Wu et al. 2017), was depleted from DL1 cells for 3 d and copper was added for the final 14 h. Using RNA-seq, we identified 409 and 49 genes that were up and down-regulated, respectively, upon IntS9 depletion (fold change >1.5 and $P < 0.001$) (Figure 3.5A). The set of up-regulated mRNAs was enriched in genes that respond to stimuli as well as gene ontology categories related to cell migration, proliferation, and cell-fate specification. In contrast, no gene ontology categories were enriched in the set of down-regulated mRNAs.

To validate the RNA-seq results, seven mRNAs that had differing magnitudes of fold change upon IntS9 depletion were selected for further analysis (genes marked in orange in Figure 3.5A). Among these genes, five contain introns and two are intron-less (CG8620 and CG6770). RT-qPCR confirmed that expression of all seven of these mRNAs increased upon IntS9 depletion regardless of whether the cells were subjected to metal stress (Figure 3.5B). Therefore, we did not induce metal stress in subsequent experiments. Upon depleting each Integrator sub-unit individually, we noted that expression of these mRNAs was often most affected by depletion of IntS4, which forms the scaffold of the Integrator cleavage module (Figure 3.5C (Albrecht et al. 2018)). Moreover, analogous to our prior results with MtnA (Figure 3.1E), depletion of many non-catalytic Integrator subunits (notably IntS1, IntS2, IntS5, IntS6, IntS7, and IntS8) also caused large increases in the expression of these mRNAs (Figure 3.5C).

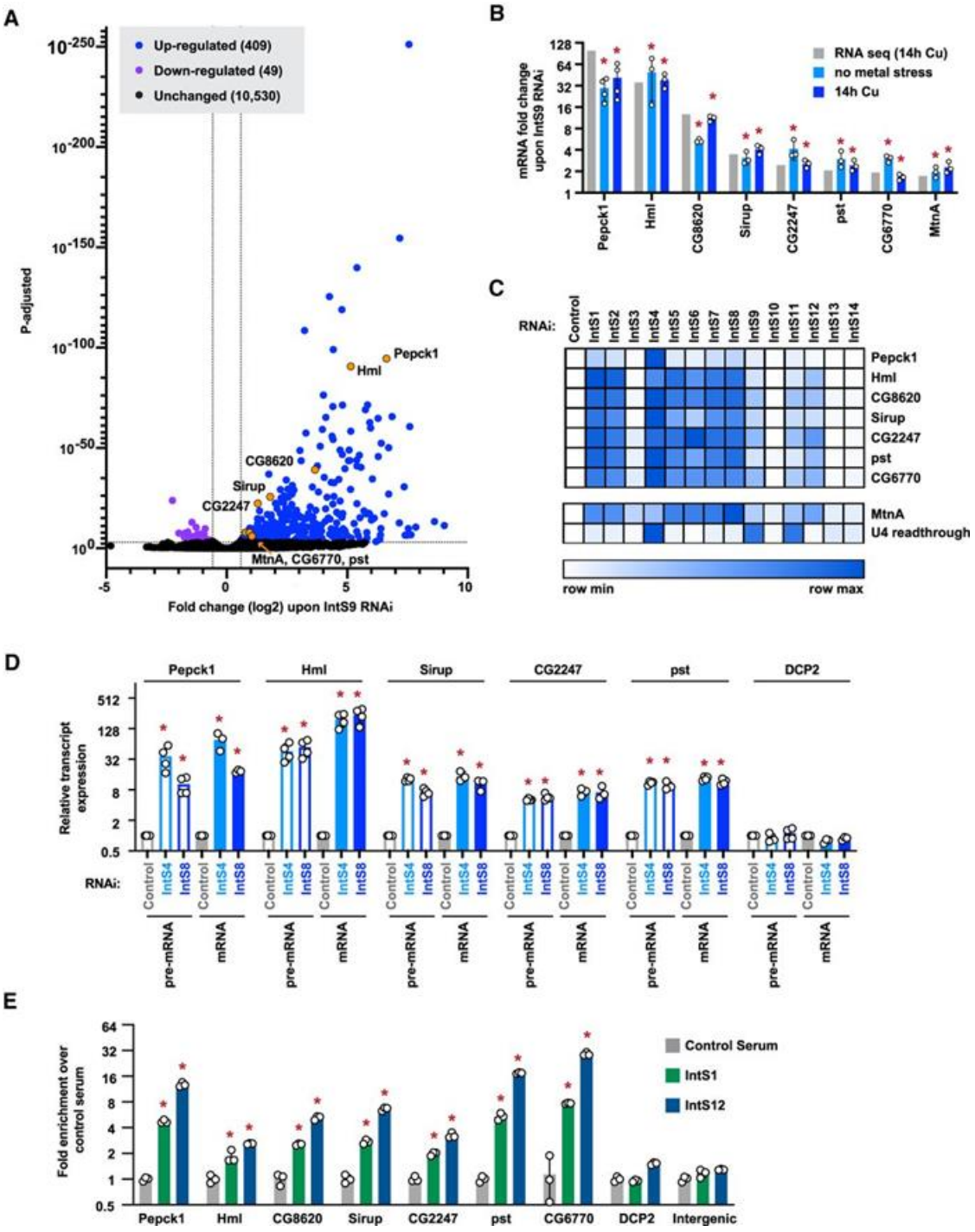


Figure 3.5: Integrator depletion results in up-regulation of many protein-coding genes

(A) DL1 cells were treated for 3 d with a control (βgal) dsRNA or a dsRNA to deplete IntS9, and CuSO4 was added for the last 14 h. Total RNA was isolated, depleted of ribosomal RNAs, and RNA-seq libraries prepared (three biological replicates per condition). The magnitude of change in mRNA expression compared with statistical significance (P-value) is shown as a volcano plot. Threshold used to define IntS9-affected mRNAs was fold change >1.5 and $P < 0.001$.

(B) To verify the RNA-seq results (gray), DL1 cells were treated for 3 d with a control (βgal) dsRNA or a dsRNA to deplete IntS9 with or without CuSO4 added for the last 14 h (light blue and dark blue, respectively). RT-qPCR was then used to quantify changes in mRNA expression levels. Data were normalized to RpL32 mRNA expression and are shown as mean ± SD compared with treatment with a control dsRNA, $N \geq 3$. (*) $P < 0.05$.
(C) DL1 cells were treated with dsRNA to induce RNAi of the indicated factor, and RT-qPCR was then used to quantify changes in mRNA expression levels. CuSO4 was added for the last 14 h only when measuring MtnA mRNA levels. Northern blotting was used to quantify readthrough transcription downstream from the U4:39B snRNA as described in Figure 3.1E. Data are summarized as a heat map using Morpheus (Broad Institute) with darker shades representing increased transcript expression compared with treatment with a control (βgal) dsRNA.
(D) RT-qPCR was used to measure the mRNA and pre-mRNA levels of the indicated transcripts. Data were normalized to RpL32 mRNA expression and are shown as mean ± SD, $N \geq 3$. (*) $P < 0.05$.
(E) ChIP-qPCR was used to measure IntS1 and IntS12 occupancy at the indicated promoter regions. Data are shown as fold change relative to the IgG control serum (mean ± SD, $N = 3$). (*) $P < 0.05$.
(Figure and legend text reprinted with permission from Tatomer DC, Elrod ND, Liang D, Jonathan M, Wagner EJ, Cherry S, Wilusz JE: The Integrator complex cleaves nascent mRNAs to attenuate transcriptional attenuation of Methallothionein. Genes and Development Sep;33/21-21/1525 doi:10.1101/gad.330167.119 2019.)

The Integrator complex cleaves many nascent mRNAs to trigger transcription termination

To determine whether Integrator controls the outputs of these protein-coding genes transcriptionally or post-transcriptionally, we measured pre-mRNA levels from the intron-containing genes. Expression of these pre-mRNAs increased upon depletion of Integrator subunits, and the observed fold changes are similar to the increases in mature mRNA levels (Figure 3.5D). These results mirror our findings at MtnA (Figure 3.2A) and strongly suggest that the observed increases in mature transcript levels are due to transcriptional control by Integrator, and not due to indirect effects of Integrator functioning in snRNA processing. Moreover, ChIP-qPCR confirmed that multiple Integrator subunits are recruited to the 5′ ends of these gene loci (Figure 3.5E). As an additional control, we monitored the DCP2 locus and observed minimal Integrator binding to the gene (Figure 3.5E) as well as no change in DCP2 pre-mRNA or mRNA levels upon Integrator depletion (Figure 3.5D).

To further confirm that Integrator regulation of these genes was driven by their promoters, we cloned each of these regions (along with a portion of the 5′ UTRs) up-stream of an eGFP reporter (Figure 3.6A). Indeed, the expression of eGFP mRNA driven from each of the examined promoters was sensitive to the levels of Integrator sub-units, including those in the Integrator cleavage module (especially IntS4) as well as many of the non-catalytic sub-units (Figure 3.6A). Similar results were obtained with the MtnA-driven eGFP reporter (Figure 3.6A), whereas a reporter plasmid that monitors Integrator activity downstream from an snRNA (Chen et al. 2013) displayed a distinct sensitivity pattern (Figure 3.6B). For example, IntS6 depletion caused up-regulation of the output from all the Integrator regulated protein-coding gene promoters (Figure 3.6A) but had minimal effect on the U4:39B snRNA readthrough reporter (Figure 3.6B). As an additional control, we confirmed that expression of an eGFP reporter driven by the ubiquitin-63E (Ubi-p63e) promoter did not increase upon depletion of any of the Integrator subunits (Figure 3.6A).

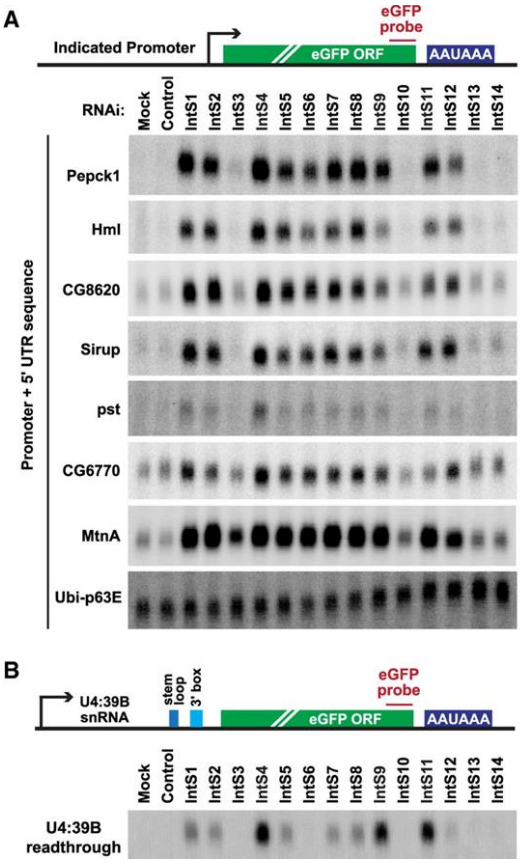


Figure 3.6: eGFP reporter genes driven by the example promoters are regulated by Integrator

(A) The promoter and 5′ UTR of each of the indicated protein-coding genes was cloned upstream of an eGFP reporter. The plasmids were then individually transfected into DL1 cells that had been treated with the indicated dsRNAs. CuSO4 was added for the last 14 h only when measuring eGFP production from the MtnA promoter. Northern blots were used to quantify expression of each eGFP reporter mRNA. Representative blots are shown.
(B) DL1 cells were treated with the indicated dsRNAs and then transfected with a reporter plasmid that produces eGFP when the encoded U4:39B snRNA fails to be properly processed at its 3′ end. Northern blots were used to quantify eGFP mRNA expression that is a result of U4:39B readthrough. A representative blot is shown.
(Figure and legend text reprinted with permission from Tatomer DC, Elrod ND, Liang D, Jonathan M, Wagner EJ, Cherry S, Wilusz JE: The Integrator complex cleaves nascent mRNAs to attenuate transcriptional attenuation of Methallothionein. Genes and Development Sep;33/21-21/1525 doi:10.1101/gad.330167.119 2019.)

This is consistent with the RNA-seq results that showed endogenous Ubi-p63e mRNA levels do not change upon Integrator depletion.

Next, we tested whether Integrator catalyzes premature transcription termination at these genes in a manner analogous to how it controls MtnA. We first investigated whether the IntS11 endonuclease activity is required. Depletion of the endogenous IntS11 protein using a dsRNA targeting the IntS11 3′ UTR (Figure 3.3A) resulted in increased expression of each of the examined Integrator-dependent mRNAs (Figure 3.7A). Expression of a wild-type (WT) IntS11 transgene restored mRNA expression to levels similar to control treated cells, whereas the catalytically dead IntS11 E203Q mutant did not (Figure 3.7A). The IntS11 endonuclease activity is thus indeed required for regulation of each of these genes and, notably, the presence of the E203Q mutant protein exacerbated the changes in expression of the CG6770, pst, and Sirup mRNAs, potentially indicative of a dominant negative effect.

Northern blots were then used to detect premature termination products from each of the Integrator-regulated genes (Figure 3.7B). Given that the MtnA cleavage products are rapidly degraded by the RNA exosome (Figure 3.4), we reasoned that small RNAs generated from other loci would likewise be unstable. Depletion of the exosome-associated RNA helicase Mtr4 (Lubas et al. 2011) enabled small RNAs to be detected from the 5′ ends of the Integrator-dependent genes, and these transcripts were lost upon Integrator co-depletion (RNAs marked in orange; Figure 3.7B). Interestingly, these small RNAs were of defined lengths and often 50–110 nt, roughly mirroring the sizes of cleavage products observed at the MtnA locus (Figure 3.4). We thus conclude that the Integrator complex is recruited to a number of protein-coding genes where it cleaves nascent RNAs and facilitates premature transcription termination (Figure 3.7C).

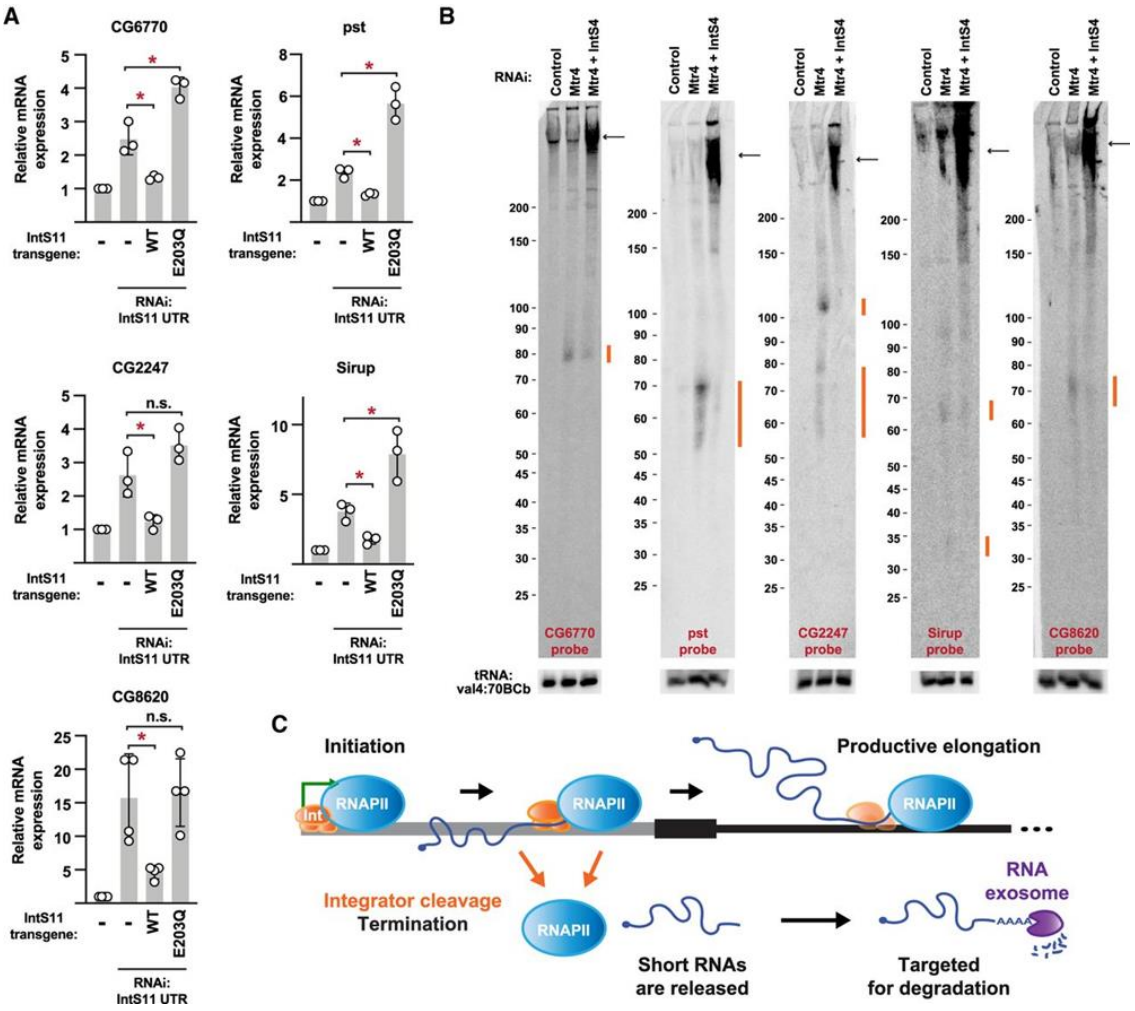


Figure 3.7: The Integrator complex cleaves many nascent mRNAs to catalyze premature transcription termination

(A) Parental DL1 cells (denoted –) or DL1 cells stably expressing WT or catalytically inactive (E203Q) IntS11 transgenes were treated with a dsRNA to the IntS11 3′ UTR, thereby depleting endogenous IntS11 but not IntS11 made from the transgenes. RT-qPCR was then used to quantify expression of the indicated mRNAs. Data were normalized to RpL32 mRNA expression and are shown as mean ± SD, N = 3. (*) $P < 0.05$.
(B) DL1 cells were treated with dsRNAs for 3 d to induce RNAi and depletion of the indicated factors. Northern blotting using 50 µg of total RNA was then used to analyze transcripts from the 5′ ends of the indicated protein-coding loci. Integrator-dependent small RNAs (orange) and full-length mRNAs (black arrows) are noted. Representative blots are shown.
(C) Schematic of a protein-coding locus, highlighting the presence of Integrator (Int, orange) and possible fates of RNAPII (blue). After transcription initiation, RNAPII can transition into productive elongation to generate the mature mRNA. Alternatively, Integrator can cleave the nascent RNA, thereby enabling transcription termination and degradation of the short RNA by the nuclear RNA exosome (purple).

(Figure and legend text reprinted with permission from Tatomer DC, Elrod ND, Liang D, Jonathan M, Wagner EJ, Cherry S, Wilusz JE: The Integrator complex cleaves nascent mRNAs to attenuate transcriptional attenuation of Methallothionein. Genes and Development Sep;33/21-21/1525 doi:10.1101/gad.330167.119 2019.)

Integrator cleavage of nascent mRNAs does not require a 3′ box sequence

At snRNA gene loci, a conserved but relatively degenerate sequence known as the 3′ box is located 9–19 nt down-stream from the 3′ ends of mature snRNA transcripts and is required for Integrator cleavage (Hernandez 1985; Baillat and Wagner 2015). A similar 3′ box-like sequence is not immediately recognizable within MtnA or the other transcripts we studied in detail, and we thus introduced deletions into the MtnA 5′ UTR upstream of the eGFP reporter in an attempt to alter the cleavage product sizes. Notably, this analysis revealed that Integrator-dependent small RNAs derived from the MtnA promoter appear to be largely 70–90 nt in length, regardless of the mRNA sequence. This suggests that Integrator may cleave nascent mRNAs at a set distance from the TSS in a manner independent of local DNA or RNA sequence content, perhaps at positions of RNAPII pausing/stalling or nucleosomes (Chiu et al. 2018) (Figure 3.7B).

SUMMARY

Altogether, our data indicate that the Integrator complex can attenuate the expression of protein-coding genes by catalyzing premature transcription termination (Figure 3.7C). The IntS11 endonuclease cleaves a subset of nascent mRNAs, which ultimately triggers degradation of the transcripts by the RNA exosome along with RNAPII termination. We suggest that many protein-coding genes are negatively regulated via this attenuation mechanism, and the Drosophila MtnA promoter highlights context-specific regulation by Integrator. Transcription of MtnA is induced by copper or cadmium stress, and yet we find that Integrator is robustly recruited to the MtnA promoter only under copper stress conditions (Figure 3.2).

Chapter 4. The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes

This chapter contains text and figures reprinted with permission from Elrod ND, Henriques T*, Huang K.L., Tatomer DC, Wilusz JE, Wagner EJ, Adelman K. The Integrator complex terminates promoter-proximal transcription at protein-coding genes. Mol. Cell Nov;76(5):738-752.e7 doi: 10.1016/j.molcel.2019.10.034 2019.

INTRODUCTION

Dysregulated gene activity underlies a majority of developmental defects and many diseases including cancer, immune and neurological disorders. Accordingly, the transcription of protein-coding messenger RNA (mRNA) is tightly controlled in metazoan cells, and can be regulated at the steps of initiation, elongation or termination. During initiation, transcription factors (TFs) cooperate with coactivators such as Mediator to recruit the general transcription machinery and Pol II to a gene promoter. The polymerase then initiates RNA synthesis and moves downstream from the transcription start site (TSS) into the promoter-proximal region. However, after generating a short, 25-60 nt-long RNA, Pol II pauses in early elongation (Adelman and Lis 2012). Pausing by Pol II is manifested by the DSIF and NELF complexes, which collaborate to stabilize the paused conformation (Henriques et al. 2013; Vos et al. 2018; Core and Adelman 2019). Release of paused Pol II into productive elongation requires the kinase P-TEFb, which phosphorylates DSIF, NELF and the Pol II C-terminal domain (CTD), removing NELF from the elongation complex and allowing Pol II to resume transcription into the gene body, with enhanced elongation efficiency (Peterlin and Price 2006).

Release of paused Pol II into productive RNA synthesis is essential for formation of a mature, functional mRNA. If promoter-paused Pol II becomes permanently arrested or dissociates from the DNA through premature termination, then the process of gene

expression is short-circuited, and the gene will not be expressed. Thus, the stability and fate of paused Pol II at a given promoter will have profound effects on gene output. Interestingly, work from a number of laboratories has highlighted that the stability of paused Pol II can differ substantially among genes (Henriques et al. 2013; Buckley et al. 2014; Chen et al. 2015; Krebs et al. 2017; Shao and Zeitlinger 2017; Erickson et al. 2018). In particular, recent studies of paused Pol II in *Drosophila* revealed a surprising diversity of behaviors following treatment of cells with Triptolide (Trp), an inhibitor of TFIIH that prevents new transcription initiation (Vispe et al. 2009; Krebs et al. 2017; Shao and Zeitlinger 2017; Henriques et al. 2018). At ~20% of genes, inhibition of transcription initiation with Trp caused a dramatic reduction of promoter Pol II levels within <2.5 minutes (Henriques et al. 2018). Thus, these genes consistently require new transcription initiation in order to maintain appropriate levels of promoter Pol II. As such, it has been proposed that Pol II undergoes multiple iterative cycles of initiation, early elongation and premature termination at these genes, each time releasing a short, non-functional RNA (Krebs et al. 2017; Nilson et al. 2017; Erickson et al. 2018; Steurer et al. 2018; Kamieniarz-Gdula et al. 2019). In contrast, a majority of genes were found to harbor a more stable Pol II, with paused polymerase levels persisting after Trp treatment. In fact, after inhibiting transcription initiation, the median half-life of paused Pol II was ~10 minutes in both mouse and *Drosophila* systems (Jonkers et al. 2014; Chen et al. 2015; Shao and Zeitlinger 2017; Henriques et al. 2018). Critically, the distinct stabilities of Pol II observed at different promoters suggests that the lifetime of paused polymerase is modulated to tune gene expression levels. However, the factors that mediate this regulation have yet to be elucidated.

Regulation of promoter-proximal termination is well-described in bacteria, where it is termed *attenuation* (Yanofsky 1981). Attenuation serves to tightly repress gene activity, even under conditions where the polymerase is recruited to a promoter and initiates RNA synthesis at high levels. Mechanistically, bacterial attenuation often involves

destabilization of the RNA-DNA hybrid within the polymerase through RNA structures and/or termination factors with RNA helicase activity (Yanofsky 1981; Gollnick and Babitzke 2002; Henkin and Yanofsky 2002). Similar termination mechanisms are recognized in the yeast *Saccharomyces cerevisiae,* where the Nrd1-Nab3-Sen1 (NNS) complex directs termination using coordinated RNA binding and helicase activities (Bresson and Tollervey 2018). Intriguingly, the NNS complex, which predominantly drives termination of non-coding RNAs, has also been implicated in premature termination at select mRNA loci (Porrua and Libri 2015; Merran and Corden 2017; Sohrabi-Jahromi et al. 2019). However, despite the regulatory potential of promoter-proximal attenuation, no similar strategies or factors have yet been described in metazoan cells. In particular, it remains unclear whether higher eukaryotes possess a termination machinery that promotes dissociation of paused early elongation complexes.

Elongating Pol II is typically extremely stable, with formation of a mature mRNA often involving transcription of many kilobases without Pol II dissociation from DNA. Termination at mRNA 3'-ends involves recognition of specific sequences by cleavage and polyadenylation (CPA) factors and slowing of Pol II elongation. CPSF73, a component of the CPA complex, utilizes a β-lactamase/β-CASP domain (Mandel et al. 2006) to cleave pre-mRNA, producing both a substrate for polyadenylation and a free 5' end on the nascent RNA still engaged with Pol II. This 5' end lacks the protective 7-methy-G cap, allowing it to be targeted by the Xrn2 exonuclease, which ultimately leads to termination (Eaton et al. 2018). Hence, cleavage of the nascent RNA is coupled to the termination of elongation and dissociation of Pol II from template DNA, as well as degradation of the associated short RNA. Although the CPA machinery typically functions at gene 3' ends, there are examples of premature cleavage and polyadenylation (PCPA) occurring within gene bodies, especially within intronic regions (Kamieniarz-Gdula et al. 2019; Venters et al. 2019). However, whether this machinery is involved in RNA cleavage and termination of promoter proximal Pol II remains unknown.

We set out to determine the causes of differential stability of paused Pol II across mRNA genes. In particular, we were interested in defining factors that might render promoter Pol II susceptible to premature termination and the release of short, immature RNAs (Krebs et al. 2017; Nilson et al. 2017; Shao and Zeitlinger 2017; Erickson et al. 2018; Henriques et al. 2018; Steurer et al. 2018). Strikingly, we discovered that the Integrator complex is enriched at mRNA promoters with unstable Pol II pausing. The 14-subunit, metazoan-specific, Integrator complex was initially reported to be exclusively required for cleavage and 3'-end formation of small nuclear RNAs (snRNAs) involved in splicing (Baillat et al. 2005). However, subsequent work has suggested a broader role, including at signal-responsive mammalian genes (Gardini et al. 2014; Stadelmayer et al. 2014; Lai et al. 2015; Skaar et al. 2015). Our work elucidates this role and reveals that Integrator targets paused Pol II at selected protein-coding genes and enhancers, to mediate premature termination. Notably, the Integrator complex, like the CPA machinery, possesses an RNA endonuclease, and we find that this activity is critical for gene repression. Thus, our findings unearth transcription attenuation as a conserved, broad mode of gene control in metazoan cells.

RESULTS AND DISCUSSION

The underlying cause for the short lifetime of paused Pol II at a subset (~20%) of *Drosophila* protein coding genes is not understood (Buckley et al. 2014; Krebs et al. 2017; Shao and Zeitlinger 2017; Henriques et al. 2018). One potential explanation for the brief lifetime of Pol II near these promoters is that paused polymerase is quickly released into productive elongation. This model would predict that such genes would generally have lower levels of Pol II near their promoters, and more Pol II elongating within gene bodies. An alternative possibility is that fast Pol II turnover at these genes results from rapid transcription termination of promoter-paused Pol II. The key prediction of this latter model

is that these genes would display lower levels of productively elongating Pol II within gene bodies.

To evaluate these possibilities, we compared nascent RNA profiles, determined by PRO-seq, a single-nucleotide resolution method for mapping active and transcriptionally engaged Pol II (Kwak et al. 2013). Genes were stratified into four clusters based on their Pol II decay rate following Trp treatment (Krebs et al. 2017; Henriques et al. 2018) and were analyzed for PRO-seq signals near the promoter or within the gene body. We found that genes with short-lived promoter Pol II occupancy (defined as half-life upon Trp-treatment <2.5 min) have significantly lower elongating Pol II levels than other gene classes (Figure 4.1a), despite modestly higher promoter Pol II signals. These data are thus consistent with a model wherein Pol II is efficiently recruited to these promoters, but fails to enter productive elongation, possibly due to premature termination (Krebs et al. 2017).

To evaluate this prediction and define factors that might contribute to this behavior, we computationally assessed a comprehensive repertoire of ChIP-seq data (mod et al. 2010; Ho et al. 2014; Weber et al. 2014; Baumann and Gilmour 2017; Henriques et al. 2018; Kaye et al. 2018). Specifically, we sought to identify factors enriched (or de-enriched) at gene promoters where pausing is unstable as compared to other promoters. Chromatin accessibility was observed to be consistent across Pol II decay classes (as assessed by ATAC-seq, Figure 4.1b), consistent with the similar promoter Pol II levels observed. However, reduced levels of tri-methylated H3 Lysine 36 (H3K36me3) were noted within genes harboring unstable promoter Pol II (Figure 4.1c). The H3K36me3 mark is deposited during productive elongation, and H3K36me3 levels typically correlate with transcription activity (Wagner and Carpenter 2012; Venkatesh and Workman 2015). Thus, the observed, low H3K36me3 signal indicates weak transcription elongation at genes with unstable Pol II, consistent with PRO-seq data. Conversely, genes with stable pausing exhibited stronger transcription activity and higher levels of H3K36me3 (Figures 4.1a and 4.1b), in agreement with recent work (Tettey et al. 2019).

Genes with unstable Pol II also displayed a significant enrichment in H3K4 mono-methylation (H3K4me1) and lower tri-methylation of H3K4 (H3K4me3) and as compared to genes with more stable pausing (Figure 1b). This finding suggests that H3K4 methylation levels increase near promoters as Pol II stability and residence time increases, in agreement with a recent study in yeast (Soares et al. 2017). Intriguingly, elevated H3K4me1 levels, with deficiencies in H3K36me3, H3K4me3 and productive RNA elongation are considered to be characteristics of enhancers (ENCODE-Project-Consortium 2012; Kim and Shiekhattar 2015). Enhancers are also characterized by unstable Pol II and the production of short RNAs (Henriques et al. 2018), suggesting a connection between the chromatin signatures typical of enhancers and defective or inefficient transcription elongation.

To define additional factors that could contribute to the transcriptional properties of these genes, we analyzed ChIP-seq profiles of non-chromatin proteins. We found the Integrator subunit 1 (IntS1) among the most significantly enriched factors at genes with unstable Pol II (Figures 4.1d and 4.1e). This is an interesting finding, given that Integrator is implicated in the biogenesis of enhancer-derived RNAs (eRNAs) in human cells (Lai et al. 2015), and further underscores the similarity between this class of genes and enhancers. To confirm these results, we conducted ChIP-seq using an antibody raised against another *Drosophila* Integrator subunit, IntS12, and found a highly similar enrichment at genes with unstable Pol II).

In summary, genes with unstable promoter Pol II display high levels of Pol II recruitment and promoter DNA accessibility, but significantly diminished Pol II elongation. Further, these genes display chromatin features reminiscent of enhancers, suggestive that a lack of stable pausing has considerable consequences on local chromatin modifications (Figures 4.1e and 4.1f). Interestingly, these genes also show elevated occupancy by Integrator, a factor known to mediate RNA cleavage and Pol II termination at non-coding RNA loci.

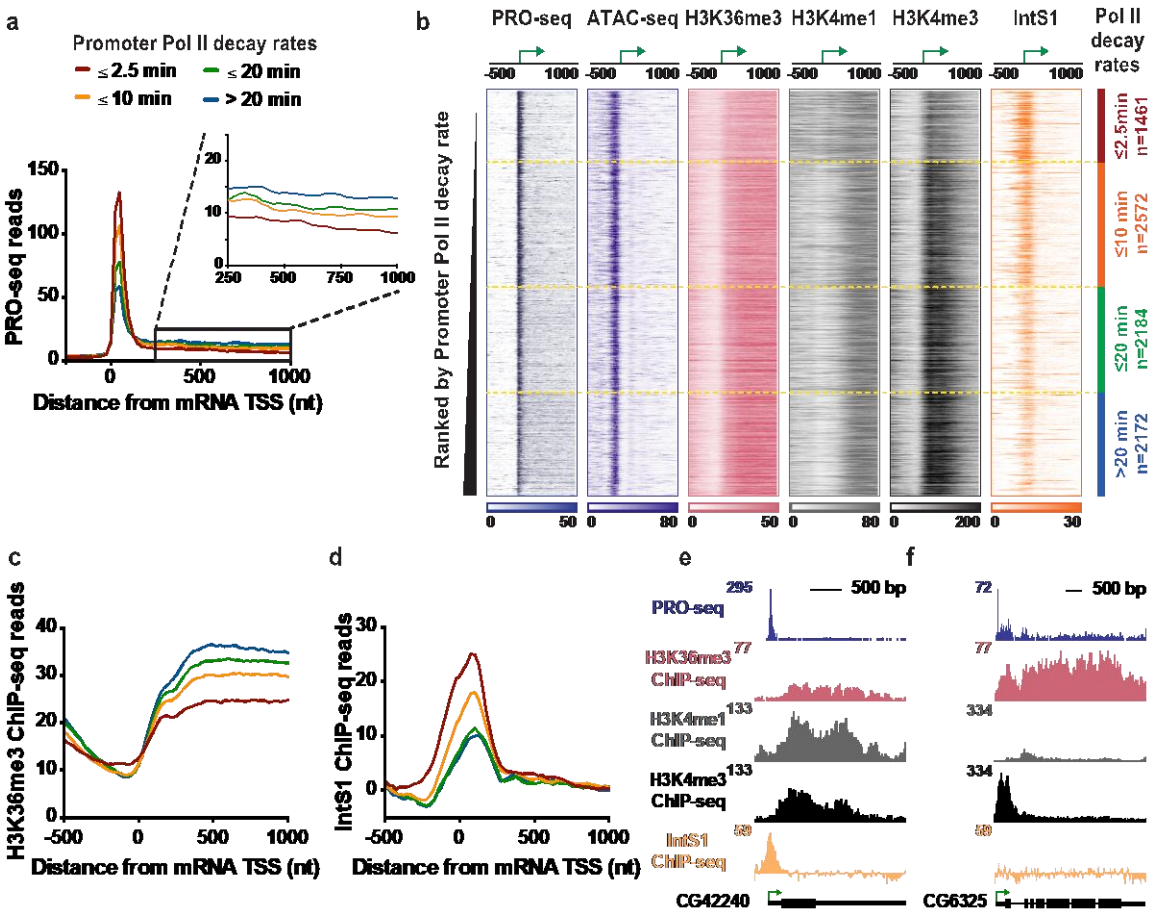


Figure 4.1: Genes with highly unstable promoter Pol II are characterized by poor transcription elongation and enriched binding of Integrator

(a) The average distribution of PRO-seq signal is shown at mRNA transcription start sites (TSSs), with genes divided into four groups based on Pol II promoter decay rates following Triptolide treatment (groups defined in Henriques *et al*, 2018). Inset shows the gene body region. Read counts are summed in 25-nt bins.
(b) Heatmap representations of PRO-seq and ATAC-seq signal, along with ChIP-seq reads for H3K36me3, H3K4me1 and H3K4me3 histone modifications and the Integrator subunit 1 (IntS1). Data are aligned around mRNA TSSs, shown as a green arrow (n=8389). Data are ranked by Promoter Pol II decay rate, where promoters with fastest decay rates (≤2.5 min) are on top. Dotted line separates each group of genes.
(c and d) Average distribution of (c) H3K36me3, and (d) IntS1, ChIP-seq signal is shown, aligned around TSSs and divided into groups based on Pol II decay rate, as in A.
(e and f) Example gene loci, representative of genes in the (e) fast, or (f) slow, Pol II promoter decay groups, displaying profiles of PRO-seq and ChIP-seq signals, as indicated.
(Figure and legend text reprinted with permission from Elrod ND, Henriques T, Huang K.L., Tatomer DC, Wilusz JE, Wagner EJ, Adelman K. The Integrator complex terminates promoter-proximal transcription at protein-coding genes. Mol. Cell Nov;76(5):738-752.e7 doi: 10.1016/j.molcel.2019.10.034 2019.

Loss of Integrator leads to loss of promoter-proximal termination and upregulation of gene expression

Two Integrator subunits, IntS11 and IntS9, are paralogs of the CPA proteins CPSF73 and CPSF100, respectively. IntS11, like CPSF73, has a β-lactamase/β-CASP domain and harbors endonuclease activity. Moreover, similar to CPSF73/100, IntS11 forms a heterodimer with IntS9 and this association is essential for function (Wu et al. 2017). This similarity suggests that Integrator might be capable of mediating transcription termination at protein-coding genes using a mechanism related to that of the CPA machinery. To evaluate this possibility, IntS9 was depleted using RNA interference (RNAi) for 60 hours, followed by polyA-selected RNA-seq to identify mRNA expression changes. Consistent with the reported stability of snRNAs, their steady-state levels were not perturbed during the relatively short time course of RNAi, and very few differences in splicing events were observed in IntS9-depleted cells. Thus, short-term loss of Integrator has minimal effects on snRNA functionality or splicing patterns. Nonetheless, genes with any evidence of altered splicing in IntS9-depleted cells were removed from all further analyses, enabling us to solely focus on transcriptional targets of Integrator.

Our analysis revealed 723 upregulated and 163 downregulated mRNAs upon IntS9 depletion (Figure 4.2s), suggesting that *Drosophila* Integrator is predominantly a transcriptional repressor. The expression changes observed upon IntS9 RNAi were validated using RT-qPCR at selected genes. Gene Ontology analysis of upregulated transcripts shows significant enrichment in signal-responsive pathways, including metabolic, receptor and oxidoreductase activities, as well as Epidermal Growth Factor (EGF)-like protein domains. Consistently, work on mammalian Integrator has implicated this complex in EGF-responsive gene activity (Gardini et al. 2014).

To probe the mechanisms by which Integrator regulates gene expression, we directly monitored nascent RNA synthesis using PRO-seq in control or IntS9-depleted cells. Critically, PRO-seq is amenable to spike-in normalization, allowing us to ensure that

quantitative differences between samples can be accurately measured. PRO-seq in control cells revealed that Pol II is effectively recruited to IntS9-repressed promoters, but the polymerase often fails to transition into productive elongation (Figures 4.2b and 4.2c). In fact, genes upregulated upon IntS9 depletion exhibited significantly higher PRO-seq signal at promoters, yet lower PRO-seq signal within gene bodies and lower mRNA expression than unaffected genes. These data demonstrate that Integrator does not repress transcription initiation but rather prevents the transition of promoter-proximal Pol II into productive RNA synthesis, perhaps by mediating transcription termination. Consistent with this possibility, depletion of IntS9 relieved the strong block to productive elongation at upregulated genes, allowing a significant, median 3-fold, increase of PRO-seq signal within gene bodies (Figures 4.2c and 4.2d).

There was highly significant overlap between transcripts deemed significantly upregulated in PRO-seq and RNA-seq experiments, confirming that the upregulated mRNA production observed upon IntS9 depletion generally results from increased transcription elongation at these genes (Figure 4.2e). In contrast, decreases in RNA-seq signal were not well-reflected in PRO-seq levels, with fold-changes between the assays correlating poorly (Figure 4.2e). Indeed, only 29 transcripts were defined as downregulated by IntS9-depletion in both the RNA-seq and PRO-seq assays. We thus conclude that the dominant transcriptional effect of *Drosophila* Integrator at protein-coding genes is in transcription repression.

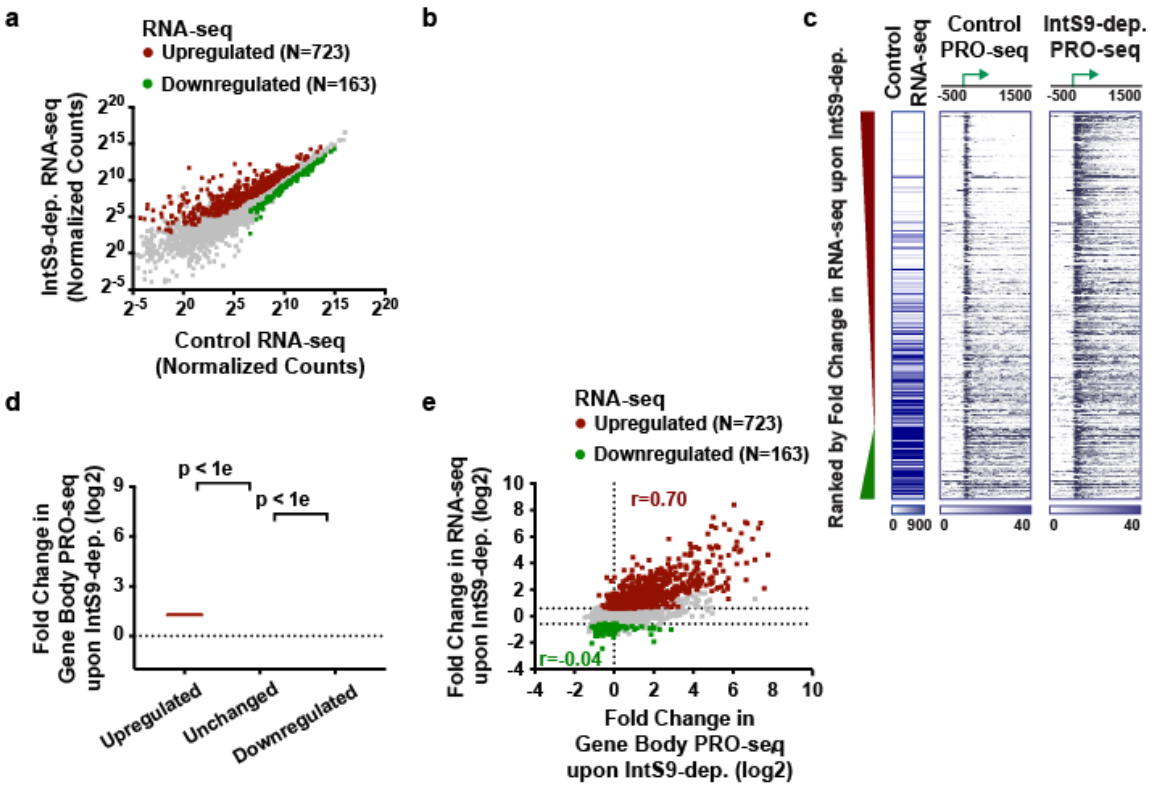


Figure 4.2: The Integrator complex attenuates expression of protein-coding genes

(a) Drosophila cells were treated for 60 h with control dsRNA, or dsRNA targeting IntS9 (N=3). Normalized RNA-seq signal is shown, with significantly affected genes defined as P<0.0001 and fold change >1.5.
(b) even skipped (eve) locus displaying profiles of RNA-seq and PRO-seq in control and IntS9-depleted cells.
(c) Heatmap representations of RNA-seq levels are shown, along with PRO-seq reads from control and IntS9-depleted cells (treated as in a). The location of mRNA TSSs is indicated by an arrow. Genes that are upregulated or downregulated upon IntS9-depletion in RNA-seq are shown, ranked from most upregulated to most downregulated.
(d) Violin plots depict the change in gene body PRO-seq signal upon IntS9-depletion for each group of genes. IntS9-affected genes are defined as in A, as compared to 8613 unchanged genes. Plots show the range of values, with a line indicating median. P-values are calculated using a Mann-Whitney test.
(e) Comparison of fold changes in RNA-seq and PRO-seq signals upon IntS9-depletion is shown. Pearson correlations are shown separately for upregulated and downregulated genes, indicating good agreement between steady-state RNA-seq and nascent PRO-seq signals for upregulated genes, but little correspondence for downregulated genes.
(Figure and legend text reprinted with permission from Elrod ND, Henriques T, Huang K.L., Tatomer DC, Wilusz JE, Wagner EJ, Adelman K. The Integrator complex terminates promoter-proximal transcription at protein-coding genes. Mol. Cell Nov;76(5):738-752.e7 doi: 10.1016/j.molcel.2019.10.034 2019.

The Integrator RNA endonuclease is required for transcriptional repression

The above data suggest that Integrator might use its endonuclease activity to catalyze transcription termination of paused Pol II. To test this model and determine whether IntS11 catalytic function is required for gene repression, we took advantage of a previously described mutant (IntS11 E203Q; Figure 4.3a) that abrogates endonuclease function yet retains the integrity of the Integrator complex (Baillat et al. 2005) . We treated *Drosophila* cells for 60hrs with either control RNAi or with RNAi targeting the IntS11 UTRs and re-expressed either wild-type IntS11 or the E203Q mutant in cells depleted of endogenous IntS11. RNA from these cells was isolated and subjected to poly(A)-enriched RNA-seq. As with IntS9 depletion, the major effect of IntS11 knockdown was upregulation of transcription, and mature snRNA levels are not perturbed. Further, the levels of gene upregulation observed upon depletion of IntS9 or IntS11 were highly concordant (Figure 4.3b). In contrast, there was less agreement and smaller effect sizes observed at downregulated genes (Figure 4.3b), again suggesting that Integrator is predominantly a transcriptional repressor.

The vast majority of gene expression changes observed in IntS11-depleted cells were restored to control levels upon expression of the wild-type IntS11 (Figures 4.3b and 4.3c). In contrast, expression of the E203Q mutant not only failed to rescue the IntS11 depletion but exacerbated the knockdown phenotype, supportive of a dominant negative effect of the catalytically inactive IntS11 protein (Figures 4.3b and 4.3c). The results observed by RNA-seq (e.g. Figure 3d) were confirmed by RT-qPCR. Together, these data indicate that depletion of either IntS9 or IntS11 lead to alteration of a similar set of protein-coding genes and that the IntS11 endonuclease activity is essential for the function of Integrator at these loci.

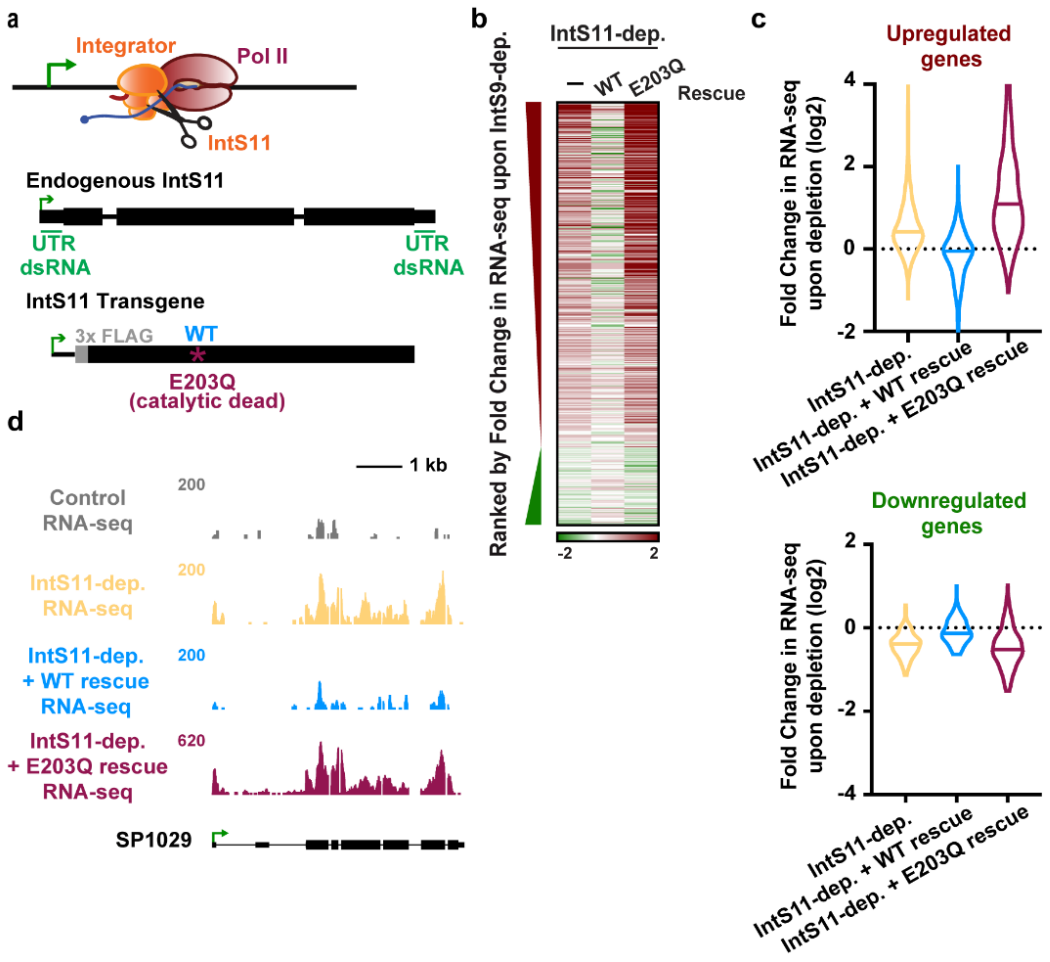


Figure 4.3: Integrator subunit 11 (IntS11) endonuclease activity is essential for altered protein-coding gene expression

(a) The IntS11 subunit of Integrator harbors RNA endonuclease activity (depicted as scissors). To test the importance of this activity, cells were depleted of IntS11 and rescued using a stably integrated transgene expressing WT IntS11, or IntS11 with a mutation that disrupts endonuclease activity (E203Q). To specifically deplete endogenous IntS11 from the rescue cell lines, a dsRNA targeting the untranslated (UTR) regions of endogenous IntS11 (green) was used. Cells were treated for 60 h with control or IntS11 UTR RNAi (N=3), and RNA harvested for RNA-seq.
(b) Heatmap representations of RNA-seq fold changes in IntS11-depleted cells, as compared to cells rescued with WT or E203Q mutant. Genes shown are those affected upon IntS9-depletion, ranked by fold-change as in Figure 4.2c.
(c) Fold Change in RNA-seq signal upon IntS11-depletion at genes (top) upregulated (N=723) or (bottom) downregulated by IntS9-depletion (N=163). Changes in RNA-seq levels as compared to the parental cell line are shown in IntS11-depleted cells, and those rescued by WT or E203Q mutant IntS11. Violin plots show range of values, with a line indicating median.

(d) Example locus (SP1029) showing an upregulated gene whose expression is rescued by WT IntS11, but not by the catalytic dead mutant (E203Q mutation). RNA-seq tracks are shown in control cells and each of the treatments.
(Figure and legend text reprinted with permission from Elrod ND, Henriques T, Huang K.L., Tatomer DC, Wilusz JE, Wagner EJ, Adelman K. The Integrator complex terminates promoter-proximal transcription at protein-coding genes. Mol. Cell Nov;76(5):738-752.e7 doi: 10.1016/j.molcel.2019.10.034 2019.

Integrator attenuates mRNA transcription

The critical involvement of the IntS11 endonuclease in gene repression by Integrator supports a model wherein RNA cleavage triggers premature termination. To further evaluate this model, we defined the full repertoire of transcriptional targets of Integrator, by comparing spike normalized PRO-seq signals in gene bodies between control and IntS9-depleted samples. We found 1204 transcripts with significantly more elongating Pol II upon depletion of Integrator (Figure 4.4a), and 210 with reduced gene-body Pol II signal. This reveals that transcription of ~15% of active *Drosophila* genes is upregulated upon loss of Integrator activity.

Gene ontology analyses of the genes upregulated in PRO-seq agreed well with those from RNA-seq, highlighting metabolic, oxidoreductase and EGF pathways. In contrast, enriched pathways for the downregulated genes in PRO-seq overlapped little with those enriched among RNA-seq downregulated genes, in agreement with the lack of concordance between nascent transcription and steady-state RNA levels within the downregulated gene sets (only 29 genes downregulated in both PRO-seq and RNA-seq). Thus, we focused our attention on the much larger set of upregulated loci.

The increase in gene body PRO-seq signal upon IntS9-depletion was substantial at upregulated genes, with a median increase of over 3.3-fold (Figure 4.4b). As anticipated, the majority of this increase in actively engaged Pol II is evident in PRO-seq signal near TSSs. Thus, we conclude that Integrator typically acts on promoter-proximal Pol II, and that loss of Integrator results in increased levels of engaged polymerase that successfully transition from promoter regions into productive elongation. We then wished to distinguish between models wherein Integrator catalyzes promoter-proximal termination vs. those wherein Integrator prevents escape of promoter-associated Pol II into productive elongation. We evaluated the PRO-seq signal at genes upregulated upon depletion of IntS9. If Integrator stabilizes Pol II pausing, then IntS9 depletion should release this paused Pol

II into gene bodies, resulting in less promoter-proximal PRO-seq signal and an increase in signal downstream. In contrast, if Integrator stimulates termination and dissociation of paused Pol II, then IntS9 depletion should increase both promoter-proximal PRO-seq signals and signals within genes. In support of a termination model, we observed that IntS9 depletion resulted in increased PRO-seq signal near promoters, as well as in gene bodies (Figure 4.4c). Strikingly, the increase in PRO-seq signal from IntS9-depleted cells localized precisely at the position of Pol II pausing, in the window from 25-60 nt into the gene (Figure 4.4d). This finding supports that Integrator targets promoter-paused Pol II and prevents its transition into productive RNA synthesis, likely through premature termination.

To determine whether Integrator similarly targets paused Pol II at enhancers, we made use of a comprehensive set of *Drosophila* enhancer transcription start sites (eTSSs) we recently defined (Henriques et al. 2018). We note that these sites were rigorously defined both functionally, in enhancer plasmid-based reporter assays (Arnold et al. 2013; Zabidi et al. 2015) and spatially, with the TSSs of enhancer RNAs (eRNAs) mapped at single-nucleotide resolution (Henriques et al. 2018). This dataset thus allows for a high-resolution analysis of Integrator activity at functionally confirmed, transcriptionally active enhancer loci at the genome-level. We focused on 1498 intergenic eTSSs, to avoid confounding signals from enhancers within annotated genes, and defined differentially transcribed loci using PRO-seq data as we had for mRNA genes. We observed increased transcription at ~15% of eTSSs in IntS9-depleted cells (N=228), a similar fraction to mRNAs and find only 38 eTSSs with downregulated transcription. Thus, at enhancers, like at protein-coding genes, Integrator plays a generally repressive role in transcription elongation, and targets only selected loci. Importantly, many eRNA loci are not affected by loss of Integrator, consistent with work implicating CPA and other machineries in eRNA 3’ end formation (Austenaa et al. 2015; Ogami et al. 2017).

The parallel in the behavior of Integrator at protein-coding and non-coding loci is further emphasized by the profile of PRO-seq at upregulated eTSSs (compare Figures 4.4e and 4.4C), where loss of Integrator causes an increase of PRO-seq signal precisely in the region of Pol II pausing (compare Figures 4.4f and 4.4d). We conclude that the function of Integrator is highly similar at coding and non-coding RNA loci: a comparable subset of TSSs are affected by Integrator, and Integrator depletion causes increased Pol II near TSSs and higher levels of release downstream into productive elongation.

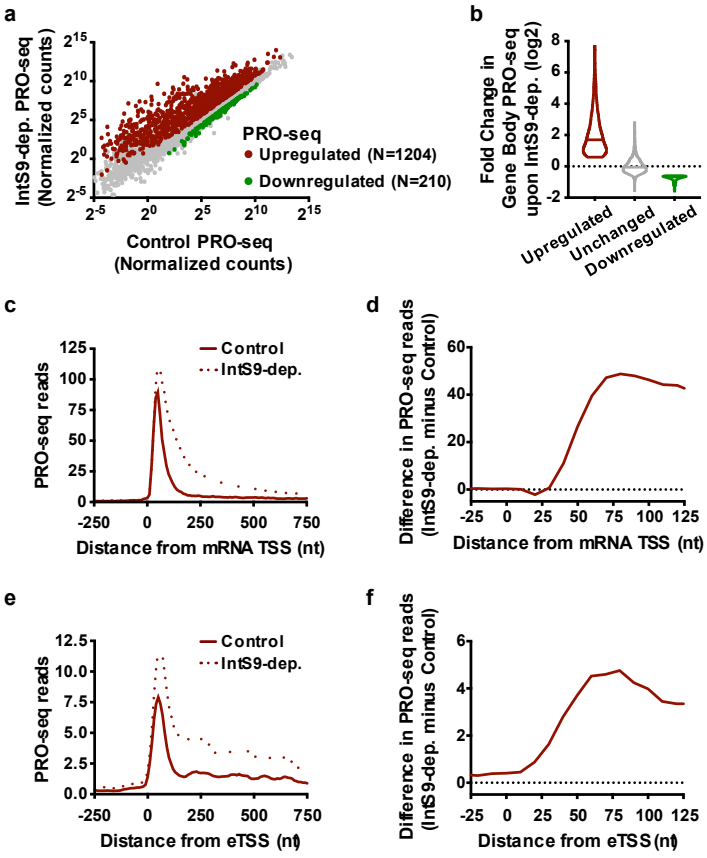


Figure 4.4: Integrator represses productive elongation by Pol II at genes and enhancers

(a) Drosophila cells were treated for 60 h with control or IntS9 RNAi (N=3). Normalized PRO-seq signal across gene bodies is shown, with IntS9-affected genes defined as P<0.0001 and fold change >1.5.
(b) Violin plots depict the change in gene body PRO-seq signal upon IntS9-depletion for each group of genes. IntS9-affected genes are defined as in A, as compared to unchanged genes (N=8085). Violin plots show range of values, with a line indicating median.
(c) Average distribution of PRO-seq signal in control and IntS9-depleted cells is shown at upregulated genes.
(d) The difference in PRO-seq signal between IntS9-depleted and control cells for upregulated genes is shown. Increased signal in IntS9-depleted cells is consistent with the position of Pol II pausing, from +25 to +60 nt downstream of the TSS.
(e) Average distribution of PRO-seq reads from control and IntS9-depleted cells are displayed, centered on enhancer transcription start sites (eTSS) that are upregulated upon IntS9 RNAi (N=228).
(f) Difference in PRO-seq signal between IntS9-depleted and control cells for IntS9-upregulated enhancer RNAs. Note that signal increases at enhancers in the same interval (+25-60 nt from TSS) as at coding loci.
(Figure and legend text reprinted with permission from Elrod ND, Henriques T, Huang K.L., Tatomer DC, Wilusz JE, Wagner EJ, Adelman K. The Integrator complex terminates promoter-proximal transcription at protein-coding genes. Mol. Cell Nov;76(5):738-752.e7 doi: 10.1016/j.molcel.2019.10.034 2019.

Integrator is widely associated with mRNA promoter regions

The mechanism for Integrator-mediated 3' end formation at snRNA loci involves both selective recruitment of Integrator to snRNA promoters and recognition of a degenerate motif near snRNA 3' ends that promotes IntS11 cleavage activity (Hernandez 1985; Hernandez and Weiner 1986; Baillat and Wagner 2015). Interestingly, several factors implicated in recruiting Integrator to snRNA genes are also found at protein coding loci, such as the pause-inducing factors DSIF and NELF (Stadelmayer et al. 2014; Yamamoto et al. 2014), and phosphorylation on the Pol II C-terminal domain (CTD) repeats at Serine 7 residues (Ser7-P;(Egloff et al. 2007; Kim et al. 2010). Consistent with this, Integrator has been observed to associate with some mRNA promoters in human systems (Gardini et al. 2014; Stadelmayer et al. 2014; Skaar et al. 2015). However, it has not been fully explored how well the localization of Integrator at promoters corresponds to its gene regulatory activities at a genome-wide level.

To address this question, we investigated the global localization of Integrator using our ChIP-seq datasets. We find that IntS1 and IntS12 subunits showed highly correlated localization across snRNA (r=0.99) and mRNA promoters (r=0.89), with a strong enrichment near mRNA transcription start sites (Figure 4.5a). However, Integrator signal at promoters correlated only weakly with levels of paused Pol II as determined by promoter PRO-seq signal (r=0.39). Whereas these findings are consistent with Pol II, DSIF and NELF representing interaction surfaces for Integrator, they also indicate that association of Integrator with mRNA promoters is not strictly tied to paused Pol II levels. We thus asked whether there was enrichment of IntS1 or IntS12 occupancy at genes that are upregulated upon depletion of IntS9. Indeed, genes repressed by Integrator were significantly enriched in both IntS1 and IntS12 ChIP-seq signal as compared to genes unaffected by Integrator depletion (Figures 4.5b, 4.5c, and 4.5d). In fact, levels of Integrator observed at IntS9-repressed promoters were even higher than levels at snRNAs (Figure 4.5d). We noted,

however, that Integrator ChIP-seq signals at genes with unchanged expression upon IntS9 RNAi were well above background levels, suggesting that Integrator is also recruited to promoters where it remains inactive.

To further investigate the relationship between Integrator binding and activity, we rank ordered all active mRNA promoters by their IntS1 ChIP-seq signal and calculated cumulative distributions of Integrator-repressed and unchanged genes across this ranking (Figure 4.5e). This analysis demonstrated that Integrator exhibits the full spectrum of binding levels at unchanged genes. However, IntS9-repressed genes were clearly and significantly biased towards higher IntS1 occupancy (Figure 4.5e, >50% of IntS9-repressed genes fall within the top 20% of IntS1 levels, whereas only 15% of unchanged genes fall in this group). Thus, like at the snRNAs, Integrator recruitment to an mRNA promoter is not sufficient to dictate function, but high-level Integrator occupancy is predictive of activity.

To determine whether increased recruitment of Integrator was also related to functional outcomes at enhancers, we identified eTSSs that exhibited significant peaks of IntS1/IntS12 signal. Comparing PRO-seq at these loci in control vs. IntS9-depleted conditions demonstrated that Integrator-bound eTSSs showed increased transcription elongation upon IntS9 RNAi (Figure 4.5f). In contrast, no significant change in PRO-seq signal was observed at Integrator-unbound eTSSs upon depletion of IntS9. We conclude that functional mRNA and eRNA targets of Integrator display greater recruitment of this complex. Although the factors governing this elevated recruitment of Integrator at snRNA or other loci remain to be elucidated, our results underscore a common behavior for Integrator at coding and non-coding loci.

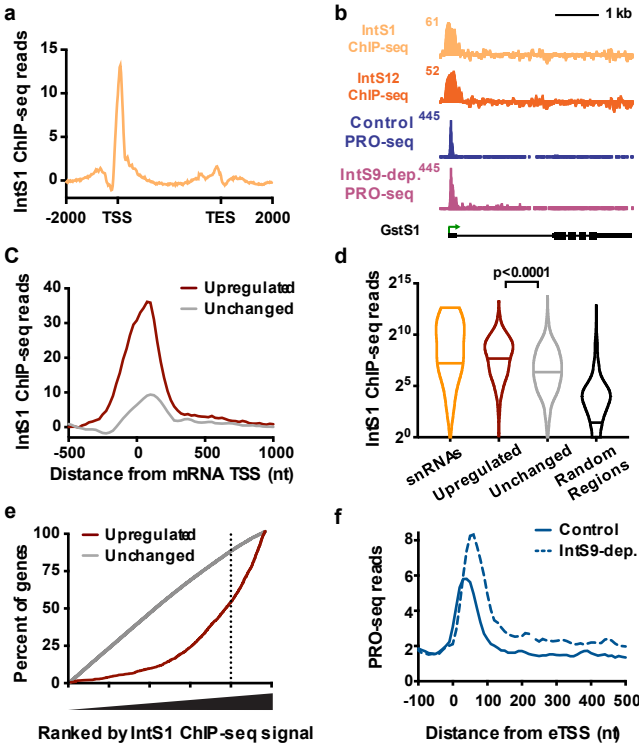


Figure 4.5: Integrator binding is enriched at promoters of target genes

(a) Distribution of IntS1 ChIP-seq signal along the transcription units of all active mRNA genes (N=9499). Windows are from 2 kb upstream of the TSS to 2 kb downstream of the transcription end site (TES). Bin size within genes is scaled according to gene length.
(b) Example locus (GstS1) of an upregulated gene upon IntS9-dep. showing PRO-seq and Integrator ChIP-seq.
(c) Metagene analysis of average IntS1 ChIP-seq signal around promoters of upregulated (N=1204) and unchanged (N=8085) mRNA genes in IntS9-depleted cells. Data are shown in 25 bp bins.
(d) Promoter-proximal IntS1 ChIP-seq reads for each group of sites: snRNAs (N=31), upregulated or unchanged genes, and randomly-selected intergenic regions (N=5000). Violin plots show range of values, with a line indicating median. P-values are calculated using a Mann-Whitney test.
(e) All active genes (N=9499) were rank ordered by increasing IntS1 ChIP-seq signal around promoters (± 250bp), and the cumulative distribution of upregulated or unchanged genes across the range of IntS1 signal is shown. IntS1 levels at unchanged genes show no deviation from the null model, but upregulated genes display a significant bias towards elevated IntS1 ChIP-seq signal.
(f) Average distribution of PRO-seq signal at eTSSs bound by the Integrator complex (N=691) in control and IntS9-depleted cells is shown.
(Figure and legend text reprinted with permission from Elrod ND, Henriques T, Huang K.L., Tatomer DC, Wilusz JE, Wagner EJ, Adelman K. The Integrator complex terminates promoter-proximal transcription at protein-coding genes. Mol. Cell Nov;76(5):738-752.e7 doi: 10.1016/j.molcel.2019.10.034 2019.

Integrator mediates cleavage of nascent RNA and promoter-proximal termination

Taken together, our results are most consistent with Integrator serving as a promoter-proximal cleavage and termination factor for a set of protein-coding genes. To definitively test this possibility, we investigated the short, TSS-associated RNAs that would accompany Pol II termination. In particular, we used Start-seq (Nechaev et al. 2010; Henriques et al. 2018) to identify RNAs under 100 nt in length that were 3′ oligoadenylated, a modification that can be detected on a minor fraction of RNAs released by Pol II during termination (Figure 4.6a). Such oligoadenylated termination products are subject to degradation, and normally are very short-lived, but are stabilized in cells depleted of the RNA Exosome. Accordingly, following depletion of the Exosome subunit Rrp40, we observed significantly more oligoadenylated short RNAs from IntS9-repressed genes than unchanged genes (Figure 4.6b). Strikingly, the 3′ ends of these oligoadenylated RNAs are highly and specifically enriched within the region of Pol II pausing (Figure 4.6c).

We considered that Integrator-mediated RNA cleavage should occur on nascent RNA that has exited the polymerase. The structure of paused elongation complexes (Henriques et al. 2013; Vos et al. 2018; Core and Adelman 2019), indicates that RNA emerges from the exit channel and is available for binding ~15-20 nt upstream of the 3' end position of the nascent RNA. Accordingly, the peak of oligoadenylated RNA 3' end locations at upregulated genes is +35 nt (Figure 4.6c), which is 20nt upstream of the peak of paused Pol II at these genes, at +55nt. From these data, we conclude that Integrator-repressed genes undergo markedly higher levels of Pol II termination as compared to non-Integrator target genes, and that promoter-proximally paused Pol II is the predominant target of this activity.

We next compared the stability of promoter-associated Pol II at Integrator-repressed genes after treatment with Triptolide. Based on increased premature termination at these genes, and our identification of Integrator enrichment at genes with unstable Pol II

(Figure 4.1d), we predicted that Integrator-repressed genes would exhibit reduced promoter Pol II stability as compared to Integrator-unaffected genes. In agreement with this, we observed that Pol II was lost quickly at a majority of IntS9-repressed genes, with half-lives <10 minutes (Figures 4.6d and 4.6e). In contrast, genes whose expression is unchanged by IntS9-depletion presented a Pol II that is stable after Trp treatment, indicative of long-lived pausing (Figure 4.6e). Furthermore, genes upregulated by IntS9-depletion exhibited lower levels of H3K36me3 and H3K4me3 (Figures 4.6f and 4.6g) and higher levels of H3K4me1 than unchanged genes, consistent with defects in productive elongation. Thus, based on many independent lines of evidence we conclude that genes with unstable Pol II recruit Integrator, rendering them susceptible to promoter-proximal termination, and resulting in reduced productive RNA synthesis and chromatin features that accompany transcription elongation.

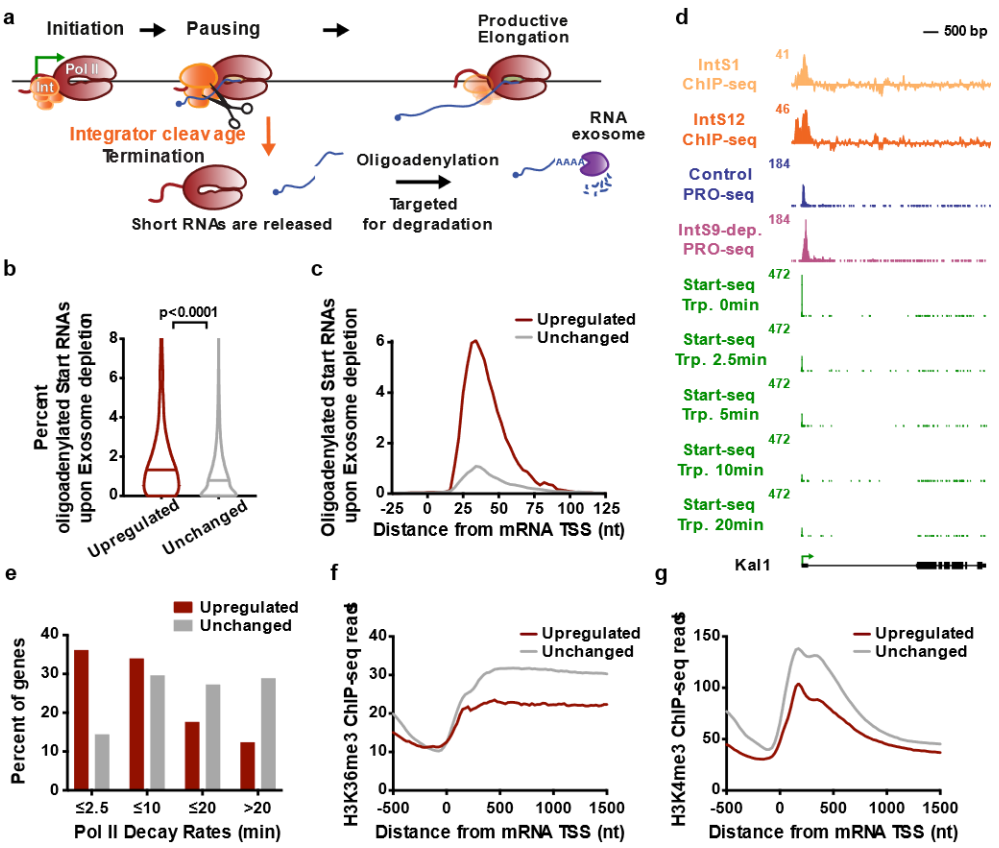


Figure 4.6: Integrator attenuates mRNA expression through promoter-proximal termination

(a) Schematic of transcription cycle with possible fates of Pol II. Paused Pol II can enter into productive elongation or terminate and release a short RNA. A small fraction of released RNA is oligoadenylated to facilitate degradation by the RNA exosome.
(b) The percent of Start RNA reads bearing oligoadenylated 3' ends in exosome- depleted (Rrp40 subunit) cells is shown for each gene group. Violin plots indicate range of values, with a bar at median. P-value is calculated using a Mann-Whitney test.
(c) The 3′ end locations of oligoadenylated RNAs identified in exosome-depleted cells are shown at mRNA genes that are upregulated or unchanged by IntS9-depletion.
(d) Kal1 (CG6173) locus displaying profiles of ChIP-seq for Integrator subunits, PRO-seq, and Start-seq following a time course of Triptolide treatment.
(e) Decay rates for promoter Pol II were determined using Start-seq over a Triptolide treatment time course, and the percentage of upregulated or unchanged genes in each group is shown.
(f-g) Average distribution of (f) H3K36me3 and (g) H3K4me3 ChIP-seq signal is shown, aligned around mRNA TSSs. Genes shown are those upregulated or unchanged in the PRO-seq assay upon IntS9-depletion.
(Figure and legend text reprinted with permission from Elrod ND, Henriques T, Huang K.L., Tatomer DC, Wilusz JE, Wagner EJ, Adelman K. The Integrator complex terminates promoter-proximal transcription at protein-coding genes. Mol. Cell Nov;76(5):738-752.e7 doi: 10.1016/j.molcel.2019.10.034 2019.

Integrator-mediated gene repression is conserved in human cells

Our data in *Drosophila* indicate a mechanistically conserved role for Integrator in promoter-proximal termination of mRNA and eRNA synthesis. Although our model is in agreement with data from mammalian systems as regards eRNA biogenesis (Lai et al. 2015), it differs considerably from any of the proposed roles of Integrator at mammalian protein-coding genes (Gardini et al. 2014; Stadelmayer et al. 2014; Lai et al. 2015; Skaar et al. 2015; Barbieri et al. 2018). In particular, a majority of models posit that mammalian Integrator is an activator of transcription, and none of the proposed functions involve the IntS11 endonuclease in termination. For example, based on genomic studies of Integrator localization and activity in HeLa cells, it was proposed that Integrator stabilizes paused Pol II and facilitates both processive transcription elongation and RNA processing (Stadelmayer et al. 2014). Alternatively, other work in HeLa cells has implicated Integrator as critical for the rapid, EGF-mediated induction of ~100 'immediate early' genes, including *JUNB* and *FOS*. At these genes, Integrator was found to stimulate gene activity through recruitment of the Super Elongation Complex (Gardini et al. 2014). However, a detailed analysis of *JUNB* and several other immediate early genes gene in Integrator-depleted HeLa cells prior to EGF stimulation indicated that these genes were upregulated by loss of Integrator. Thus, it was suggested that Integrator inhibits expression of EGF-responsive genes under basal conditions (Skaar et al. 2015). Thus, it remains an open question whether, in the absence of a stimulus, mammalian Integrator plays a repressive role similar to that uncovered for the *Drosophila* complex.

To investigate whether loss of mammalian Integrator led to upregulation of gene transcription, as we observed for *Drosophila*, we analyzed previously published chromatin-associated RNA-seq from control and IntS11-depleted HeLa cells harvested prior to EGF stimulation. While chromatin-associated RNA-seq lacks the spatial resolution of PRO-seq, it is a significantly better indicator of ongoing transcription than is steady-state RNA-seq.

Thus, we probed for differentially transcribed genes following IntS11-depletion in chromatin RNA-seq, using the same strategies employed for analysis of PRO-seq. Strikingly, we found a substantial number of genes upregulated in IntS11-depleted cells (N=667; Figure 4.7a), comparable to the number of genes downregulated under these conditions (N=616). Thus, mammalian Integrator appears capable of repressing as well as activating gene transcription. Importantly, despite the lower resolution of chromatin RNA-seq, increased transcript levels in Integrator-depleted cells are apparent within the initially transcribed region (Figure 4.7a), as observed in the *Drosophila* system.

The *JUNB* gene, which is a defined target of Integrator (Gardini et al. 2014), is strongly upregulated in un-stimulated HeLa cells (Figure 4.7b), consistent with earlier work (Skaar et al. 2015). Moreover, many characterized immediate early genes experience elevated transcription under these conditions and enriched Gene Ontology categories for upregulated transcripts include receptor and EGF pathways. Interestingly, there is a concordance between upregulated pathways in *Drosophila* and human cells, supporting a functional conservation of Integrator activity within specific pathways. Importantly, these findings suggest that basal upregulation of stimulus-responsive genes upon Integrator depletion may be linked to the defective induction of these genes upon activation of signaling cascades.

To further probe the parallels between Integrator-mediated gene repression in *Drosophila* and human cells, we determined whether Integrator-repressed human genes also displayed chromatin features indicative of defective transcription elongation, such as reduced H3K36me3 and H3K4me3. As is seen in *Drosophila* (Figure 4.1b), both of these histone modifications were significantly lower at human genes upregulated upon Integrator depletion as compared to unchanged genes (Figures 4.7c and 4.7d). In addition, these genes showed enrichment in H3K4me1, a feature of both *Drosophila* Integrator gene targets and enhancers (Figure 4.7e). Thus, the significant commonalities among *Drosophila* and human genes repressed by Integrator, suggest a conserved mechanism across metazoan

species (Figure 4.7f), wherein Integrator targets promoter-proximal elongation complexes at a set of genes to repress gene activity.

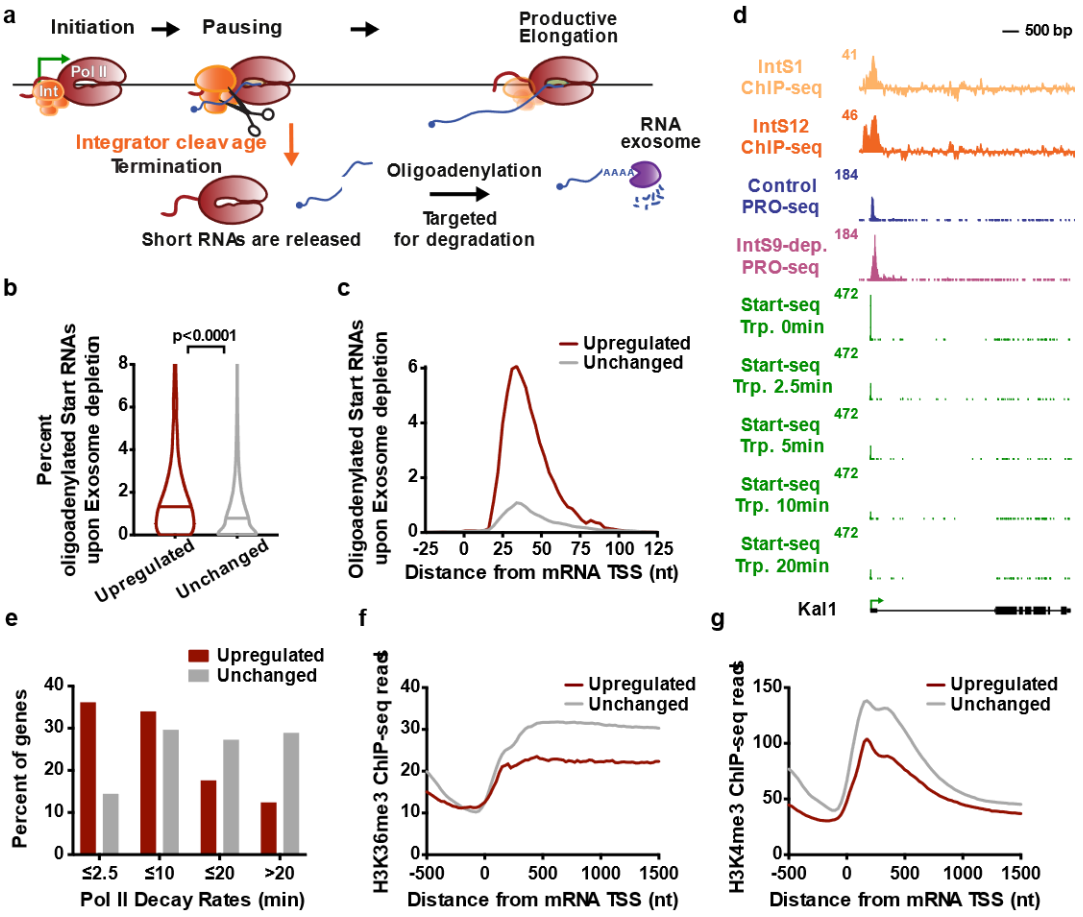


Figure 4.7: The Integrator complex represses expression of mammalian protein-coding genes

(a) Average distribution of chromatin RNA-seq reads in control and IntS11-depleted HeLa cells is shown for genes upregulated upon IntS11-depletion (data from Lai et al, 2015).
(b) JUN locus showing upregulation of transcription upon IntS11-depletion. Shown are profiles of chromatin RNA-seq in control and IntS11-depleted HeLa cells (data from Lai et al, 2015).
(c-d) Average distribution of (c) H3K36me3 and (d) H3K4me3 histone modifications (data from ENCODE project) is shown around mRNA TSSs for Upregulated (N=667) and unchanged (N=15979) genes.
(e) H3K36me3, H3K4me3 and H3K4me1 ChIP-seq levels are shown for upregulated and unchanged genes. Violin plots show range of values, with a line indicating median. P-values are calculated using a Mann-Whitney test.
(f-g) Schematic representation of the effect of the Integrator complex at protein-coding and enhancer loci.
(Figure and legend text reprinted with permission from Elrod ND, Henriques T, Huang K.L., Tatomer DC, Wilusz JE, Wagner EJ, Adelman K. The Integrator complex terminates promoter-proximal transcription at protein-coding genes. Mol. Cell Nov;76(5):738-752.e7 doi: 10.1016/j.molcel.2019.10.034 2019.

SUMMARY

Collectively, our results demonstrate that the Integrator complex mediates transcription attenuation in metazoan cells. This activity involves the association of Integrator with promoter-proximally paused Pol II, cleavage of nascent mRNA transcripts by the Integrator endonuclease, and promoter-proximal termination (Figure 4.7f). This inhibitory function is broad: 15% of *Drosophila* genes and enhancers are impacted by Integrator, with receptor, growth and proliferative pathways particularly affected. Furthermore, the mammalian Integrator complex targets genes in similar pathways for transcriptional repression, underlining the conserved nature of this behavior.

Chapter 5. Integrator 11 is Inhibited by Its Interaction with a Novel Binding Protein

INTRODUCTION

One of the key questions that we faced upon the investigation of IntS11's role in transcriptional attenuation is if that activity is regulated and, if so, how that activity might be regulated. Given the diversity of genes involved, we felt that there had to be at least one regulatory method to control IntS11. Initial searches for possible RNA motifs that might suggest a cleavage site or secondary structure did not reveal any defined consensus sequences. This led us to then ask if there were other factors that could be involved, which in turn led to the study described below. We found that Drosophila CG7044 forms a stable complex with IntS11, and our structure of this complex, determined by cryo-electron microscopy (cryo-EM) at 3.54 Å resolution, suggests that CG7044 is an inhibitor of IntS11 nuclease activity.

RESULTS

CG7044 is a binding partner of IntS11

To gain insight into factors that may regulate IntS11, we analyzed the components of the Drosophila Integrator complex using affinity purification followed by liquid chromatography-mass spectrometry (LC-MS) analysis of their tryptic peptides. We purified Drosophila Integrator from S2 cell nuclear extracts derived from independent cell lines stably expressing FLAG-IntS11, FLAG-IntS5, or not expressing any exogenous FLAG protein (control). As expected, we observed strong enrichment of all 14 Integrator subunits in both FLAG-IntS11 and FLAG-IntS5 purifications relative to control, as well as other factors commonly associated with Integrator, including PP2A subunits (Figure 5.1a). Unexpectedly, there was one additional protein, CG7044, present in high quantities and

unique to FLAG-IntS11 purifications. To further probe CG7044, we expanded the analysis to include purifying complexes from cells stably expressing FLAG-IntS1 and FLAG-IntS8. Analysis of these purified complexes by LC-MS (Figure 5.1b), and Western blotting confirmed that CG7044 appears to be associated exclusively with IntS11. To definitively determine whether any Integrator subunits beyond IntS11 are associated with CG7044, we generated nuclear extracts from S2 cells expressing FLAG-CG7044 and found high levels of IntS11 associated with CG7044, and, except for a small amount of IntS9, no other Integrator subunit was detected (Figure 5.1c). Altogether, these results indicate that a previously uncharacterized protein, CG7044, associates with IntS11, and this complex is distinct from Integrator.

IntS11 is in an inactive conformation in the CG7044 complex

To gain molecular insight into the IntS11-CG7044 association, we co-expressed and purified the *Drosophila* IntS11-CG7044 complex using baculovirus-infected insect cells, confirming that the two proteins form a stable complex. Further, we could also observe that purified human IntS11 forms a stable complex with Brat1 (Aglipay et al. 2006; Van Ommeren et al. 2018), the human ortholog of CG7044. We determined the structure of the *Drosophila* IntS11-CG7044 complex at 3.54 Å resolution by cryo-EM. Most of the residues of CG7044 could be identified (Figure 5.1d), with good sidechain density.

EM density for the metallo-β-lactamase and β-CASP domains of IntS11 is observed, while no density was observed for the two CTDs of IntS11 (Figures 5.1e,f), suggesting that they are disordered in the complex with CG7044. The metallo-β-lactamase and β-CASP domains are in contact with each other and there is no canyon between them that would allow the RNA substrate to reach the active site, as was observed for the active form of CPSF73 (Sun et al. 2020). Therefore, IntS11 appears to be in an inactive state in this structure.

The structure also reveals that CG7044 primarily forms a large Arm/HEAT domain, covering residues 1-932 (Figures 5.1e,f). The 15 pairs of anti-parallel helices are arranged in the shape of a horseshoe that wraps around IntS11 (Figure 5.1e). The connection within the last repeat is exceptionally long, covering residues 830-920 (Figure 5.1d). This extended connection forms a 'lasso' structure, with the two helices at the tip of this lasso being projected more than 30 Å away from the body of the CG7044, where they interact with the β-CASP domain of IntS11 (Figure 5.1f). The C-terminal extension (CTE) of CG7044 beyond the Arm/HEAT domain traverses the open end of the horseshoe, giving the overall structure a circular shape.

A total of 3,900 $Å^2$ of the surface area of CG7044 is buried in the complex with IntS11, while 3,500 $Å^2$ of the surface area of IntS11 is buried in the complex. This extensive buried surface area suggests that the IntS11-CG7044 complex is very stable. The Arm/HEAT domain of CG7044 contributes 1,750 $Å^2$ to the buried surface area of CG7044, and residues in the two helices of the lasso (αL1 and αL2) contribute 960 $Å^2$ to the buried surface area. The helices contact the 'back' face of IntS11, interacting primarily with the β-CASP domain (Figure 5.1f). Most of the residues from the two helices that interact with IntS11 are highly conserved among CG7044 homologs. On the other hand, the linkers from these two helices to the Arm/HEAT domain are poorly conserved, suggesting that this structural feature is likely highly dynamic.

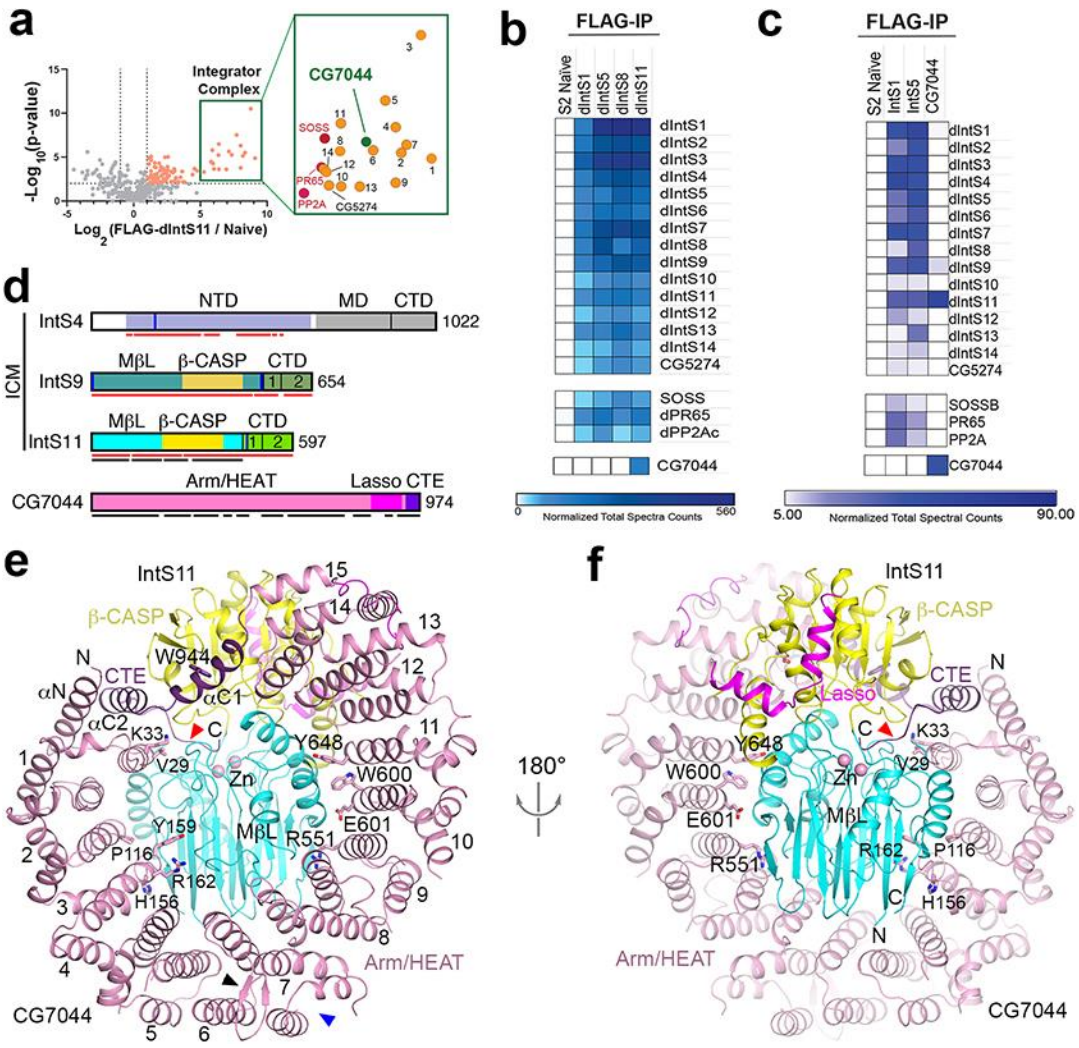


Figure 5.1: CG7044 uniquely associates with IntS11 and the overall structure of the IntS11-CG7044 complex\

(a) Volcano plot of purifications using FLAG-dIntS11 relative to naive control. Integrator subunits are labeled in orange, Integrator associated proteins are in red. Proteins shown in pink are statistically significant, and gray are not significant or unenriched.
(b,c) Heatmap derived from immunoprecipitation (IP) LC-MS analysis of FLAG IP from S2 cells expressing the indicated FLAG-tagged proteins. Heatmaps reflect normalized spectral counts observed from analysis of samples performed in triplicate. Control cells lack any exogenously expressed FLAG-tagged protein.
(d) Domain organizations of *Drosophila* IntS4, IntS9, IntS11, and CG7044. The domains are named and given different colors. The domains of IntS9 are shown in slightly darker colors as compared to IntS11. Residues observed in the structure of the IntS11-CG7044 complex are indicated with the black lines, while those observed in the IntS4-IntS9-IntS11 complex are indicated with the red lines. The vertical bars in blue represent

positively charged residues in the IP_6 binding site. ICM: Integrator cleavage module; NTD: N-terminal domain; MD: middle domain; CTD: C-terminal domain; CTE: C-terminal extension.
(e) The overall structure of the IntS11-CG7044 complex. The domains are colored as in Figure 5.1d. The repeats in the Arm/HEAT domain of CG7044 are labeled. The red arrowhead indicates the C-terminal end of CG7044 in the active site of IntS11. The black arrowhead indicates the β-hairpin linker between the two helices of repeat 7, and the blue arrowhead indicates the insert of two anti-parallel helices in the linker between repeats 7 and 8. CG7044 residues with >50 $Å^2$ buried surface area in the interface with IntS11 are shown in stick models and labeled. MβL: metallo-β-lactamase.
(f) The overall structure of the IntS11-CG7044 complex, viewed after 180° rotation around the vertical axis. The back face of IntS11 and the lasso of CG7044 are visible. The structure figures were produced with PyMOL (www.pymol.org) unless otherwise indicated.

CG7044 inhibits IntS11 through residues at its C-terminus

IntS11 is in an inactive conformation in the complex with CG7044. In fact, even if IntS11 could assume an active conformation in this complex, CG7044 would block the two ends of the canyon and it would still be unlikely for the RNA to be able to access the active site. Moreover, the structure unexpectedly reveals that the C-terminal end of CG7044 is inserted into the active site region of IntS11 (Figure 5.1e), contributing 800 $Å^2$ to the buried surface area. The last three residues of CG7044, Asp972-Cys-Tyr974 (DCY), are in a deep pocket at the interface between the metallo-β-lactamase and β-CASP domains (Figures 5.1e, 2a). These DCY residues are conserved among all known CG7044 and Brat1 homologs (Figure 5.2b), underscoring their importance in the interaction. The side chain of Asp972 is hydrogen-bonded to His392 of IntS11 (Figure 5.2c), the general acid for the nuclease reaction, which is activated by Glu203 (Mandel et al. 2006; Sun et al. 2020). The side chain of Cys973 is directly coordinated to both zinc ions in the active site, which would not allow the scissile phosphate of the RNA substrate to coordinate to the zinc ions (Sun et al. 2020). Finally, Tyr974 is surrounded by residues in the metallo-β-lactamase and β-CASP domains of IntS11. The carboxylate group at the C-terminus of CG7044 has ion-pair interactions with Arg244 in the β-CASP domain.

Based upon these observations, the C-terminal residues of CG7044 would directly compete against the RNA substrate, further ensuring the inhibition of IntS11. In fact, the binding mode of DCY has extensive overlap with that of the RNA substrate observed in the structure of CPSF73 (Figure 5.2d) (Sun et al. 2020). Especially, the side chain of Tyr974 is located in generally the same position as the base of the nucleotide just 3′ to the cleavage site (+1 nucleotide). Overall, the structural observations predict that CG7044 is an inhibitor of IntS11.

To assess the structural observations, we deleted various portions of the CG7044 CTE and lasso and tested their impact on complex formation with IntS11. Deleting the C-

terminal end (Δ966-974) or the CTE (Δ947-974) could not abolish the complex with IntS11, consistent with the structural observations that the Arm/HEAT domain of CG7044 has extensive interactions with IntS11. The expression level of the mutant lacking the lasso was too low and could not be studied.

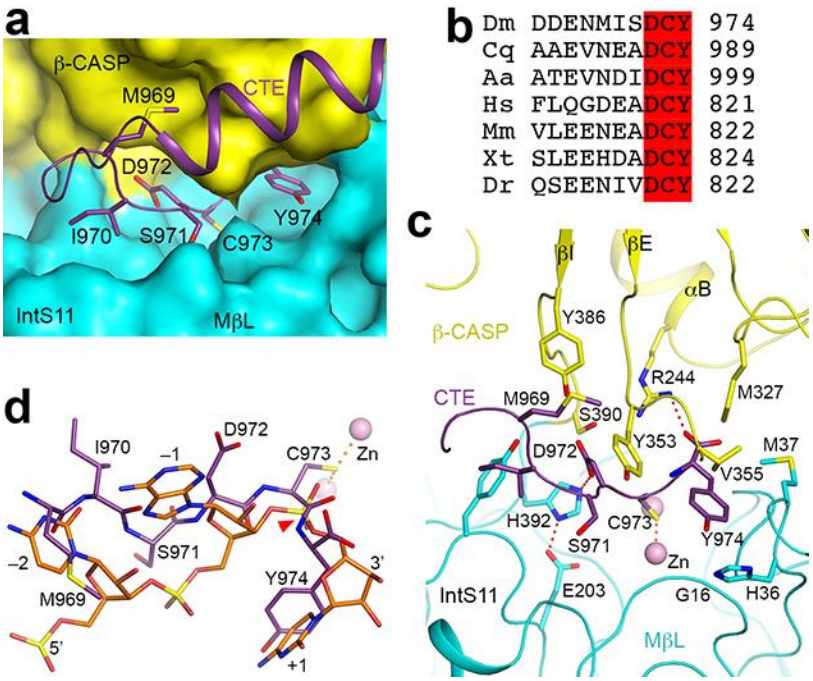


Figure 5.2: The C-terminal residues of CG7044 are located in the active site of IntS11

(a) Residues 972-974 of CG7044 CTE (violet) are located in a pocket at the interface of the metallo-β-lactamase and β-CASP domains of IntS11.
(b) Alignment of the C-terminal residue sequences of selected CG7044 and Brat1 homologs. Conserved residues are highlighted in red. Dm: *D. melanogaster*; Cq: *Culex quinquefasciatus*; Aa: *Aedes aegypti*; Hs: *Homo sapiens* (human); Mm: *Mus musculus* (mouse); Xt: *Xenopus tropicalis* (frog); Dr: *Danio rerio* (zebrafish).
(c) Detailed interactions between the C-terminal end of CG7044 and IntS11. Hydrogen-bonding interactions are indicated with the dashed lines in red. The secondary structure elements in IntS11 are named according to those in CPSF73 (Mandel et al. 2006).
(d) Overlay of the binding mode of the C-terminal end of CG7044 (violet) with that of the histone pre-mRNA substrate in human CPSF73 (orange) (Sun et al. 2020)

To probe the function of the IntS11-CG7044 complex, we used RNAi to deplete either protein from S2 cells and then monitored the impact on Integrator function. We observed upon effective depletion of CG7044 that there was also significant co-depletion of IntS11, suggesting that IntS11 accumulation is highly dependent on available CG7044. The reverse effect was not observed, nor did we find other Integrator subunits whose accumulation was impacted by CG7044 depletion. We measured the degree of either U2snRNA or U4snRNA misprocessing and found that while depletion of IntS11 led to significant increases, the depletion of CG7044 did not have any effect. Similarly, depletion of CG7044 did not affect the expression of mRNAs subject to Integrator attenuation. It may be possible that the amount of IntS11 expression that remains upon CG7044 depletion is sufficient to support its function in cells, and further studies are needed to characterize this complex fully.

IntS11 is in a semi-open state in complex with CG7044

To gain further insight into whether there are conformational changes in IntS11 upon complex formation with CG7044, we co-expressed and purified the IntS4-IntS9-IntS11 complex (the *Drosophila* ICM (Albrecht et al. 2018)) and determined its structure at 2.74 Å resolution by cryo-EM. Most of the modeled residues have good sidechain density (Figure 5.1d). The overall structure of *Drosophila* ICM is generally similar to that of human ICM (Zheng et al. 2020; Pfleiderer and Galej 2021).

The structure shows that the C-terminal segments of IntS9 and IntS11 contain two separate domains, CTD1 and CTD2 (Figures 5.1d and 5.3a), similar to their paralogs CPSF100 and CPSF73 (Sun et al. 2020). The two CTD2 domains have weak EM density, and their atomic models were guided by the structure of the human IntS9-IntS11 CTD2 complex (Wu et al. 2017). The CTDs have extensive interactions with each other, which should facilitate the association of IntS9 and IntS11. The metallo-β-lactamase and β-CASP domains of IntS9 and IntS11 form a pseudo-dimer in this structure (Figure 5.3a),

remarkably similar to the pseudo-dimer for the equivalent domains of CPSF100 and CPSF73 in the active U7 machinery (Sun et al. 2020). The N-terminal domain (NTD) of IntS4 contacts the metallo-β-lactamase domain of IntS9 and the back face of IntS11 metallo-β-lactamase and β-CASP domains (Figure 5.3a), which may promote the formation of this pseudo-dimer.

IntS11 is in a closed, inactive state in our structure of ICM, as well as those reported recently (Zheng et al. 2020; Pfleiderer and Galej 2021). Remarkably, compared to the structure of IntS11 in the CG7044 complex, which is also in an inactive state, there is a sizeable conformational difference for the β-CASP domain, corresponding to a rotation of 6.3° relative to the metallo-β-lactamase domain (Figure 5.3b). This change is distinct from that for the open-closed transition of CPSF73 (Sun et al. 2020) and does not create a canyon for RNA binding.

Therefore, IntS11 assumes a new state in the CG7044 complex, and we will refer to it as a semi-open state, which is stabilized by CG7044. Significantly, the first helix of CG7044 repeat 1 is positioned directly at the interface between the two IntS11 domains, whereas the last helix of the CTE (αC2) clashes with the position of the β-CASP domain in the closed state (Figure 5.3b). These interactions occur near the binding pocket for the C-terminal end of CG7044, promoting the formation of this pocket (Figure 5.2a). The pocket does not exist in the closed state of IntS11 in the ICM, as it is occupied by residues in the β-CASP domain. Conformational changes in several other regions of IntS11 are also observed.

There is also a significant conformational change for helices αC and αD of the β-CASP domain upon binding the lasso of CG7044. Strikingly, the αL2 helix of the lasso invades the β-CASP domain and displaces the αD helix, which becomes disordered. Overall, the structural analysis suggests that the lasso of CG7044 has more robust interactions with the β-CASP domain compared to IntS4.

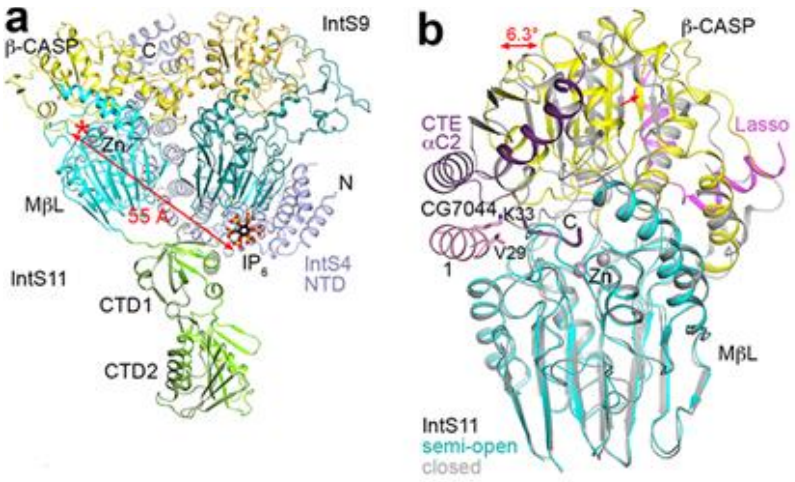


Figure 5.3: An CG7044 binding sites in the IntS4-IntS9-IntS11 complex

(a). The overall structure of the IntS4-IntS9-IntS11 complex, also known as the ICM. The domains are colored as in Figure 5.2a and labeled. The IP6 molecule is shown as stick models (black for carbon atoms).

(b). Overlay of the metallo-β-lactamase and β-CASP domains of IntS11 in the CG7044 complex (in color) with those in the closed state in ICM. Part of the CTE (violet), the lasso (magenta), and a helix of the first repeat of the Arm/HEAT domain (pink) of CG7044 are also shown.

SUMMARY

These findings show that IntS11 has a different binding partner than the canonical Integrator complex in the form of CG7044. More importantly, CG7044 might act as a stabilizer and inhibitor of IntS11. There are still questions that remain mainly in the form of how CG7044 and IntS11's interaction is regulated to inactivate IntS11's transcriptional attenuation.

Chapter 6. Discussion

Our data indicate that the Integrator complex attenuates the expression of protein-coding genes by catalyzing premature transcription termination (Figure 3.7C). The IntS11 endonuclease cleaves a subset of nascent mRNAs, which ultimately triggers degradation of the transcripts by the RNA exosome along with RNAPII termination. We also found that IntS11 binds closely with a novel partner in a manner that would inhibit its catalytic function.

We suggest that many protein-coding genes are negatively regulated via this attenuation mechanism, and the Drosophila MtnA promoter highlights context-specific regulation by Integrator. In addition to cleaving MtnA transcripts, Integrator cleaves multiple other RNA classes in metazoan cells, including enhancer RNAs (Lai et al. 2015), snRNAs (Baillat et al. 2005), telomerase RNA (Rubtsova et al. 2019), and some herpesvirus microRNA precursors (Cazalla et al. 2011; Xie et al. 2015). Using RNA-seq, we expanded this list of Integrator target loci and identified hundreds of additional protein-coding genes that are negatively regulated by Integrator (Figure 3.5). We focused on a set of Integrator-dependent genes and found that Integrator catalyzes pre-mature transcription termination of these genes (Figures 3.6, 3.7), which is consistent with prior studies that suggested roles for Integrator in termination (Skaar et al. 2015; Shah et al. 2018; Gomez-Orte et al. 2019). Some of these genes (CG8620, Pepck1, and Sirup) have promoter-proximal RNAPII that rapidly turns over (Shao and Zeitlinger 2017), which may indicate that Integrator can aid in clearing paused or stalled RNAPII. Once Integrator has cleaved the nascent mRNAs, we find that they are rapidly degraded from their 3′ ends by the RNA exosome (Figure 3.7C).

Collectively, our results demonstrate that the Integrator complex mediates transcription attenuation in metazoan cells. This activity involves the association of Integrator with promoter-proximally paused Pol II, cleavage of nascent mRNA transcripts

by the Integrator endonuclease, and promoter-proximal termination (Figure 4.7F). This inhibitory function is broad: 15% of *Drosophila* genes and enhancers are impacted by Integrator, with receptor, growth and proliferative pathways particularly affected. Furthermore, the mammalian Integrator complex targets genes in similar pathways for transcriptional repression, underlining the conserved nature of this behavior.

These data resolve long-standing questions about the intrinsic stability of promoter-proximal Pol II. We demonstrate that genes that harbor highly unstable promoter Pol II are those in which there is an active process of termination, catalyzed by the Integrator complex. Our data support a model wherein the paused polymerase is inherently stable in the absence of termination factors, consistent with a wealth of biochemical characterization of elongation complexes (Wilson et al. 1999; Kireeva et al. 2000). Thus, we propose that rapid turnover of promoter Pol II at specific genes results from a regulated process of Integrator-mediated RNA cleavage and active dissociation of Pol II from the DNA template.

The attenuation activity we uncover here for Integrator at protein-coding genes and enhancers parallels that described at snRNA genes, where Integrator cleaves the nascent RNA and promotes Pol II termination (Hernandez 1985; Cazalla et al. 2011; Baillat and Wagner 2015; Xie et al. 2015). Therefore, our model for Integrator function is parsimonious with its previously defined biochemical activities. Moreover, consistent with IntS9 and IntS11 subunits being paralogs of CPSF100 and CPSF73, respectively, there are many similarities between premature Pol II termination caused by Integrator, and mRNA cleavage and termination by the CPA machinery. We note that mRNA cleavage and termination at gene ends is coupled with polyadenylation to protect the released mRNA. Likewise, Integrator-catalyzed cleavage of snRNAs is coupled to proper 3' end biogenesis. In contrast, termination driven by Integrator at protein-coding and enhancer loci would typically be followed by RNA degradation (Ogami et al. 2017). These results indicate that the Integrator endonuclease activity can be deployed for different purposes at different loci,

with the outcome governed by the locus-specific recruitment of RNA processing or RNA decay machineries. Therefore, probing the interplay between Integrator and the complexes that govern RNA fate is an area that merits future study.

It has been established that cleavage and termination by the CPA machinery is greatly facilitated by pausing of Pol II (Proudfoot 2016), as is snRNA 3'end formation by Integrator (Guiro and Murphy 2017). Current models invoke a kinetic competition between Pol II elongation and termination, wherein slowed transcription elongation provides a greater window of opportunity for termination to occur (McDowell et al. 1994; Fong et al. 2015). Consistent with these models, we find that promoter-proximally paused Pol II is an optimal target for Integrator-mediated cleavage and termination at mRNA and eRNA loci. Our findings thus suggest a novel function for Pol II pausing in early elongation, wherein pausing provides a regulatory opportunity that enables gene attenuation.

It is interesting that Integrator-repressed genes, which exhibit very low levels of productive elongation, have chromatin characteristics that are common at enhancers. In particular, these genes display low levels of active histone modifications H3K4me3 and H3K36me3, with an enrichment in H3K4me1. Like at Integrator-repressed genes, transcription at enhancers is known to be non-productive, with a highly unstable Pol II that yields only short, rapidly degraded RNAs (Kim and Shiekhattar 2015; Henriques et al. 2018). Thus, our data support models wherein these chromatin features reflect the level and productivity of transcription at the locus, rather than specifically demarcating the coding vs. non-coding potential of the region (Core et al. 2014; Andersson et al. 2015; Soares et al. 2017; Henriques et al. 2018).

Taken together, the role we describe here for Integrator in determining the fate of promoter Pol II sheds new light on Integrator function in development and disease states. Mutations in Integrator have been associated with a myriad of diseases (Rienzo and Casamassimi 2016), with each of the 14 Integrator subunits implicated in one or more disorders. Intriguingly, many of these disease states are not characterized by defects in

splicing and are often associated with disruption in normal development (Rienzo and Casamassimi 2016). Thus, human genetics foretold that Integrator functions extend well beyond snRNA processing. Accordingly, we find that Integrator targets a set of stimulus- and developmentally-responsive genes to potently repress their activity. It will be interesting in future work to tease out the specific roles of the individual Integrator subunits in gene regulation, in the hopes of exploiting this knowledge for therapeutic benefit.

IntS11 is in an inactive conformation in the complex with CG7044. In fact, even if IntS11 could assume an active conformation in this complex, CG7044 would block the two ends of the canyon and it would still be unlikely for the RNA to be able to access the active site. Moreover, the structure unexpectedly reveals that the C-terminal end of CG7044 is inserted into the active site region of IntS11 (Figure 5.1e), contributing 800 $Å^2$ to the buried surface area. The last three residues of CG7044, Asp972-Cys-Tyr974 (DCY), are in a deep pocket at the interface between the metallo-β-lactamase and β-CASP domains (Figures 5.1e, 5.2a). These DCY residues are conserved among all known CG7044 and Brat1 homologs (Figure 5.2b), underscoring their importance in the interaction. The side chain of Asp972 is hydrogen-bonded to His392 of IntS11 (Figure 5.2c), the general acid for the nuclease reaction, which is activated by Glu203 (Mandel et al. 2006; Sun et al. 2020). The side chain of Cys973 is directly coordinated to both zinc ions in the active site, which would not allow the scissile phosphate of the RNA substrate to coordinate to the zinc ions (Sun et al. 2020). Finally, Tyr974 is surrounded by residues in the metallo-β-lactamase and β-CASP domains of IntS11. The carboxylate group at the C-terminus of CG7044 has ion-pair interactions with Arg244 in the β-CASP domain.

Based upon these observations, the C-terminal residues of CG7044 would directly compete with the RNA substrate, further ensuring the inhibition of IntS11. In fact, the binding mode of DCY has extensive overlap with that of the RNA substrate observed in the structure of CPSF73 (Figure 5.2d) (Sun et al. 2020). Especially, the side chain of Tyr974 is located in generally the same position as the base of the nucleotide just 3′ to the

cleavage site (+1 nucleotide). Overall, the structural observations demonstrate that CG7044 is an inhibitor of IntS11.

All together, we have demonstrated a novel mechanism of promoter-proximal termination of RNAPII through the endonuclease activity of IntS11 and have begun to discover a method of control for IntS11. However, many questions remain to be explored in this exciting new mechanism of mRNA regulation. How does IntS11 get recruited to these particular mRNA and what is the signal for its cleavage? Is it just a spatial-temporal effect of general RNAPII recruitment of Integrator or is there a finer control has yet to be elucidated? What signals the inhibition of IntS11 by CG7044 and by what means is this inhibition release? These and many more questions remain for IntS11 and or further understanding of RNAPII regulation.

References

Adelman K, Lis JT. 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**: 720-731.

Aglipay JA, Martin SA, Tawara H, Lee SW, Ouchi T. 2006. ATM activation by ionizing radiation requires BRCA1-associated BAAT1. *J Biol Chem* **281**: 9710-9718.

Albrecht TR, Shevtsov SP, Wu Y, Mascibroda LG, Peart NJ, Huang KL, Sawyer IA, Tong L, Dundr M, Wagner EJ. 2018. Integrator subunit 4 is a 'Symplekin-like' scaffold that associates with INTS9/11 to form the Integrator cleavage module. *Nucleic acids research* **46**: 4241-4255.

Albrecht TR, Wagner EJ. 2012. snRNA 3' end formation requires heterodimeric association of integrator subunits. *Mol Cell Biol* **32**: 1112-1123.

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

Andersson R, Sandelin A, Danko CG. 2015. A unified architecture of transcriptional regulatory elements. *Trends Genet* **31**: 426-433.

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**: 1074-1077.

Austenaa LM, Barozzi I, Simonatto M, Masella S, Della Chiara G, Ghisletti S, Curina A, de Wit E, Bouwman BA, de Pretis S. 2015. Transcription of mammalian cis-regulatory elements is restrained by actively enforced early termination. *Molecular cell* **60**: 460-474.

Baillat D, Hakimi MA, Naar AM, Shilatifard A, Cooch N, Shiekhattar R. 2005. Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell* **123**: 265-276.

Baillat D, Wagner EJ. 2015. Integrator: surprisingly diverse functions in gene expression. *Trends Biochem Sci* **40**: 257-264.

Barbieri E, Trizzino M, Welsh SA, Owens TA, Calabretta B, Carroll M, Sarma K, Gardini A. 2018. Targeted Enhancer Activation by a Subunit of the Integrator Complex. *Mol Cell* **71**: 103-116 e107.

Baumann DG, Gilmour DS. 2017. A sequence-specific core promoter-binding transcription factor recruits TRF2 to coordinately transcribe ribosomal protein genes. *Nucleic acids research* **45**: 10481-10491.

Beckedorff F, Blumenthal E, daSilva LF, Aoi Y, Cingaram PR, Yue J, Zhang A, Dokaneheifard S, Valencia MG, Gaidosh G et al. 2020. The Human Integrator

Complex Facilitates Transcriptional Elongation by Endonucleolytic Cleavage of Nascent Transcripts. *Cell Rep* **32**: 107917.

Beltran T, Pahita E, Ghosh S, Lenhard B, Sarkies P. 2021. Integrator is recruited to promoter-proximally paused RNA Pol II to generate Caenorhabditis elegans piRNA precursors. *EMBO J* **40**: e105564.

Berg MG, Singh LN, Younis I, Liu Q, Pinto AM, Kaida D, Zhang Z, Cho S, Sherrill-Mix S, Wan L et al. 2012. U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150**: 53-64.

Brannan K, Kim H, Erickson B, Glover-Cutter K, Kim S, Fong N, Kiemele L, Hansen K, Davis R, Lykke-Andersen J et al. 2012. mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. *Mol Cell* **46**: 311-324.

Bresson S, Tollervey D. 2018. Surveillance-ready transcription: nuclear RNA decay as a default fate. *Open Biol* **8**.

Buckley MS, Kwak H, Zipfel WR, Lis JT. 2014. Kinetics of promoter Pol II on Hsp70 reveal stable pausing and key insights into its regulation. *Genes & development* **28**: 14-19.

Burley SK, Roeder RG. 1996. Biochemistry and structural biology of transcription factor IID (TFIID). *Annu Rev Biochem* **65**: 769-799.

Cazalla D, Xie M, Steitz JA. 2011. A primate herpesvirus uses the integrator complex to generate viral microRNAs. *Mol Cell* **43**: 982-992.

Chalamcharla VR, Folco HD, Dhakshnamoorthy J, Grewal SI. 2015. Conserved factor Dhp1/Rat1/Xrn2 triggers premature transcription termination and nucleates heterochromatin to promote gene silencing. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 15548-15555.

Chapman RD, Heidemann M, Hintermair C, Eick D. 2008. Molecular evolution of the RNA polymerase II CTD. *Trends Genet* **24**: 289-296.

Chen F, Gao X, Shilatifard A. 2015. Stably paused genes revealed through inhibition of transcription initiation by the TFIIH inhibitor triptolide. *Gene Dev* **29**: 39-47.

Chen J, Ezzeddine N, Waltenspiel B, Albrecht TR, Warren WD, Marzluff WF, Wagner EJ. 2012. An RNAi screen identifies additional members of the Drosophila Integrator complex and a requirement for cyclin C/Cdk8 in snRNA 3'-end formation. *RNA* **18**: 2148-2156.

Chen N, Zheng Y, Yin J, Li X, Zheng C. 2013. Inhibitory effects of silver nanoparticles against adenovirus type 3 in vitro. *J Virol Methods* **193**: 470-477.

Chiu AC, Suzuki HI, Wu X, Mahat DB, Kriz AJ, Sharp PA. 2018. Transcriptional Pause Sites Delineate Stable Nucleosome-Associated Premature Polyadenylation Suppressed by U1 snRNP. *Mol Cell* **69**: 648-663 e647.

Conaway RC, Conaway JW. 1993. General initiation factors for RNA polymerase II. *Annu Rev Biochem* **62**: 161-190.

Consortium GT, Laboratory DA, Coordinating Center -Analysis Working G, Statistical Methods groups-Analysis Working G, Enhancing Gg, Fund NIHC, Nih/Nci, Nih/Nhgri, Nih/Nimh, Nih/Nida et al. 2017. Genetic effects on gene expression across human tissues. *Nature* **550**: 204-213.

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

Core LJ, Lis JT. 2008. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science* **319**: 1791-1792.

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature genetics* **46**: 1311-1320.

Darnell JE, Jr., Kerr IM, Stark GR. 1994. Jak-STAT pathways and transcriptional activation in response to IFNs and other extracellular signaling proteins. *Science* **264**: 1415-1421.

Dominski Z, Yang X-C, Purdy M, Wagner E, Marzluff W. 2005. *A CPSF-73 Homologue Is Required for Cell Cycle Progression but Not Cell Growth and Interacts with a Protein Having Features of CPSF-100*.

Dower K, Kuperwasser N, Merrikh H, Rosbash M. 2004. A synthetic A tail rescues yeast nuclear accumulation of a ribozyme-terminated transcript. *RNA* **10**: 1888-1899.

Eaton JD, Davidson L, Bauer DLV, Natsume T, Kanemaki MT, West S. 2018. Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity. *Gene Dev* **32**: 127-139.

Eaton JD, Francis L, Davidson L, West S. 2020. A unified allosteric/torpedo mechanism for transcriptional termination on human protein-coding genes. *Genes & Development* **34**: 132-145.

Egloff S, O'Reilly D, Chapman RD, Taylor A, Tanzhaus K, Pitts L, Eick D, Murphy S. 2007. Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. *Science* **318**: 1777-1779.

Egloff S, Szczepaniak SA, Dienstbier M, Taylor A, Knight S, Murphy S. 2010. The integrator complex recognizes a new double mark on the RNA polymerase II carboxyl-terminal domain. *J Biol Chem* **285**: 20564-20569.

Elrod ND, Henriques T, Huang KL, Tatomer DC, Wilusz JE, Wagner EJ, Adelman K. 2019. The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes. *Mol Cell* **76**: 738-752.e737.

Emsley P, Cowtan K. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* **60**: 2126-2132.

ENCODE-Project-Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57.

Erickson B, Sheridan RM, Cortazar M, Bentley DL. 2018. Dynamic turnover of paused Pol II complexes at human promoters. *Gene Dev* **32**: 1215-1225.

Ezzeddine N, Chen J, Waltenspiel B, Burch B, Albrecht T, Zhuo M, Warren WD, Marzluff WF, Wagner EJ. 2011. A subset of Drosophila integrator proteins is essential for efficient U7 snRNA and spliceosomal snRNA 3'-end formation. *Mol Cell Biol* **31**: 328-341.

Flanagan PM, Kelleher RJ, 3rd, Sayre MH, Tschochner H, Kornberg RD. 1991. A mediator required for activation of RNA polymerase II transcription in vitro. *Nature* **350**: 436-438.

Fong N, Brannan K, Erickson B, Kim H, Cortazar MA, Sheridan RM, Nguyen T, Karp S, Bentley DL. 2015. Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Mol Cell* **60**: 256-267.

Fraser NW, Sehgal PB, Darnell JE. 1978. DRB-induced premature termination of late adenovirus transcription. *Nature* **272**: 590-593.

Furuichi Y. 2015. Discovery of m(7)G-cap in eukaryotic mRNAs. *Proc Jpn Acad Ser B Phys Biol Sci* **91**: 394-409.

Fury MG, Zieve GW. 1996. U6 snRNA maturation and stability. *Exp Cell Res* **228**: 160-163.

Gardini A, Baillat D, Cesaroni M, Hu D, Marinis JM, Wagner EJ, Lazar MA, Shilatifard A, Shiekhattar R. 2014. Integrator regulates transcriptional initiation and pause release following activation. *Mol Cell* **56**: 128-139.

Gariglio P, Bellard M, Chambon P. 1981. Clustering of RNA polymerase B molecules in the 5' moiety of the adult beta-globin gene of hen erythrocytes. *Nucleic acids research* **9**: 2589-2598.

Gollnick P, Babitzke P. 2002. Transcription attenuation. *Biochim Biophys Acta* **1577**: 240-250.

Gomez-Orte E, Saenz-Narciso B, Zheleva A, Ezcurra B, de Toro M, Lopez R, Gastaca I, Nilsen H, Sacristan MP, Schnabel R et al. 2019. Disruption of the Caenorhabditis elegans Integrator complex triggers a non-conventional transcriptional mechanism beyond snRNA genes. *PLoS Genet* **15**: e1007981.

Grayhack EJ, Yang X, Lau LF, Roberts JW. 1985. Phage lambda gene Q antiterminator recognizes RNA polymerase near the promoter and accelerates it through a pause site. *Cell* **42**: 259-269.

Green M, Schuetz TJ, Sullivan EK, Kingston RE. 1995. A heat shock-responsive domain of human HSF1 that regulates transcription activation domain function. *Mol Cell Biol* **15**: 3354-3362.

Guiro J, Murphy S. 2017. Regulation of expression of human RNA polymerase II-transcribed snRNA genes. *Open Biol* **7**: 170073.

Gunther V, Lindert U, Schaffner W. 2012. The taste of heavy metals: gene regulation by MTF-1. *Biochim Biophys Acta* **1823**: 1416-1425.

Henkin TM, Yanofsky C. 2002. Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. *Bioessays* **24**: 700-707.

Henriques T, Gilchrist DA, Nechaev S, Bern M, Muse GW, Burkholder A, Fargo DC, Adelman K. 2013. Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. *Mol Cell* **52**: 517-528.

Henriques T, Scruggs BS, Inouye MO, Muse GW, Williams LH, Burkholder AB, Lavender CA, Fargo DC, Adelman K. 2018. Widespread transcriptional pausing and elongation control at enhancers. *Genes & development* **32**: 26-41.

Hernandez N. 1985. Formation of the 3' end of U1 snRNA is directed by a conserved sequence located downstream of the coding region. *EMBO J* **4**: 1827-1837.

Hernandez N, Weiner AM. 1986. Formation of the 3' end of U1 snRNA requires compatible snRNA promoter elements. *Cell* **47**: 249-258.

Ho JW, Jung YL, Liu T, Alver BH, Lee S, Ikegami K, Sohn KA, Minoda A, Tolstorukov MY, Appert A et al. 2014. Comparative analysis of metazoan chromatin organization. *Nature* **512**: 449-452.

Ho SN, Biggar SR, Spencer DM, Schreiber SL, Crabtree GR. 1996. Dimeric ligands define a role for transcriptional activation domains in reinitiation. *Nature* **382**: 822-826.

Huang J, Gong Z, Ghosal G, Chen J. 2009. SOSS complexes participate in the maintenance of genomic stability. *Mol Cell* **35**: 384-393.

Huang KL, Jee D, Stein CB, Elrod ND, Henriques T, Mascibroda LG, Baillat D, Russell WK, Adelman K, Wagner EJ. 2020. Integrator Recruits Protein Phosphatase 2A to Prevent Pause Release and Facilitate Transcription Termination. *Mol Cell* **80**: 345-358 e349.

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

Kadonaga JT, Courey AJ, Ladika J, Tjian R. 1988. Distinct regions of Sp1 modulate DNA binding and transcriptional activation. *Science* **242**: 1566-1570.

Kagey MH, Newman JJ, Bilodeau S, Zhan Y, Orlando DA, van Berkum NL, Ebmeier CC, Goossens J, Rahl PB, Levine SS et al. 2010. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**: 430-435.

Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, Dreyfuss G. 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**: 664-668.

Kamieniarz-Gdula K, Gdula MR, Panser K, Nojima T, Monks J, Wisniewski JR, Riepsaame J, Brockdorff N, Pauli A, Proudfoot NJ. 2019. Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination. *Mol Cell* **74**: 158-172 e159.

Kao S-Y, Calman AF, Luciw PA, Peterlin BM. 1987. Anti-termination of transcription within the long terminal repeat of HIV-1 by tat gene product. *Nature* **330**: 489-493.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**: 1484-1488.

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

Kaye EG, Booker M, Kurland JV, Conicella AE, Fawzi NL, Bulyk ML, Tolstorukov MY, Larschan E. 2018. Differential Occupancy of Two GA-Binding Proteins Promotes Targeting of the Drosophila Dosage Compensation Complex to the Male X Chromosome. *Cell Rep* **22**: 3227-3239.

Kelleher RJ, Flanagan PM, Kornberg RD. 1990. A novel mediator between activator proteins and the RNA polymerase II transcription apparatus. *Cell* **61**: 1209-1215.

Kim H, Erickson B, Luo W, Seward D, Graber JH, Pollock DD, Megee PC, Bentley DL. 2010. Gene-specific RNA polymerase II phosphorylation and the CTD code. *Nat Struct Mol Biol* **17**: 1279-1286.

Kim T-K, Shiekhattar R. 2015. Architectural and functional commonalities between enhancers and promoters. *Cell* **162**: 948-959.

Kireeva ML, Komissarova N, Waugh DS, Kashlev M. 2000. The 8-nucleotide-long RNA: DNA hybrid is a primary stability determinant of the RNA polymerase II elongation complex. *J Biol Chem* **275**: 6530-6536.

Kirstein N, Gomes Dos Santos H, Blumenthal E, Shiekhattar R. 2021. The Integrator complex at the crossroad of coding and noncoding RNA. *Curr Opin Cell Biol* **70**: 37-43.

Kline MP, Morimoto RI. 1997. Repression of the heat shock factor 1 transcriptional activation domain is modulated by constitutive phosphorylation. *Mol Cell Biol* **17**: 2107-2115.

Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L, Schübeler D. 2017. Genome-wide single-molecule footprinting reveals high RNA polymerase II turnover at paused promoters. *Molecular cell* **67**: 411-422. e414.

Krumm A, Meulia T, Brunvand M, Groudine M. 1992. The block to transcriptional elongation within the human c-myc gene is determined in the promoter-proximal region. *Genes & development* **6**: 2201-2213.

Kwak H, Fuda NJ, Core LJ, Lis JT. 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**: 950-953.

LaCava J, Houseley J, Saveanu C, Petfalski E, Thompson E, Jacquier A, Tollervey D. 2005. RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell* **121**: 713-724.

Lai F, Gardini A, Zhang A, Shiekhattar R. 2015. Integrator mediates the biogenesis of enhancer RNAs. *Nature* **525**: 399-403.

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

Larschan E, Bishop EP, Kharchenko PV, Core LJ, Lis JT, Park PJ, Kuroda MI. 2011. X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. *Nature* **471**: 115-118.

Lavender CA, Shapiro AJ, Burkholder AB, Bennett BD, Adelman K, Fargo DC. 2017. ORIO (Online Resource for Integrative Omics): a web-based platform for rapid integration of next generation sequencing data. *Nucleic acids research* **45**: 5678-5690.

Li J, Ma X, Banerjee S, Baruah S, Schnicker NJ, Roh E, Ma W, Liu K, Bode AM, Dong Z. 2021. Structural basis for multifunctional roles of human Ints3 C-terminal domain. *J Biol Chem* **296**: 100112.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)* **30**: 923-930.

Liebschner D, Afonine PV, Baker ML, Bunkoczi G, Chen VB, Croll TI, Hintze B, Hung LW, Jain S, McCoy AJ et al. 2019. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr D Struct Biol* **75**: 861-877.

Lim SJ, Boyle PJ, Chinen M, Dale RK, Lei EP. 2013. Genome-wide localization of exosome components to active promoters and chromatin insulators in Drosophila. *Nucleic acids research* **41**: 2963-2980.

Lubas M, Christensen MS, Kristiansen MS, Domanski M, Falkenby LG, Lykke-Andersen S, Andersen JS, Dziembowski A, Jensen TH. 2011. Interaction profiling identifies the human nuclear exosome targeting complex. *Mol Cell* **43**: 624-637.

Mandel CR, Kaneko S, Zhang H, Gebauer D, Vethantham V, Manley JL, Tong L. 2006. Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature* **444**: 953.

Manley JL, Di Giammartino DC, Zaher H. 2021. Transcription | mRNA Polyadenylation in Eukaryotes. in *Encyclopedia of Biological Chemistry III (Third Edition)* (ed. J Jez), pp. 443-448. Elsevier, Oxford.

Marr MT, 2nd, Isogai Y, Wright KJ, Tjian R. 2006. Coactivator cross-talk specifies transcriptional output. *Gene Dev* **20**: 1458-1469.

Mascibroda LG, Shboul M, Elrod ND, Colleaux L, Hamamy H, Huang K-L, Peart N, Singh MK, Lee H, Merriman B et al. 2020. INTS13 Mutations Causing a Developmental Ciliopathy Disrupt Integrator Complex Assembly. *Nature Communications*.

McDowell JC, Roberts JW, Jin DJ, Gross C. 1994. Determination of intrinsic transcription termination efficiency by RNA polymerase elongation rate. *Science* **266**: 822-825.

Mendoza-Figueroa MS, Tatomer DC, Wilusz JE. 2020. The Integrator Complex in Transcription and Development. *Trends Biochem Sci* **45**: 923-934.

Merran J, Corden JL. 2017. Yeast RNA-Binding Protein Nab3 Regulates Genes Involved in Nitrogen Metabolism. *Mol Cell Biol* **37**: e00154-00117.

Min IM, Waterfall JJ, Core LJ, Munroe RJ, Schimenti J, Lis JT. 2011. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Gene Dev* **25**: 742-754.

mod EC, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L et al. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330**: 1787-1797.

Nacheva GA, Guschin DY, Preobrazhenskaya OV, Karpov VL, Ebralidse KK, Mirzabekov AD. 1989. Change in the pattern of histone binding to DNA upon transcriptional activation. *Cell* **58**: 27-36.

Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K. 2010. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **327**: 335-338.

Nguyen T, Nioi P, Pickett CB. 2009. The Nrf2-antioxidant response element signaling pathway and its activation by oxidative stress. *J Biol Chem* **284**: 13291-13295.

Nilson KA, Lawson CK, Mullen NJ, Ball CB, Spector BM, Meier JL, Price DH. 2017. Oxidative stress rapidly stabilizes promoter-proximal paused Pol II across the human genome. *Nucleic acids research* **45**: 11088-11105.

Ogami K, Richard P, Chen Y, Hoque M, Li W, Moresco JJ, Yates JR, 3rd, Tian B, Manley JL. 2017. An Mtr4/ZFC3H1 complex facilitates turnover of unstable nuclear RNAs to prevent their cytoplasmic transport and global translational repression. *Gene Dev* **31**: 1257-1271.

Ong C-T, Corces VG. 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics* **12**: 283-293.

Peterlin BM, Price DH. 2006. Controlling the elongation phase of transcription with P-TEFb. *Mol Cell* **23**: 297-305.

Pfleiderer MM, Galej WP. 2021. Structure of the catalytic core of the Integrator complex. *Mol Cell* **81**: 1246-1259 e1248.

Plet A, Eick D, Blanchard JM. 1995. Elongation and premature termination of transcripts initiated from c-fos and c-myc promoters show dissimilar patterns. *Oncogene* **10**: 319-328.

Porrua O, Libri D. 2015. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nature reviews Molecular cell biology* **16**: 190-202.

Proudfoot NJ. 2016. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* **352**: aad9926.

Punjani A, Rubinstein JL, Fleet DJ, Brubaker MA. 2017. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat Methods* **14**: 290-296.

Rasmussen EB, Lis JT. 1993. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proceedings of the National Academy of Sciences of the United States of America* **90**: 7923-7927.

Ren W, Chen H, Sun Q, Tang X, Lim SC, Huang J, Song H. 2014. Structural basis of SOSS1 complex assembly and recognition of ssDNA. *Cell Rep* **6**: 982-991.

Rienzo M, Casamassimi A. 2016. Integrator complex and transcription regulation: Recent findings and pathophysiology. *Biochim Biophys Acta* **1859**: 1269-1280.

Roeder RG, Rutter WJ. 1969. Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature* **224**: 234-237.

Rohou A, Grigorieff N. 2015. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J Struct Biol* **192**: 216-221.

Rosa-Mercado NA, Zimmer JT, Apostolidi M, Rinehart J, Simon MD, Steitz JA. 2021. Hyperosmotic stress alters the RNA polymerase II interactome and induces readthrough transcription despite widespread transcriptional repression. *Mol Cell* **81**: 502-513.e504.

Rougvie AE, Lis JT. 1988. The RNA polymerase II molecule at the 5′ end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell* **54**: 795-804.

Rougvie AE, Lis JT. 1990. Postinitiation transcriptional control in Drosophila melanogaster. *Mol Cell Biol* **10**: 6041-6045.

Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**: 725-738.

Rubtsova MP, Vasilkova DP, Moshareva MA, Malyavko AN, Meerson MB, Zatsepin TS, Naraykina YV, Beletsky AV, Ravin NV, Dontsova OA. 2019. Integrator is a key component of human telomerase RNA biogenesis. *Sci Rep* **9**: 1701.

Sabath K, Staubli ML, Marti S, Leitner A, Moes M, Jonas S. 2020. INTS10-INTS13-INTS14 form a functional module of Integrator that binds nucleic acids and the cleavage module. *Nat Commun* **11**: 3422.

Sari D, Gupta K, Thimiri Govinda Raj DB, Aubert A, Drncova P, Garzoni F, Fitzgerald D, Berger I. 2016. The MultiBac Baculovirus/Insect Cell Expression Vector System for Producing Complex Protein Biologics. *Advances in experimental medicine and biology* **896**: 199-215.

Shah N, Maqbool MA, Yahia Y, El Aabidine AZ, Esnault C, Forne I, Decker TM, Martin D, Schuller R, Krebs S et al. 2018. Tyrosine-1 of RNA Polymerase II CTD Controls Global Termination of Gene Transcription in Mammals. *Mol Cell* **69**: 48-61 e46.

Shao W, Zeitlinger J. 2017. Paused RNA polymerase II inhibits new transcriptional initiation. *Nat Genet* **49**: 1045-1051.

Shuman S. 2015. RNA capping: progress and prospects. *RNA* **21**: 735-737.

Skaar JR, Ferris AL, Wu X, Saraf A, Khanna KK, Florens L, Washburn MP, Hughes SH, Pagano M. 2015. The Integrator complex controls the termination of transcription at diverse classes of gene targets. *Cell research* **25**: 288-305.

Smirnova IV, Bittel DC, Ravindra R, Jiang H, Andrews GK. 2000. Zinc and cadmium can promote rapid nuclear translocation of metal response element-binding transcription factor-1. *J Biol Chem* **275**: 9377-9384.

Soares LM, He PC, Chun Y, Suh H, Kim T, Buratowski S. 2017. Determinants of Histone H3K4 Methylation Patterns. *Mol Cell* **68**: 773-785 e776.

Sogaard TM, Svejstrup JQ. 2007. Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator. *J Biol Chem* **282**: 14113-14120.

Sohrabi-Jahromi S, Hofmann KB, Boltendahl A, Roth C, Gressel S, Baejen C, Soeding J, Cramer P. 2019. Transcriptome maps of general eukaryotic RNA degradation factors. *Elife* **8**: e47040.

Stadelmayer B, Micas G, Gamot A, Martin P, Malirat N, Koval S, Raffel R, Sobhian B, Severac D, Rialle S et al. 2014. Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. *Nat Commun* **5**: 5531.

Steurer B, Janssens RC, Geverts B, Geijer ME, Wienholz F, Theil AF, Chang J, Dealy S, Pothof J, van Cappellen WA et al. 2018. Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. *Proceedings of the National Academy of Sciences of the United States of America* **115**: E4368-E4376.

Strobl LJ, Eick D. 1992. Hold back of RNA polymerase II at the transcription start site mediates down-regulation of c-myc in vivo. *The EMBO Journal* **11**: 3307-3314.

Suloway C, Pulokas J, Fellmann D, Cheng A, Guerra F, Quispe J, Stagg S, Potter CS, Carragher B. 2005. Automated molecular microscopy: the new Leginon system. *J Struct Biol* **151**: 41-60.

Sun Y, Zhang Y, Aik WS, Yang XC, Marzluff WF, Walz T, Dominski Z, Tong L. 2020. Structure of an active human histone pre-mRNA 3'-end processing machinery. *Science* **367**: 700-703.

Tettey TT, Gao X, Shao W, Li H, Story BA, Chitsazan AD, Glaser RL, Goode ZH, Seidel CW, Conaway RC et al. 2019. A Role for FACT in RNA Polymerase II Promoter-Proximal Pausing. *Cell Rep* **27**: 3770-3779 e3777.

Thomas QA, Ard R, Liu J, Li B, Wang J, Pelechano V, Marquardt S. 2020. Transcript isoform sequencing reveals widespread promoter-proximal transcriptional termination in Arabidopsis. *Nat Commun* **11**: 2589.

Van Ommeren RH, Gao AF, Blaser SI, Chitayat DA, Hazrati LN. 2018. BRAT1 Mutation: The First Reported Case of Chinese Origin and Review of the Literature. *J Neuropathol Exp Neurol* **77**: 1071-1078.

Venkatesh S, Workman JL. 2015. Histone exchange, chromatin structure and the regulation of transcription. *Nat Rev Mol Cell Biol* **16**: 178-189.

Venters CC, Oh JM, Di C, So BR, Dreyfuss G. 2019. U1 snRNP Telescripting: Suppression of Premature Transcription Termination in Introns as a New Layer of Gene Regulation. *Cold Spring Harb Perspect Biol* **11**.

Vispe S, DeVries L, Creancier L, Besse J, Breand S, Hobson DJ, Svejstrup JQ, Annereau JP, Cussac D, Dumontet C et al. 2009. Triptolide is an inhibitor of RNA polymerase I and II-dependent transcription leading predominantly to down-regulation of short-lived mRNA. *Mol Cancer Ther* **8**: 2780-2790.

Vos SM, Farnung L, Urlaub H, Cramer P. 2018. Structure of paused transcription complex Pol II-DSIF-NELF. *Nature* **560**: 601-606.

Wada T, Takagi T, Yamaguchi Y, Ferdous A, Imai T, Hirose S, Sugimoto S, Yano K, Hartzog GA, Winston F. 1998. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes & development* **12**: 343-356.

Wagner EJ, Carpenter PB. 2012. Understanding the language of Lys36 methylation at histone H3. *Nature reviews Molecular cell biology* **13**: 115-126.

Wagschal A, Rousset E, Basavarajaiah P, Contreras X, Harwig A, Laurent-Chabalier S, Nakamura M, Chen X, Zhang K, Meziane O et al. 2012. Microprocessor, Setx, Xrn2, and Rrp6 co-operate to induce premature termination of transcription by RNAPII. *Cell* **150**: 1147-1157.

Weber CM, Ramachandran S, Henikoff S. 2014. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol Cell* **53**: 819-830.

Wilson KS, Conant CR, von Hippel PH. 1999. Determinants of the stability of transcription elongation complexes: interactions of the nascent RNA with the DNA template and the RNA polymerase. *Journal of molecular biology* **289**: 1179-1194.

Wilusz JE, JnBaptiste CK, Lu LY, Kuhn CD, Joshua-Tor L, Sharp PA. 2012. A triple helix stabilizes the 3' ends of long noncoding RNAs that lack poly(A) tails. *Gene Dev* **26**: 2392-2407.

Wu Y, Albrecht TR, Baillat D, Wagner EJ, Tong L. 2017. Molecular basis for the interaction between Integrator subunits IntS9 and IntS11 and its functional importance. *Proceedings of the National Academy of Sciences* **114**: 4394.

Xie M, Zhang W, Shu MD, Xu A, Lenis DA, DiMaio D, Steitz JA. 2015. The host Integrator complex acts in transcription-independent maturation of herpesvirus microRNA 3' ends. *Gene Dev* **29**: 1552-1564.

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Munster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**: 1033-1037.

Yamaguchi Y, Takagi T, Wada T, Yano K, Furuya A, Sugimoto S, Hasegawa J, Handa H. 1999. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97**: 41-51.

Yamamoto J, Hagiwara Y, Chiba K, Isobe T, Narita T, Handa H, Yamaguchi Y. 2014. DSIF and NELF interact with Integrator to specify the correct post-transcriptional fate of snRNA genes. *Nat Commun* **5**: 4263.

Yanofsky C. 1981. Attenuation in the control of expression of bacterial operons. *Nature* **289**: 751-758.

Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, Stark A. 2015. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**: 556-559.

Zawel L, Reinberg D. 1993. Initiation of Transcription by RNA Polymerase II: A Multi-step Process. in *Progress in Nucleic Acid Research and Molecular Biology* (eds. WE Cohn, K Moldave), pp. 67-108. Academic Press.

Zawel L, Reinberg D. 1995. Common themes in assembly and function of eukaryotic transcription complexes. *Annu Rev Biochem* **64**: 533-561.

Zhang F, Ma T, Yu X. 2013. A core hSSB1-INTS complex participates in the DNA damage response. *J Cell Sci* **126**: 4850-4855.

Zhang W, Liu HT. 2002. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res* **12**: 9-18.

Zhang Y, Sun Y, Shi Y, Walz T, Tong L. 2020. Structural insights into the human pre-mRNA 3'-end processing machinery. *Mol Cell* **77**: 800-809.

Zheng H, Qi Y, Hu S, Cao X, Xu C, Yin Z, Chen X, Li Y, Liu W, Li J et al. 2020. Identification of Integrator-PP2A complex (INTAC), an RNA polymerase II phosphatase. *Science* **370**.

Zinder JC, Lima CD. 2017. Targeting RNA for processing or destruction by the eukaryotic RNA exosome and its cofactors. *Gene Dev* **31**: 88-100.

Zivanov J, Nakane T, Forsberg BO, Kimanius D, Hagen WJ, Lindahl E, Scheres SH. 2018. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife* **7**: e42166.

Vita

Nathan David Elrod was born on July 23, 1979 in Fort Campbell, Kentucky to Diane C. and Anthony L. Elrod. He graduated from James Madison High School in 1998. Nathan then spent the next 10 years working as a Licensed Marine Engineer aboard a

variety of vessels in the Gulf of Mexico and Pacific Ocean. He then attended Tarleton State University from 2009 to 2015 earning a BS in Animal Science, a BS in Molecular Biology, and a MS in Biology graduating Summa Cum Laude. He then attended graduate school at the University of Texas Medical Branch from 2016-2021, defending his doctoral work in September 2021 and graduating with a Ph.D. in Biochemistry and Molecular Biology in December 2021. At time of this writing, Nathan has the following 14 peer-reviewed publications with 174 citations along with 2 more articles under review or revision.

1. Cao J, Verma, SK, Jaworski E, Mohan S, Nagasawa CK, Rayavara K, Sooter A, Miller SN, Holcomb RJ, Ji P, Elrod ND, Yildirim E, Wagner EJ, Popov V, Garg NJ, Routh AL, Kuyumcu-Martinez MN. RBFOX2 is Critical for Maintaining Alternative Polyadenylation Patterns and Mitochondrial Health in Rat Myoblasts. ***Cell Reports***. (In press) 2021.

2. Yalamanchili HK, Elrod ND, Jensen MK, Ji P, Lin A, Wagner EJ, Liu Z. A computational pipeline to infer alternative poly-adenylation from 3'sequencing data. ***Methods in Enzymology***. 2021;655:185-204.

3. Li L, Huang KL, Gao Y, Peng F, Elrod ND, Ji P, Wagner EJ, Li W. Genetic Basis of Alternative Polyadenylation is an Emerging Molecular Phenotype for Human Traits and Diseases. ***Nature Genetics*****.** 2021:1-12.

4. Jensen MK, Elrod ND, Yalamanchili HK, Ji P, Lin A, Liu Z, Wagner EJ. Application and design considerations for 3′-end sequencing using click-chemistry. ***Methods in Enzymology***: Academic Press Inc.; 2021

5. Enwerem II, Elrod ND, Chang CT, Lin A, Ji P, Bohn JA, Levdansky Y, Wagner EJ, Valkov E, Goldstrohm AC. Human Pumilio proteins directly bind the CCR4-NOT deadenylase complex to regulate the transcriptome. ***RNA***. 2021;27(4):445-64. Epub 2021/01/06. doi: 10.1261/rna.078436.120.

6. Huang KL, Jee D, Stein CB, Elrod ND, Henriques T, Mascibroda LG, Baillat D, Russell WK, Adelman K, Wagner EJ. Integrator Recruits Protein Phosphatase 2A to Prevent Pause Release and Facilitate Transcription Termination. ***Molecular Cell***.Volume 80, Issue 2, Pages 345-358.e9. doi: 10.1016/j.molcel.2020.08.016 2020.

7. Alcott CE, Yalamanchili HK, Ji P, van der Heijden ME, Saltzman A, Elrod ND, Lin A, Leng M, Bhatt B, Hao S, Wang Q, Saliba A, Tang J, Malovannaya A, Wagner EJ, Liu Z, Zoghbi HY. Partial loss of CFIm25 causes learning deficits and aberrant neuronal alternative polyadenylation. ***Elife***. Apr;9:e50895. doi: 10.7554/eLife.50895 2020.

8. Elrod ND*, Henriques T*, Huang K.L., Tatomer DC, Wilusz JE, Wagner EJ, Adelman K. The Integrator complex terminates promoter-proximal transcription at protein-coding genes. ***Mol. Cell*** Nov;76(5):738-752.e7 doi: 10.1016/j.molcel.2019.10.034 (*co-first authors) 2019.

9. Tatomer DC, Elrod ND, Liang D, Jonathan M, Wagner EJ, Cherry S, Wilusz JE: The Integrator complex cleaves nascent mRNAs to attenuate transcriptional attenuation of Methallothionein. ***Genes and Development*** Sep;33/21-21/1525 doi:10.1101/gad.330167.119 2019.

10. Chu Y, Elrod ND, Wang C, Li L, Chen T, Routh A, Xia Z, Li W, Wagner EJ, and Ji P: *Nudt21* regulates the alternative polyadenylation of *Pak1* and is predictive in the prognosis of glioblastoma patients. ***Oncogene*** May;38(21):4154-4168 doi: 10.1038/s41388-019-0714-9 2019.

11. Elrod ND*, Jaworski E*, Ji P, Wagner EJ, Routh A: Development of Poly(A)-Click-seq as a Tool Enabling Simultaneous Genome-wide Poly(A)-site identification and Differential Expression Analysis. ***Methods*** Feb;15;155:20-29 doi: 10.1016/j.ymeth.2019.01.002 (*co-first authors), 2019.

12. Elrod N, Brady J, Rathburn H, Speshock J.L. Alteration of Cytokine Profiles Inhibits Efficacy of Silver Nanoparticle-based Neutralization of arenaviruses. ***Toxicology: Open Access*** 3(2). doi: 10.4172/2476-2067.1000124 2017.

13. Speshock JL, Elrod N, Sadoski DK, Maurer E, Braydich-Stolle LK, Brady J, & Hussain SM Differential organ toxicity in the adult zebrafish following exposure to acute sub-lethal doses of 10 mm silver nanoparticles. ***Frontiers in Nanoscience and Nanotechnology*** 2(3), 114-120. doi: 10.15761/FNN.1000119 2016.

14. Elrod N, Harp RM, Bryan KG. Effect of calcium ion supplementation on swine parturition. ***The Texas Journal of Agriculture and Natural Resources*** 28, 12-17. 2015

Permanent address: 217 County Road, Iredell, TX 76649

This dissertation was typed by Nathan D. Elrod.