UNIVERSITY OF CALIFORNIA,
IRVINE

The Role of mRNA 3' end Processing Factors in Regulating Global RNA Pol II Transcription and its Termination

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Sciences

by

Nabila Haque

Dissertation Committee:
Professor Yongsheng Shi, Chair
Professor Klemens Hertel
Professor Marian Waterman

2019

DEDICATION

To

my mother,

my father

&

my husband

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

First, I would like to thank my PI Dr. Yongsheng Shi, for taking me into his lab and training me to be the scientist I am today. I am grateful for his support and mentorship, his passion for understanding how things work at the most basic level, and his continuous striving for excellence, whether in science or as a mentor. He is a role model in my own quest to overcome the challenges in my way. I would also like to thank him for picking the best lab mates one could ask for.

I would like to thank my committee members Dr. Marian Waterman and Dr. Klemens Hertel for their thoughtful discussions, constructive feedback and words of support throughout the years. My committee members provided me with the boosts of confidence along the way that are so vital for persevering with a lightness of heart, and they are my role models for how to nurture and mentor graduate students.

I would like to thank my funding sources for making this work possible. First, the UCI MSTP grant T32GM008620 allowed me to enter this program and pursue medical scientist training. Funding from various NIH grants in Dr. Shi's lab supported me through most of my PhD training. I would also like to thank funding from the ARCS Scholars Foundation, the Stanley Behrens Recognition Award, and the Gazzaniga Family Award, through which I met wonderful inspiring people who truly believed in the power of science to make the world a better place. Their funding allowed me to travel and broaden my horizons and purchase resources that made the journey a little easier.

I would like to thank the Medical Scientist Training Program at UCI and Dr. Alan Goldin, who took me in and provided a home at UCI. Stacey Sanchez and Joanne Wu at the MSTP office have always provided a supportive and cheery source of help. My fellow MSTP classmates and colleagues have been a source of support, strength and inspiration through this journey and have become family.

I would like to thank the MMG department for providing a friendly close-knit environment in which to train. Many graduate students have provided mentorship, support and reagents over the years, particularly from the Raffatellu and Hertel labs. The office staff have always been helpful and friendly, especially Janet Horwitz and Kimberly Smith-Lyons.

I would like to thank my lab mates, who became my family through the years and helped me get through PhD training intact. I would like to thank Elmira Forouzmand, who performed the bioinformatics analyses in this work, and without whom this work would not have been possible. She has also been a great gym buddy. Yong Zhu and Xiuye Wang were like big brothers to me and taught me most of the techniques I used; Yong has seen me through two pregnancies, and was always there to help with heavy lifting and talk through ideas. Kristianna Sarkan has been the best of friends and an invaluable source of support personally and in the lab in the past few years, and especially through my last pregnancy. Lindsey Soles has been a wonderful friend and I will miss our conversations about politics and science and everything in between.

Most importantly, I would like to thank my family. Without them, none of this would be possible. My mother, Maharuna Begum, ingrained in me the importance of gaining an education and being independent and has been my inspiration all these years. My father, Mohd. Anowarul Haque, has shown me through his example how to work hard, take big risks and persevere courageously and with integrity despite seemingly insurmountable

odds. And even now, they sacrifice to take care of me and their grandchildren so I can pursue my dreams. Most of all, they love me with their whole hearts. I will never have enough words to thank them.

I would like to thank my sister, Samiah Haque, who has been my bright light and my best friend and has always believed in me and sent me prescient presents along the way to aid in my journey.

I would like to thank my babies, Zakariya and Maryam, who give me joy and renew my spirit.

I would like to thank my husband, Mahir Haroon Rabbi, who has been my rock all these years and has been there to support me through family crises and undergraduate training and graduate school and labor and babies and kept me supplied with boba and who made sure I never gave up. He has been the best husband and the best father to our children and I am grateful for him every day.

CURRICULUM VITAE

Nabila Haque

2011 B.S.E. in Biomedical Engineering, Duke University

2019 Ph.D. in Biomedical Sciences, University of California, Irvine

FIELD OF STUDY

3' end Processing of mRNAs and eRNAs

PUBLICATIONS & POSTERS

Haque, N., Forouzmand, E. & Shi, Y. (August 2017) *A novel role for CPSF100 in the 3' processing of enhancer RNAs.* Eukaryotic mRNA Processing, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Haque, N., Yuan, M. & Leslie, F.M. (October 2013) *Functional changes in D2 and 5HT1A receptors following drug exposure in adolescent rats.* UCI MSTP Annual Retreat, Lake Arrowhead, CA.

Haque, N. & Reichert, W. (October 2012) *Characterizing Endothelial Progenitor Cells for Small-Diameter Vascular Grafts.* UCI MSTP Annual Retreat, Lake Arrowhead, CA.

In vitro functional testing of endothelial progenitor cells that overexpress thrombomodulin. Stroncek JD, Xue Y, Haque N, Lawson JH, Reichert WM. Tissue Eng Part A. 2011 Aug 17; (15-16):2091-100. Epub 2011 May 25.

GRADUATE LEADERSHIP & SERVICE

2017-2019	'Women in MSTP' Founding Member and Chair, UC Irvine
2017-2018	Graduate Student Representative, Dept. of Medical Microbiology and Molecular Genetics, UC Irvine
2017-2018	Seminar Committee Member, Dept. of Medical Microbiology and Molecular Genetics, UC Irvine
2016-2017	Associated Medical Student Government Representative to Associated Graduate Students, UC Irvine
2013-2014	Crescent Clinic President and Board Member, UC Irvine
2013-2016	MSTP Distinguished Lecture Series Committee Member, UC Irvine School of Medicine
2014	MSTP IDP Working Group Lead, UCI School of Medicine, UC Irvine
2013-2014	President & Founder, Muslim Medical Student Association, UC Irvine School of Medicine

2013-2014	Social Medicine Elective Coordinator, Physicians for Human Rights at UCI School of Medicine
2013-2014	Research Coordinator , UCI Outreach Clinics, UC Irvine
2013-2014	Editor-in-Chief , Plexus Journal of Arts and Humanities at UCI School of Medicine, UC Irvine
2013-2014	Design and Layout Editor, Plexus Journal of Arts and Humanities at UCI School of Medicine, UC Irvine

GRADUATE EMPLOYMENT

2017-2018	Medical Microbiology Lab Teaching Assistant, UCI School of Medicine, UC Irvine
2014	Small Group Tutor, UCI School of Medicine, UC Irvine

HONORS

2017	Gazzaniga Family Medical Research Award, UC, Irvine
2017	Stanley Behrens Recognition Award, UC, Irvine
2016-2018	ARCS Scholar Foundation Award, 2016, UC Irvine
2011	Howard G. Clark Award, Duke University

ABSTRACT OF THE DISSERTATION

The Role of mRNA 3' end Processing Factors in Regulating Global RNA Pol II Transcription and its Termination

By

Nabila Haque

Doctor of Philosophy in Biomedical Sciences

University of California, Irvine, 2019

Professor Yongsheng Shi, Chair

Not only is mRNA 3' processing essential for gene expression, it is crucial in gene regulation as well. The regulation of 3' end processing is important in maintaining normal processes like neural activity, T-cell activity and stem cell differentiation and renewal, whereas its dysregulation can result in many diseases such as cancers. To study how 3' processing factors regulate gene expression, comprehensive ChIP-seq (Chromatin Immunoprecipitation-Seq) analyses of subunits of the 3' processing complex were performed. Unexpectedly, one subunit, CPSF100, showed minimal peaks at the 3' ends of genes, with maximum signal at the *promoters* of a large subset of genes and another large subset of strong peaks at *enhancers,* both regulatory elements that can be dysregulated to precipitate aberrant gene expression patterns. To analyze the effect of CPSF100 and the CPSF complex on transcription and its termination at protein coding genes and enhancers, we performed RNA PolII ChIP-Seq, 4SU-Seq and PolyA Site (PAS)-Seq in HeLa cells where CPSF73 and CPSF100 were knocked down using an shRNA-mediated lentiviral delivery method. We found that knocking down CPSF factors led to bidirectional transcription

termination defects at enhancers and protein coding genes. In addition, it led to an aberrant response to transcription activation by reducing the ratio of PolII elongating on the gene body to the amount of PolII at the promoter. Finally, depletion of CPSF led to the formation of long and abundant noncoding RNAs from enhancers, promoters of genes, and at permissive intergenic regions. These findings advance the field of RNA 3' end processing in several ways. We show that CPSF factors are important for facilitating productive elongation. In addition, the effect of depleting them on transcription termination far surpasses the effect of depleting the exonuclease that is thought to enable termination, suggesting that another exonuclease-independent model of termination may be the major mechanism. Finally, our findings show that CPSF is primarily responsible for mediating the cleavage and termination of eRNA transcripts, which was previously thought to occur via another factor. These findings open up new avenues of research and highlight the importance of the 3'end processing machinery in regulating transcription.

Chapter 1: Introduction

Cleavage and polyadenylation (CPA) of messenger RNA (mRNA) at the 3' end [Fig. 2] is not only an essential step of eukaryotic gene expression, but also a critical mechanism for gene regulation. Regulation of CPA is important for many physiological processes, and aberrant CPA has been associated with many diseases, including many cancers (Fig 1).(Danckwardt, Hentze, & Kulozik, 2008) Following mRNA cleavage, transcription must terminate to prevent widespread dysregulation of downstream genes and allow recycling of the transcription machinery. The following steps describe the current model of how the CPA factors assemble and function and how transcription is terminated:

1. CPSF (**c**leavage and **p**olyadenylation **s**pecificity **f**actor) subunits associate with the general transcription factor TFIID(Dantonel, Murthy, Manley, & Tora, 1997) at the promoter; however maximal recruitment of the few CPA factors studied still occurs at the 3' end.(Glover-Cutter, Kim, Espinosa, & Bentley, 2008)

2. After RNA Polymerase II (Pol II) is recruited to the promoter, CPA factors present are passed to the C-terminal domain (CTD) of elongating Pol II.(Dantonel et al., 1997) The CPA factors are thought to play no role in initiation and passively ride with Pol II until they reach the 3' end.

3. Once Pol II transcribes the polyadenylation signal AAUAAA (PAS) and other potential regulatory cis-elements, the full 3' end processing machinery assembles. This consists of 4 core multi-unit complexes, CPSF, CstF (**c**leavage **st**imulation **f**actor), CFI (**C**leavage **F**actor I) and CF II (Figure3), as well as PAP and others.(Shi, Di Giammartino, Taylor, Sarkeshik, Rice, Yates, Frank, & Manley, 2009) Each complex contains at least 2-6 proteins and the whole 3' end processing machinery may contain ~85 proteins.

4. 3' end processing of mRNA occurs through the following steps: 1. CPSF specifically binds and recognizes the PAS (Takagaki Y1, Ryner LC, & Manley JL, 1988) on pre-mRNA. 2. CstF binds to the downstream U/GU rich sequence and stimulates CPSF73(Corey R. Mandel et al., 2006a) to carry out endonucleolytic cleavage,(C. R. Mandel, Bai, & Tong, 2008) 3. PAP (**p**oly(**A**) **p**olymerase) adds a poly(A) tail which is about 200nt long.

After Pol II passes the PAS, transcription terminates; there are two predominant models (not mutually exclusive) for how this may occur: 1. The allosteric model(Zhang, Rigo, & Martinson, 2015a) proposes that when Pol II passes over the PAS, a conformational change in the transcription machinery results in transcription termination. 2. The torpedo model(Baejen et al., 2017) posits that after cleavage, an exonuclease such as XRN2 digests the 5' end of the nascent RNA on the still-transcribing Pol II, catches up with Pol II, and 'torpedoes' it off the DNA template. This allows transcription to terminate [Fig. 2]. In either case, these models suggest that the cleavage and the polyadenylation machinery is crucial for the essential step of termination transcription, whether through recognizing the PAS and promoting a conformational change at Pol II, or through cleaving the pre-mRNA off and exposing the nascent RNA to an exonuclease.

Despite the importance of 3' end processing, the specific functions, compositions and mechanisms of assembly of the mammalian CPA and transcription termination machinery and its components are still not well understood. Many questions remain: How are these factors recruited for 3' end processing? Is it an RNA Polymerase II (Pol II) dependent process, or does maximal recruitment occur at the 3' end on to pre-mRNA? Are all the factors recruited at once, and if not, what order are they recruited in? Is the composition of the complex the same at all genes? How do these factors affect transcription and

transcription termination? Do they only act at protein-coding genes or do they have roles at other Pol II transcripts? Understanding these details is the first step towards unraveling how the disruption of each factor leads to dysregulation of distinct pathways important for health and disease. As more studies are published linking different subunits of the 3' processing complex to specific diseases and cancers, it becomes more imperative to unravel the role each factor plays in gene expression.

Biological Process	Change	Reference
Cancer development	APA –UTR shortening	*Mayr et al. 2009, Cell*
Stem cell differentiation	APA – UTR lengthening	*Shepard et al. 2011, RNA*
Stem cell renewal	Fip1	*Lackford, Yao et al 2014 EMBO J.*
T cell activation	APA – UTR shortening	*Sandberg et al. 2008, Science*
Neural activity	Truncated mRNAs	*Flavell et al. 2008, Neuron*
Chronic Lymphocytic Leukemia	Mutated CPSF100	*Wang et al. 2011, NEJM*

Figure 1. Examples of regulation of 3' end processing in health and disease

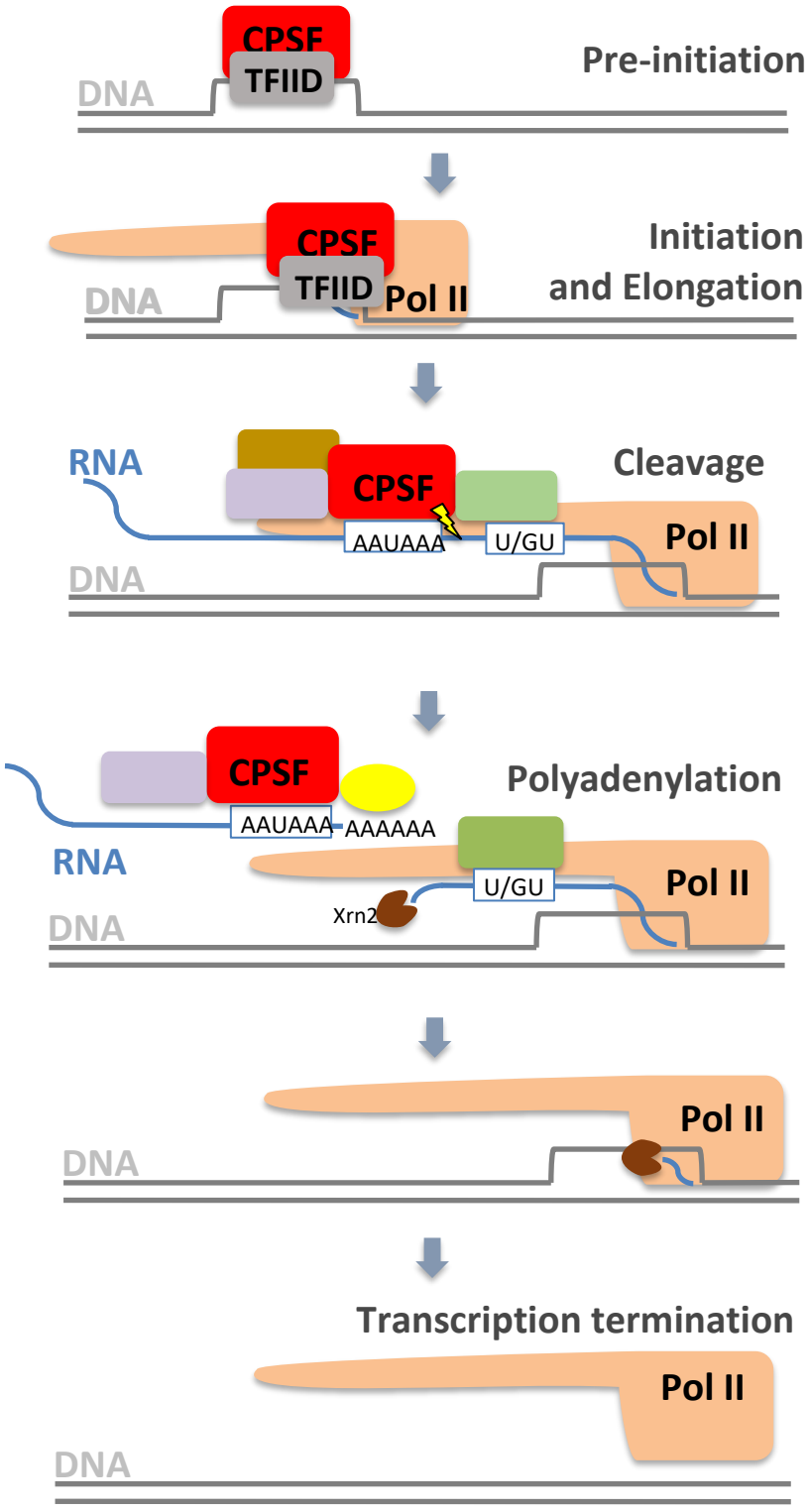


Figure 2. Transcription and mRNA 3' processing

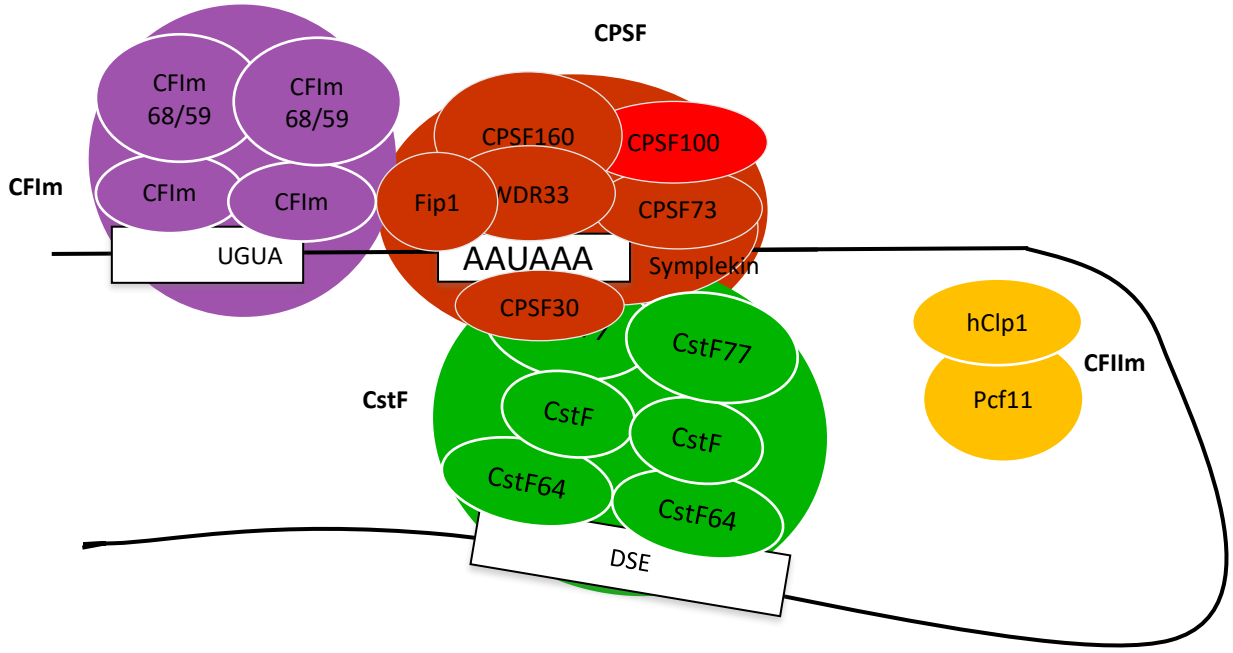


Figure 3: The core mRNA 3' end processing machinery. Contains 4 subcomplexes: CPSF, CstF, CFIm, CFIIm

Methods

ChIP-Seq

7-20x10^6 HeLa cells were crosslinked with 1% formaldehyde for 9 min at room temperature, quenched with 2.5M Glycine at final concentration of 0.125M Glycine for 5min, washed with 1x PBS, and harvested in Farnham Lysis Buffer (5 mM PIPES pH 8.0, 85 mM KCl, 0.5% NP-40). The pellet was resuspended in fresh Farnham Lysis Buffer and passed through a needle at least 20 times on ice. The cells were pelleted and resuspended in RIPA buffer (1X PBS / 1% NP-40 / 0.5% sodium deoxycholate / 0.1% SDS) at a concentration of 9x 10^6 cells /300uL. The cell lysate was sonicated with a Diagenode Bioruptor Pico for 8 to 10 cycles, 30 sec ON, 30 sec OFF, to obtain fragments 200-300 bp, cleared by centrifugation for 15 min at 13000g, and used immediately for ChIP or snap-frozen in liquid N2 and stored at -80C. Immunoprecipitation was performed overnight with 5-10ug antibody per 20x10^6 cells and Dynabeads Protein A or Protein G were added the next morning and incubated for 4 hours. Beads were washed 5 times with LiCl wash buffer (100 mM Tris pH 7.5 / 500 mM LiCl / 1% NP-40 / 1% sodium deoxycholate), once with TE and the chromatin eluted and decrosslinked overnight in 1% SDS and 0.1 M NaHCO3 solution at 65C. DNA was extracted using DNA purification columns (Denville) and Illumina sequencing libraries were prepared using NEXTFlex ChIP-Seq kit (Biooscientific) according to manufacturer's instructions. Libraries were validated by qPCR and sequenced on a HiSeq 2500 or 4000 (Illumina). Every ChIP-Seq experiment was performed in at least two independent biological replicates.

ChIP-qPCR

ChIP eluates and input were prepared as for ChIP-Seq (described above). They were assayed by real-time quantitative PCR in a 10 µl reaction containing 0.2 µM of each primer, 5 µl of PowerUp SYBR Green Master Mix (ThermoFisher Scientific), and 1 µl of template (out of 30 uL eluate) using a LightCycler 480 system (Roche). Thermal cycling parameters were: 3 minutes at 95°C, followed by 40 cycles of 10 seconds at 95°C, 15 seconds at 65°C followed by 30 seconds at 72°C. Enrichment was calculated as a percentage of input at each locus.

4SU-Seq

24 hours before labeling, complete media was replaced with serum free media. 2 hours before labeling, media was replaced with fresh serum free media. For EGF-treated cells, XM EGF was added 30 min before harvest. 10x10^6 HeLa cells were incubated with 0.5mM 4sU (Sigma, T4509 4-thiouridine) for 30 min, harvested in Trizol and the RNA purified. 60-100ug total RNA was then biotinylated using 2ug Biotin HPDP (Pierce) in 1uL DMF per ug RNA, in 10uL Biotinylation Buffer (10 mM Tris pH 7.4, 1 mM EDTA) per ug RNA, and rotated in the dark for 4 hours. RNA was purified using two chloroform extractions and isopropanol precipitation and resuspended in RNAse-free water at 1ug/uL. 100ug biotinylated RNA was added to 100uL Dynabeads MyOne Streptavidin C1 mix or a mixture of equal parts of the four types of Dynabeads in the Dynabeads Streptavidin Trial Kit (Thermo Fisher Scientific) washed according to manufacturer instructions. RNA and beads were rotated for 15 min up to 1 hr at room temperature and washed 3x with B&W buffer according to manufacturer's instructions. RNA was eluted twice with 100uL 10mM DTT and isolated by ethanol precipitation. Library preparation of 4sU labeled RNA was

performed by the UCI GHTF in a strand specific manner with no polyA selection nor ribosomal depletion

4SU-Seq Data Analysis

4sU-seq data was mapped against hg19 using STAR version 2.5.2a (Dobin et al., 2013). No multimapping was allowed in alignment. Bigwig files were then generated using deepTools v3.0.2 with RPKM normalization (Ramírez, Dündar, Diehl, Grüning, & Manke, 2014).

Read distribution data around termination sites or gene body was generated using deepTools. Peak distribution data was extracted by Homer software (Heinz et al., 2010). The visualization and any postprocessing step then were done in Python.

Termination defection was studied based on the ratio of reads mapped on 5kb downstream of the termination site to the expression of the gene. This ratio was calculated for each gene and in all conditions and its change was used as measure of termination defect. Bedtools (Quinlan & Hall, 2010) was used to extract the downstream region, and featureCounts (Liao, Smyth, & Shi, 2014) generated the read counts. The read counts were normalized to RPKM values before calculating the ratios.

Bedtools was used to extract the sequence for each region under study when necessary. Any post processing then was done in Python. Motif analysis was done using DREME (Bailey, 2011).

PAS-seq

Total RNA was extracted with Trizol (Life technologies) and 10 µg total RNA was incubated with fragmentation reagent (Ambion) at 70°C for 10 minutes followed by ethanol precipitation. Reverse transcription was performed with PASSEQ7-2 RT oligo: [phos]NNNNAGATCGGAAGA

GCGTCGTGTTCGGATCCATTAGGATCCGAGACGTGTGCTCTTCCGATCTTTTTTTTTTTTTTTTT TTT[V-Q] and Superscript III. The cDNA produced was purified and size-selected for 120-200 nucleotides using an 8% Urea-PAGE gel. Recovered cDNA was circularized with Circligase II (Epicenter) at 60°C overnight. The cDNA was heated with Buffer E (Promega) at 95°C for 2 minutes and then cooled to 37°C slowly. Circularized cDNA was linearized by BamH I (Promega) and purified by ethanol precipitation. PCR was carried out with primers PE1.0 and PE2.0, which contain indices. PCR products 200bp long were gel-purified and submitted for sequencing (single read 100 nucleotides).

PAS-Seq Data Analysis

The raw PAS-seq reads were first filtered by discarding reads with no polyA tail (less than 15 consecutive "A"s). The remaining reads were trimmed and mapped to hg19 genome using TopHat (v2.1.0) with -g 1 and strand specificity parameters (D. Kim et al., 2013). If 6 consecutive "A"s or more than 7 "A"s were observed in the 10 nucleotides downstream of poly(A) (PAS) for a reported alignment, it was marked as a possible internal priming event and that read was removed. Bigwig files were generated with the remaining reads using deepTools (v2.4), using the parameters "normalizeUsingRPKM" and "ignoreDuplicates" (Ramírez et al., 2016).

The locations of 3′ ends of the aligned reads were then extracted and those within 40nt of each other were merged into one to provide a list of potential poly(A) sites. This list was then annotated based on the canonical transcripts for known genes. The final count table was created using the reads with their 3′ ends within −40nt to 40nt of these potential PASs.

PASs with significant changes in different experimental conditions were identified using diffSpliceDGE and topSpliceDGE from the edgeR package(v3.8.5) (Robinson, McCarthy, & Smyth, 2010).

Cell Culture and RNAi

CPSF73 and CPSF100 knockdown cell lines were generated from HeLa cells using lentiviral infection of the pLKO vector containing shRNA template, followed by selection using 1.6mg/mL puromycin. Cells were harvested for assays 7 days after infection. For rescue assays, plasmids containing mouse CPSF73 or CPSF100 CDS sequences in pCMV-3x-Flag or pcDNA 3.0 vectors respectively were transfected 2 days after lentiviral infection.

Sequences for producing shRNA inserts:

hCPSF100-pLKO-F1

CCGGCCCTCAGATTCTAGCGTTATACTCGAGTATAACGCTAGAATCTGAGGGTTTTTG

hCPSF100-pLKO-R1

AATTCAAAAACCCTCAGATTCTAGCGTTATACTCGAGTATAACGCTAGAATCTGAGGG

hCPSF73-pLKO-F1

CCGGGCTGAGATTGATCTCCTATTACTCGAGTAATAGGAGATCAATCTCAGCTTTTTG

hCPSF73-pLKO-R1

AATTCAAAAAGCTGAGATTGATCTCCTATTACTCGAGTAATAGGAGATCAATCTCAGC

Sequences for primers used to clone CPSF73 or CPSF100:

mCPSF100-EcoR1-F CAGGAATTCATGACATCTATCATCAAGTT

mCPSF100-Xba1-R ACATCTAGA CACAATGGCATACTGTTCAT

mCPSF2_H67A_F TGTCTCATCCTGATCCACTCgcCCTCGGTGCCCTCCCATTCGC

mCPSF2_H67A_R GCGAATGGGAGGGCACCGAGGGCGAGTGGATCAGGATGAGACA

mCPSF73-NotI-F	CAGGCGGCCGCATGTCTGCGATTCCT
mCPSF73-BglII-R	ACAAGATCTATGTGCACCGGCGTCA
mCPSF3_75DK76HA_F	TGATCAGTCATTTCCATTTGaaggcCTGTGGAGCCCTGCCCTGGT
mCPSF3_75DK76HA_R	ACCAGGGCAGGGCTCCACAGGCCTTCAAATGGAAATGACTGATCA
Mut_CPSF2_201DQ_F	GACCCTCTCTACTTATCACAcAgTCATTTAATGCTACTTACGT
Mut_CPSF2_201DQ_R	ACGTAAGTAGCATTAAATGACTGTGTGATAAGTAGAGAGGGTC
Mut_CPSF3_204EQ_F	AGCCAGACATCCTGATCATTcAGTCTACGTATGGGACCCATAT
Mut_CPSF3_204EQ_R	ATATGGGTCC CATACGTAGA CTGAATGATC AGGATGTCTG GCT

Antibodies

Chromatin immunoprecipitation was performed with antibodies against CPSF100, CPSF73 (Bethyl, A301-581A, A301-091A) and RNAPII (Santa Cruz, N-20, CTD).

Antibodies used for Western Blotting analyses were: Ints11 (Bethyl, A301-274A), GAPDH (GeneTex, GT239), CPSF100, CPSF73, Pcf11, *Symplekin, WDR33,* CPSF160, *CPSF30, CFIm68, CFIm59, CstF64* (Bethyl, A301-581A, A301-091A, A303-706A, A301-464/5A, A301-151/2A, A301-580A, A301-584/5A, A301-356/7/8A, A301-359/60A, A301-092/3A).

Chapter 2: Role of CPSF in regulating transcription and its termination at protein coding genes

Introduction

3' end processing of mRNA is crucial for gene expression and requires the assembly of a large complex comprising more than 80 proteins (Shi, Di Giammartino, Taylor, Sarkeshik, Rice, Yates, Frank, Manley, et al., 2009) at certain cis-elements to function. **Cis-elements in pre-mRNAs**

For 3' end processing to occur, four main cis-elements are usually present. First, the polyadenylation signal (PAS) is a highly conserved hexamer sequence that usually takes the form AAUAAA, but can vary slightly, especially in the first three nucleotides(Hu, Lutz, Wilusz, & Tian, 2005). Second, cleavage typically occurs 10 to 35 nt downstream of the PAS ; there is no consensus site for endonucleolytic cleavage, but it often follows a CA dinucleotide(Sheets, Ogg, & Wickens, 1990). Third, about 30nt downstream of the cleavage site, there may be two downstream elements (DSEs), which are not as well conserved as the PAS; these DSEs consists of GU-rich(McLauchlan, Gaffney, Whitton, & Clements, 1985) and U-rich sequences(Chou, Chen, & Wilusz, 1994). Finally, multiple UGUA elements may be present about 50nt upstream of the cleavage site. These cis-elements are recognized and bound by different members of the 3' end processing machinery, which will be discussed in the following section.

Protein Factors for 3' end processing

The mRNA 3' end processing complex contains 4 main subcomplexes: cleavage and polyadenylation specificity factor (CPSF) , cleavage stimulation factor (CstF), cleavage factor I (CFI) and cleavage factor II (CFII)(Takagaki, Ryner, & Manley, 1989) as well as

poly(A) polymerase (PAP). Other components include poly(A)-binding protein 1 (PABPN1) and the Pol II CTD (Hirose & Manley, 1998). More proteins associate which may help in regulation or in coupling polyadenylation to other processes, but many functions remain uncharacterized.

CPSF

CPSF is known to be important for recognizing and binding the PAS, carrying out cleavage, and associating with the transcription complex from the initiation stage. CPSF consists of six major protein subunits: WDR33, CPSF160, CPSF100, CPSF73, Fip1 and CPSF30, as well as a scaffold protein Symplekin which also bridges to other factors.

CPSF consists of two functional subcomplexes. The first is a recognition complex that consists of CPSF160, WDR33, CPSF30 and Fip1, and is necessary and sufficient for PAS recognition and polyadenylation (along with PAP)(Chan et al., 2014)(Schönemann et al., 2014)(Sun et al., 2018). CPSF160 acts as a scaffold that preorganizes WDR33 and CPSF30 for binding to the PAS. Fip1 binds to RNA upstream of the PAS and together with CPSF160 helps to recruit PAP.(Kaufmann, Martin, Friedlein, Langen, & Keller, 2004)

The second part of CPSF consists of the core cleavage complex. This contains CPSF73 (the putative endonuclease), CPSF100 and the scaffold protein Symplekin, and seems to form a minimum core cleavage unit that can act in a modular fashion with either the rest of the CPSF recognition complex or in other 3' end processing complexes such as by partnering with SLBP to process histone mRNAs(Sullivan, Steiniger, & Marzluff, 2009)(Hill, Kumar, Girbig, Skehel, & Passmore, 2019).

CPSF73 and CPSF100 are both members of the metallo-β-lactamase (MBL) family of proteins (Aravind, 1999)(Callebaut, Moshous, Mornon, & de Villartay, 2002)(Dominski,

2008) that act on nucleic acids. They are part of a subgroup of the MBL family called the β-CASP (metallo-β-lactamase-associated CPSF Artemis SNM1/PSO2) group (Callebaut et al, 2002). Its other members include Int11 and Int9, which process 3′-ends of small nuclear RNAs (snRNAs; Baillat et al, 2005), RNase J, which acts on mRNAs in bacteria (Mathy et al., 2007)(de la Sierra-Gallay, Zig, Jamalli, & Putzer, 2008) and other nucleases (Callebaut et al, 2002; Dominski, 2007). MBL proteins have five conserved motifs (1–5) containing histidine and aspartate residues (Aravind, 1999) that cogordinate two metal ions (usually Zn2+), the members of the β-CASP group lack motif 5. Instead, they contain three other motifs, A–C (Callebaut et al, 2002).

Crystal structures have been solved for human CPSF73, but human CPSF100 could not be crystallized, and instead the structure of yeast (*S. cerevisiae*) CPSF100/Ydh1 was solved concurrently (Mandel et al, 2006). The sequence that forms the MBL domain is interrupted by a large segment that forms the β-CASP domain. The active site was found to be situated deep inside between the MBL and β-CASP domains, with no obvious access to an RNA substrate. In CPSF73, two bound Zn2+ ions were found, leading to the conclusion that CPSF73 must be the endonuclease.

Most of the signature residues of the MBL motifs are not conserved in yeast (*S. cerevisiae*) CPSF100/Ydh1 (Aravind, 1999) and no bound metal atoms were observed in the solved structure (Mandel et al, 2006). However, these residues are significantly conserved in other organisms, from plants to vertebrates, even in *S. pombe*. In addition, point mutations in conserved residues of the mammalian CPSF73 and CPSF100 led to an abolishment of the endonuclease activity that creates the histone mRNA 3′-end (Kolev et al., 2008), suggesting that CPSF100 in other species where the MBL residues are conserved

may also participate in enzymatic activity. CPSF73 and CPSF100 are tightly associated in a heterodimer and it is possible both are needed together for catalysis activity.

CstF

CstF binds to the DSEs and stimulates cleavage, and influences the choice of the exact cleavage site. (MacDonald, Wilusz, & Shenk, 1994) Similarly to CPSF, CstF also associates with Pol II during transcription elongation and may facilitate cotranscriptional processing.(McCracken et al., 1997)(Glover-Cutter et al., 2008) Three proteins comprise CstF: CstF77, CstF64, and CstF 50, each of which exists as a dimer in the complex. (Takagaki, Manley, MacDonald, Wilusz, & Shenk, 1990)

CstF77 bridges both CstF64 and CstF50 and also interacts with Symplekin(Legrand, Pinaud, Minvielle-Sebastia, & Fribourg, 2007). It has also been found in the histone 3' end processing complex. CstF64 binds to RNA(MacDonald et al., 1994) and influences PAS selection. It contains an isoform τCstF64 which may act redundantly.(Yao et al., 2013) CstF50 interacts with CstF77, and it has an N-terminal dimerization domain which together with CstF77 helps it give CstF its hexameric architecture. It has been recently shown to have a role in regulating recognition of G/U sequences based on length and content (W. Yang, Hsu, Yang, Song, & Varani, 2018).

CFIm

Along with CPSF and CstF, the CFIm complex participates in the cooperative binding of the poly(A) site and also associates with them early in the transcription process(Venkataraman, Brown, & Gilmartin, 2005). CFIm binds to the UGUA upstream of the PAS and stabilizes the binding of CPSF (Q. Yang, Gilmartin, & Doublie, 2010). UGUA acts as an enhancer sequence for the PAS it precedes. CFIm regulates alternative

polyadenylation by activating such sites (Zhu et al., 2018). The CFIm complex consists of a hetratetramer of two CFIm25 subunits, and two of a combination of CFIm59 and CFIm68 (Rüegsegger, Blank, & Keller, 1998). All the subunits have been shown to crosslink to RNA(Rüegsegger, Beyer, & Keller, 1996); the activator function of CFIm is mediated by the binding of CFIm68/59 to Fip1 in the CPSF complex through their arginine/serine (RS) domains (Zhu et al., 2018).

CFIIm

CFIIm, especially in humans, is poorly characterized and consists of at least two proteins, Pcf11 and Clp1(de Vries et al., 2000). Pcf11 binds to the Pol II CTD and enhances cleavage and transcription termination(Kamieniarz-Gdula et al., 2019). Clp1 may tether CPSF and CFIm to CFII, and otherwise has a role in tRNA splicing (Weitzer & Martinez, 2007).

In this study, we first aimed to study the recruitment of 3' end processing factors. We performed ChIP-Seq on all the major protein subunits except for Clp1. The results of our unbiased series of ChIP-Seq experiments revealed a striking binding profile for CPSF100. The rest of our studies focused on the effect of CPSF100 and its binding partner CPSF73, the putative endonuclease, on transcription and its termination.

Results

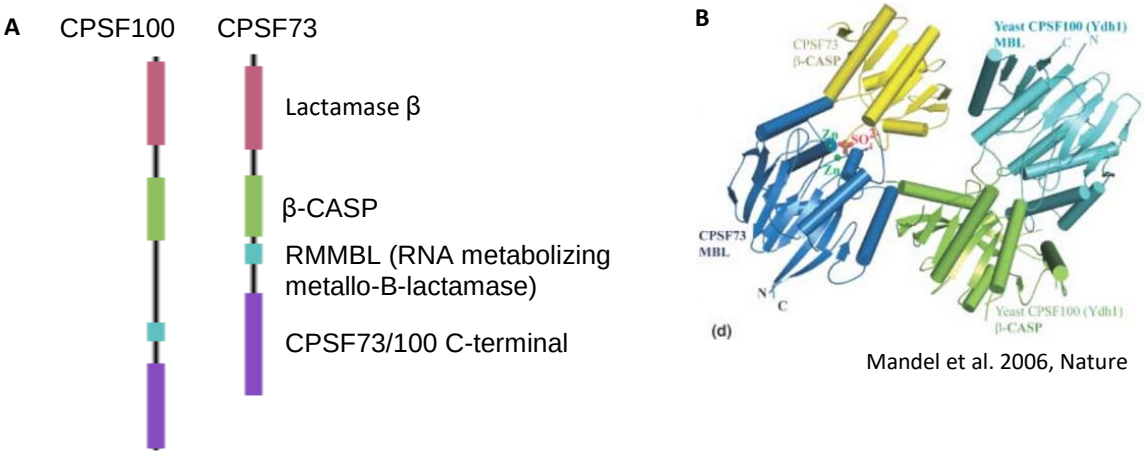
Binding pattern of CPA factors on Protein Coding Genes

To start answering the question of when CPA factors are recruited and whether the same pattern applies at all genes, we began by performing ChIP-Seq of thirteen 3' end processing factors in HeLa cells. These were: CPSF subunits CPSF160, WDR33, CPSF100, CPSF73, Symplekin, CPSF30 and Fip1, CstF subunits: CstF 77, CstF 64 and CstF 50, as well as CFIm25, CFIm59 and CFIm68, hPcf11 and RNA Polymerase II (Pol II).

In accordance with previous findings (Glover-Cutter et al., 2008) and consistent with their role in 3' end processing of mRNA, most of the CPSF and CstF factors localized on chromatin with maximal enrichment at the 3' ends of genes (Fig 5A, B); some CPSF and CstF factors also showed very modest enrichment at the promoter region (Fig 5B), consistent with a model of co-transcriptional recruitment. However, maximum binding at the 3'end suggests that the most CPA factors do not consistently travel with elongating Pol II from promoter to 3'end and may only transiently bind during elongation.

Strikingly, one CPSF factor –CPSF100 – showed a distinctly different binding pattern that was unique in multiple respects compared to the other CPA factors that mostly localized at the 3' ends of genes. While it shared the same maximal 3' recruitment binding pattern as other CPSF factors at previously studied genes such as GAPDH, MYC and histone genes (Glover-Cutter et al., 2008), confirmed in Fig 5B, CPSF100 was unique in showing maximum enrichment at the promoters of most genes, with minimal signal at the 3' end (Fig 6A-D). Of all the CPSF100 binding sites identified, 47% were at the promoter. In fact, the global CPSF100 chromatin binding pattern looked remarkably similar to that of Pol II and other transcription factors, suggesting its maximal recruitment occurred early in association with Pol II (Fig 6A). At many genes, such as those shown in Fig 6D, CPSF100 ChIP signal was found mainly at the promoter, even if other 3' end processing factors like Pcf11 were mostly detectable at the 3' end (Fig 6D). However, a visual comparison of the CPSF100 and PolII signals at various genes (Fig 5B and 6D) showed different patterns of enrichment, suggesting that CPSF100 was at least partly recruited independently of PolII, and perhaps even before PolII. On certain paused genes like HAUS5 (Fig 6D), the CPSF100

signal was an order of magnitude larger than the PolII signal, while at other paused genes, they were comparable (FOS) or much lower (GAPDH, Fig 5B).



Mandel et al. 2006, Nature

C

Kolev et al. 2008, EMBO reports

Figure 4. Comparison of the structures of CPSF73 and CPSF100 **A.** Schematic of domain organization in CPSF73 and CPSF100. **B.** Comparison of crystal structure of CPSF73 and yeast homolog of CPSF100, Ydh1 **C.** Protein amino acid sequence alignment of MBL motifs of CPSF73 and CPSF100

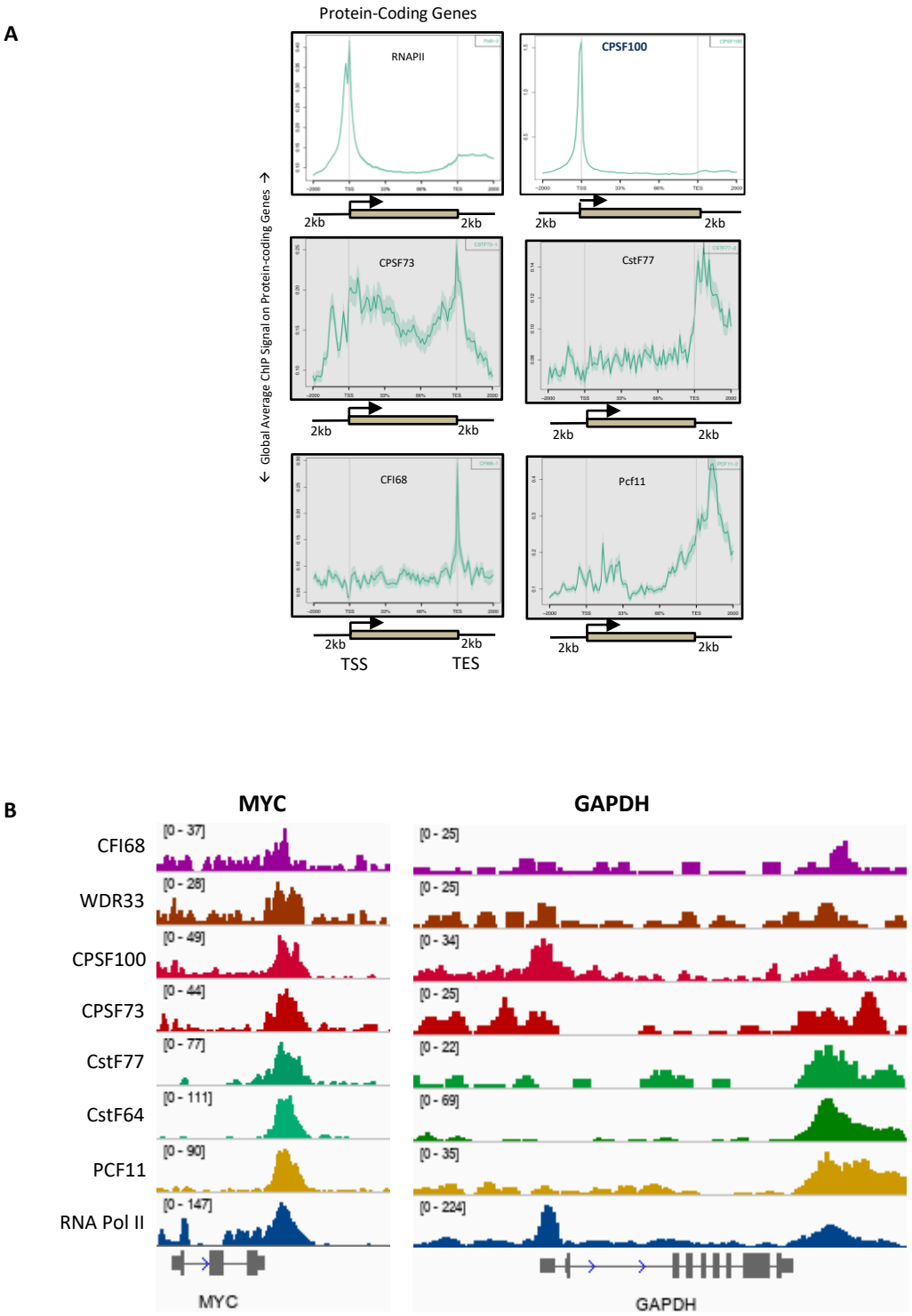


Figure 5. Genomic pattern of mRNA cleavage and Polyadenylation (CPA) factor binding.
A. Meta-gene analysis of select CPA factor binding on protein coding genes. Average ChIP-Seq signal on protein-coding genes (n=20,8005 for RNAPII and CPSF100 plots, n=30-12,000 for other factors) relative to input, normalized to gene length. **B.** Genomic profile of 3' end processing factor binding to MYC and GAPDH

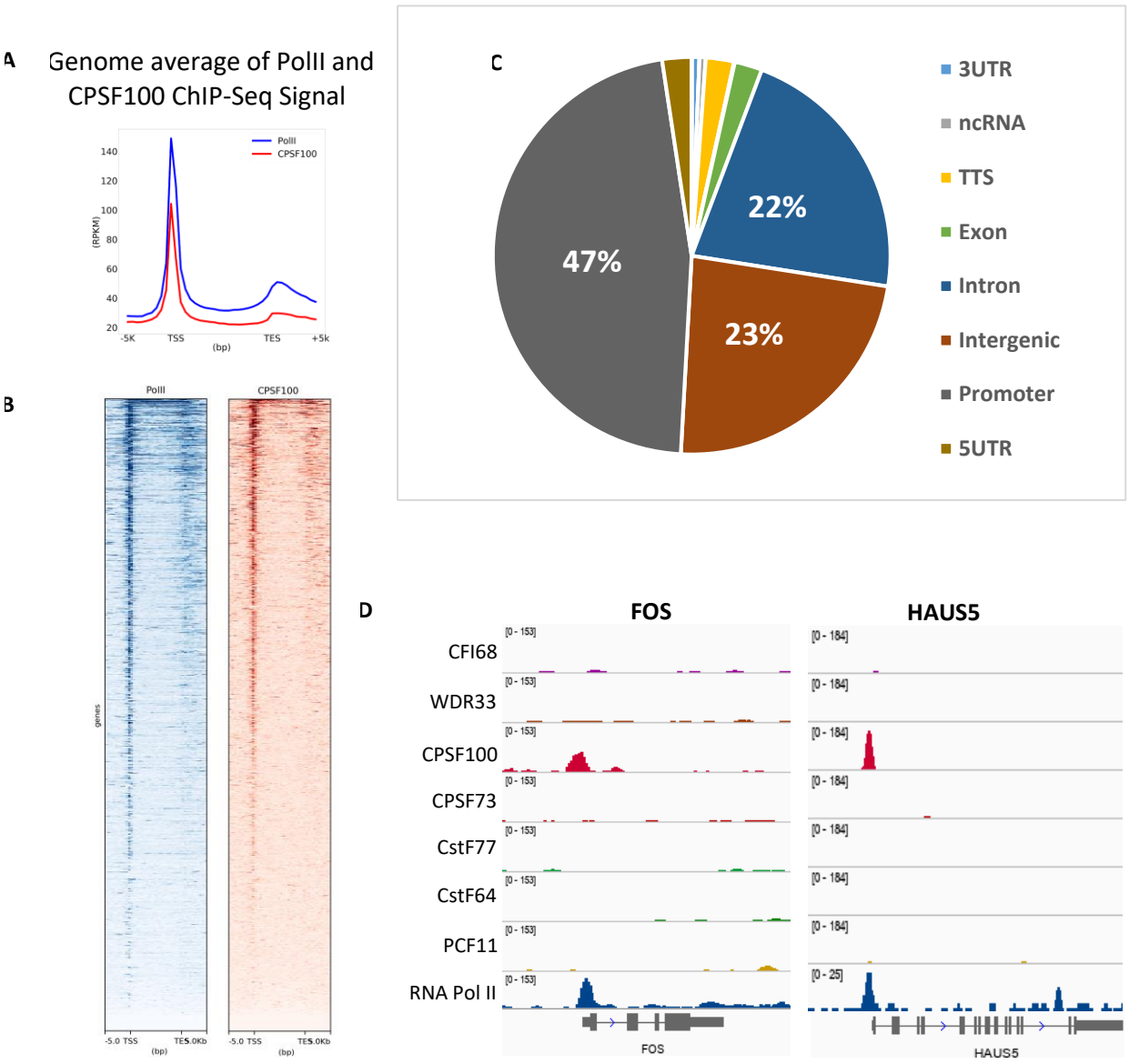


Figure 6: CPSF100 binds to promoters. **A**. Metaplot of CPSF100 and PolII binding on protein-coding genes.. **B**: Heatmap analysis of CPAF100 and PolII binding on protein-coding genes. **C**. Pie chart representation of CPSF100 binding site locations **D**. Genomic profile of 3' end processing factor binding at FOS and HAUS5. **E-F.** Metascape gene ontology analysis of CPSF100 (F) and PolII (E) binding sites

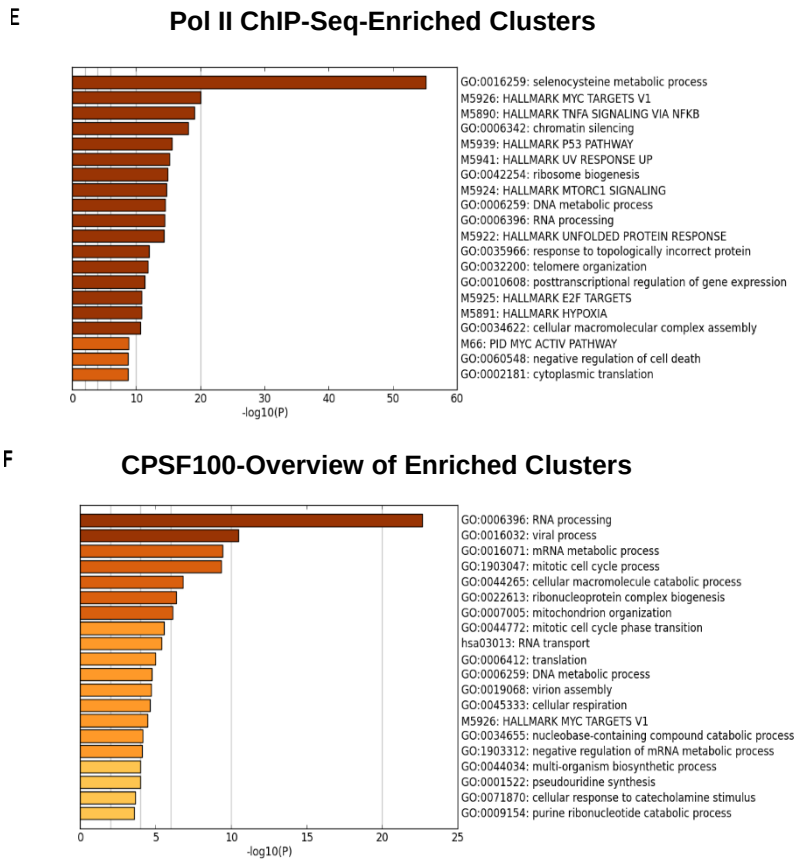
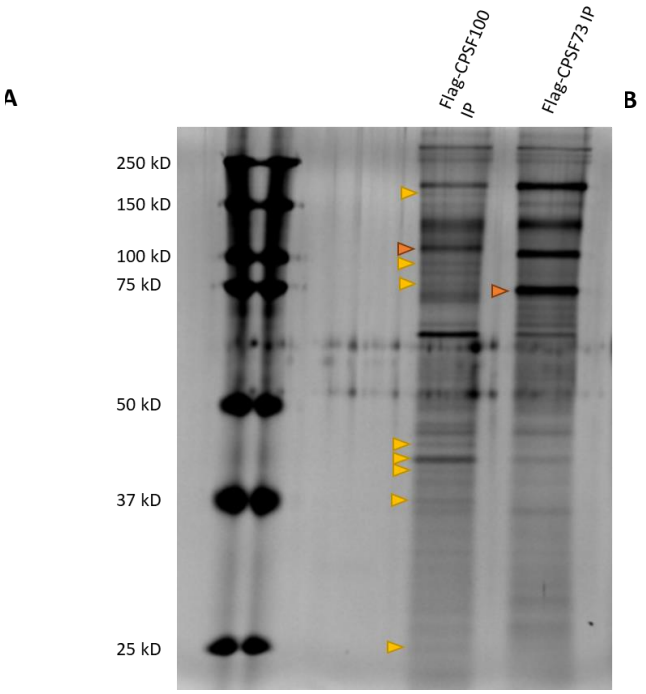


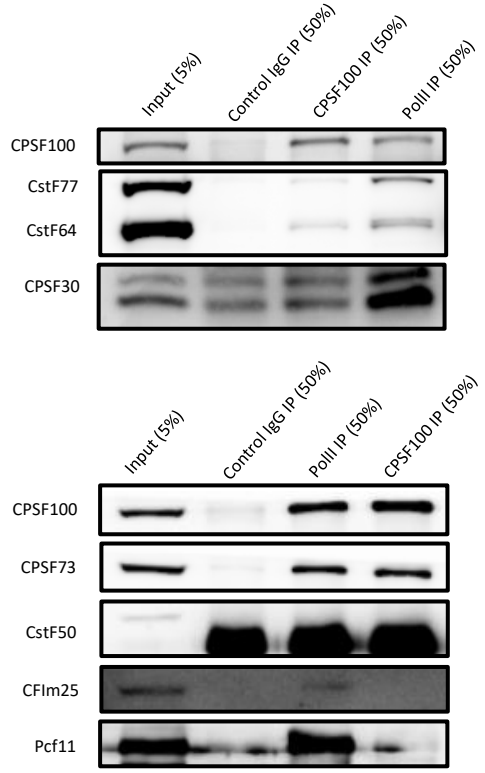
Figure 6: CPSF100 binds to promoters.

To further test whether the CPSF100 ChIP signal at promoters correlated directly with Pol II binding or showed an independent pattern, and to determine which classes of genes CPSF100 may regulate, we performed gene enrichment analysis of the PolII and CPSF100 peaks that were significantly above background (1% false discovery rate, FDR) using Metascape (Fig 6 E and F). It was clear that PolII and CPSF100 were enriched at different sets of genes, suggesting that they are not co-recruited. Notably, CPSF100 was enriched at genes related to RNA processing, suggesting that in addition to directly participating in 3'end processing, CPSF100 may also transcriptionally re gulate other RNA processing factors.

We then examined how closely PolII associates with different 3'end processing factors, and whether the striking correlation with the CPSF100 binding pattern would reveal a stronger interaction with CPSF100 compared to other 3'end processing factors (Fig 7B). Immunoprecipitation of PolII using an antibody to its body (N-20, Santa Cruz) pulled down CPSF100, CPSF73, CPSF30, Pcf11 and CFIm25, as well as CstF77 and CstF64 in smaller amounts compared to input. Immunoprecipitation of CPSF100, however, did not pull down CFIm25 and Pcf11. These results confirmed the notion of co-transcriptional recruitment of CPA factors but CPSF100 was not overwhelmingly co-IP'd with PolII. CPSF30 and Pcf11 seemed to show the strongest associations with PolII compared to input; this confirms previous findings that CPSF30 may be the bridge connecting PolII to the CPSF complex (Zhang et al., 2015a) and the known binding of PCF11 to the PolII CTD (Hollingworth, Noble, Taylor, & Ramos, 2006).



Flag-CPSF100 and Flag CPSF73 co-precipitate different set of proteins



CPSF100 and Pol II IP with PolyA factors

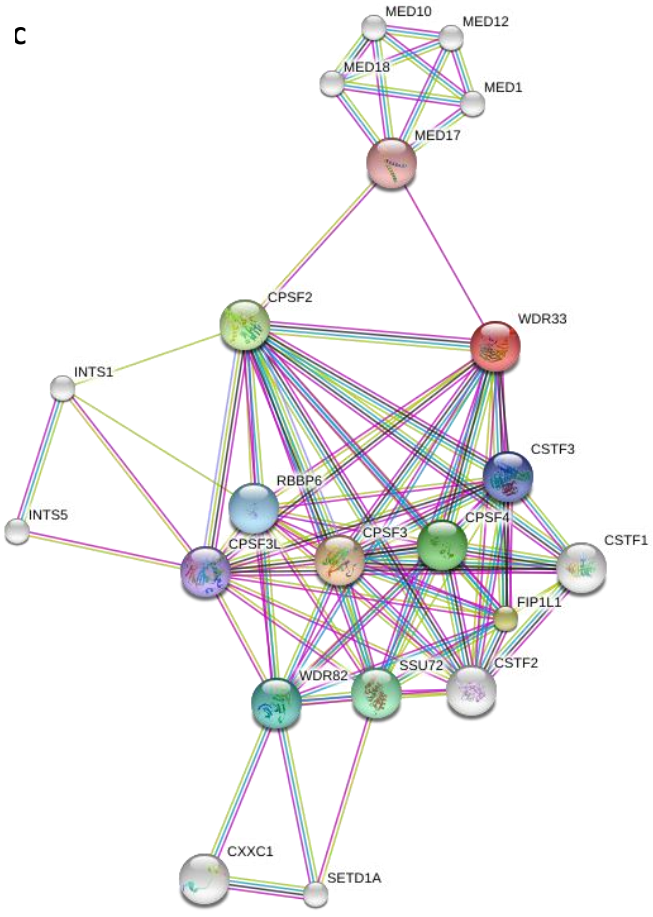


Figure 7: CPSF100 recruitment to promoter **A**. Silver stained SDS-PAGE gel comparing FLAG-IP of CPSF73 and CPSF100 **B**. Immunoprecipitation of CPSF100 and PolII **C.** STRING-dB analysis of CPSF100 interacting factors

Since CPSF100 did not pull down PCF11 or CFIm 25, it would suggest that CPSF and CstF members interact more closely and CFI and CFII may be recruited separately, interact more loosely, or have a different affinity to CPSF100 than to CPSF73 (Shi, Di Giammartino, Taylor, Sarkeshik, Rice, Yates, Frank, Manley, et al., 2009). Indeed it appeared that CPSF100 may have a different interactome than CPSF73; comparing the Flag-IP products of C-terminal Flag-CPSF100 and C-terminal Flag-CPSF73 on an SDS-PAGE gel followed by silver staining (Fig 7A) showed that these two proteins pulled down different sets of proteins and identical proteins in different ratios. Therefore it is possible that CPSF100 interacts with other complexes at the promoter independent of the core CPSF complex. A STRING-dB analysis (Fig 7C) of the CPSF100 interactome shows that CPSF100 may interact with members of the Mediator complex and; the only other CPA factor that is shown to possibly interact with Mediator as well is WDR33, which was the only other CPA factor that showed striking enrichment at the promoter region, albeit at only a handful of genes, compared to the thousands of genes where CPSF100 was enriched at the promoter. CPSF100 also uniquely seems to interact with WDR82, a protein that is important for chromatin remodeling and also localizes to the promoter (Austenaa et al., 2015).

To explore why CPSF100 appeared to be maximally enriched at the promoters of some genes and the 3' ends of others, we examined the subset of genes where a CPSF100 3' end peak was found; these genes were all highly expressed, like GAPDH or MYC (Fig 5B). This led to the hypothesis that CPSF100 is recruited to the promoter at a basal level, where it pauses with Pol II; when transcription is activated, it moves to the 3' end to help carry out processing. To test this hypothesis, we first looked at the relationship between CPSF100 signal at the promoter and gene expression level. Genes were clustered based on the

CPSF100 binding pattern using K-means clustering. The average Pol II and CPSF100 ChIP signal in each cluster is shown in Fig 8A. The average expression level of genes in each cluster was calculated (not shown); as transcription levels increased, CPSF100 binding at the promoter increased (Fig 8A); at the highest gene expression levels, CPSF100 binding demonstrated a shift to the 3' end. To test if this hypothesis held true within the same gene, we added EGF to wildtype empty-vector infected (WT) cells to stimulate transcription above the basal level at EGF target genes and performed CPSF100 and Pol II ChIP-Seq. CPSF100 bound mainly to the promoter of the FOS gene (Fig 8D) in unstimulated cells; upon adding EGF, recruitment of CSF100 to the promoter increased and a robust peak also appeared at the 3' end (Fig 8C,D). As expected upon EGF stimulation, Pol II levels at the target gene promoter proximal region (PP) and over the gene body also increased (Fig 8B,D). These findings demonstrate that CPSF100 is recruited to protein coding genes in response to transcription activation. It is also notable that this results in more CPSF100 enrichment at the 3'end than at the promoter, which is different from the pattern of PolII binding, suggesting that additional recruitment may occur at the 3' end.

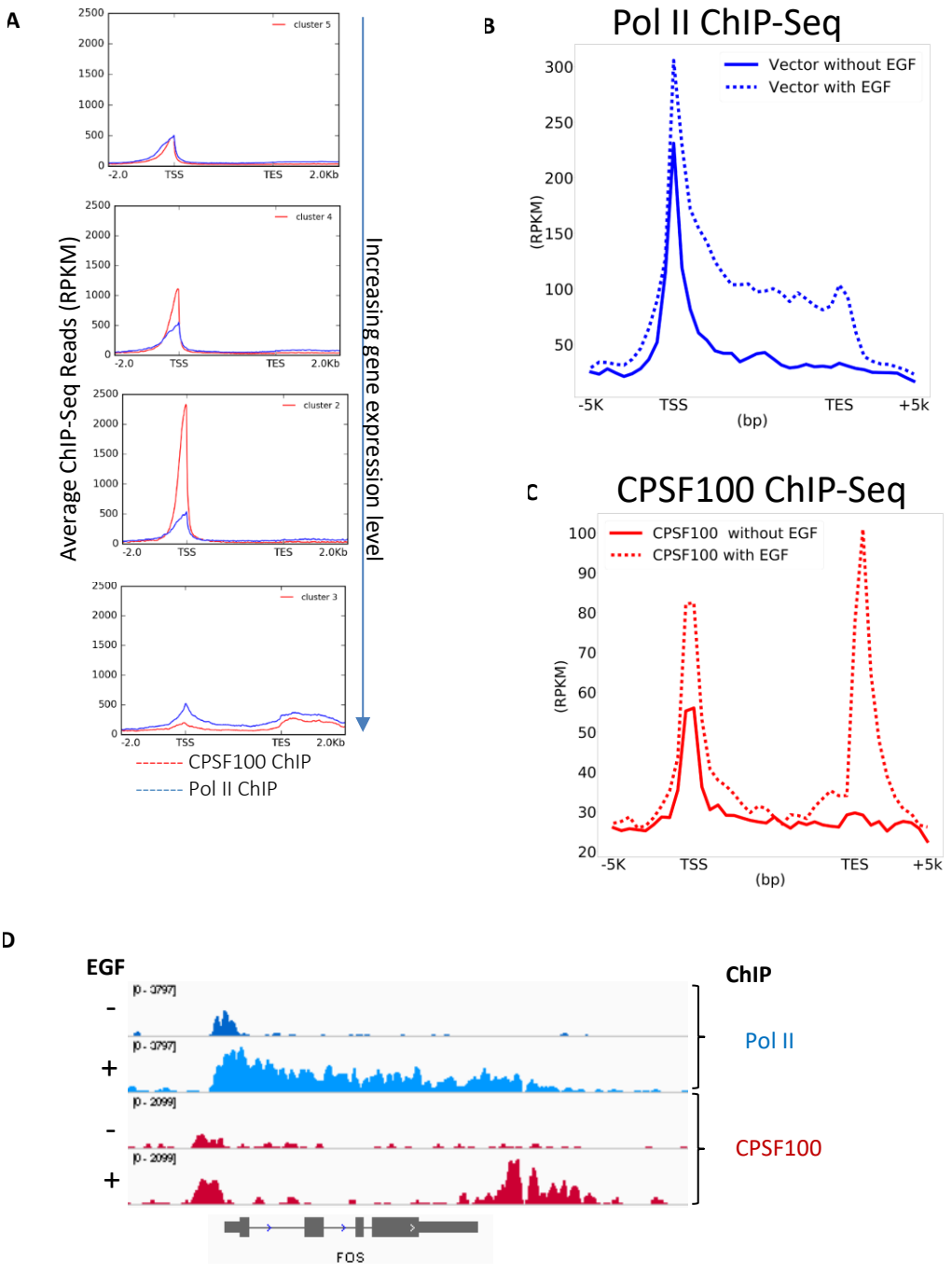


Figure 8. CPSF100-chromatin association pattern depends on gene expression. A. Meta-plot analysis of CPSF100 and PolII binding on subsets of genes grouped by K-means clustering., ordered by average gene expression levels. **B, C**: Meta-plot analysis of PolII (B) and CPSF100 (C) binding at EGF target genes before and after activation by EGF. **D**: Genomic profile of PolII and CPSF100 ChIP-Seq on the FOS gene before and after EGF stimulation.

Effects of CPSF depletion on the 3' end processing complex and the transcription response

Although it appeared that CPSF100 had a unique enrichment at the promoter, it was possible that the other CPA factors were equally present but that other factors such as epitope availability or variability in ChIP efficiency masked their presence. In addition, CPSF100 exists with CPSF73, the known endonuclease, and symplekin in a core complex responsible for cleavage. CPSF100 function is not clear, but some studies (Kolev, Yario, Benson, & Steitz, 2008) have shown that mutating the homologous putative active site residues in CPSF100 as compared to CPSF73 also results in a loss of cleavage at histone substrates. To test whether CPSF100 has an effect separate from CPSF73, and to clarify whether CPSF73 also binds CPSF100-bound locations where there is no detectable CPSF73 ChIP signal, we knocked down CPSF100 and CPSF73 in HeLa cells using lentivirus-mediated shRNA delivery. We then sequenced nascent RNA by labeling with 4-thiouridine(4SU) for 30 min, and also sequenced polyadenylated 3'ends (PolyA-Site Sequencing, PAS-Seq) and performed PolII ChIP Seq in these cells.

Knocking down CPSF100 resulted in a decrease in protein levels of both CPFS73 and Symplekin (Fig 9A), the other two members of the core cleavage complex, while mRNA levels did not appear to change significantly for CPSF73, and even showed a slight increase for Symplekin. Knocking down CPSF73 to a comparable level as in the CPSF100 knockdown cells resulted in a similar decrease in Symplekin but the decrease in CPSF100 did not reach the same level as in the CPSF100 knockdown cells. These results suggest that both CPSF73

and CPSF100 are needed for the stability of the CPSF core complex, with CPSF100 having a greater impact on complex stability.

In addition, there appeared to be a significant increase in PCF11 protein levels (Fig 9A) after CPSF73 and CPSF100 depletion. PCF11 is part of the CFII complex and aids in cleavage and termination; the increase in PCF11 could be part of a self-regulating mechanism of the CPA complex to compensate for deficiencies in its factors. The increase in PCF11 seems to occur due to a decrease in an intronic polyadenylation site that results in more full length product (Fig 9D); this mechanism is similar to the autoregulation of PCF11 levels (ref).

It was in fact a general trend that there was a net upregulation of gene expression after knocking down CPSF73 and CPSF100. After CPSF100 KD, 4046 genes were upregulated and 1298 genes downregulated; after CPSF73KD, 1259 genes were upregulated and 108 were downregulated. Since CPSF is a general factor required for processing nearly all mRNAs, these results would suggest that the CPSF core complex serves to restrict expression of proteins under normal conditions.

Effect of CPSF100 depletion on transcription initiation and elongation: Because CPSF100 was found to bind at the promoter, we hypothesized that depleting CPSF100 would cause defects in transcription around the promoter. To understand how CPSF100 affects Pol II recruitment and transcription elongation, we performed Pol II ChIP-Seq in CPSF100KD cells with or without EGF, and in CPSF73KD cells as a control. We also performed PAS-Seq and 4SU-Seq with and without the addition of EGF.

Our results showed that fewer genes were upregulated in the CPSF100KD condition after EGF stimulation compared to vector infected cells (Fig 9E). Of the genes that showed a response in the 100KD cells, the magnitude of response was significantly lower. Thus,

CPSF100 depletion reduces the transcriptional response at EGF target genes after stimulation.

Our results also showed that CPSF100 may play a role in recruiting Pol II to the promoter. After addition of EGF, control cells showed an increase in Pol II ChIP signal at the promoter as well as increased PolII signal throughout the body of the promoter, indicating release of paused PolII into the gene body (Fig 9 F,H-J), as expected. Upon CPSF100 and CPSF73 depletion however, there was little or no additional recruitment of PolII to the promoter. In addition, there appeared to be a defect in pause release or transcription elongation as measured by the traveling ratio (TR). The TR is a measure of Pol II pause release, which is the density of Pol II in the gene body relative to the Pol II density in the PP region [Fig.9F]. While the global profile as compared to WT did not show a pause release defect after CPSF100KD (Fig 9F, G), visual inspection of EGF dependent genes (Fig 9H-J) showed a clearly reduced ratio of elongating PolII in the gene body as compared to the promoter in CPSF73 and CPSF100KD cells as compared to WT upon EGF stimulation.

While PolII seemed to be recruited at comparable amounts in both WT and CPSF73KD, it appeared that less PolII was recruited overall in CPSF100KD; this led to a marked decreased level of transcription as measured by 4SU-Seq in CPSF100KD that was not seen in CPSF73KD (Fig 9H-J). These results demonstrate that while knocking down CPSF73 and CPSF100 have similar effects on PolII recruitment, albeit at different scales, CPSF100KD leads to far greater defects in transcription.

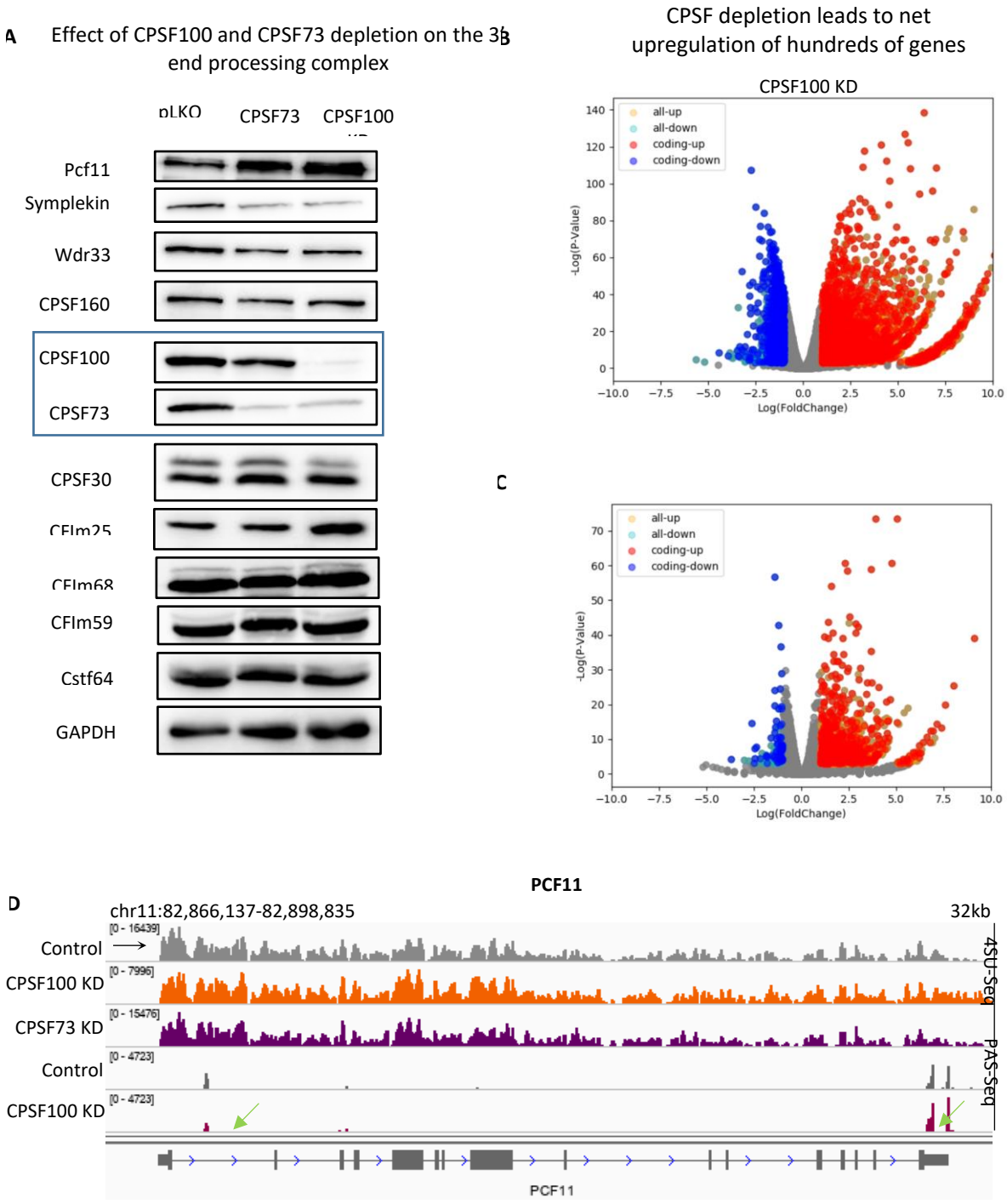
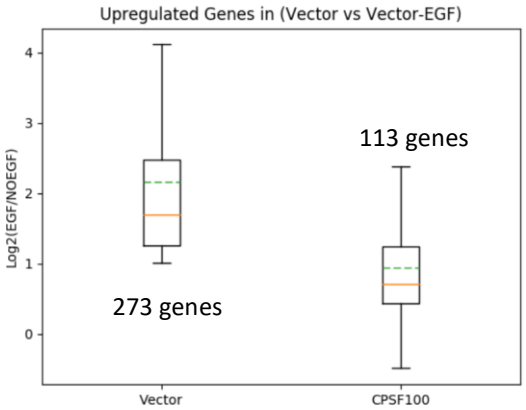
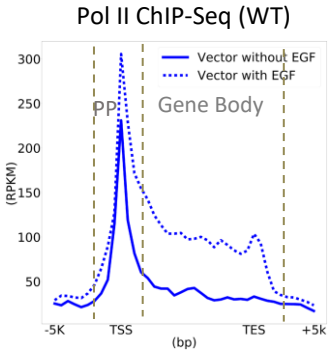


Figure 9. **Global effects of CPSF100 depletion on CPA factors and gene expression A.** Western Blot of members of the members of the 3' end processing complex after CPSF73 and CPSF100KD **B,C.** Volcano Plot showing genes differentially expressed after (**B**) CPSF100 knockdown [4046 upregulated; 1298 downregulated] and (**C**) CPSF73 knockdown [1259 upregulated; 108 downregulated]. n=12,000. **D.** Genomic profile of 4SU-Seq and PAS-Seq signal at PCF11. **E.** Box plot showing response to EGF in control and 100KD cells. Y axis shows log (RPKM in +EGF/RPKM in –EGF). **F,G.** Meta-plot analysis of PolII EGF target genes before and after activation by EGF in control (**F**) and CPSF100KD (**G**) cells. **H-J**. Genomic profile of 4SU-Seq and ChIP-Seq signal before and after EGF stimulation at (**H**) JUN (**I**) DUSP1 (**J**) FOS

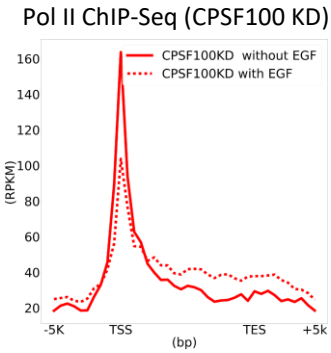
E



F



G



Traveling Ratio = <u>Pol II density at promoter</u>

H

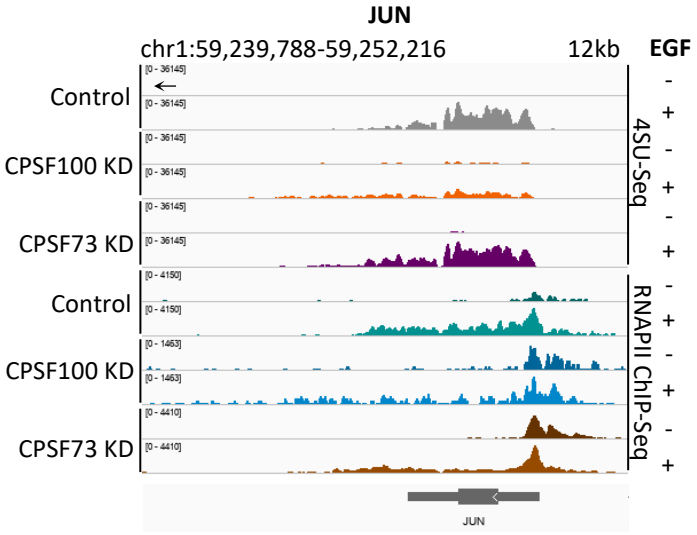


Figure 9. Global effects of CPSF100 depletion on CPA factors and gene expression

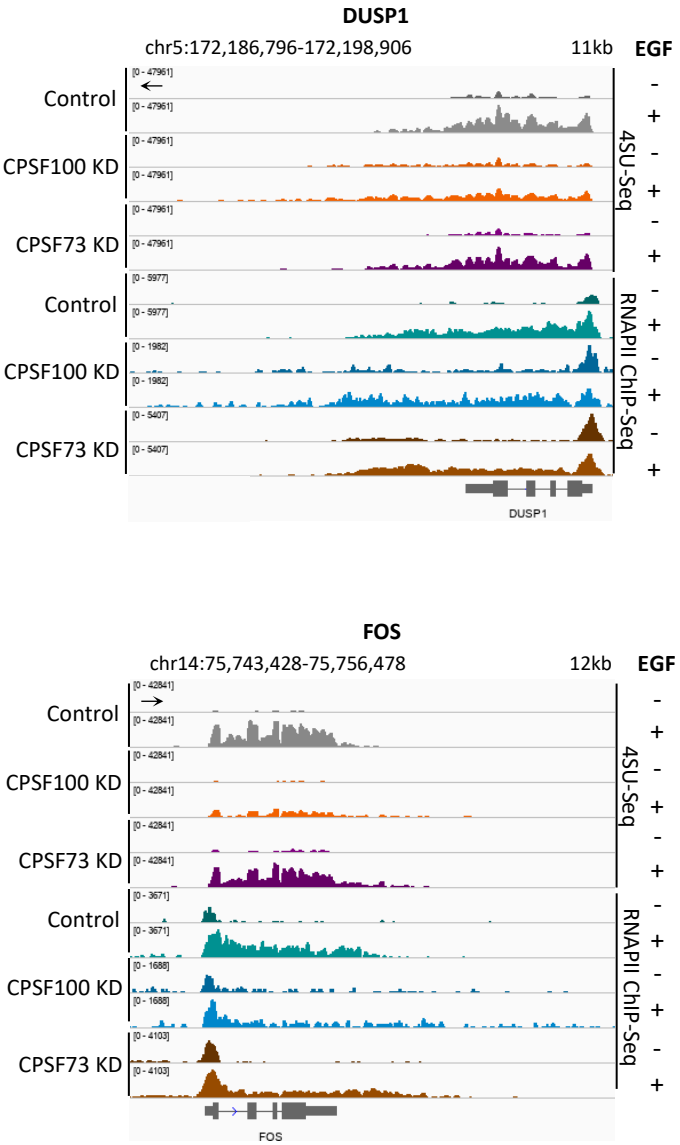


Figure 9. Global effects of CPSF100 depletion on CPA factors and gene expression

CPSF73 and CPSF100 depletion lead to Transcription Termination Defects

The mechanism of transcription termination is still a subject of debate. Recent studies have shown that while the torpedo model with XRN2 as the 5' to 3' exonuclease seems to play a role in transcription termination (Fong et al., 2015; West, Gromak, & Proudfoot, 2004), only a severe degron-based depletion of XRN2 protein showed appreciable transcription termination defects (Eaton et al., 2018), suggesting that other mechanisms might be at play. We next decided to study the effect of CPSF73 and CPSF100 on transcription termination.

In a preliminary study, we examined the termination zone in a readily available Fip1KD HeLa stable cell line and in freshly infected CPSF100KD cells via 4SU-Seq. Fip1KD cells were used a control to determine whether the effects of CPSF100KD were unique or a consequence of altering the CPSF complex.

The termination zone represents the distance between the cleavage site, which was approximated by the location of the PAS-Seq signal, near a polyA site, and the transcription end site (Fig 10A). In both knockdown conditions, the prevalence of short terminations zones (0-2.5kb) decreased and the frequency of longer termination zones increased. The prevalence of the longest termination zones (>20Kb) increased more in CPSF100KD cells as compared to Fip1KD cells.

We next compared the effect of knocking down CPSF73 to CPSF100 by 4SU-Seq and PolII ChIP-Seq on transcription termination. A metaplot of the 4SU-labeled RNA-Seq signal normalized at the transcription end site reveals a termination defect upon CPSF73 and CPSF100KD, which was more severe in the case of CPSF100KD, where the average 4SU-Seq signal did not return to normal even 20kb after the transcription end site (Fig 11A).

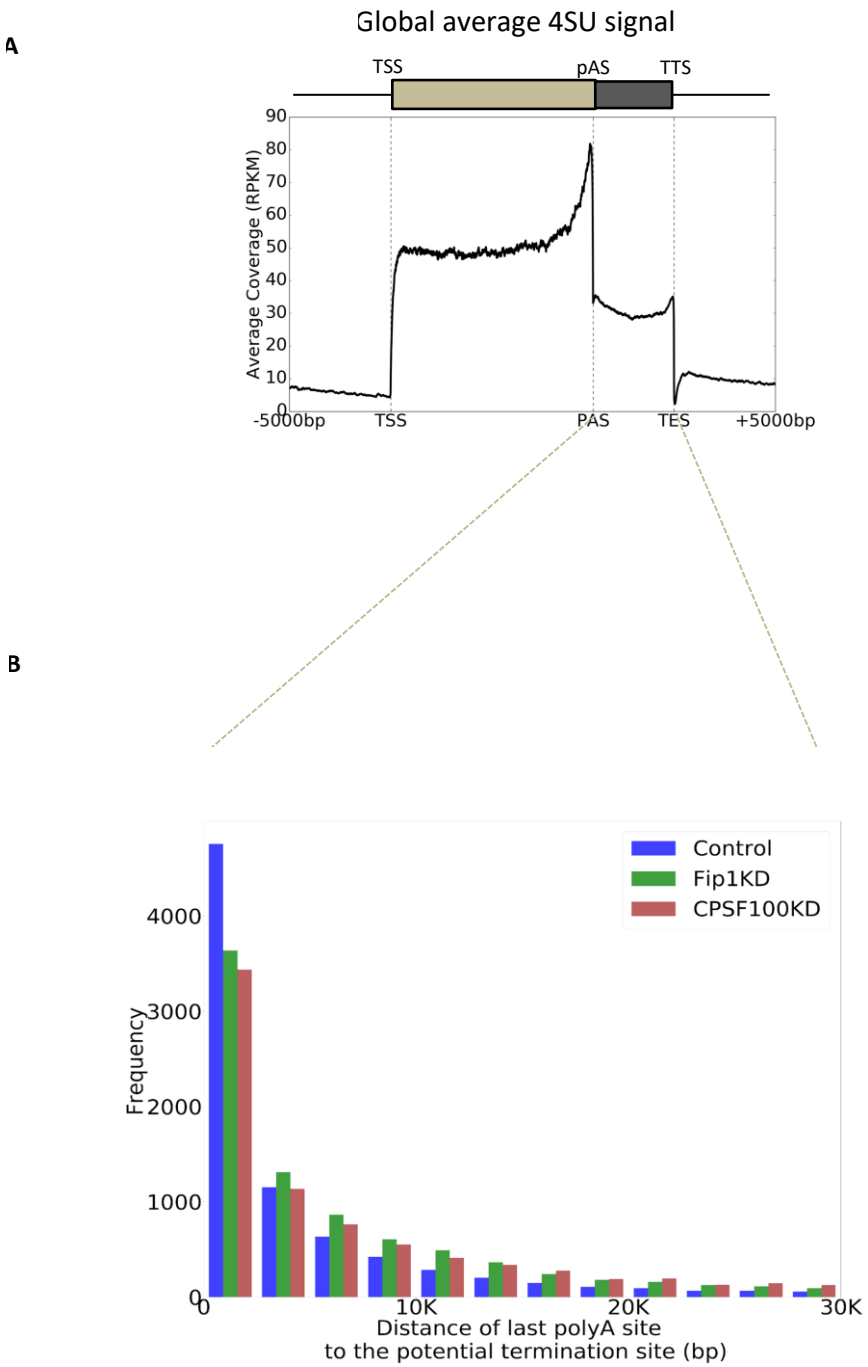


Figure 10: Analysis of the termination window in mammalian cells. **A**. Global average 4SU-Seq signal over genes showing PolyA site (PAS) and Transcription End Site (TES) to show termination zone. **B**. Histogram showing frequency of termination zones binned at 2.5kb intervals in control, Fip1KD, and CPSF100KD cells.

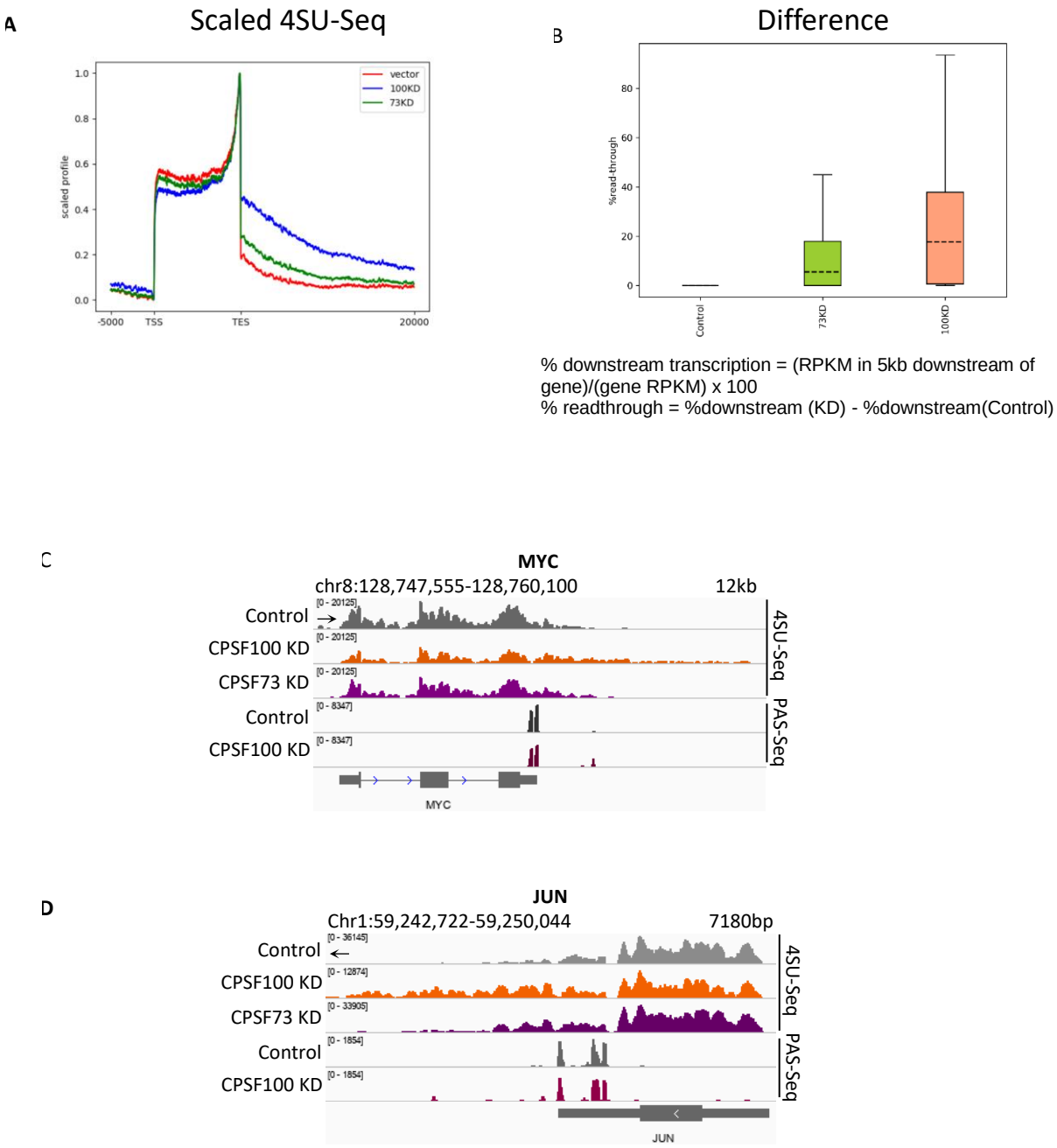


Figure 11. CPSF100 KD leads to termination defects of protein coding genes. **A**. Scaled profile of 4SU-Seq signal at protein coding genes -5kb to +20kb from gene body in Control, CPSF100KD and CPSF73KD conditions. **B**. Boxplot showing percentage difference of readthrough after knockdown; % downstream transcription = (RPKM in 5kb downstream of gene)/(gene RPKM) x 100; % readthrough = % downstream (KD) - % downstream (Control) **C-G**. Genomic profile of 4SU-Seq and PAS-Seq signal at (C) MYC (D) JUN (E) Histones (F) PDE3A (G) RPS29 **H**. Flag-Immunoprecipitation of CPSF73 mutants **I-J**. Western Blot showing CPSF73 and CPSF100 levels after knockdown and rescue **K**. RT-qPCR using primers on the MYC gene exon at position 5071 from the start site and in the readthrough region at 7028 bp after the TSS.

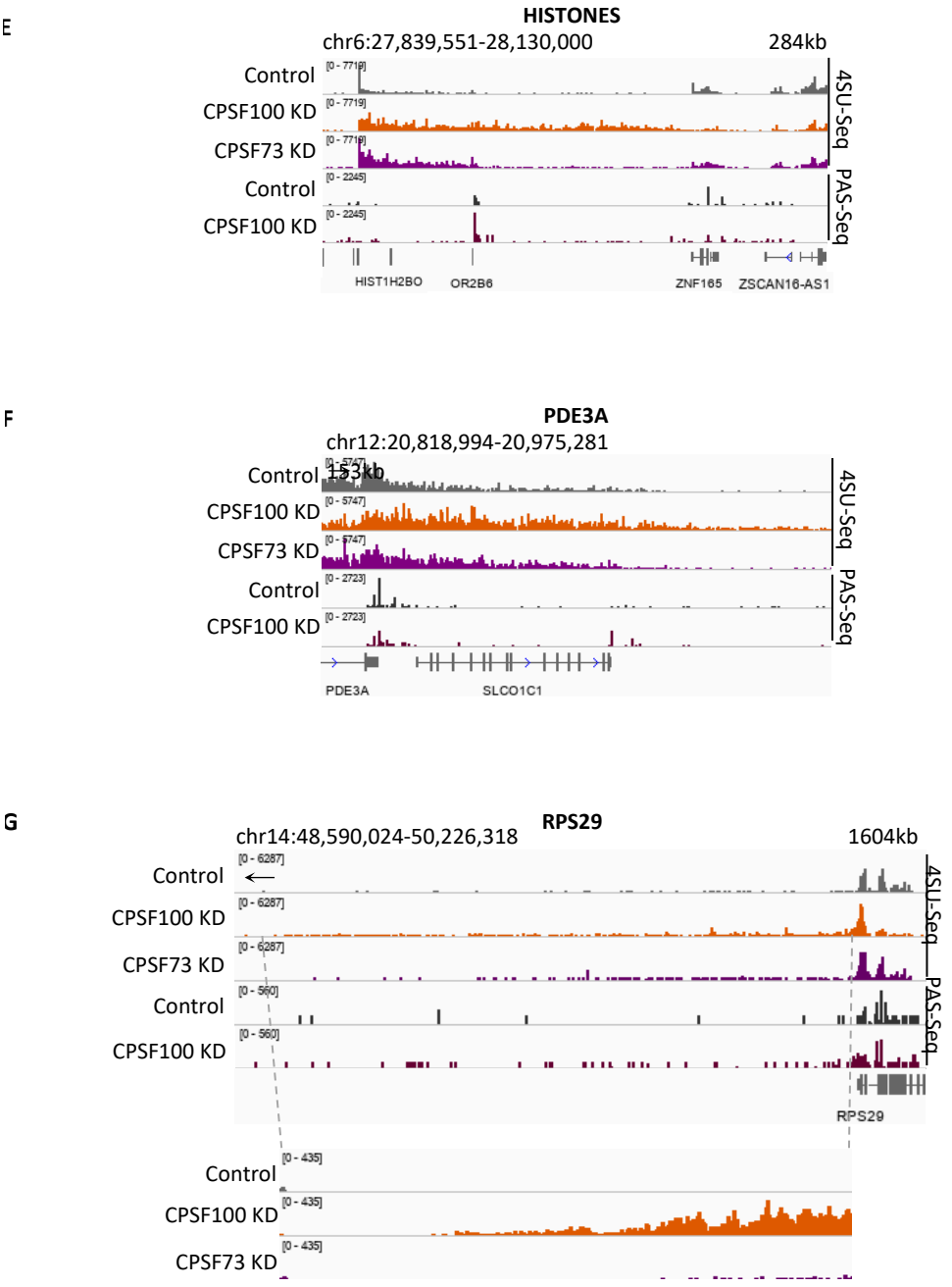


Figure 11. CPSF100 KD leads to termination defects of protein coding genes.

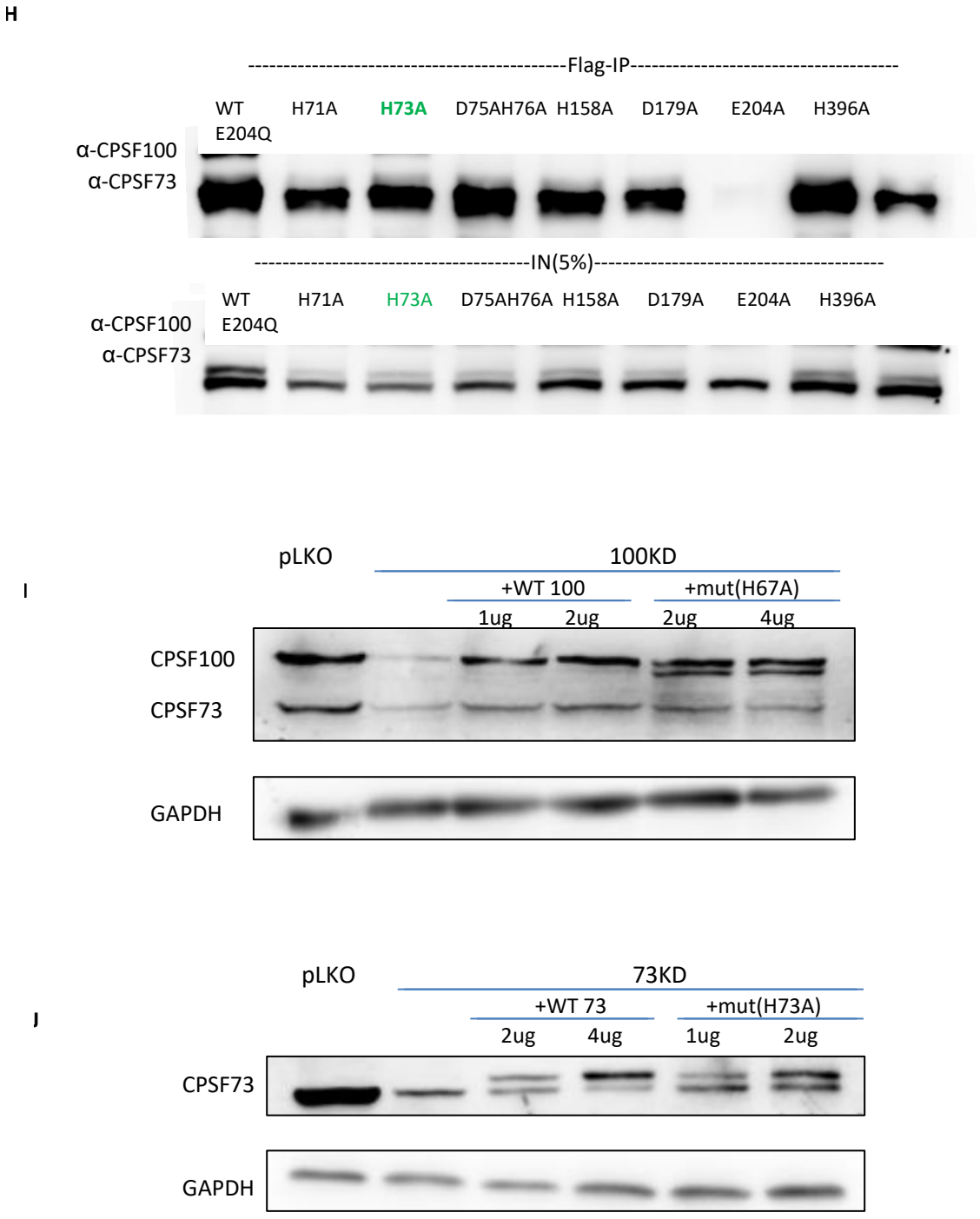


Figure 11. CPSF100 KD leads to termination defects of protein coding genes.

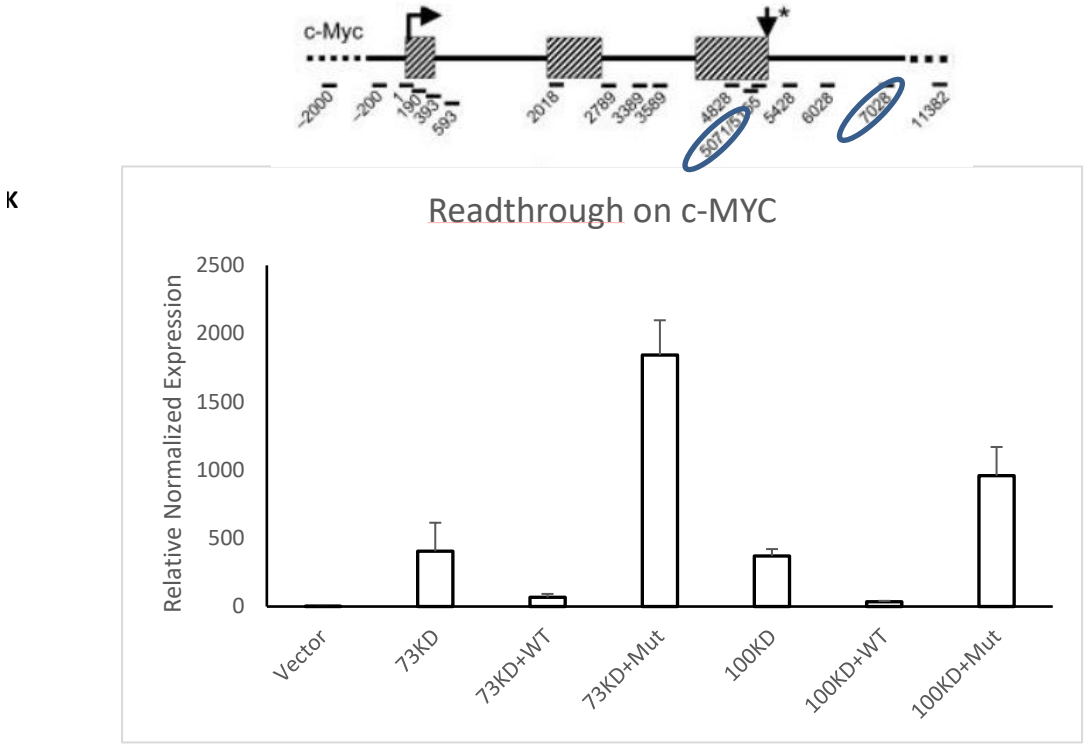


Figure 11. CPSF100 KD leads to termination defects of protein coding genes.

The readthrough defect was quantified by taking the percentage of the readthrough signal 5kb downstream of a gene to the signal over the gene body; the percentage of readthrough was then compared to control cells. In CPSF73KD cells, the difference in percentage of readthrough from control cells differed by up to 50 percentage points; in CPSF100 KD cells, the difference reached up to a hundred percentage points, indicating that there was no reduction in signal from the gene body even 5kb downstream of the transcription end site (Fig 11B).

The termination defects are clearly observed at multiple genes of varying lengths, basal expression levels, and proximities to other genes (Fig. 11C-G). The presumably inefficient cleavage that results in the transcriptional readthrough in CPSF73 and CPSF100KD cells does not result in a global shift to usage of a major distal polyadenylation site, examples of which can be seen in the PAS-Seq tracks of the genes MYC and JUN (Fig 11 C,D). Instead almost identical levels of stable, polyadenylated transcripts are found upon CPSF100KD, suggesting that most transcripts are cleaved. Yet there appears to be robust transcription after the major PAS sites; transcription fails to terminate and results in the appearance of numerous small PAS-Seq peaks until transcription finally terminates (Fig 11C-G), sometimes millions of bases later as in the case of RPS29 (Fig 11G) in the CPSF100KD condition. Thus, cleavage and polyadenylation occur at many minor distal polyA sites during the attempt to terminate transcription. Examining the termination zone of RPS29(Fig 11G) also reveals that the level of transcripts do not continue to fall evenly; there are points of rise and fall which may suggest that there could be points along the termination zone where transcription starts anew.

Another result of the transcription termination defect is that transcription reads through into downstream genes and activates transcription of genes that would otherwise not be transcribed. For example, the readthrough from PDE3A (Fig 11F) results in lower PAS-Seq signal for PDE3A and increased PAS-Seq signal at the PAS of the downstream gene. Whether these are separate transcripts of each gene or longer chimeric transcripts is not clear. It is possible that both populations exist. While transcription of downstream genes can increase, especially if they were previously quiet, readthrough can also result in transcriptional interference and result in lower levels of transcripts in the downstream gene, especially if the gene was already being actively transcribed. This phenomenon can be observed in Fig 11E, where readthrough from a histone gene upregulates transcription of the normally inactive OR2B6 olfactory receptor gene; when transcription terminates at this gene in CPSF73KD, the downstream gene ZNF165 is transcribed close to normal levels; however, when the readthrough transcription continues in the CPSF100KD condition for another 100kb, transcription levels have started to level off by the time the ZNF165 gene is reached and transcription does not increase once the gene has been reached.

To test the impact of the catalytic activity of CPSF73 on transcription termination, we developed C-terminal Flag-tagged wildtype and mutant mouse CPSF73 constructs so they would be refractory to the shRNA used. We made point mutations in the beta-casp residues, expressed the wildtype and mutant CPSF73 in HeLa cells, and made nuclear extract to perform Flag-immunoprecipitation. We found that most of the mutants did not co-immunoprecipitate CPSF100 in the same ratio as WT, suggesting that the mutation disrupted the complex. However, one mutant, H73A showed the strongest association with CPSF100 compared to the other mutants, and we decide to use that one for subsequent

experiments. We next decided to check if we could rescue the transcription termination defect in CPSF73 knockdown cells by reintroducing WT and mutant CPSF73. We measured readthrough on c-Myc downstream of the cleavage site (indicated by an arrow) using qRT-PCR (Fig 11K) and found that reintroducing CPSF73 rescued the termination defect. However, the catalytic mutant did not rescue the defect, and in fact showed a dominant negative effect and increased the level of readthrough. This would suggest that the catalytic cleavage activity of CPSF73 is important for transcription termination. However, the lack of rescue could also be because the mutant does not associate with CPSF100 in the same ratio as the wildtype.

The termination defect upon CPSF100KD was far more severe than in the case of CPSF73 alone. CPSF73 is the putative endonuclease, but because CPSF73 and CPSF100 share similar protein structures and are part of the Beta-CASP family of proteins that normally consist of nucleases, we wanted to test whether potentially enzymatically active sites in CPSF100 may also contribute to transcription termination. Similarly, we constructed mouse WT and mutant (H67A) CPSF100 mutants and performed a rescue experiment in CPSF100KD cells. Rescue was seen at the protein level (Fig 11 I) where CPSF73 levels increased in both the reintroduction of WT and mutant CPSF100, suggesting that the mutation allows for CPSF73 levels to come back to normal, presumably by at least partially stabilizing the complex again; however, previous studies (Kolev et al., 2008) have shown that the CPSF100 H67A mutant does not associate with CPSF73 in the same stoichiometry as wildtype CPSF100. WT CPSF100 was able to rescue the termination defect but mutant CPSF100 was not able to, suggesting that the potentially active sites of CPSF100 may

indeed be active in cleavage. However, we cannot rule out that the rescue of CPSF73 levels accounted for the readthrough rescue, and not any catalytic activity of CPSF100; this will have to be demonstrated by careful in vitro experiments using recombinant CPSF73 and CPSF100 proteins.

CPSF knockdown leads to termination defects of bidirectional transcription termination.

Transcription can occur bidirectionally at promoters, especially at newly evolved promoters where DNA cis-regulatory elements and perhaps the associated trans-acting factors have not yet evolved to promote productive transcription in the sense direction. In the mammalian genome, directionality of transcription and stability of transcripts may be controlled by the relative density and position of PAS and U1 snRNP recognition sites downstream of the transcription start site [TSS](Andersen, Lykke-Andersen, & Jensen, 2012),(Jin, Eser, Struhl, & Churchman, 2017),(Almada, Wu, Kriz, Burge, & Sharp, 2013). In protein coding genes, there is a higher density and precedence of U1 snRNP sites compared to PASs in the sense direction, since U1 snRNP inhibits premature cleavage and polyadenylation, whereas there is a high density of PASs close to the TSS in the antisense direction. Because of the observed CPSF100 binding at promoters, we hypothesized that CPSF100 played a central role in suppressing transcription of the antisense ***PROMoter uPstream Transcripts (PROMPTs).*** This would suggest that if the polyA complex were disturbed, there would be increased transcription levels in the antisense direction.

Results

A global analysis of the 4SU-Seq data shows that after CPSF100 depletion, a modest termination defect results in PROMPTs, but there is no relative increase in PROMPT transcription levels (Fig 12A) compared to the level of transcription in the sense direction. However, what the genome-wide analysis does not show is that at many promoters (Fig 12B,C) PROMPTs are indeed upregulated after CPSF73 and CPSF100 depletion. A number of PROMPTs appeared to be more upregulated after CPSF73KD compared to CPSF100KD after EGF stimulation (Fig 12B), perhaps because the PolII recycling is less severely affected in CPSF73KD. Notably, as seen in the EGF target gene DUSP5 (Fig 12C), PROMPTS are expressed even without EGF stimulation, suggesting CPSF100 is involved in suppressing 'leaks' in transcription. After EGF stimulation in CPSF100KD cells, transcription in the anti-sense direction is increased more than in control, and also shows a termination defect; transcription in the sense direction is less than in control, suggesting a loss in directionality and a redistribution of PolII away from the sense direction, causing a dampening effect in the response to EGF at the DUSP5 gene. These findings suggest that the role of PAS's in promoter directionality is also mediated by availability of the CPSF complex.

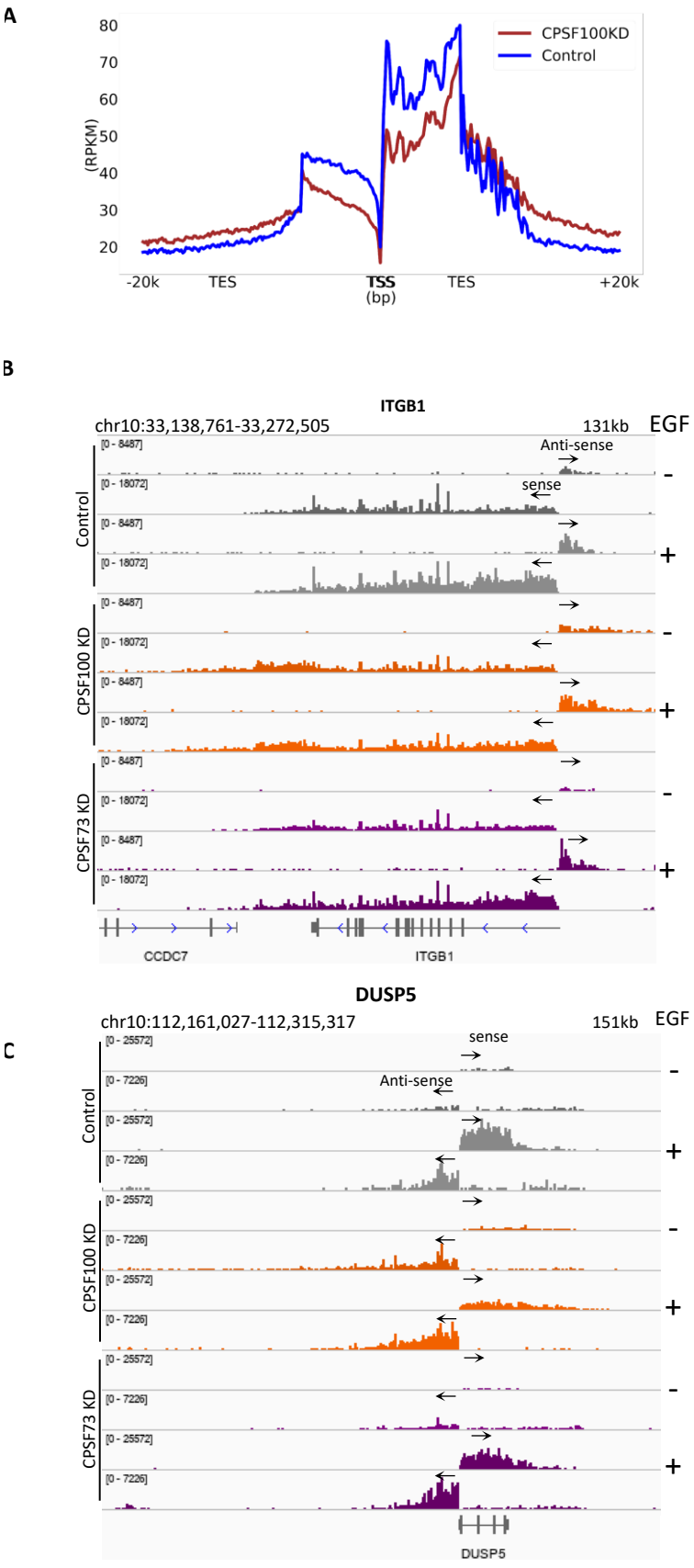


Figure 12. CPSF100 KD leads to termination defects of bidirectional transcription termination. A. 4SU-seq meta-gene profile of control and CPSF100 KD cells showing upstream promoter region **B-C**. Genomic profile of 4SU-Seq signal in both sense and anti-sense directions at (B) ITGB1 and (C) DUSP5

PolyA factors mediate transcription termination at transcripts with non-adenylated 3' ends

There are several classes of PolII transcripts that are not polyadenylated, and have alternative 3' end processing mechanisms. One such class is the small nuclear RNAs, or snRNAs, many of which are involved in splicing. They usually contain a 3' box element and a stem-loop (SL) element and are cleaved by the Integrator complex, which is also a member of the Beta-CASP family of proteins, like CPSF73/100 (Dominski, 2008) (Chen & Wagner, 2010). Another class is the histone mRNAs, which are protein-coding but are unique in not usually possessing polyA tails. Instead of a PAS, they contain an evolutionarily conserved stem-loop sequence that binds SLBP (stem-loop binding protein) and a histone downstream element (HDE) that base pairs to the U7 snRNA; the U7 snRNP is involved in recruiting the core CPSF73/CPSF100/Symplekin complex to the histone mRNA 3' end (Marzluff & Koreski, 2017).

Interestingly, while the role of the polyA complex members has not been defined for these two classes of non-polyadenylated genes, our ChIP-Seq analyses (Figure 14) revealed that all the different polyA subcomplexes, CPSF, CstF, CFIm, and CFIIm are recruited to snRNAs and histones. Previous reports have shown the presence of CstF(ref) at some histones and our results confirm these findings and show that other members like CFIm68 and WDR33 are also recruited. Perhaps if the standard cleavage mechanism fails, the polyA machinery stands by as backup to complete cleavage at a downstream PAS.

When CPSF73 and CPSF100 are knocked down, there is a significant readthrough effect at histones (Fig 13B) especially under CPSF100KD conditions and an increase in polyadenylated transcripts, suggesting that the normal cleavage mechanisms are disrupted. At an snRNA gene, there is also appearance of readthrough transcription in CPSF100 and CPSF73KD, as well as a small increase in polyadenylated transcript levels.

These results suggest that the polyA machinery is ubiquitous at PolII transcripts and serve as an important mechanism to control transcription termination.

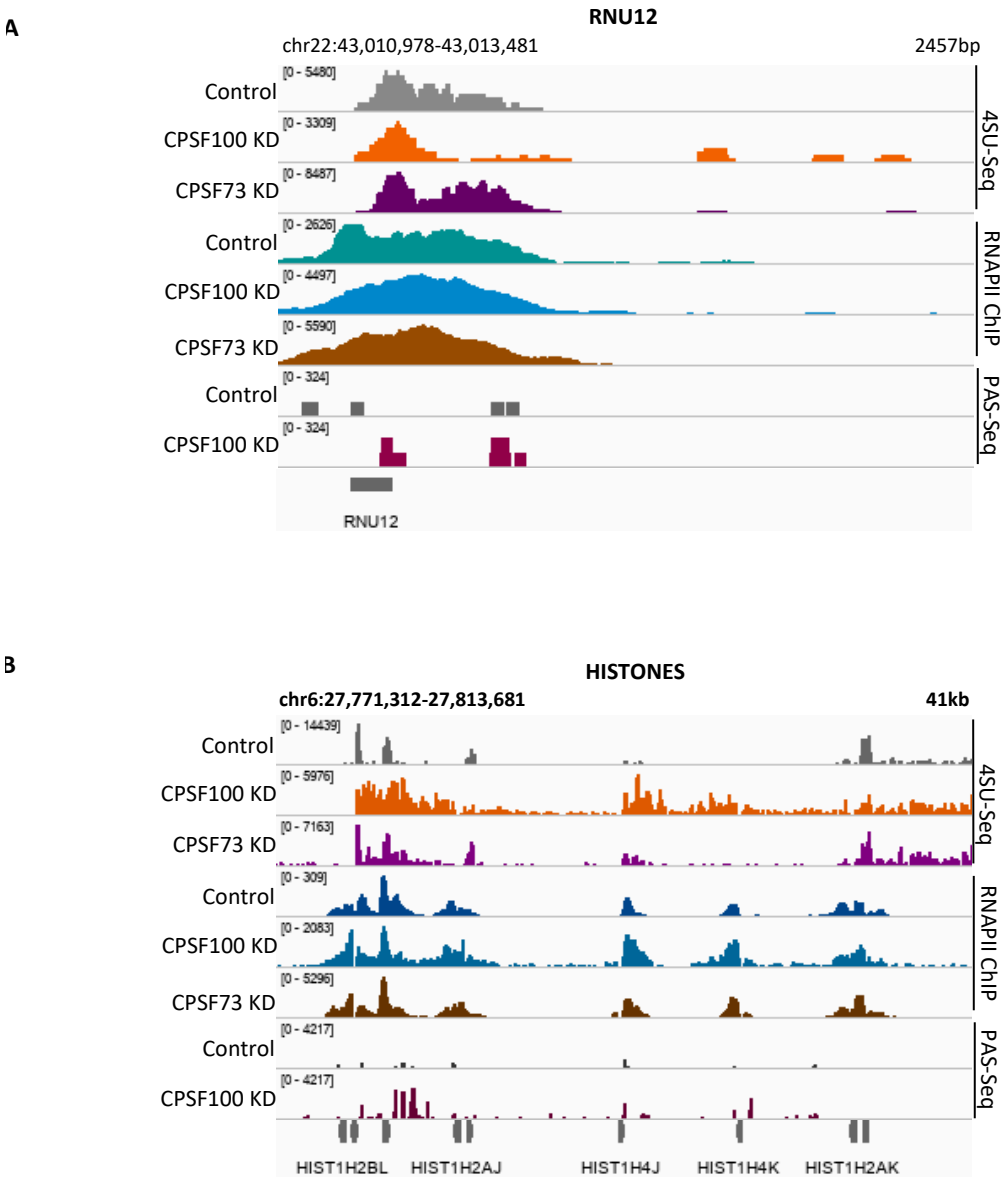


Figure 13. CPSF100 depletion inhibits termination of transcripts with alternative 3' end processing mechanisms. A-B. Genomic profile of 4SU-Seq, ChIP-Seq and PAS-Seq signal at an snRNA gene (**A**) and at a histone cluster (**B**).

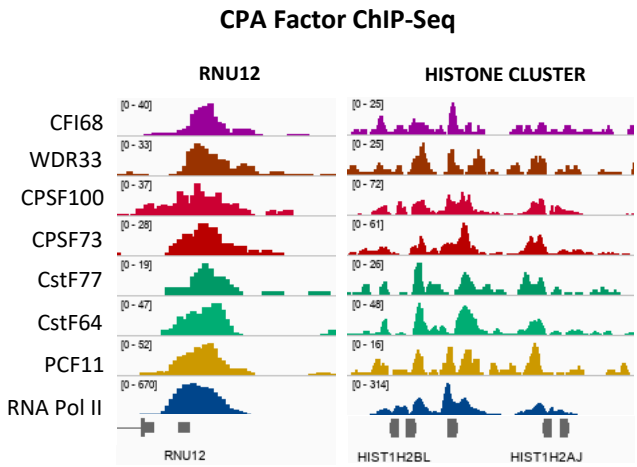


Figure 14. CPSF100 localizes to 3' ends of genes with non-adenylated transcripts. Binding profile of 3' end processing factors on RNU12 and a histone cluster.

Discussion

Knocking down CPSF factors lead to striking defects in transcription termination. How is it possible that the amount of polyadenylated product remained almost unchanged(Fig 11C,D) yet there was a significant amount of transcriptional readthrough compared to control? An unchanged level of polyA products implied that cleavage was efficient enough to cleave most transcripts; if cleavage were very inefficient, as would be expected when the main endonucleases are severely depleted, one would expect lower PAS-Seq signal at the proximal site, followed by more product at the major distal polyA site. One explanation could be that both CPSF73 and CPSF100 have cleavage activity and they can compensate for one another. However, in the CPSF100KD condition (Fig 9A), there is also very little CPSF73 left, which implies that even at low concentrations, CPSF73 can localize and efficiently cleave at certain polyA sites like those of MYC and JUN. If we conclude that that cleavage is complete at the polyA site, as the equal amount of PAS-Seq signal and 4SU signal at the polyA sites between control and CPSF100KD conditions seem to imply; the torpedo model would suggest that XRN2 would degrade the remaining nascent transcript by 5' to 3' exoribonuclease activity. The levels of XRN2 remain unchanged, yet it seems that XRN2 cannot compensate. This means that either the transcripts were not cleaved or that if they were cleaved, the CPSF complex has a significant role in transcription termination. Another explanation could be that while the PAS-Seq signals appear similar, there are a low level of transcripts that escape cleavage and because the polyA sites downstream of the main ones are weaker, it takes passing through many polyA sites before they can be terminated. This is especially clear in the case of RPS29 (Fig 11G). It is possible that both cleaved and

uncleaved transcripts contribute to the dramatic termination defects seen upon CPSF knockdown. When cleavage is inefficient, readthrough transcripts accumulate dramatically due to the pooling of the available and scarce CPSF factors at the strong gene-associated PAS; this doesn't allow enough CPSF factors to be recruited to PAS that are distant from the normal cleavage sites of the gene. If the transcript is cleaved, CPSF may have other roles in helping the transcribing PolII terminate. It may have a role in recruiting other factors like XRN2, which doesn't happen as efficiently in CPSF73 and CPSF100 KD; or perhaps it must bind to the CTD in sufficient amounts to cause a conformational shift that leads to destabilization of the PolII complex; or perhaps CPSF73 or CPSF100 have exoribonuclease activity, as proposed by some studies on histone mRNA processing (X. Yang, Sullivan, Marzluff, & Dominski, 2009). Perhaps the PolII CTD contains two populations of CPSF complex; one bound more loosely, which associates with RNA and cleaves it, and leaves PolII with the RNA, and another population that remains bound and either participates as an exoribonuclease, causes a conformational shift after passing a polyA site, or recruits other factors to help in termination. When CPSF is destabilized due to low levels of CPSF73 or 100 (along with the concomitant decrease in Symplekin and the factor that was not directly targeted for knockdown), it cannot remain bound to PolII, to aid in termination or the few CPSF complexes that bind are taken up by the RNA.

It is also clear that the termination defect is far more severe in the case of CPSF100KD than CPSF73KD. In the CPSF73KD case, there is still a significant amount of CPSF100 left; when CPSF100 is knocked down, both CPSF73 and CPSF100 are depleted, leading to a greater disruption of CPSF function. Since the level of CPSF73 protein seems comparable by Western Blot in both CPSF100 and CPSF73 KD, the additional dysfunction seen in the case

of CPSF100KD seems to suggest that CPSF100 itself has an important role in the function of the CPSF complex. The exact nature of the role is unclear. There are several possibilities that don't exclude one another. One possibility is that CPSF100 also has nuclease activity. Previous work has shown that mutating the MBL motifs of CPSF73 and CPSF100 both result in defects in cleavage activity at a histone mRNA substrate(Kolev et al., 2008). Another possibility is that both CPSF100 and CPSF73 are required for nuclease activity; a recent paper showed that the yeast homolog of CPSF73, Ysh1, requires incorporation into an 8-unit complex to demonstrate nuclease activit y(Hill, Boreikaitė, et al., 2019). Another possibility could be that CPSF73 and CPSF100 form a dimer and one of them binds RNA while the other cleaves it. Previous studies in our lab showed that CPSF100 appears to bind RNA while CPSF73 does not. Another possibility is that CPSF100 is the component that is responsible for interfacing with the rest of the CPSF complex, or PolII or other factors that may aid in termination. In fact, our collaborators recently solved a low resolution structure of the core CPSF73-CPSF100-Symplekin core complex and have identified a domain of CPSF100 that appears to interact with the other members of the CPSF complex that are responsible for identifying and binding to polyA sites. If that is the case, then depleting CPSF100 would result in a severe defect in function of the CPSF complex, since the endonuclease core complex would be very inefficiently recruited to polyA sites. This mechanism would explain our findings, but still does not explain why mutating the potential active sites would result in a termination defect.

Finally it is clear that the polyA factors are ubiquitous at PolII transcription sites, and they may serve as a failsafe termination mechanism even at non-polyadenylated transcripts. Whether mRNAs, PROMPTs, histone mRNAs, or snRNAs, CPSF73/100 play a

crucial role in controlling transcription at these sites and ensuring that readthrough transcripts are recognized at PAS, cleaved and polyadenylated (Fig 15). This ensures that readthrough PolII does not upregulate silent genes or engage in transcriptional interference at active genes.

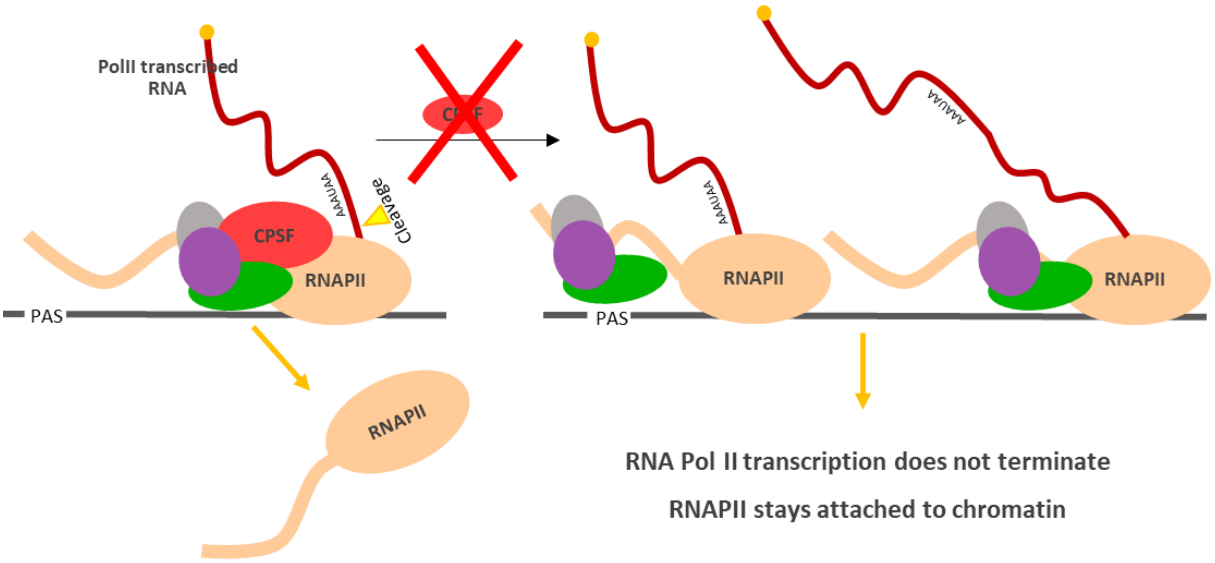


Figure 15. Working Model of CPSF action. After PAS is transcribed, RNA is cleaved and transcription can terminate.

Chapter 3: eRNA Transcription and Processing Regulation by CPSF

Introduction

Two main regulatory elements, promoters and enhancers, control gene expression and are frequently mutated in cancer(Kron, Bailey, & Lupien, 2014),(T.-K. Kim & Shiekhattar, 2015). Promoters define the transcription start site where transcription factors (TFs) can bind to control expression of genes. Enhancers are distal elements that loop to contact promoters and increase promoter activity. They are key to maintaining cell identity(Kron et al., 2014) and control the precise spatiotemporal gene transcription programs required for development, cell cycle control and cell fate; this is important because the loss of cell fate commitment and gain in pluripotency are also key features of carcinogenesis.(Ben-Porath et al., 2008) Recently, it has been shown that many enhancers are transcribed into mainly non-polyadenylated short RNAs called enhancer RNAs (eRNAs) (De Santa et al., 2010),(T.-K. Kim et al., 2010), which have a role in facilitating looping to the promoter(Y. Yang et al., 2016), recruiting Pol II to promoters and even regulating transcription pause release.

Like mRNAs, eRNAs are capped at the 5' end, but are overwhelmingly unspliced and are mostly non-polyadenylated(Djebali et al., 2012). Enhancers that show bidirectional transcription also usually have shorter transcripts and are not polyadenylated, while others are unidirectional, have longer transcripts and are polyadenylated. Significantly, eRNA induction is an independent marker of functionally active enhancers(Landt et al., 2012),(Liu et al., 2014),(Core et al., 2014) and correlates with the formation of enhancer-promoter loops(Sanyal, Lajoie, Jain, & Dekker, 2012). Because correct enhancer activation

is crucial for normal regulation of cell-type specific gene transcription programs, it is essential to understand the regulation of eRNA transcription and processing.

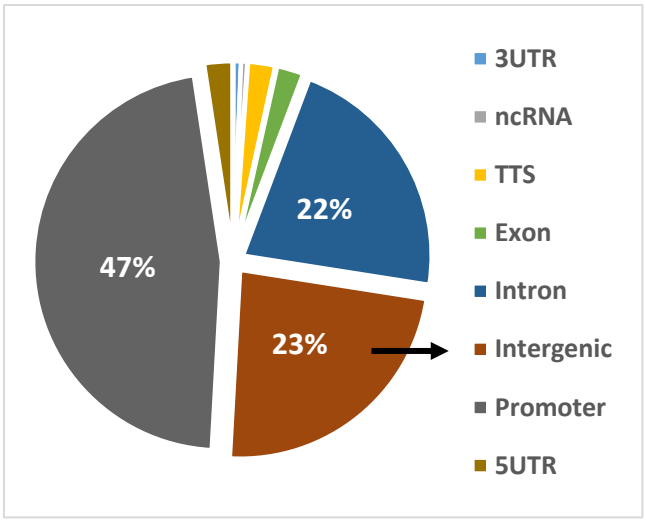
3' end processing of eRNA is just beginning to be understood. Several computational analyses have shown that enhancers are enriched with poly-A sites (PAS) above background genomic levels. However, what role the PAS and Cleavage and Polyadenylation (CPA) factors that recognize the PAS play in termination of eRNA transcription is unclear and has not been described. To date, two factors have been implicated in eRNA termination: 1. Integrator, the snRNA 3' processing factor, which has been proposed to catalyze eRNA 3' end processing(Lai, Gardini, Zhang, & Shiekhattar, 2015). 2. WDR82, which targets a histone modifier to chromatin(Austenaa et al., 2015). However, the mechanisms through which they act remain undefined. Integrator proteins Ints9 and Ints11 are structurally similar to Cleavage and Polyadenylation Specificity Factor (CPSF) subunits CPSF100 and CPSF73 respectively, but target the BoxB element in snRNAs. These BoxB elements have not been shown to be enriched in eRNAs, and it is not clear how Integrator plays a role in cleavage before or after the PAS. Because defects in termination of eRNA can lead to disruption in the normal transcription levels and stability of enhancer transcripts, termination of eRNA transcription is an important regulatory step that needs to be studied further.

Due to the implication of various different proteins in eRNA processing, it also seems likely that at different enhancers, different trans-acting protein factors play starring roles. These factors recognize different cis-elements, which may then be enriched at different levels between different enhancers. These cis-elements might have risen during evolution to tighten the regulation of enhancer activation and transcriptional programs.

Results

While the CPSF complex is known to be mainly responsible for 3' end processing of mRNAs, we were surprised to find in our ChIP-Seq studies that at steady state the CPSF complex was mainly enriched at promoters of protein-coding genes and in intronic and intergenic regions that overlapped with sites of annotated ncRNAs, including eRNAs (Fig 16A, B). To assess a potential non-canonical role of CPSF in processing eRNAs, we stimulated transcription of immediate early genes (IEGs) using EGF (see Methods) and evaluated changes in recruitment of the CPSF complex to EGF-responsive enhancer sites. To determine transcriptional activity and enrich for short-lived eRNA species, we labeled cells with 4-thiouridine (4sU) for 30 min and isolated the nascent labeled RNA for sequencing. We started by examining a set of about 900 enhancers identified based on ENCODE data of H3K27ac and H3K4me1/2 ChIP-Seq and DNAseI hypersensitivity and retained those that were also identified in the Enhancer Atlas and annotated in multiple databases (see Methods). We then kept only those enhancers in our analysis that were transcriptionally active in at least two replicates of our 4-thiouridine (4sU)-labeled nascent RNA-sequencing datasets (~765 enhancers). To assess the polyadenylation state of eRNAs, total RNA was enriched for polyadenylated RNA and the poly-A sites were subjected to high-throughput sequencing (PAS-Seq, see Methods).

We started by examining the presence of CPSF at enhancers before and after EGF stimulation. While CPSF100 could be found at enhancers before stimulation by EGF (Fig 16B), EGF induction resulted in increased CPSF 100 ChIP-Seq signal at EGF-responsive enhancers(Fig 16C-F), which were determined by their transcriptional response as measured by 4sU-Seq signal. RNA Pol II also showed a similar pattern of increased



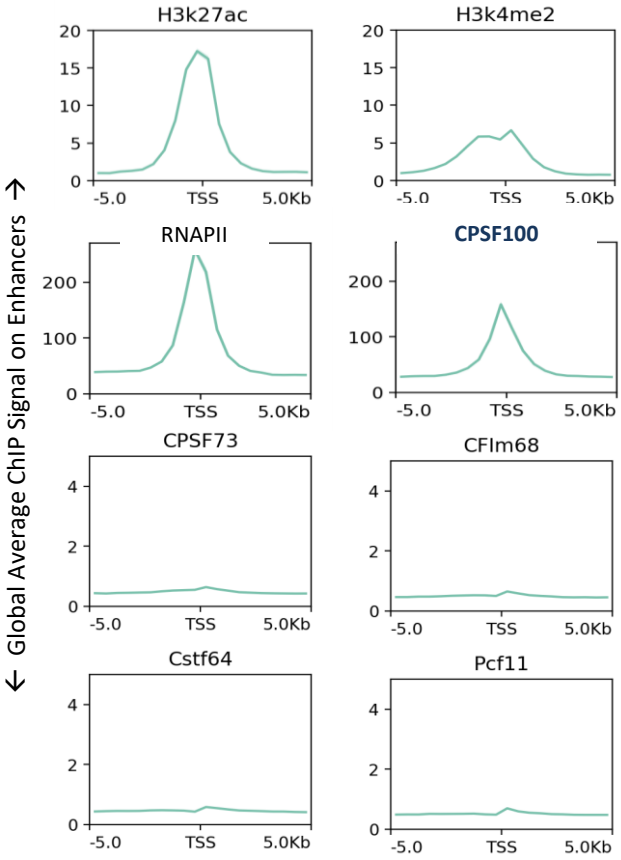


Figure 16: CPSF100 is recruited to enhancers. A. Pie chart showing percentage of CPSF100 binding sites at different genomic regions. **B.** Metaplot of binding profiles of 3' end processing factors at enhancer regions as defined by Pol II binding, H3K27ac and DNAseI hypersensitivity. **C-F**. Response to EGF activation at EGF target genes. Genomic profiles of 4SU-Seq, Pol II and CPSF100 binding before and after EGF activation at (C)DUSP1 (D)CCNL1 (E)RSPH3 (F)CLMP enhancers.

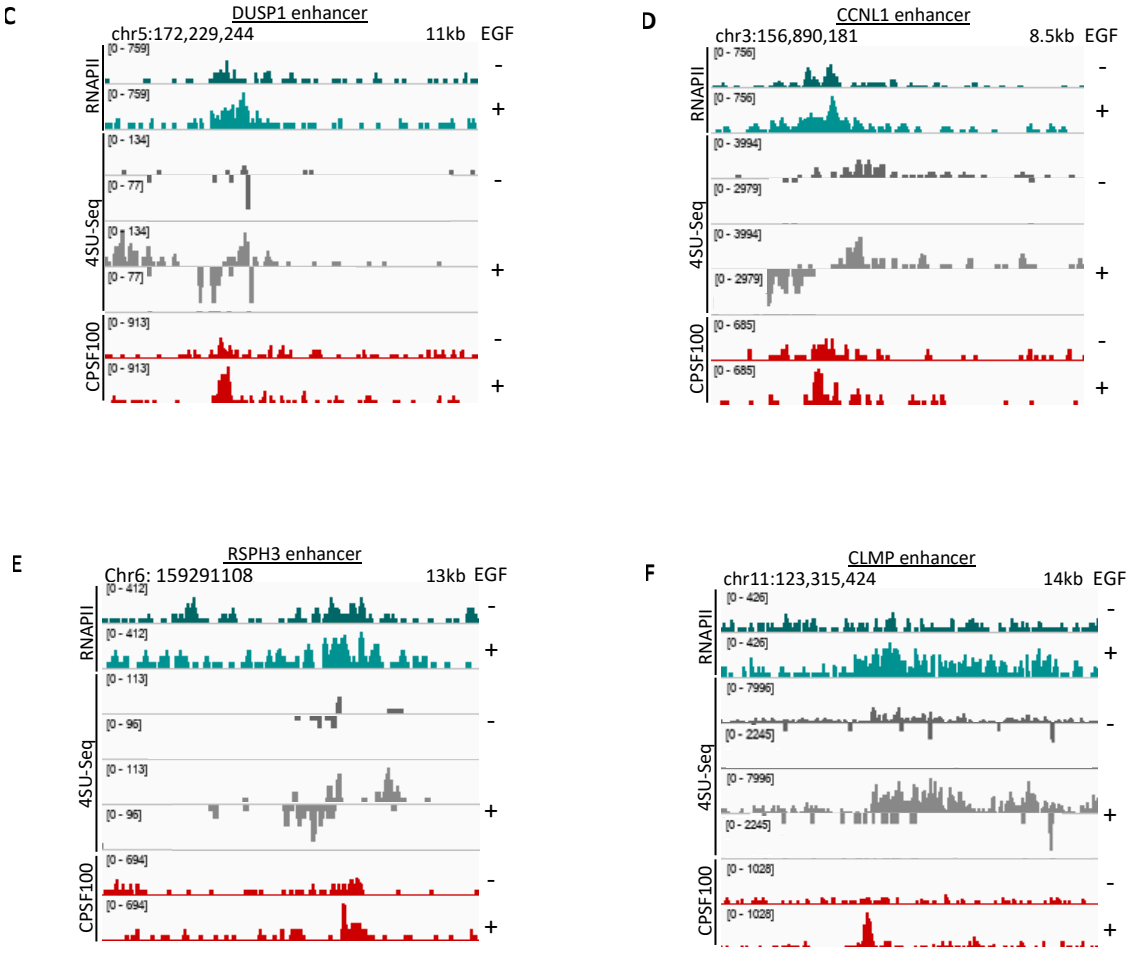


Figure 16: CPSF100 is recruited to enhancers.

recruitment to EGF-responsive enhancers after transcription activation. These results demonstrate that CPSF100, and presumably the rest of the CPSF complex is actively recruited to enhancers following their stimulation.

In order to analyze the functional impact of the CPSF complex at enhancers, we used short hairpin RNAs (shRNAs) to knock down the CPSF73 and CPSF100 subunits of the CPSF complex in HeLa cells. The other members of the CPSF complex were not significantly perturbed in each case, except for the core third Symplekin subunit. The level of Integrator was also not affected as measured by immunoblotting against Ints11.

Depleting CPSF100 and, to a lesser extent CPSF73, resulted in reduction in transcription at enhancers at steady state (no EGF condition) at many enhancers, such as that shown at the CCNL1 enhancer (Fig 17B). At these enhancers, the response to EGF treatment was dampened in CPSF100KD cells, but enhanced in CPSF73KD cells, compared to control. This could be due to reduced availability of PolII in these regions or transcriptional interference from preexisting upstream transcription._Notably, CPSF100 depletion led to increased eRNA transcripts at some EGF-dependent enhancers even without EGF stimulation. At such enhancers (Fig 17 C-E), EGF treatment enhanced the increase in transcription markedly in CPSF100KD cells, and to a lesser extent in CPSF73KD cells. It was also interesting that both CPSF73 and CPSF100KD had effects on eRNA transcription responses, suggesting that both members of the complex were involved in regulating eRNA transcription and processing, even though CPSF73 could not be detected by ChIP. This suggest that both CPSF73, CPSF100 and perhaps other members of the CPSF and polyA complex were also present, even if undetectable in our ChIP-Seq results.

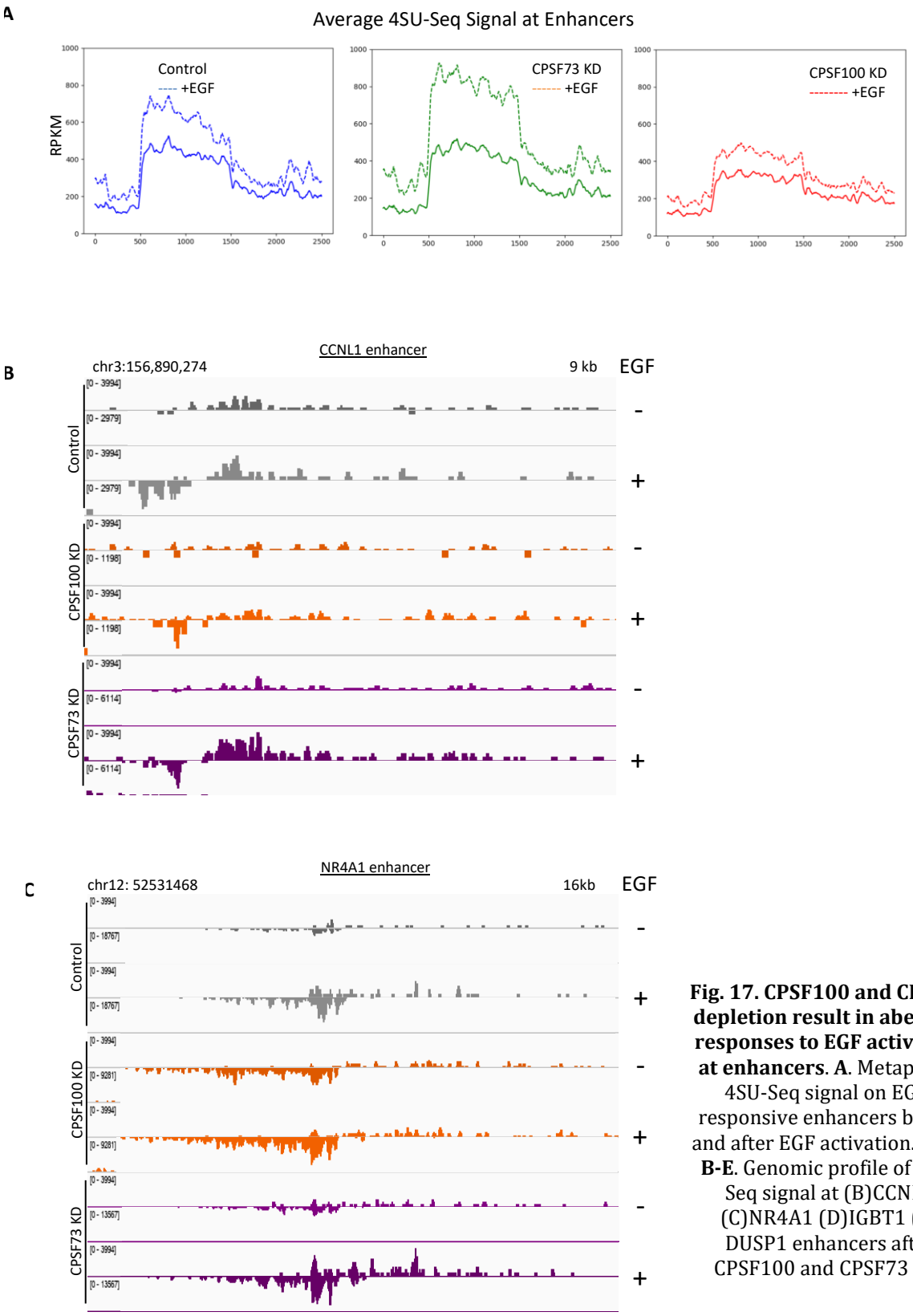


Fig. 17. CPSF100 and CPSF73 depletion result in aberrant responses to EGF activation at enhancers. **A**. Metaplot of 4SU-Seq signal on EGF-responsive enhancers before and after EGF activation. n=99 **B-E**. Genomic profile of 4SU-Seq signal at (B)CCNL1 (C)NR4A1 (D)IGBT1 (E) DUSP1 enhancers after CPSF100 and CPSF73 KD.

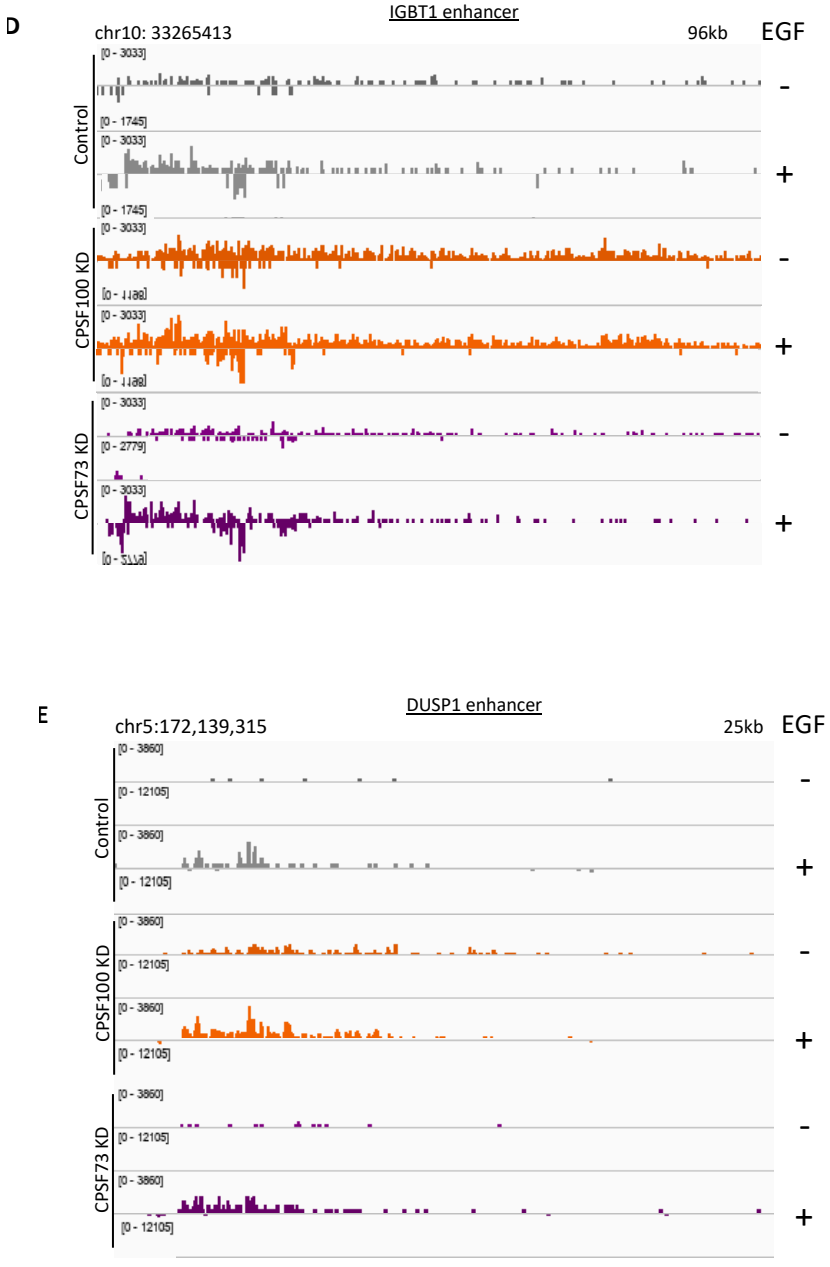


Fig. 17. CPSF100 and CPSF73 depletion result in aberrant responses to EGF activation at enhancers

In addition to disrupting responses to transcription activation, knocking down CPSF100 and CPSF73 led to transcription defects in eRNA transcription (Fig 17B, C, E), with the most striking changes seen in CPSF100KD cells (Fig 18. A-D). These effects were most notable at clusters of enhancers, or superenhancers, where CPSF100KD led to transcriptional readthrough of hundreds of kilobases (Fig 18A, B), producing large amounts of lengthy noncoding RNAs from enhancer sites. In addition, CPSF100KD lead to activation of transcription at normally quiet enhancers such as those in the DUSP1 and TMEM18 enhancers (Fig 18 C,D). Thus CPSF100KD led to both increased transcription from enhancer regions as well as termination defects of eRNA transcription. A metaplot of 4SU-signal from constitutively expressed enhancers (n=365) with signal normalized over the transcription unit showed the termination defect was most severe in CPSF100KD cells, with CPSF73KD also showing an effect (Fig 18E). This effect was quantified by measuring the readthrough as a percentage of the signal on the enhancer body and subtracting from the readthrough in control cells (Fig 18 F). The increased readthrough was statistically significant in both CPSF73KD ($p=8x10^{-13}$) and CPSF100KD ($p=8.5x10^{-19}$). A heatmap of the 4SU signal at these enhancers in control and CPSF100KD cells illustrates that the clear TES (Transcription End Site) boundary seen in the control disappears into a more diffuse boundary due to transcriptional readthrough. These results suggest CPSF100 is very important for controlling the amount and length of transcription from enhancer sites.

The transcription termination defect observed upon CPSF73 and CPSF100 depletion led us to hypothesize that the polyA complex may play a role in the cleavage and 3' end processing of eRNAs. Our analysis of the transcription end site of enhancer RNAs revealed an enrichment of the AATAAA hexamer about 20nt before the end site; this hexamer serves

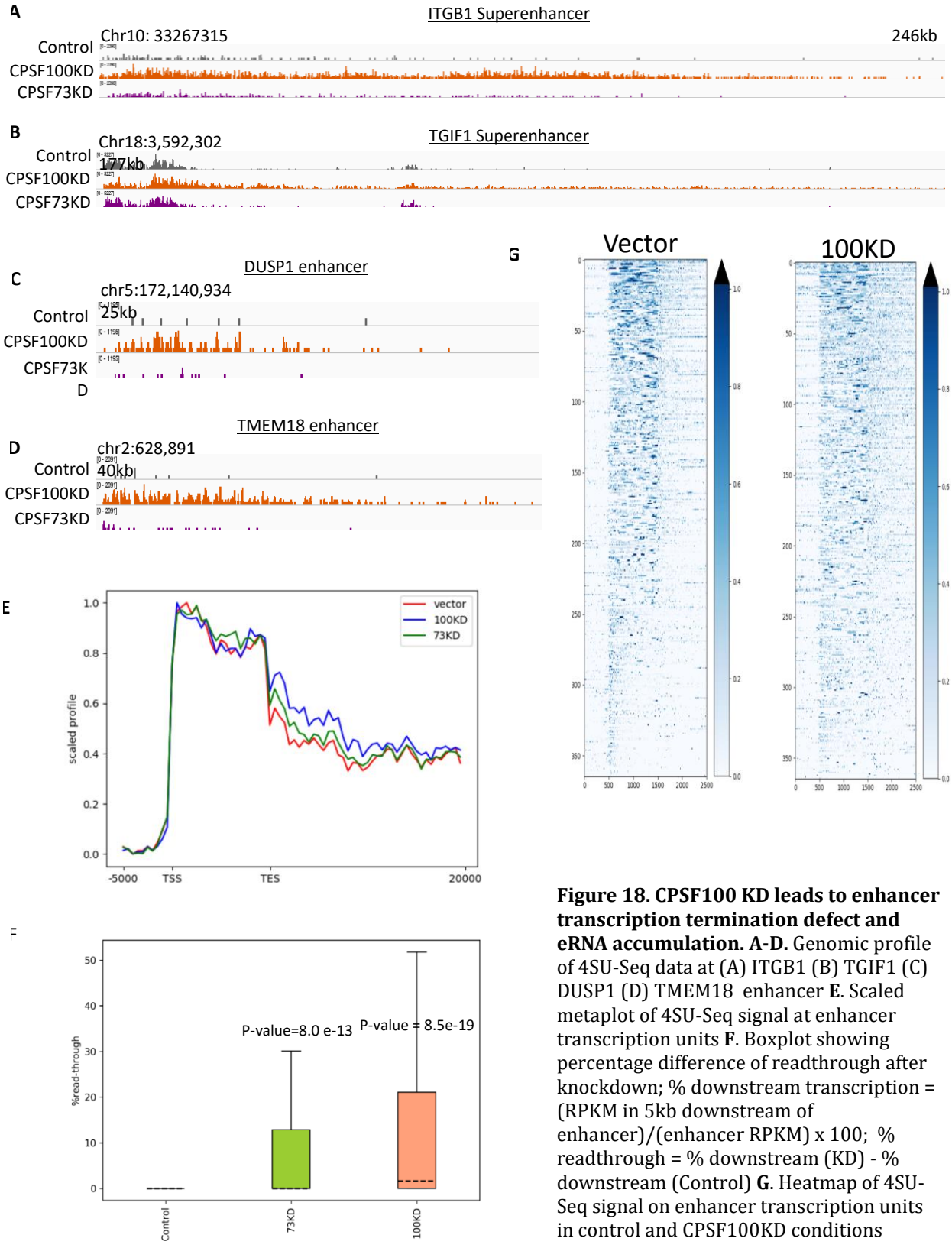


Figure 18. CPSF100 KD leads to enhancer transcription termination defect and eRNA accumulation. A-D. Genomic profile of 4SU-Seq data at (A) ITGB1 (B) TGIF1 (C) DUSP1 (D) TMEM18 enhancer **E**. Scaled metaplot of 4SU-Seq signal at enhancer transcription units **F**. Boxplot showing percentage difference of readthrough after knockdown; % downstream transcription = (RPKM in 5kb downstream of enhancer)/(enhancer RPKM) x 100; % readthrough = % downstream (KD) - % downstream (Control) **G**. Heatmap of 4SU-Seq signal on enhancer transcription units in control and CPSF100KD conditions

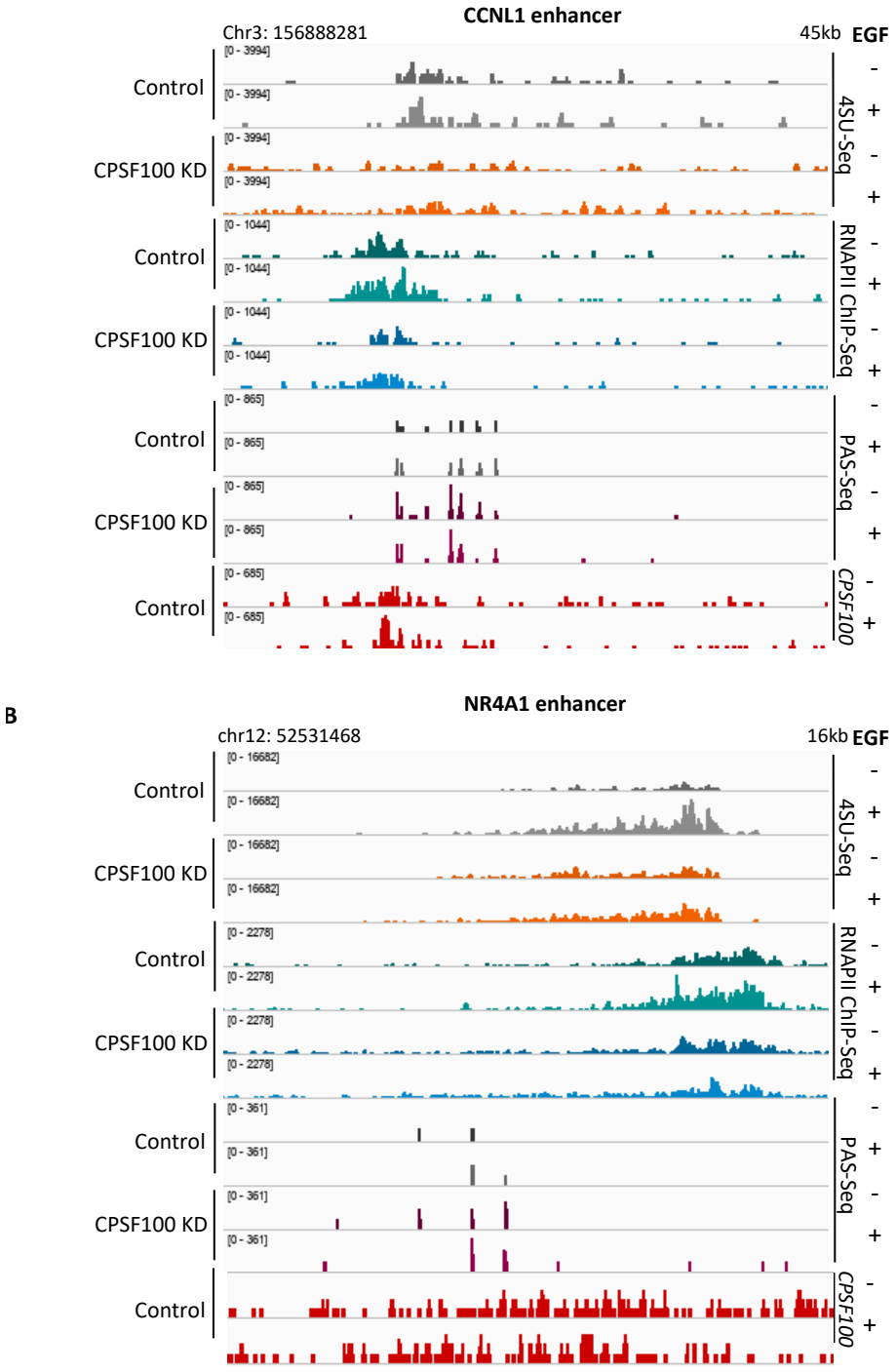


Figure 19. The CPSF complex mediates 3' processing of eRNAs. A-D. Genomic profiles of 4SU-Seq, PolII and CPSF100 ChIP-Seq, PAS-Seq in control and CPSF100KD cells at (**A**) CCNL1 (**B**) NR4A1 (**C**) DUSP1 (**D**) TMEM18 enhancers **E.** Left two panels represent metaplots at genes with CPSF100 bound; right two panels represent metaplots at genes with no CPSF100 peaks detected compared to input. Top two panels represent average 4SU-Seq signal and bottom two panels represent average PAS-Seq signal. **F-G.** Metaplot of distribution of the

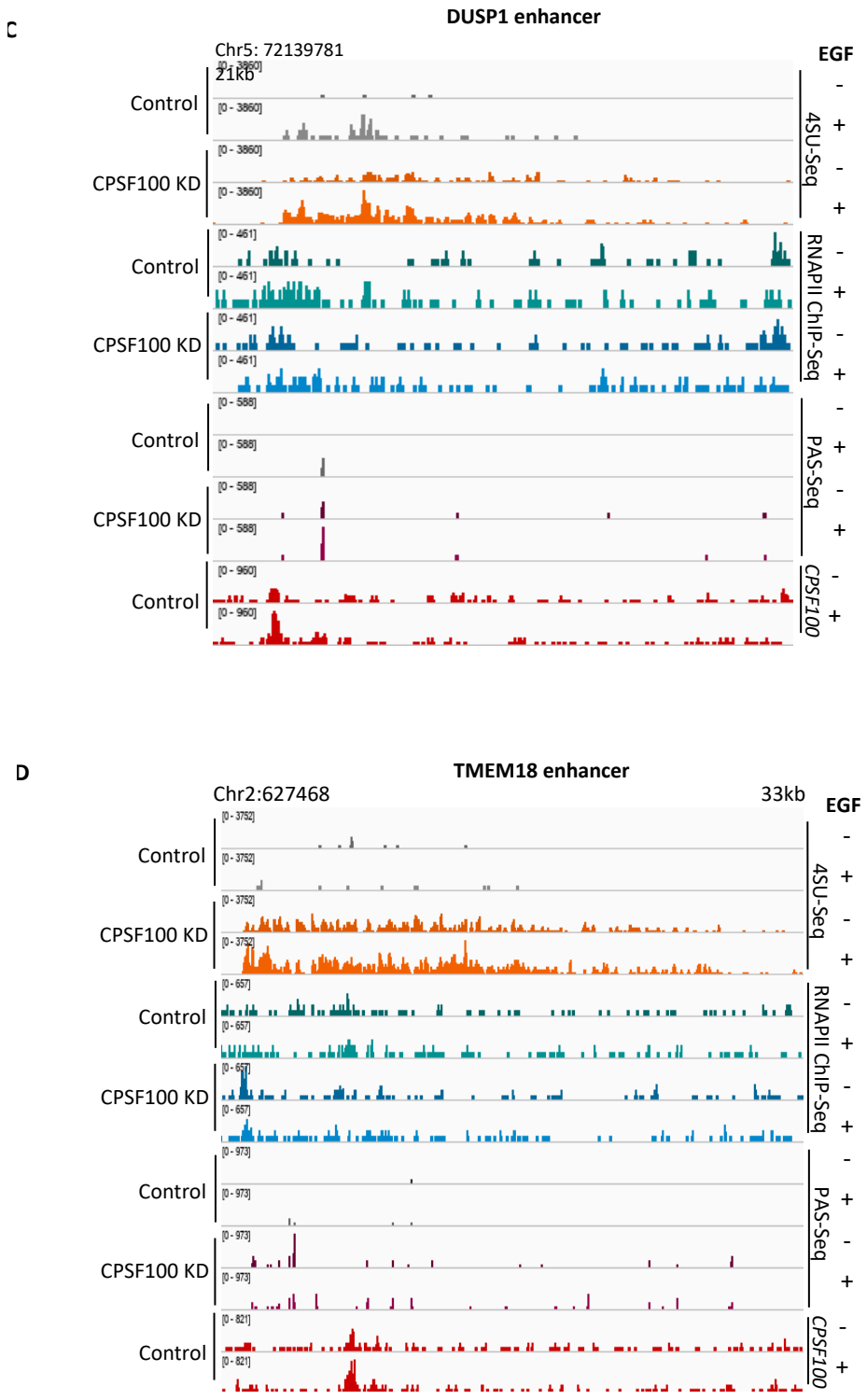
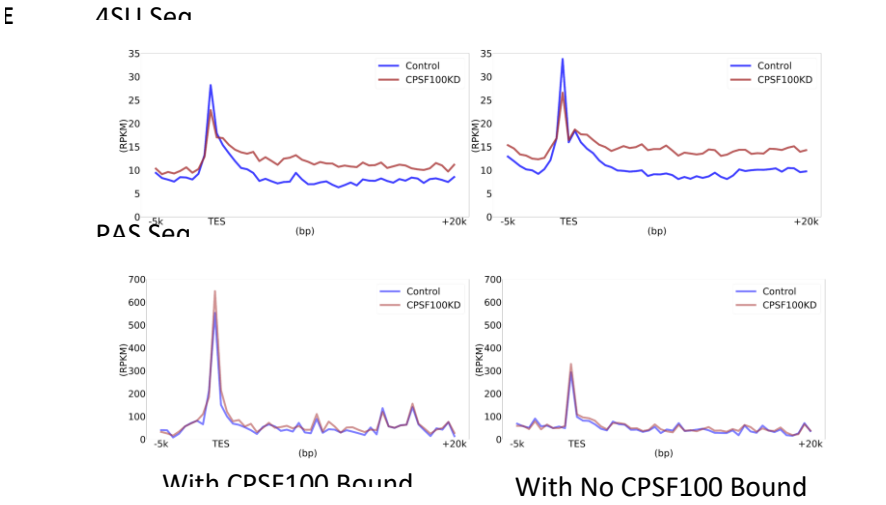


Figure 19. The CPSF complex mediates 3' processing of eRNAs.



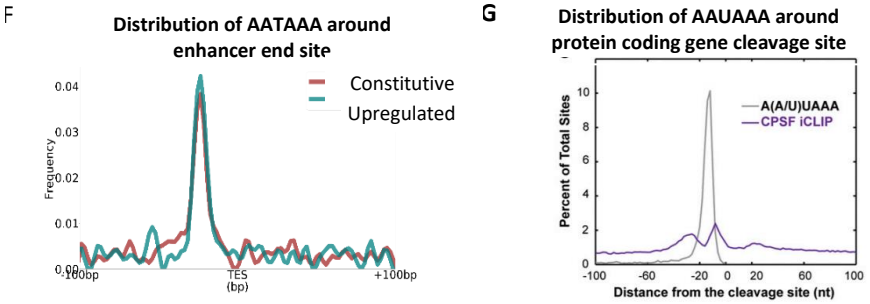


Figure 19. The CPSF complex mediates 3' processing of eRNAs.

as a binding site and signal for the CPSF complex to cleave RNA and allow for further processing such as polyadenylation (Fig 19F). In addition, the level of CPSF100 bound to enhancers correlated with a higher overall level of polyadenylated transcripts (Fig 19E); there was also a slight negative correlation between the level of CPSF100 bound and the amount of 4SU-Seq signal, i.e. there was a slightly higher level of average 4SU-Seq signal at enhancers that had sub-threshold levels of CPSF100 bound; this may indicate that the CPSF serves to control the amount of eRNAs being transcribed. Indeed, knocking down eRNAs led to increased amounts of polyadenylated eRNA transcripts as indicated by PAS-Seq data (Fig. 19A-D; Fig 19E, bottom left panel). In normally polyadenylated eRNAs such as those produced at the CCNL1 enhancer (Fig 19A), the level of polyadenylation increased, while those that normally do not express eRNAs, showed an appearance of PAS-Seq signal along the length of the enhancer derived RNA being produced after CPSF100KD (Fig 19D).

Discussion

It may be that the increased levels of eRNAs upon CPSF73 and CPSF100KD represent an accumulation of eRNAs instead of an increase in transcription. Because eRNAs themselves can have a stimulatory effect on target gene expression, an abnormal accumulation of eRNAs due to defects in 3'end processing may pose a threat to the homeostasis of the transcriptional program required to maintain cell identity. CPSF is widely dysregulated in many diseases, especially cancer(He et al., 2016)(Chang, Yeh, & Yong, 2017), and our work provides a starting point to examine how the dysregulation of CPSF disrupts enhancer function in specific diseases, and provide avenues for therapeutic intervention directly at the eRNA level, or by manipulating the CPSF complex.

Chapter 4: Summary and Conclusions

Despite the fact that cleavage and polyadenylation is an essential RNA processing event, there remains a lot to be understood about this process. Our findings here advance the understanding of the role of the cleavage and polyadenylation machinery in several important ways. First, we show that CPSF complex, particularly its members CPSF73 and CPSF100, are required for efficient transcription termination after a gene has been transcribed. Second, the role of CPSF in cleaving and terminating transcripts acts to regulate gene expression from the promoter in both positive and negative ways. Third, CPSF plays an important role in regulating the expression and processing of enhancer RNAs and other noncoding RNAs. Finally, a poorly studied member of the CPSF complex, CPSF100, is shown to be crucial for the integrity of the complex and its function. In conclusion, wherever Pol II transcribes, processing by CPSF is a general genome-wide mechanism for transcription termination and regulation.

Role of CPSF100

The function of CPSF100 in the essential CPSF complex is not well understood. Here, we show that CPSF100 is integral to the stability of the CPSF complex as well as its function. CPSF100 is homologous to CPSF73 and contains many of the conserved presumably catalytic residues(Dominski, 2008). However, it has been difficult to crystallize and functionally test in vitro(Corey R. Mandel et al., 2006b). Our results show that depleting CPSF100 also depletes Symplekin and CPSF73; when we deplete CPSF73 to the same level as CPSF73 is in the CPSF100KD cells, we find striking additional defects of transcription termination that are especially pronounced at enhancer RNAs and histone mRNAs, as well as at genes with long termination regions. This suggests that CPSF100 may have catalytic

activity; our catalytic mutant of CPSF100 failed to rescue termination defects. However, it could also be because a mutated CPSF100 fails to associate with the rest of the complex in a stable stoichiometric manner. Recent evidence in yeast has shown that nuclease activity cannot be seen unless are members of the core complex are assembled (Hill, Kumar, et al., 2019) and it could be that the active site cannot take the proper conformation when both CPSF73 and CPSF100 are not present. It could also be that CPSF100 is responsible for recruiting additional proteins to the complex necessary for transcription termination. Unpublished work by our collaborators seems to suggest that CPSF100 contains a domain that is used to bind the core cleavage subcomplex of CPSF73. CPSF100 and Symplekin to the core binding complex with Fip1, CPSF30 and WDR33.

CPSF100 may also have independent function sin disease, and it is seen to be dysregulated in certain thyroid cancers (Yon et al., 2015) and mutated in certain lymphomas (Wang et al., 2011). Understanding the precise role of CPSF100 in regulating transcription and its termination will crack open a longstanding mystery in the field and potentially provide areas for therapeutic intervention.

Role of CPSF in Transcription Termination

In this study, we showed that transcription termination is severely delayed in CPSF100 knockdown cells, with defects also seen in CPSF73 knockdown cells. This defect was observed at all types of PolII transcripts genome-wide, including mRNAs, eRNAs, snRNAs, and histone mRNAs. The termination defect could not be rescued at selected genes by presumed catalytic mutants of CPSF73 (the putative endonuclease) and CPSF100 (contains most active site residues homologous to CPSF73). This suggested that PAS cleavage is

required for transcription termination, and would support a cleavage-dependent transcription termination model such as the currently prevailing 'torpedo' model. In the torpedo model, cleavage is followed by degradation of the nascent transcript still attached to Pol II by a 5' to 3' exoribonuclease such as Xrn2 in mammalian cells.

However, we also show evidence that the cleavage-dependent 'torpedo' model may not be the main mechanism. Even though the catalytic mutants of CPSF73 and potentially CPSF100 could not rescue the defect, it does not necessarily mean that the cleavage function is required for termination; the defect could be because the mutants do not associate as strongly with the rest of the complex and may lead to the formation of an unstable or incomplete complex. In addition, studies with Xrn2 depletion (Eaton et al., 2018; Fong et al., 2015) have shown mild termination defects at best, especially when compared to the striking readthrough seen in CPSF knocked down cells. This would suggest that while XRN2-dependent transcription termination may be one mechanism of action, it is probably not the only one.

In our CPSF knockdown cells, Xrn2 levels were intact; this would imply that if there were no cleavage defect, Xrn2 would terminate transcription. However, many genes with no changes in transcript levels on the gene body and no changes in the levels of cleaved and polyadenylated products still showed severe termination defects. This supports the in vitro findings that PAS cleavage is not required for termination (Zhang, Rigo, & Martinson, 2015). In addition, Xrn2-depleted cells (Eaton et al., 2018) did not show termination defects after histone genes, while in CPSF73 and especially CPSF100-depleted cells, the transcription termination defects at histone clusters were quite striking. These results suggest that perhaps another mechanism, such as the 'conformational change' model is at

play. In this model, PolII undergoes a conformational change after passing over a PAS, with or without cleavage. This conformational change might be potentially mediated by the action or assembly of the polyA machinery, including CPSF.

Our findings also leave open the possibility for other models of transcription termination. One alternative explanation of the severe termination defects upon CPSF knockdown could be that CPSF contains exoribonuclease activity, which has been proposed before in relation to histone mRNAs (X. Yang et al., 2009), or is responsible for recruiting the main exoribonuclease or exosome component to degrade nascent uncapped RNA or RNA looping out of backtracked PolII (Lemay et al., 2014). Another potential model could be a 'tug of war', where CPSF stays attached to both the RNA and PolII after PAS transcription, and PolII is decelerated by the drag of the RNA and 3' end processing machinery. When CPSF concentration is low in the environment, perhaps the kinetics of staying attached become less favorable and CPSF dissociates more easily from PolII.

While it is difficult to precisely decipher the molecular mechanism of transcription termination in cells, we provide important evidence that the action of the CPSF complex is necessary for efficient termination, more so than Xrn2. Because we see equal levels of cleaved and polyadenylated transcripts in control and knockdown cells, but a severe termination defect in knockdown cells, we also provide compelling data to support the hypothesis that cleavage is not required for termination, although it may contribute to more efficient termination. The results of our study provide a clarity to the almost 20-year old debate around the mechanisms of transcription termination.

Role of CPSF in terminating transcripts to regulate gene expression

In our study, we found that knocking down CPSF results in a net upregulation of gene expression, while some genes were upregulated. While not widely discussed in the field, transcription termination is an important mechanism for gene regulation. We show that this can be achieved in numerous ways.

CPSF-mediated transcription termination allows positive expression of genes that would otherwise be downregulated due to transcriptional interference from adjacent genes. This is especially true of closely spaced active genes. On the other hand, transcription termination prevents readthrough into downstream inactive genes; thus, when CPSF is downregulated, many inactive genes are stimulated and form polyadenylated transcripts due to readthrough transcription from upstream genes. This type of defective readthrough can also be seen under cellular stress (Vilborg et al., 2017) and in cancer cells (Grosso et al., 2015) where it can interfere with normal expression or cause expression of large chimeric transcripts.

Timely transcription termination also allows for recycling of PolII to the promoter and normal responses to EGF activation. Without this process, Pol II is engaged in nonproductive transcription over large swathes of the genome and cannot be engaged effectively to respond to transcription activation signals at targeted genes. A more focal redistribution of PolII at the promoter also affects transcription of genes; a consequence of the reduction in promoter directionality seen in CPSF knockdown cells is the redistribution of transcribing PolII from the start site into the anti-sense direction; this leaves less Pol II to transcribe the gene and results in lower gene expression. What PolII is present and paused at the start site cannot be effectively released upon EGF activation without CPSF. It is not

clear what mechanism CPSF employs to allow normal pause release; recent studies have shown that the PolII population at the promoter is not stagnant and has rapid turnover (ref) and produces short TSS-RNAs (ref); this could mean that a cleavage and termination factor is needed to cleave these short RNAs and terminate transcription and destabilize PolII binding at the promoter to allow for more Pol II to come in attempt productive transcription. We would be the first to show that CPSF depletion leads to impaired pause release.

At the promoter, premature termination also seems to play a role in gene regulation at a subset of genes. This phenomenon has been mostly studied in the context of promoter upstream transcripts and promoter directionality in mammalian cells (Nojima et al., 2015); what these studies failed to point out is that in unperturbed cells, premature transcription termination keeps transcription at check in both directions and keeps certain genes quiet. Thus, when U1 snRNP, which acts to suppress polyA factors, is depleted in cells , polyA factors can act without restraint and widespread premature transcription of protein coding genes is seen (Almada et al., 2013; Kaida et al., 2010). Recent studies in mammalian and *Drosophila* have also shown that PolII accumulation at the promoter is not solely due to PolII pausing but is also associated with premature transcription termination (Krebs et al., 2017; Nojima et al., 2015) and the formation of short transcription start site RNAs 9-20nt long (Taft et al., 2009). When transcription termination is extremely efficient at the promoter in low expressing genes, the gene stays silent; when CPSF is removed, transcription can now proceed.

Thus, depending on the level of transcription of a gene in unperturbed cells, and its proximity to other genes, transcription termination can act to either upregulate or downregulate expression.

Role of CPSF in regulating eRNA and ncRNA transcription

One of the most remarkable effects seen upon CPSF100 depletion was the appearance of extensive regions of noncoding transcription in intergenic and noncoding regions. These long ncRNAs emerged from various locations. At active enhancer regions, existing transcription failed to terminate and resulted in abnormally long eRNAs; at other normally inactive enhancers or unannotated regions, the shortage of CPSF100 resulted in the appearance of transcripts. Noncoding transcription also extended beyond the control condition or was increased in the antisense direction from promoters, forming PROMPTs that showed termination delays. Thus, it is clear that CPSF negatively regulates the pervasive PolII derived transcription seen genome-wide to ensure only the appropriate transcripts are expressed. In addition, noncoding RNA transcription is also regulated by the CPSF complex.

Along these lines, another important contribution of this finding is explaining how eRNAs are processed. Our current understanding in the field has been that Integrator, which normally processes snRNAs, cleaves eRNAs(Lai et al., 2015). However, the defects seen upon CPSF depletion appear far more severe even though Integrator levels remain unchanged; interestingly, slight readthrough transcription is also observed upon CPSF depletion at snRNAs, which are processed by Integrator. Thus, it appears that different cleavage and termination mechanism work together in a compound fashion to process eRNAs and noncoding RNAs in general.

This is important at noncoding regions due to several reasons. Changes in levels of eRNAs have been shown to have an effect on gene expression(Bose & Berger, 2017; Pnueli, Rudnizky, Yosefzon, & Melamed, 2015)(Melo et al., 2013) and many eRNAs are dysregulated in diseases such as cancer(Ding et al., 2018). Transcription is also a source of genomic instability (N. Kim & Jinks-Robertson, 2012) and the large areas of the genome that are opened up for transcription in CPSF100KD may result in aberrant recombination and other types of instability. In addition, dysregulated long noncoding RNAs may recruit chromatin remodelers in an ectopic manner (Rinn & Chang, 2012) and cause other epigenetic changes that lead to abnormal gene expression programs.

Conclusions

This work advances our field by improving our understanding of the current dogma in at least two important ways. First, it defines a central role for the CPSF complex in the 3' end processing of eRNAs. The current understanding in the field has been that eRNA 3' ends are cleaved by the Integrator complex. However, our results show that the polyA signal recognized by CPSF is enriched at 3' ends of eRNAs and that depletion of CPSF results in termination defects of eRNAs and accumulation of eRNA transcripts. These effects are more pronounced upon CPSF depletion as compared to Integrator depletion. It is possible that both complexes work in tandem, with Integrator acting as a failsafe to terminate escaped transcripts from CPSF. Second, this work challenges the notion that the currently prevailing torpedo model is the major mechanism for transcription termination. In our work, we found that CPSF-depleted cells retained similar levels of cleaved and polyadenylated transcripts of most genes, as compared to wildtype cells, but that readthrough defects were still severe. This was despite normal levels of the exonuclease

Xrn2, which plays the ‘torpedo’ in the torpedo model of termination. This implies that retaining a high concentration of the CPSF complex is essential for efficient termination, and that the conformational change model is probably the major mechanism of transcription termination, with the torpedo model acting as a failsafe to make termination more efficient.

This work shows that not only is CPSF important for transcription termination at protein coding genes, it is vital for maintaining the very integrity of the transcriptome and plays a general role in most transcription related activities genome wide, whether at promoters, transcription end sites, or noncoding regulatory regions like enhancers. Understanding how each component of the various factors that contribute to transcription termination and expression of regulatory RNAs will help us better develop the means to manipulate the gene expression program for understanding the function of our cells and producing new therapies.

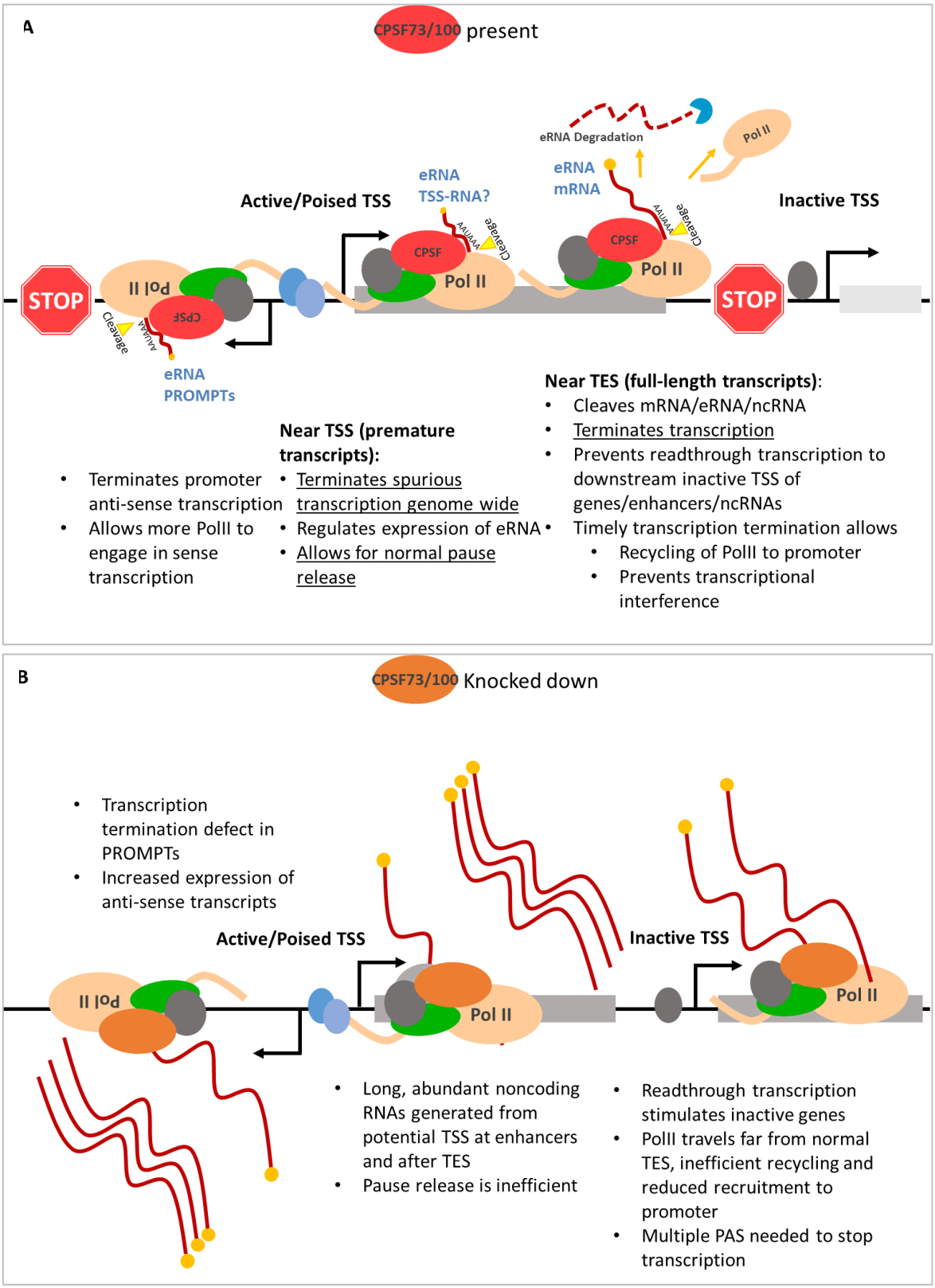


Figure 20. A. Working model of how CPSF regulates transcription by cleavage and termination at various regions of transcription **B.** Working model of how depletion of CPSF affects transcriptional activity at various regions of transcription

References

Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B., & Sharp, P. A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature*, *499*(7458), 360–363. https://doi.org/10.1038/nature12349

Andersen, P. K., Lykke-Andersen, S., & Jensen, T. H. (2012). Promoter-proximal polyadenylation sites reduce transcription activity. *Genes & Development*, *26*(19), 2169–2179. https://doi.org/10.1101/gad.189126.112

Aravind, L. (1999). An evolutionary classification of the metallo-beta-lactamase fold proteins. *In Silico Biology*, *1*(2), 69–91. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11471246

Austenaa, L. M. I., Barozzi, I., Simonatto, M., Masella, S., Della Chiara, G., Ghisletti, S., … Natoli, G. (2015). Transcription of Mammalian cis-Regulatory Elements Is Restrained by Actively Enforced Early Termination. *Molecular Cell*, *60*(3), 460–474. https://doi.org/10.1016/j.molcel.2015.09.018

Baejen, C., Andreani, J., Torkler, P., Battaglia, S., Schwalb, B., Lidschreiber, M., … Cramer, P. (2017). Genome-wide Analysis of RNA Polymerase II Termination at Protein-Coding Genes. *Molecular Cell*, *12*(0), 435–445. https://doi.org/10.1016/j.molcel.2017.02.009

Bailey, T. L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics (Oxford, England)*, *27*(12), 1653–1659. https://doi.org/10.1093/bioinformatics/btr261

Ben-Porath, I., Thomson, M. W., Carey, V. J., Ge, R., Bell, G. W., Regev, A., & Weinberg, R. A. (2008). An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nature Genetics*, *40*(5), 499–507. https://doi.org/10.1038/ng.127

Bose, D. A., & Berger, S. L. (2017). eRNA binding produces tailored CBP activity profiles to regulate gene expression. *RNA Biology*, *14*(12), 1655–1659. https://doi.org/10.1080/15476286.2017.1353862

Callebaut, I., Moshous, D., Mornon, J.-P., & de Villartay, J.-P. (2002). Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family. *Nucleic Acids Research*, *30*(16), 3592–3601. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12177301

Chan, S. L., Huppertz, I., Yao, C., Weng, L., Moresco, J. J., Yates, J. R., … Shi, Y. (2014). CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing. *Genes & Development*, *28*(21), 2370–2380. https://doi.org/10.1101/gad.250993.114

Chang, J. W., Yeh, H. S., & Yong, J. (2017). Alternative Polyadenylation in Human Diseases. *Endocrinology and Metabolism (Seoul, Korea)*, *32*(4), 413–421. https://doi.org/10.3803/EnM.2017.32.4.413

Chen, J., & Wagner, E. J. (2010). snRNA 3' end formation: the dawn of the Integrator complex. *Biochemical Society Transactions*, *38*(4), 1082–1087. https://doi.org/10.1042/BST0381082

Chou, Z. F., Chen, F., & Wilusz, J. (1994). Sequence and position requirements for uridylate-rich downstream elements of polyadenylation signals. *Nucleic Acids Research*, *22*(13), 2525–2531. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7518915

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of

nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*, *46*(12), 1311–1320. https://doi.org/10.1038/ng.3142

Danckwardt, S., Hentze, M. W., & Kulozik, A. E. (2008). 3′ end mRNA processing: molecular mechanisms and implications for health and disease. *The EMBO Journal*, *27*(3), 482–498. https://doi.org/10.1038/sj.emboj.7601932

Dantonel, J. C., Murthy, K. G., Manley, J. L., & Tora, L. (1997). Transcription factor TFIID recruits factor CPSF for formation of 3' end of mRNA. *Nature*, *389*(6649), 399–402. https://doi.org/10.1038/38763

de la Sierra-Gallay, I. L., Zig, L., Jamalli, A., & Putzer, H. (2008). Structural insights into the dual activity of RNase J. *Nature Structural & Molecular Biology*, *15*(2), 206–212. https://doi.org/10.1038/nsmb.1376

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., … Natoli, G. (2010). A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLoS Biology*, *8*(5), e1000384. https://doi.org/10.1371/journal.pbio.1000384

de Vries, H., Rüegsegger, U., Hübner, W., Friedlein, A., Langen, H., & Keller, W. (2000). Human pre-mRNA cleavage factor IIm contains homologs of yeast proteins and bridges two other cleavage factors. *The EMBO Journal*, *19*(21), 5895–5904. https://doi.org/10.1093/emboj/19.21.5895

Ding, M., Liu, Y., Liao, X., Zhan, H., Liu, Y., & Huang, W. (2018). Enhancer RNAs (eRNAs): New Insights into Gene Transcription and Disease Treatment. *Journal of Cancer*, *9*(13), 2334–2340. https://doi.org/10.7150/jca.25829

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., … Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, *489*(7414), 101–108. https://doi.org/10.1038/nature11233

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dominski, Z. (2008). Nucleases of the Metallo-β-lactamase Family and Their Role in DNA and RNA Metabolism. *Http://Dx.Doi.Org/10.1080/10409230701279118*.

Eaton, J. D., Davidson, L., Bauer, D. L. V., Natsume, T., Kanemaki, M. T., & West, S. (2018). Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity. *Genes & Development*, *32*(2), 127. https://doi.org/10.1101/GAD.308528.117

Fong, N., Brannan, K., Erickson, B., Kim, H., Cortazar, M. A., Sheridan, R. M., … Bentley, D. L. (2015). Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Molecular Cell*, *60*(2), 256–267. https://doi.org/10.1016/j.molcel.2015.09.026

Glover-Cutter, K., Kim, S., Espinosa, J., & Bentley, D. L. (2008). RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nature Structural & Molecular Biology*, *15*(1), 71–78. https://doi.org/10.1038/nsmb1352

Grosso, A. R., Leite, A. P., Carvalho, S., Matos, M. R., Martins, F. B., Vítor, A. C., … de Almeida, S. F. (2015). Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. *ELife*, *4*. https://doi.org/10.7554/eLife.09214

He, X.-J., Zhang, Q., Ma, L.-P., Li, N., Chang, X.-H., & Zhang, Y.-J. (2016). Aberrant Alternative Polyadenylation is Responsible for Survivin Up-regulation in Ovarian Cancer. *Chinese*

Medical Journal, *129*(10), 1140. https://doi.org/10.4103/0366-6999.181965
Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., … Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, *38*(4), 576–589. https://doi.org/10.1016/j.molcel.2010.05.004
Hill, C. H., Boreikaitė, V., Kumar, A., Casañal, A., Kubík, P., Degliesposti, G., … Passmore, L. A. (2019). Activation of the Endonuclease that Defines mRNA 3' Ends Requires Incorporation into an 8-Subunit Core Cleavage and Polyadenylation Factor Complex. *Molecular Cell*, *73*(6), 1217-1231.e11. https://doi.org/10.1016/j.molcel.2018.12.023
Hill, C. H., Kumar, A., Girbig, M., Skehel, M., & Passmore, L. A. (2019). Activation of the Endonuclease that Defines mRNA 3 0 Ends Requires Incorporation into an 8-Subunit Core Cleavage and Polyadenylation Factor Complex Correspondence. *Molecular Cell*, *73*. https://doi.org/10.1016/j.molcel.2018.12.023
Hirose, Y., & Manley, J. L. (1998). RNA polymerase II is an essential mRNA polyadenylation factor. *Nature*, *395*(6697), 93–96. https://doi.org/10.1038/25786
Hollingworth, D., Noble, C. G., Taylor, I. A., & Ramos, A. (2006). RNA polymerase II CTD phosphopeptides compete with RNA for the interaction with Pcf11. *RNA (New York, N.Y.)*, *12*(4), 555–560. https://doi.org/10.1261/rna.2304506
Hu, J., Lutz, C. S., Wilusz, J., & Tian, B. (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA (New York, N.Y.)*, *11*(10), 1485–1493. https://doi.org/10.1261/rna.2107305
Jin, Y., Eser, U., Struhl, K., & Churchman, L. S. (2017). The Ground State and Evolution of Promoter Region Directionality. *Cell*, *170*(5), 889-898.e10. https://doi.org/10.1016/j.cell.2017.07.006
Kaida, D., Berg, M. G., Younis, I., Kasim, M., Singh, L. N., Wan, L., & Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, *468*(7324), 664–668. https://doi.org/10.1038/nature09479
Kamieniarz-Gdula, K., Gdula, M. R., Panser, K., Brockdorff, N., Pauli, A., & Proudfoot Correspondence, N. J. (2019). Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination. *Molecular Cell*, *74*, 158–172. https://doi.org/10.1016/j.molcel.2019.01.027
Kaufmann, I., Martin, G., Friedlein, A., Langen, H., & Keller, W. (2004). Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *The EMBO Journal*, *23*(3), 616–626. https://doi.org/10.1038/sj.emboj.7600070
Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*, *14*(4), R36. https://doi.org/10.1186/gb-2013-14-4-r36
Kim, N., & Jinks-Robertson, S. (2012). Transcription as a source of genome instability. *Nature Reviews Genetics*, *13*(3), 204–214. https://doi.org/10.1038/nrg3152
Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., … Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–187. https://doi.org/10.1038/nature09033
Kim, T.-K., & Shiekhattar, R. (2015). Leading Edge Review Architectural and Functional Commonalities between Enhancers and Promoters. *Cell*, *162*, 948–959. https://doi.org/10.1016/j.cell.2015.08.008
Kolev, N. G., Yario, T. A., Benson, E., & Steitz, J. A. (2008). Conserved motifs in both CPSF73

and CPSF100 are required to assemble the active endonuclease for histone mRNA 3'-end maturation. *EMBO Reports*, *9*(10), 1013–1018. https://doi.org/10.1038/embor.2008.146

Krebs, A. R., 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*(3), 411-422.e4. https://doi.org/10.1016/j.molcel.2017.06.027

Kron, K. J., Bailey, S. D., & Lupien, M. (2014). Enhancer alterations in cancer: a source for a cell identity crisis. *Genome Medicine*, *6*(9), 77. https://doi.org/10.1186/s13073-014-0077-3

Lai, F., Gardini, A., Zhang, A., & Shiekhattar, R. (2015). Integrator mediates the biogenesis of enhancer RNAs. *Nature*, *525*(7569), 399–403. https://doi.org/10.1038/nature14906

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., … Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, *22*(9), 1813–1831. https://doi.org/10.1101/gr.136184.111

Legrand, P., Pinaud, N., Minvielle-Sebastia, L., & Fribourg, S. (2007). The structure of the CstF-77 homodimer provides insights into CstF assembly. *Nucleic Acids Research*, *35*(13), 4515–4522. https://doi.org/10.1093/nar/gkm458

Lemay, J.-F., Larochelle, M., Marguerat, S., Atkinson, S., Bähler, J., & Bachand, F. (2014). The RNA exosome promotes transcription termination of backtracked RNA polymerase II. *Nature Structural & Molecular Biology*, *21*(10), 919–926. https://doi.org/10.1038/nsmb.2893

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Liu, Z., Merkurjev, D., Yang, F., Li, W., Oh, S., Friedman, M. J., … Rosenfeld, M. G. (2014). Enhancer Activation Requires trans-Recruitment of a Mega Transcription Factor Complex. *Cell*, *159*(2), 358–373. https://doi.org/10.1016/j.cell.2014.08.027

MacDonald, C. C., Wilusz, J., & Shenk, T. (1994). The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. *Molecular and Cellular Biology*, *14*(10), 6647–6654. https://doi.org/10.1128/MCB.14.10.6647

Mandel, C. R., Bai, Y., & Tong, L. (2008). Protein factors in pre-mRNA 3′-end processing. *Cellular and Molecular Life Sciences*, *65*(7–8), 1099–1122. https://doi.org/10.1007/s00018-007-7474-3

Mandel, Corey R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J. L., & Tong, L. (2006a). Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature*, *444*(7121), 953–956. https://doi.org/10.1038/nature05363

Mandel, Corey R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J. L., & Tong, L. (2006b). Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature*, *444*(7121), 953–956. https://doi.org/10.1038/nature05363

Marzluff, W. F., & Koreski, K. P. (2017). Birth and Death of Histone mRNAs. *Trends in Genetics : TIG*, *33*(10), 745–759. https://doi.org/10.1016/j.tig.2017.07.014

Mathy, N., Bénard, L., Pellegrini, O., Daou, R., Wen, T., & Condon, C. (2007). 5′-to-3′ Exoribonuclease Activity in Bacteria: Role of RNase J1 in rRNA Maturation and 5′ Stability of mRNA. *Cell*, *129*(4), 681–692. https://doi.org/10.1016/j.cell.2007.02.051

McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblatt, J., … Bentley, D. L. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature*, *385*(6614), 357–361. https://doi.org/10.1038/385357a0

McLauchlan, J., Gaffney, D., Whitton, J. L., & Clements, J. B. (1985). The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini. *Nucleic Acids Research*, *13*(4), 1347–1368. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2987822

Melo, C. A., Drost, J., Wijchers, P. J., van de Werken, H., de Wit, E., Vrielink, J. A. F. O., … Agami, R. (2013). ERNAs Are Required for p53-Dependent Enhancer Activity and Gene Transcription. *Molecular Cell*, *49*(3), 524–535. https://doi.org/10.1016/j.molcel.2012.11.021

Nojima, T., Gomes, T., Grosso, A. R. F., Kimura, H., Dye, M. J., Dhir, S., … Proudfoot, N. J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell*, *161*(3), 526–540. https://doi.org/10.1016/j.cell.2015.03.027

Pnueli, L., Rudnizky, S., Yosefzon, Y., & Melamed, P. (2015). RNA transcribed from a distal enhancer is required for activating the chromatin at the promoter of the gonadotropin α-subunit gene. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(14), 4369–4374. https://doi.org/10.1073/pnas.1414841112

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*, *42*(Web Server issue), W187-91. https://doi.org/10.1093/nar/gku365

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., … Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, *44*(W1), W160-5. https://doi.org/10.1093/nar/gkw257

Rinn, J. L., & Chang, H. Y. (2012). Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry*, *81*(1), 145–166. https://doi.org/10.1146/annurev-biochem-051410-092902

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Rüegsegger, U., Beyer, K., & Keller, W. (1996). Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors. *The Journal of Biological Chemistry*, *271*(11), 6107–6113. https://doi.org/10.1074/JBC.271.11.6107

Rüegsegger, U., Blank, D., & Keller, W. (1998). Human pre-mRNA cleavage factor Im is related to spliceosomal SR proteins and can be reconstituted in vitro from recombinant subunits. *Molecular Cell*, *1*(2), 243–253. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9659921

Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, *489*(7414), 109–113. https://doi.org/10.1038/nature11279

Schönemann, L., Kühn, U., Martin, G., Schäfer, P., Gruber, A. R., Keller, W., … Wahle, E. (2014). Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Genes & Development*, *28*(21), 2381–2393.

https://doi.org/10.1101/gad.250985.114
Sheets, M. D., Ogg, S. C., & Wickens, M. P. (1990). Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Research*, *18*(19), 5799–5805. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2170946
Shi, Y., Di Giammartino, D. C., Taylor, D., Sarkeshik, A., Rice, W. J., Yates, J. R., … Manley, J. L. (2009). Molecular architecture of the human pre-mRNA 3' processing complex. *Molecular Cell*, *33*(3), 365–376. https://doi.org/10.1016/j.molcel.2008.12.028
Shi, Y., Di Giammartino, D. C., Taylor, D., Sarkeshik, A., Rice, W. J., Yates, J. R., … Manley, J. L. (2009). Molecular Architecture of the Human Pre-mRNA 3′ Processing Complex. *Molecular Cell*, *33*(3), 365–376. https://doi.org/10.1016/j.molcel.2008.12.028
Sullivan, K. D., Steiniger, M., & Marzluff, W. F. (2009). A core complex of CPSF73, CPSF100, and Symplekin may form two different cleavage factors for processing of poly(A) and histone mRNAs. *Molecular Cell*, *34*(3), 322–332. https://doi.org/10.1016/j.molcel.2009.04.024
Sun, Y., Zhang, Y., Hamilton, K., Manley, J. L., Shi, Y., Walz, T., & Tong, L. (2018). Molecular basis for the recognition of the human AAUAAA polyadenylation signal. *Proceedings of the National Academy of Sciences*, *115*(7), E1419–E1428. https://doi.org/10.1073/PNAS.1718723115
Taft, R. J., Glazov, E. A., Cloonan, N., Simons, C., Stephen, S., Faulkner, G. J., … Mattick, J. S. (2009). Tiny RNAs associated with transcription start sites in animals. *Nature Genetics*, *41*(5), 572–578. https://doi.org/10.1038/ng.312
Takagaki, Y., Manley, J. L., MacDonald, C. C., Wilusz, J., & Shenk, T. (1990). A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs. *Genes & Development*, *4*(12A), 2112–2120. https://doi.org/10.1101/GAD.4.12A.2112
Takagaki, Y., Ryner, L. C., & Manley, J. L. (1989). Four factors are required for 3'-end cleavage of pre-mRNAs. *Genes & Development*, *3*(11), 1711–1724. https://doi.org/10.1101/GAD.3.11.1711
Takagaki Y1, Ryner LC, & Manley JL. (1988). Separation and characterization of a poly(A) polymerase and a cleavage/specificity factor required for pre-mRNA polyadenylation. *Cell*, *52*(5), 731–742.
Venkataraman, K., Brown, K. M., & Gilmartin, G. M. (2005). Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes & Development*, *19*(11), 1315–1327. https://doi.org/10.1101/gad.1298605
Vilborg, A., Sabath, N., Wiesel, Y., Nathans, J., Levy-Adam, F., Yario, T. A., … Shalgi, R. (2017). Comparative analysis reveals genomic features of stress-induced transcriptional readthrough. *Proceedings of the National Academy of Sciences*, *114*(40), E8362–E8371. https://doi.org/10.1073/pnas.1711120114
Wang, L., Lawrence, M. S., Wan, Y., Stojanov, P., Sougnez, C., Stevenson, K., … Wu, C. J. (2011). *SF3B1* and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, *365*(26), 2497–2506. https://doi.org/10.1056/NEJMoa1109016
Weitzer, S., & Martinez, J. (2007). The human RNA kinase hClp1 is active on 3′ transfer RNA exons and short interfering RNAs. *Nature*, *447*(7141), 222–226. https://doi.org/10.1038/nature05777
West, S., Gromak, N., & Proudfoot, N. J. (2004). Human 5′ → 3′ exonuclease Xrn2 promotes

transcription termination at co-transcriptional cleavage sites. *Nature*, *432*(7016), 522–525. https://doi.org/10.1038/nature03035

Yang, Q., Gilmartin, G. M., & Doublie, S. (2010). Structural basis of UGUA recognition by the Nudix protein CFIm25 and implications for a regulatory role in mRNA 3' processing. *Proceedings of the National Academy of Sciences*, *107*(22), 10062–10067. https://doi.org/10.1073/pnas.1000848107

Yang, W., Hsu, P. L., Yang, F., Song, J.-E., & Varani, G. (2018). Reconstitution of the CstF complex unveils a regulatory role for CstF-50 in recognition of 3′-end processing signals. *Nucleic Acids Research*, *46*(2), 493. https://doi.org/10.1093/NAR/GKX1177

Yang, X., Sullivan, K. D., Marzluff, W. F., & Dominski, Z. (2009). Studies of the 5' exonuclease and endonuclease activities of CPSF-73 in histone pre-mRNA processing. *Molecular and Cellular Biology*, *29*(1), 31–42. https://doi.org/10.1128/MCB.00776-08

Yang, Y., Su, Z., Song, X., Liang, B., Zeng, F., Chang, X., & Huang, D. (2016). Enhancer RNA-driven looping enhances the transcription of the long noncoding RNA DHRS4-AS1, a controller of the DHRS4 gene cluster. *Scientific Reports*, *6*, 20961. https://doi.org/10.1038/srep20961

Yao, C., Choi, E.-A., Weng, L., Xie, X., Wan, J., Xing, Y., … Shi, Y. (2013). *Overlapping and distinct functions of CstF64 and CstF64τ in mammalian mRNA 3 ′ processing*. https://doi.org/10.1261/rna.042317.113

Yon, S., KimMijin, Yong, K., Gu, K., ParkYangsoon, Eun, S., … Bae, K. (2015). Negative Expression of CPSF2 Predicts a Poorer Clinical Outcome in Patients with Papillary Thyroid Carcinoma. *Http://Dx.Doi.Org/10.1089/Thy.2015.0079*.

Zhang, H., Rigo, F., & Martinson, H. G. (2015a). Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site. *Molecular Cell*, *59*(3), 437–448. https://doi.org/10.1016/J.MOLCEL.2015.06.008

Zhang, H., Rigo, F., & Martinson, H. G. (2015b). Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site. *Molecular Cell*, *59*(3), 437–448. https://doi.org/10.1016/j.molcel.2015.06.008

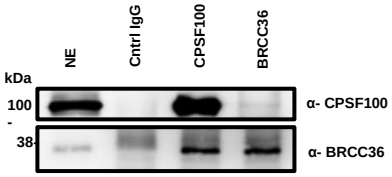
Zhu, Y., Wang, X., Forouzmand, E., Jeong, J., Qiao, F., Sowd, G. A., … Shi, Y. (2018). Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation. *Molecular Cell*, *69*(1), 62-74.e4. https://doi.org/10.1016/j.molcel.2017.11.031

APPENDIX A: Effects of CPSF 100 Knockdown

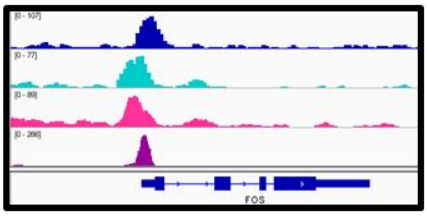
A. IP-mass-spec

ERH	**CPSF2**	ACTG2	ISY1	LUC7L2
PC4	CSTF2	TPM3	ENO1	FMNL3
BRCC36	PKM	TCEB2	YBX2	SHMT1
HSPA5	SRSF2	LDHA	YBX1	FBXO22
EIF5A	DDX17	DSTN	CSDA	CSTF1
PHF5A	RCN2	CCDC16	PUF60	RMDN1
SUMO3	LDHB	RBM39	WDR5	CDC73
ANXA2	H2BD	LUC7L3	**BUB3**	WDR57
SRSF1	TAGLN2	HIST1H1	SYMPK	HSPA8
LSM8	CFL2	T	SF3A3	TUBA1C
BABAM1	CFL1	RIF1	TPI1	FKBP15
APEX1	PPIA	RSRC2	SNRNP7	BAF57
PFN-1	PP1R8	STRAP	0	HSRP
UIMC1	PPIB	SRA1	MPG	**BRCA1**
SRP14	SMS	CPSF3	HNRNPF	**PALB2**
TCEB1	ACTB	RNPS1	DLGAP4	

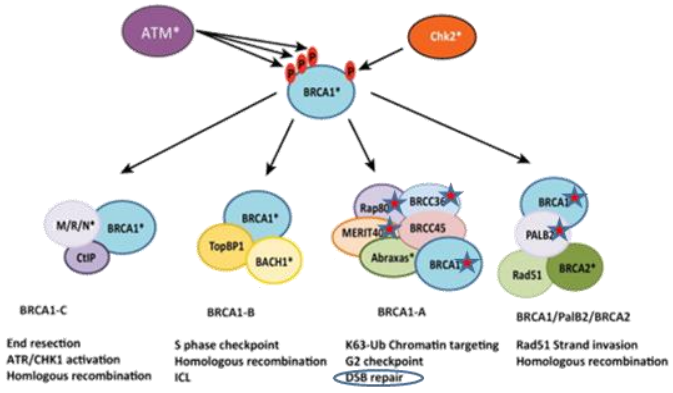
c. BRCA1 and CPSF100 Interaction



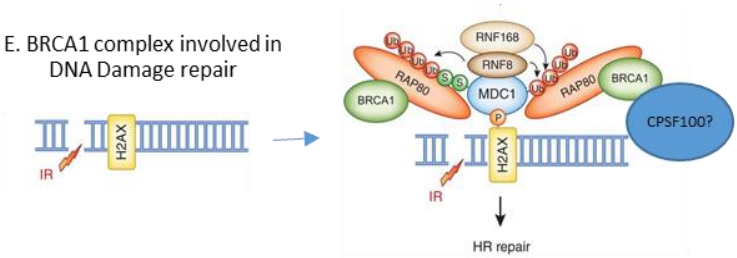
d. BRCA1 and CPSF100 co-localization



b. BRCA1 Complexes



E. BRCA1 complex involved in DNA Damage repair



F. DNA Damage Repair delayed in CPSF100 KD HeLa

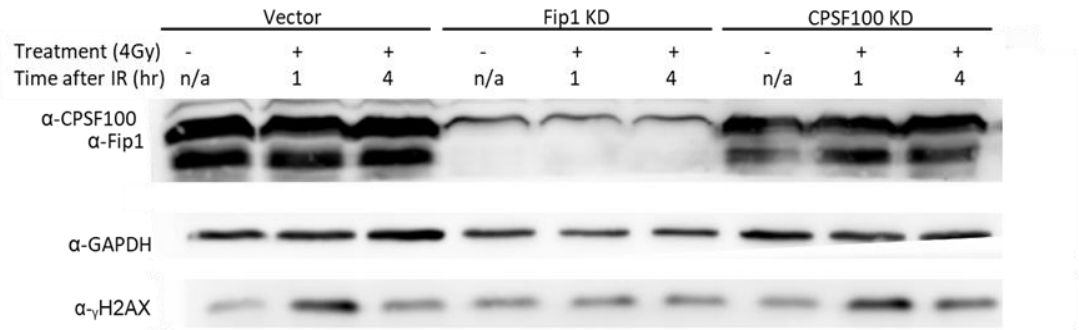
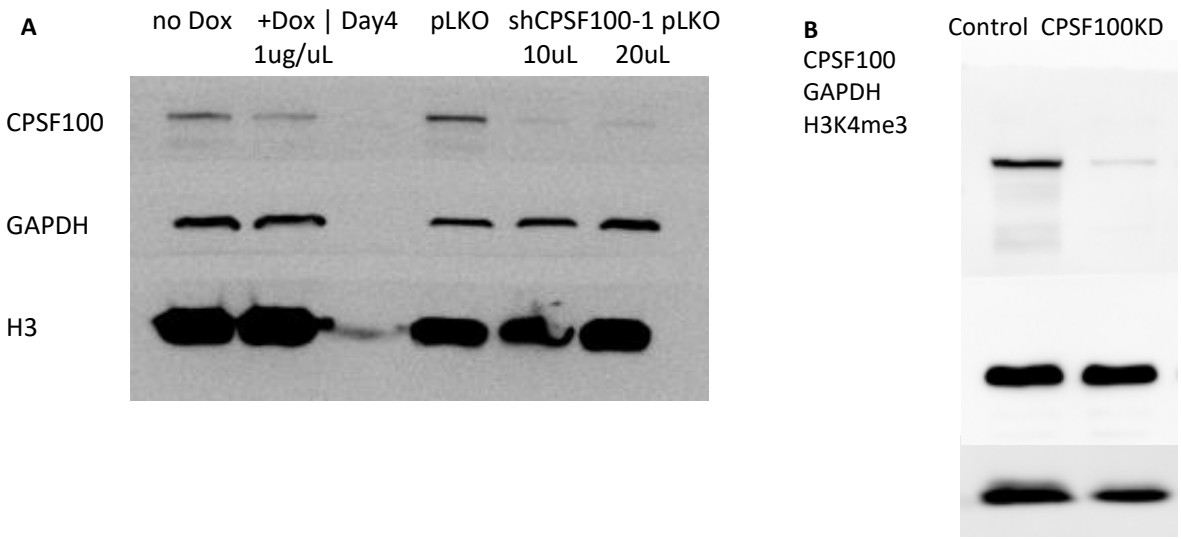
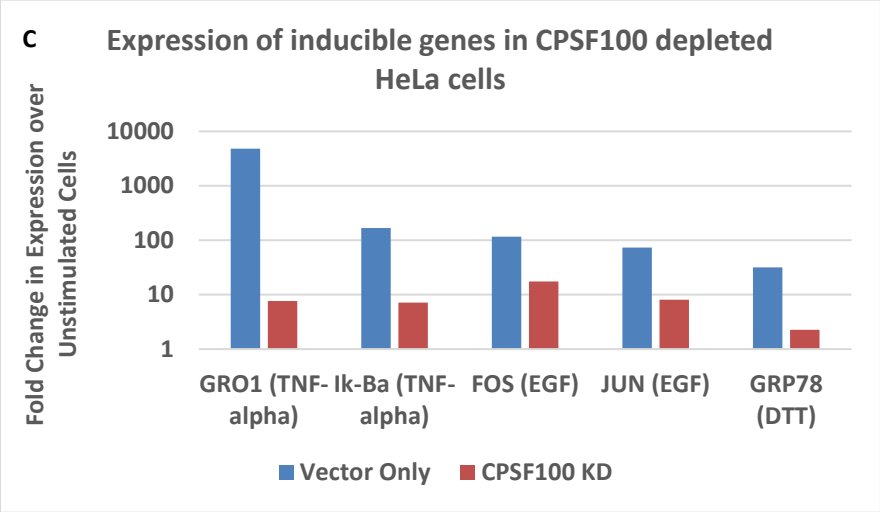


Figure 21. DNA Damage response in CPSF100KD cells. A. CPSF100 IP followed by mass spectrometry reveals members of BRCA A-1 complex members associate with CPSF100. B. Factors that co-IP'd with CPSF100 are shown in respective BRCA1 groups. C. Verification by IP and Western Blot of BRCC36 and CPSF100 interaction. D. BRCA1 and CPSF100 colocalize at FOS. E. DNA Damage stimulates increased gamma-H2AX to recruit DNA damage repair complex BRCA A-1. Gamma-H2AX levels return to normal after repair. F. Fip1KD eliminates gamma-H2AX increase, and CPSF100KD prevents return of gamma-H2AX to normal, indicating defective DNA repair.





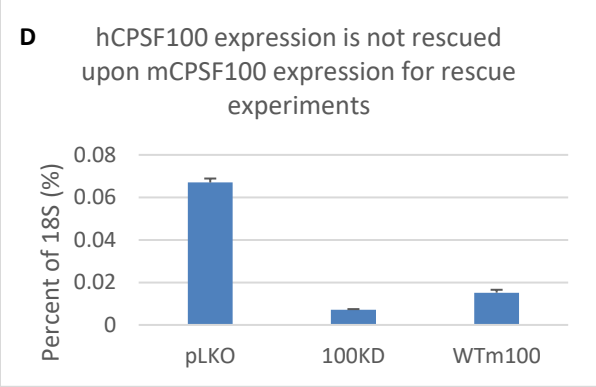


Figure 22. Effects of CPSF100 KD. A. Effect of knocking down CPSF100 on total H3 histone levels. First two lanes represent Dox-inducible CPSF100 knockdown HeLa cells. Second two lanes represent transient infection by lentiviral mediated shRNA delivery. B. Effect of CPSF100KD on H3K4me3. C. Expression of inducible genes is CPSF100KD cells shown as fold change over unstimulated cells; stimulating factor indicated in parentheses. D. Human CPSF100 expression levels in control (pLKO), 100KD, and 100KD+ WT mouse CPSF100 expressing HeLa cells.

APPENDIX B: Optimizing ChIP-Seq Parameters

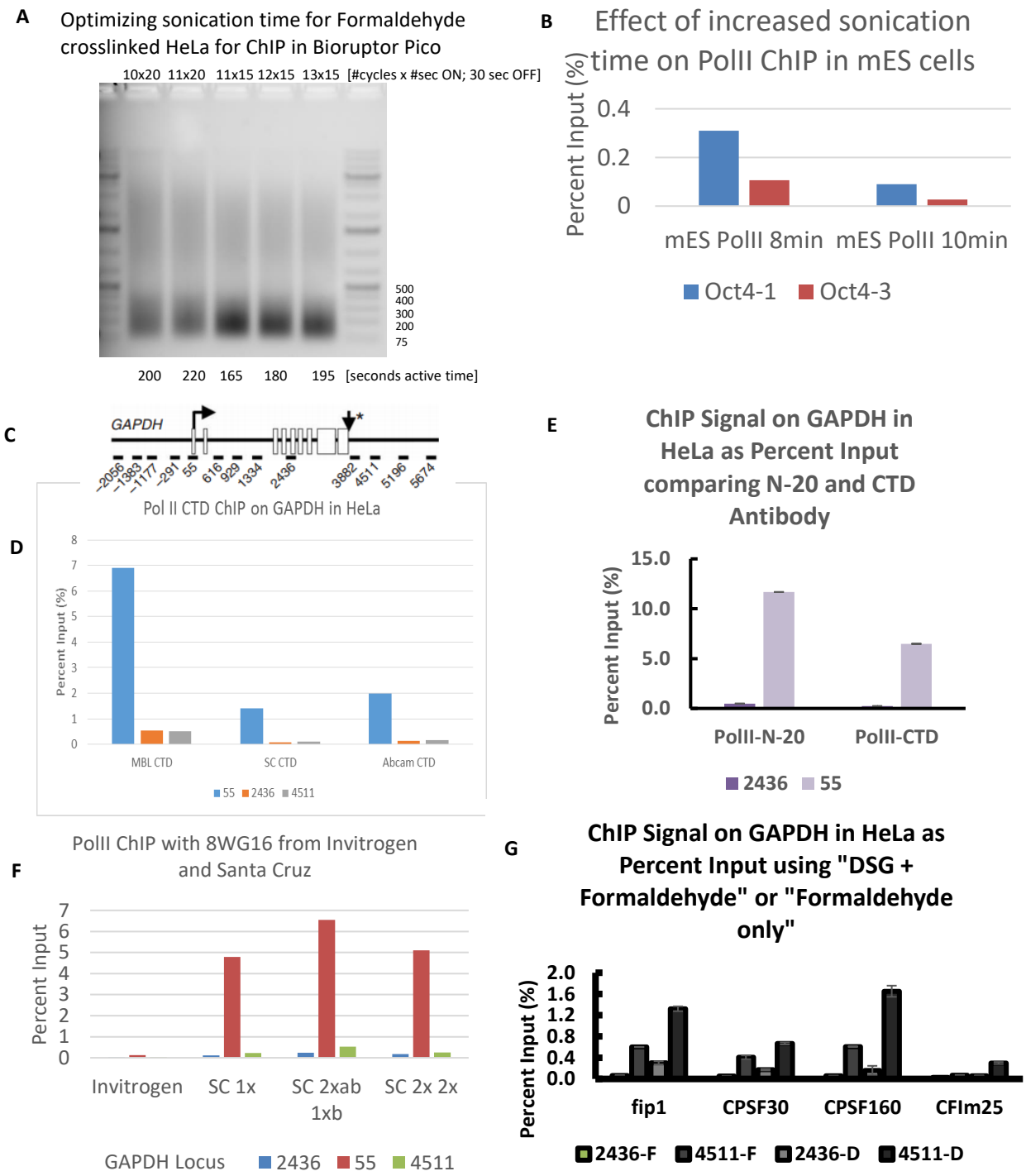


Fig. 23. Optimizing ChIP conditions. A. 1% Agarose gel showing fragmentation of chromatin after sonication in a Bioruptor Pico machine, in RIPA buffer. B. Increased sonication time decreases ChIP enrichment in mES at Oct4 loci; Oct4-1 is at promoter. C. Schematic of GAPDH and primers locations for panels D-G. From Glover-Cutter et al. 2008. D. Comparing performance of PolII CTD 8WG16 antibody from different manufacturers. SC=Santa Cruz . E. A compariosn of ChIP efficiency between the PolII N-20 antibody from Santa Cruz and the 8WG16 CTD antibody. F. ChIP comparing different amounts of beads and antibodies. Increasing the amount of antibody increased yield. G. Comparing different crosslinking reagents for ChIP, using antibodies to protein names indicated, at GAPDH loci indicated.