Abstract

The Dynamics of Early Steps in Transcription and their Dysregulation in X-Linked Dystonia Parkinsonism

Joshua Zimmer

2022

The dynamic life cycle of an RNA begins with transcription initiation and concludes when the mature transcript is degraded to monomer nucleotides. The early steps of RNA polymerase II (RNAPII) transcription, initiation and promoter-proximal pausing in metazoans, are the most complex and tightly orchestrated steps. To synthesize a single mature transcript, dozens of individual proteins must be assembled around a promoter to initiate transcription. In a reliable fashion, the initiated polymerase is halted only 20 to 60 base pairs downstream of the transcription start site (TSS) in a phenomenon called promoter-proximal pausing. Pausing is induced by a small number of additional factors, but is known to facilitate the assembly of important transcription elongation and RNA processing factors on RNAPII before entering productive elongation. RNAPII then synthesizes a nascent transcript that can be of more than a megabase in length. The transcript is processed, and, in most cases, the mature transcript is then exported from the nucleus to the cytoplasm where it is eventually degraded. Probing the dynamics of RNA synthesis and degradation can be challenging because standard RNA sequencing (RNA-seq) methods provide only a steady-state snapshot of gene expression. Metabolic labeling and nucleotide-recoding chemistry with RNA-seq (NR-seq) has proven to be a powerful tool to dissect the intricacies of the RNA synthesis pathway because it provides an extra temporal dimension to RNA-seq data.

Here I describe my work demonstrating incremental improvements in the handling of metabolically labeled RNA and analysis of RNA-seq data containing chemically induced mutations. I show that newly synthesized RNA can be specifically lost during RNA extraction, biasing NR-seq data against mutation-containing reads. In addition, I improved data analysis by demonstrating that implementation of a three-base alignment strategy improves alignment of mutation-containing reads. Furthermore, I apply these improved protocols in

several collaborative efforts using TimeLapse-seq and transient-transcriptome-TimeLapse-seq (TT-TL-seq) to characterize the dynamics of mature RNA and transcribing RNAPII.

I describe the development of Start-TimeLapse-seq (STL-seq) as the first method to directly measure the kinetics of promoter-proximal pausing in a non-perturbing, genome-wide, and TSS-specific manner. I show that STL-seq reliably quantifies the turnover of short, capped RNA transcripts associated with RNAPII at the pause site and this information accurately captures the behavior of paused RNAPII. STL-seq detects changes in paused RNAPII turnover upon a perturbation of steady-state conditions and can be used to unambiguously assign these changes to effects on pause release or premature termination. This work revealed the distinct principles of regulation of release into elongation and premature termination at the promoter-proximal pause site. Moving forward, STL-seq will be a powerful tool to dissect the mechanism and regulation of promoter-proximal pausing.

Finally, I describe work pursuing my proposed model for the disease mechanism of a rare genetic disorder, X-Linked Dystonia Parkinsonism (XDP). XDP is caused by a SINE-VNTR-Alu (SVA) retrotransposon insertion in the TATA-box binding protein (TBP) associated factor 1 gene (*TAF1*) and I present evidence that the SVA insertion gives rise to an alternative, truncated *TAF1* transcript isoform (*xTAF1*) which encodes an xTAF1 protein lacking a functional second bromodomain. I demonstrate that xTAF1 associates with promoters more strongly than canonical TAF1 (cTAF1) and induces a redistribution of the RNAPII promoter-proximal pause site. I propose that these two effects confer a dominant-negative phenotype that could ultimately lead to the neurodegenerative phenotypes observed in patients.

The Dynamics of Early Steps in Transcription and their Dysregulation in X-Linked Dystonia Parkinsonism

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

by
Joshua Zimmer

Dissertation Director: Matthew Simon

December, 2022

Contents

Acknowledgements **xii**

1 Introduction **3**

1.1 Regulation of gene expression through RNA synthesis 3

1.2 Early steps in RNAPII transcription 3

1.2.1 Initiation 5

1.2.2 Promoter-proximal pausing 6

1.2.3 Transition to productive elongation 8

1.2.4 Early steps of transcription are neither discrete nor independent . . 8

1.3 Existing methods and associated challenges to study early transcription . . 10

1.3.1 Structural works gives insight into transcription machinery 11

1.3.2 Genomics approaches capture chromatin-bound factors 11

1.3.3 Short, capped RNA are observations of paused RNAPII 12

1.3.4 RNAPII-associated nascent RNAs identify transcriptionally engaged RNAPII 13

1.3.5 Fluorescence microscopy visualizes transcription dynamics in cells . 15

1.3.6 Metabolic labeling measures transcription and RNA dynamics . . . 15

1.4 Early transcription in human disease 17

1.5 Overview of goals 18

2 Improving the study of RNA dynamics with advances in RNA-seq with nucleotide-recoding chemistry **20**

2.1 Author contributions 20

2.2 Summary 20
2.3 Introduction 21
2.4 Results 23
2.4.1 Additional care must be taken when handling s^4U-labeled RNA to avoid dropout 23
2.4.2 3-nt alignment minimizes computational dropout in NR-seq data . . 26
2.4.3 Alternative reaction conditions optimize TimeLapse chemistry with s^4U 28
2.4.4 Estimates for RNA degradation rate constants agree between all NR-seq methods and are similarly affected by dropout 32
2.5 Discussion 33

3 Case studies using RNA metabolic labeling to study transcription and RNA dynamics 35

3.1 Introduction 35
3.2 The role on lncRNA transcription in tumorigenesis 35
3.2.1 Activation of a p53-Dependent Pvt1 Isoform, Pvt1b 36
3.2.2 Pvt1b Suppresses Myc Transcriptional Activity *In Vitro* 37
3.3 Measuring readthrough transcription in the downstream of gene region . . . 38
3.3.1 Hyperosmotic stress causes widespread transcriptional repression . . 39
3.3.2 Stress-induced readthrough transcripts arise independent of gene-transcription levels 41
3.3.3 Clean DoG-producing genes are functionally enriched for transcriptional repression 42
3.3.4 Depletion of Integrator endonuclease leads to DoG production . . . 44
3.4 Probing the effect of a splicing factor mutant on RNA stability 46
3.4.1 U2AF1 mutations enhance stress granule formation improving cell fitness under stress 48
3.5 Dissecting regulatory function of *lincRNA-p21* 49

3.5.1 Development of genetic models to query the role of lincRNA-p21 transcription and transcript accumulation . . . 50
3.6 Discussion . . . 51

4 STL-seq reveals pause-release and termination kinetics for promoter-proximal paused RNA polymerase II transcripts 52

4.1 Summary . . . 52
4.2 Introduction . . . 53
4.3 Results . . . 55
4.3.1 Short, capped transcripts can be metabolically labeled using s^4U . . 55
4.3.2 STL-seq data can be used to quantify scRNA turnover accurately and robustly . . . 58
4.3.3 STL-seq reveals high turnover of scRNAs at most TSSs . . . 62
4.3.4 Termination is generally faster but less variable than release into elongation . . . 64
4.3.5 Certain histone tail modifications are associated with less permissive pausing dynamics . . . 68
4.3.6 Promoter and pause-site architecture are associated with stability of the paused complex . . . 72
4.3.7 Enhanced release into elongation is the major response to hormone stimulus . . . 73
4.3.8 Hyperosmotic stress induces premature termination . . . 75
4.4 Discussion . . . 79
4.5 Limitations of the study . . . 81

5 Probing the functional consequences of an SVA insertion in the *TAF1* gene in XDP 83

5.1 Author contributions . . . 83
5.2 Summary . . . 83
5.3 Introduction . . . 84
5.4 Results . . . 86

5.4.1 Early steps in RNA synthesis are perturbed in XDP cells 86
5.4.2 The SVA insertion gives rise to a stable, intron-retained XDP-specific *TAF1* transcript 90
5.4.3 The XDP truncation of TAF1 affects the structure and function of BD2 92
5.4.4 xTAF1 perturbs early steps of RNA synthesis 96
5.4.5 xTAF1 destabilizes and redistributes the position of promoter-proximal paused RNAPII 99
5.5 Discussion 101

6 Methods and data analysis 105
6.1 Methods 105
6.1.1 Cell lines and s^4U metabolic labeling 105
6.1.2 Generation of cell lines expressing GFP-TAF1 106
6.1.3 Immunoblots 106
6.1.4 *In vitro* histone tail pull-down assays 106
6.1.5 Drug and KCl treatments 107
6.1.6 NR-seq (TimeLapse-seq, SLAM-seq, TUC-seq) 107
6.1.7 STL-seq 108
6.1.8 TT-TL-seq 109
6.1.9 ChIP-seq 109
6.1.10 CUT&RUN 111
6.2 Data analysis 112
6.2.1 NR-seq and TT-TL-seq alignment and mutation calling 112
6.2.2 Alignment of new and previously published sequencing data 113
6.2.3 STL-seq alignment, mutational analysis, and TSS calling 113
6.2.4 Estimation of RNA decay and synthesis kinetics 114
6.2.5 Estimation of Pol II turnover with previous data under triptolide inhibition 114
6.2.6 Estimation of the new fraction of scRNA and kinetic parameters of scRNA 115

6.2.7 Estimation of the global effect of flavopiridol on premature termination 118
6.2.8 Simulation of STL-seq data 119
6.2.9 TT-TL-seq data analysis 120
6.2.10 PRO-seq data analysis 121
6.2.11 ChIP-seq and ATAC-seq data analysis 121
6.2.12 STARR-seq data analysis and eTSS identification 121
6.2.13 Identification of Promoter motifs 121

A Start-TimeLapse-seq (STL-seq) protocol 123

A.1 Important notes to be aware of before beginning 123
A.2 s^4U treatment and cell harvesting 123
A.3 RNA isolation 124
A.4 TimeLapse chemistry (25 µL volume, scalable if needed) 125
A.5 Size selection 126
A.6 Cap selection 127
A.7 3´ ligation 128
A.8 5´ ligation 128
A.9 RT and library amplification 129

List of Figures

1.1 The RNA synthesis pathway is a complex series of regulated steps 4

2.1 s^4U-labeled RNA is specifically lost during handling 24
2.2 NR-seq data prepared with different handling conditions do not correlate well with each other 26
2.3 3-base alignment recovers highly mutated reads 27
2.4 Handling and computational dropout are independent 29
2.5 Buffer conditions and reagents affect efficiency of TimeLapse chemistry . . . 30
2.6 Comparison of mutational content in TimeLapse-seq data using different reaction conditions 31
2.7 Comparison of RNA degradation rates measured by NR-seq methods and the impact of dropout 32

3.1 The *Pvt1b* isoform is induced by p53 activation 37
3.2 *Myc* and Myc target genes are downregulated following p53 activation . . . 38
3.3 TT-TL-seq reveals transcriptional profiles that accompany DoG induction after hyperosmotic stress 40
3.4 Hyperosmotic stress leads to widespread transcriptional repression 42
3.5 DoGs arise regardless of the transcriptional levels of their upstream genes upon hyperosmotic stress 43
3.6 Depletion of Integrator nuclease subunit leads to readthrough transcription 45
3.7 Int11 knockdown leads to readthrough transcription 47

3.8 Mutant U2AF1 cells show differential RNA dynamics related to stress granule content 48
3.9 Development of genetic tools to probe the contribution of lincRNA-p21 transcription and accumulation to p21 regulation 50

4.1 STL-seq captures turnover dynamics of transcripts from promoter-proximal paused polymerase 56
4.2 STL-seq captures mutation information of newly synthesized short, capped transcripts 57
4.3 Estimation of scRNA transcript turnover from STL-seq 59
4.4 Binomial modeling of STL-seq mutation data 61
4.5 Termination is fast while release into elongation explains variability of Pol II turnover at pause sites 65
4.6 Release from the pause site predicts downstream transcriptional activity . . 66
4.7 Chromatin structure defines unique profiles of pausing kinetics 68
4.8 Weak promoter architecture leads to rapid termination of paused Pol II . . 70
4.9 Hormonal stimulus by 20E preferably regulates release into elongation . . . 73
4.10 Termination is rarely regulated in response to hormone stimulus 75
4.11 Hyperosmotic stress induces premature termination at TATA-less promoters 76
4.12 Increased termination at the pause site causes widespread transcriptional repression upon hyperosmotic stress in human cells 78

5.1 RNA synthesis is perturbed in XDP patient-derived cells 87
5.2 RNA synthesis and not RNA degradation drives changes in gene expression in XDP cells 88
5.3 The SVA insertion perturbs transcription of the *TAF1* locus 91
5.4 xTAF1 binds chromatin in a similar but distinct pattern as cTAF1 93
5.5 Expressing xTAF1 does not globally perturb the chromatin landscape . . . 94
5.6 Perturbations in early transcription caused by xTAF1 expression are associated with distinct chromatin states 96
5.7 Characterizing the transcriptome of cells expressing cTAF1 and xTAF1 . . 97

5.8 xTAF1 perturbs RNAPII promoter-proximal pausing 100
5.9 Model for XDP pathogenesis and TAF1 BD2 function 103

List of Tables

1 List of abbreviations 2

6.1 *Drosophila* conensus motifs 122

6.2 Human conensus motifs 122

Acknowledgements

My development as a scientist before Yale and during my time in the Simon lab has been impacted by a large and incredible group of people, each of whom have taught me something new and unique. My PhD may be awarded to me as an individual but is earned by the village that has supported throughout my scientific career.

Joining the Simon lab allowed me to collaborate with many brilliant scientists across different fields and at different institutions. My collaborators during my time in the Simon lab include Dr. Joan Steitz, Dr. Nicolle Rosa-Mercado, Dr. Daniele Canzio, Sandy Rajkumar, Dr. Nadya Dimitrova, Dr. Christian Olivero, Dr. Lauren Winkler, Dr. Stephanie Halene, Dr. Giulia Biancon, Dr. Toma Tebaldi, Dr. Yimeng Gao, Dr. Cristopher Bragg, Dr. Christine Vaine, Shivangi Shah, Dr. Sherman Weissman, Dr. Anna Szekely, Dr. Sarah Slavoff, Dr. Xiongwen Cao, Dr. Haomiao Su, Dr. Josien van Wolfswinkel, Carmen Maria Conrow, Dr. Craig Crews, Dr. Michael Bond, Dr. Jake Swartzel, Dr. Shervin Takyar, Dr. Asawari Korde, Dr. Cigall Kadoch, Kevin So, Dr. S. Aidan Quinn, Dr. Manoj Pillai, Dr. Prajwal Boddu, and Dr. Abhishek Gupta. Each brought a new perspective to science that I could learn from.

I would like to thank the instructors, teaching fellows, and students of MB&B 300 from 2018, 2019, and 2021 who fostered a welcoming learning environment and made teaching a pleasure each and every week.

To my first scientific mentor, Dr. T. Glen Lawson, who set me off on my the path that has brought me to this point. Research in your lab introduced me to the beautiful world of science. Thank you for giving me the freedom to discover my passion.

To my committee members, Dr. Joan Steitz and Dr. Nadya Dimitrova, who inspire

me as scientists and pushed me to think critically about my own data. I will be forever indebted to you for your support and guidance over the past five years. I have had the unique opportunity to be a coauthor and collaborator with each of you twice. I consider this an extreme privilege that is a highlight of my PhD.

To my thesis advisor, Dr. Matthew Simon, who has been the best mentor I could have asked for during graduate school. You were always in my corner when I needed it and you allowed me to develop my own scientific voice as I matured in the Simon lab. I will always look up to your brilliance and dedication to the people in your lab, and I will strive to be the mentor to others that you were to me.

To past and present members of the Simon lab who have made the second floor of MIC such a special place to do science every day. Wherever I go next, I will hope to find people and a lab culture similar to those found in the Simon Lab. Each of you are an important piece of our Simon lab community and I will always remember the Simon lab being filled with beautiful, kind, and smart people. To my lab buddy from day 1, Michelle, talking you into rotating in the Simon lab was maybe the best contribution I ever made to the group.

The community at the local climbing gym, City Climb, has been a major part of my experience outside of the science. The many, many hours spent in the gym have been an important outlet and brief escape when needed. Although I will always hope to climb at a gym with better facilities, I don't think I will ever find as strong of a community.

To Dr. Kenneth Donohue, the surgeon who operated on my left hand after I broke my scaphoid and ruptured my ring finger A2 and A4 pulleys. Thank you for putting me back together twice to return to the bench and the rocks.

To my "Forced Friends" who began graduate school with me as first years in the BQBS track. I know each one of us will do great things and I will miss you all. To Carolyn, Cat, and Lukas who I somehow managed to convince climbing every week with me was a good idea. The weekly conventions have meant the world to me. I'm sorry we never went on that backpacking trip.

To my family who have been endlessly supportive and patient for the past five years. Mom and Dad, some might say that it is poetic that your careers are ending (happy retirement!) just as mine is beginning; however, I have never been one to believe in anything

more than coincidence. I will move forward in life with the guiding principles of kindness, compassion, and empathy that you instilled in me. I promise will tirelessly pursue my passions in science and medicine so long as it does not come in conflict with these principles. Matthew, my identical twin, you will always be a PhD for two more years than me, but I will try not to hold that against you. In a form of friendly twin rivalry, your mere existence has always pushed me towards self-betterment. Thank you for existing and for coming up with fun hobbies that I can keep copying.

I couldn't have made it this far without the people mentioned here. Thank you.

Abbreviation	Definition
BD	bromodomain
brdU	bromodeoxyuridine
CDS	coding sequence
ChIP-seq	chromatin immunoprecipitation sequencing
ChRO-seq	chromatin run-on sequencing
csRNA-seq	capped, short RNA-seq
cTAF1	canonical TAF1
CTD	C-terminal domain
DoG	downstream-of-gene
DRB	5,6-dichloro-1-β-d-ribofuranosylbenzimidazole
DSIF	DRB sensitivity-inducing factor
EC	elongation complex
eTSS	enhancer TSS
FP	flavopiridol
FRAP	fluorescence recovery after photobleaching
GRO-seq	global run-on sequencing
GTF	general transcription factor
iPSC	induced pluripotent stem cell
lncRNA	long noncoding RNA
MCMC	Markov Chain Monte Carlo
NDR	nucleosome-depleted region
NELF	negative elongation factor
NET-seq	nascent elongation transcript sequencing
NPC	neural progenitor cell
NR-seq	nucleotide recoding with RNA-seq
NRO	nuclear run-on
NSC	neural stem cell
PAS	polyadenylation signal
PIC	preinitiation complex
PRO-seq	precision run-on sequencing
P-TEFb	positive transcription elongation factor b
RNA-seq	RNA sequencing
RNAPII *or* Pol II	RNA polymerase II
scaRNA-seq	short, capped RNA-seq
scRNA	short, capped RNA
SINE	short interspersed nuclear element
STL-seq	Start-TimeLapse-seq
SVA	SINE-VNTR-Alu
s^4U	4-thiouridine
s^6G	6-Short interspersed nuclear elements
TAF	TBP-associated factor
TAF1	TBP-associated factor 1
TBP	TATA-box binding protein
TF	transcription factor

TL-seq	TimeLapse-seq
Trp	triptolide
TSS	transcription start site
TT-TL-seq	transient-transcriptome-TimeLapse-seq
TWI	twister
VNTR	variable number tandem repeats
XDP	X-Linked Dystonia Parkinsonism
xTAF1	XDP-specific TAF1
3-nt	3-nucleotide
20E	20-hydroxyecdysone

Table 1: List of abbreviations used in this dissertation

Chapter 1

Introduction

1.1 Regulation of gene expression through RNA synthesis

Distinct gene expression patterns allow living cells to appear entirely different in form and function despite containing identical genomes; however, cells must also have the plasticity to respond and adapt to environmental changes. This demonstrates that regulation of gene expression must be strict enough to maintain cell type fidelity and functional specialization while also allowing for some flexibility in response to stimuli. The first major step of gene expression is the transcription of DNA into RNA by an RNA polymerase. In eukaryotes, the products of genes transcribed by RNA polymerase II (RNAPII) are important for cell type diversity. Therefore, regulation RNAPII activity must be tightly controlled. The RNAPII RNA synthesis pathway is a complex series of highly regulated steps which proceeds through transcription and processing of the transcript (Figure 1.1). Regulatory input is often incorporated at early steps of transcription, making initiation and pause release important regulatory steps in gene expression. However, the complexities of these steps and their regulation are an active field of investigation.

1.2 Early steps in RNAPII transcription

Transcription begins with initiation at a transcription start site (TSS), the single base pair position encoding the first nucleotide of an RNA transcript. Canonically, initiation was

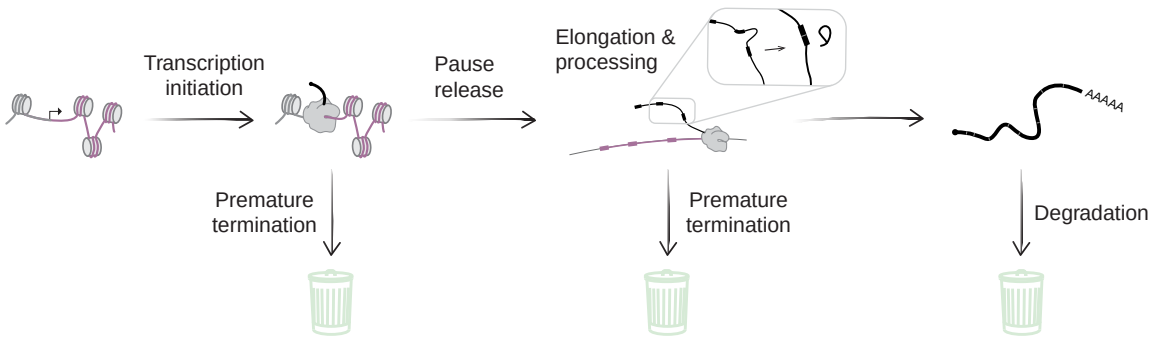

Figure 1.1: The RNA synthesis pathway is a complex series of regulated steps

thought to define sites of active promoters, but it is now appreciated that transcription initiation by RNAPII is also a hallmark of active enhancers [1,2]. In many eukaryotes RNAPII initiates and immediately proceeds into productive elongation, but, in metazoans, RNAPII only transcribes a short distance before completely arresting while remaining transcriptionally engaged with the DNA. This step in early transcription, called promoter-proximal pausing, occurs at the vast majority of genes and enhancers transcribed by RNAPII [3–6]. Except for modulating the stability of the mature transcript, little opportunity exists to change transcript copy number after RNAPII enters productive elongation. Therefore, early transcription is important in regulating gene expression by tuning the RNA synthesis rates.

These early steps of transcription are thought to occur on rapid timescales of just seconds to minutes [6]. This presents a challenge in furthering the mechanistic understanding of transcription and its regulation because it is difficult to experimentally observe and quantify these kinetics. Furthermore, a large number of transcription factors (TFs) and other complexes are involved in initiation and promoter-proximal pausing. While some important functions for certain factors are well-characterized and understood, such as the general transcription factors (GTFs), there is still ongoing research to fully understand the transcription machinery. It is also possible that factors known to be essential for one purpose may have other functions which are equally important but are currently undescribed, particularly in the more complex systems of higher eukaryotes. Together, these barriers leave the field with an incomplete understanding of how early transcription is regulated.

1.2.1 Initiation

Transcription initiation culminates in RNAPII incorporating the first nucleotide of a nascent RNA transcript, but requires assembly of the entire multi-subunit preinitiation complex (PIC) which includes GTFs (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH) and Mediator. The first regulatory opportunity for initiation is through promoter activation or silencing via modification of the local chromatin environment. DNA methylation is effective at repressing a promoter and specific sets of histone tail modifications on the nucleosomes near initiation sites are installed depending on whether the region is to be transcriptionally active or silent [7–9]. An active region is typically marked by histone tail lysine acetylation (Kac), trimethylation of histone H3 at lysine 4 (H3K4me3), and a nucleosome depleted region (NDR) that acts as a landing pad for TFs and RNAPII [9, 10].

TFIID, a megadalton-sized complex composed of the TATA-box binding protein and thirteen TBP-associated factors (TAFs) [11, 12], is the first GTF to associate with DNA during assembly of the PIC [13, 14]. The major described function of TFIID in PIC assembly is to deliver TBP to the promoter region. TBP then introduces a kink into the promoter DNA and nucleates formation of the rest of the PIC. GTFs TFIIA and TFIIB associate with TFIID, RNAPII is then recruited with TFIIF, followed by TFIIE, and finally TFIIH consumes ATP to unwind promoter DNA and facilitate the incorporation of the first nucleotides into a nascent RNA transcript [10, 14]. Mediator is a large, multi-subunit complex characterized as a component of the PIC which primarily associates with enhancers rather than promoters [15–18]. Although Mediator is not canonically described as part of the sequential assembly of the PIC, it is essential and is thought to transduce signals from enhancer-bound TFs to promoters.

Most regulatory opportunity during PIC assembly is through recruitment of PIC subunits, particularly RNAPII and TFIID which recognizes specific DNA sequence motifs and promoter-associated chromatin modifications and associates with other coactivating factors [10, 19–23]. The presence of TFIID DNA motifs in the promoter region may strengthen TFIID recruitment, and Mediator further facilitates assembly of the PIC. However, it has been suggested that the rate of PIC assembly and RNAPII recruitment are not acutely

regulated, but rather fine-tuned over long developmental timescales to settle on a stable expression profile [24, 25].

After PIC assembly and RNAPII initiation from the TSS, transcription to the promoter-proximal pause site only takes a few seconds, leaving a small temporal and spatial window to incorporate regulatory signals during initiation. Despite this, initiation is suggested to be a highly inefficient process where 80-90% of RNAPIIs abort transcription before reaching the promoter-proximal pause site [26, 27]. Abortive initiation is not well characterized in eukaryotic cells and it is unclear if this behavior represents a form of active regulation; however, the efficiency of this process could reasonably be targeted as a manner to affect gene expression.

1.2.2 Promoter-proximal pausing

In metazoans, RNAPII extends the nascent transcript to 20-60 nucleotides before reaching the promoter-proximal pause site where transcription is temporarily arrested [28–31]. The hallmark of RNAPII paused at the promoter-proximal site is the adoption of the "tilted state" by the RNA-DNA hybrid in the active site of the polymerase [32]. In the tilted state, DNA-RNA base pairing is maintained, but the RNA adopts and post-translocated state while the DNA adopts a pre-translocated state. This results in a conformation in which the DNA template base in the active site of RNAPII is bound to the last nascent RNA base, ultimately making the addition of another nucleotide impossible without exiting from this conformation. 5,6-dichloro-1-β-d-ribofuranosylbenzimidazole (DRB) sensitivity-inducing factor (DSIF) and negative elongation factor (NELF) are complexes typically associated with paused RNAPII and may induce the paused conformation [33–36]. Structural work revealed that NELF "tentacles" function as a clamp to maintain the tilted conformation [32], but interestingly, TFIID was recently shown to be sufficient to induce a similar promoter-proximal pausing effect *in vitro* [37].

Entry into productive elongation from the pause site is promoted by the kinase activity of the Cdk9 subunit of positive transcription elongation factor b (P-TEFb) [38–42]. P-TEFb phosphorylates DSIF, NELF, and the C-terminal domain (CTD) of RNAPII, promoting the dissociation of NELF from the paused complex, causing a conformational shift in the

active site of RNAPII, and allowing RNAPII to be released into elongation. Being that P-TEFb transitions RNAPII out of the paused state, direct modulation of P-TEFb activity is a common mechanism for regulation of gene expression [6]. In addition, recent work has shown that dephosphorylation of P-TEFb targets is an effective strategy to attenuate expression [43, 44]. The PP2A phosphatase acts in opposition to P-TEFb by dephosporylating the RNAPII CTD and DSIF subunit Spt5, thereby preventing RNAPII from entering productive elongation. PP2A phosphatase activity in the promoter-proximal region induces endonucleolytic cleavage of the transcript by the Integrator complex and causes premature termination of transcription [43–48].

Promoter-proximal pausing is the rate-limiting step of transcription at most genes and is therefore a good target for regulatory signals to be incorporated into the gene expression pathway. In attempts to better understand the behavior of paused RNAPII, several groups developed non-perturbing approaches to measure the duration of pausing in cells [26, 27, 49, 50]. This work demonstrated that Pol II waits at the pause site for about five minutes and 80-90% of paused complexes prematurely terminate transcription rather than enter productive elongation. It was also shown that rates of pause release are preferred targets of regulation rather than those of premature termination, suggesting that while premature termination is ubiquitous, it is not a major effector of gene expression in cells [24,49,50]. Instead, premature termination is thought to contribute to the maintenance of active promoters and prevent incompetent elongation complexes from synthesizing a misprocessed transcript. Paused RNAPII occupies the promoter-proximal region and prevents nucleosomes from occluding the TSS where TFs and the PIC must associate with DNA [51]. A continuous initiation/termination cycle helps maintain an accessible promoter state even if expression from a particular gene is not required. Premature termination at the pause site also presents a final quality control opportunity to check if the paused complex has properly matured into an competent elongation complex (EC) [46, 50]. If the EC is recognized as incompetent and may produce a misprocessed transcript, it can be evicted to allow a new complex to initiate.

These studies have made significant advances in our understanding of RNAPII kinetics at the promoter-proximal pause site and will complement future work dissecting the role of

known and unknown factors involved in the regulation of this step of early transcription.

1.2.3 Transition to productive elongation

Upon P-TEFb phosphorylation and release into elongation, the transcription complex rearranges and RNAPII is further phosphorylated at Ser2 on the CTD by Cdk12 and Cdk13 [52]. This phosphorylation and dissociation of NELF recruits elongation factors important for processivity, splicing, and 3´ end formation which mature the paused complex into a competent elongation complex (EC) [53]. The transition of a paused complex to a mature EC in the first 5 kb of elongation is the final opportunity for modulation of expression. When RNAPII stalls in the early gene body as a result of non-productive elongation, transcription is prematurely terminated either by TFIIS or the Integrator complex [54–56]. It is unknown if this trimming of ECs early in the gene body is responsive to stimuli or is just a final fail-safe for transcriptional quality control, but it seems to be the last critical stage of transcription before RNAPII is committed to synthesizing a full-length transcript.

1.2.4 Early steps of transcription are neither discrete nor independent

While early transcription can be described as distinct stages, the reality is that the boundaries between each are not so clear. Significant overlap exists between the sets of factors involved in each step, suggesting that one factor can play important roles at more than one step in the process. This presents a challenge in distinguishing a factor's role at each stage and contributes to why investigation into the mechanism and regulation of initiation and pausing is still ongoing.

Transcription initiation could be described as the series of events leading up to incorporation of the first nucleotide by RNAPII, yet the definition of abortive initiation necessitates that the incorporation of the first few nucleotides is part of initiation. To this end, it may be better to define initiation as transcription up until the promoter-pause site, but this varies by promoter because RNAPII does not always transcribe the same distance before pausing [57]. Furthermore, defining the beginning of promoter-proximal pausing is not necessarily trivial. NELF locks RNAPII in the tilted conformation at the pause site but the mechanism of promoter-proximal pausing is unknown. With existing data, we cannot dis-

tinguish between several possible models of pausing induction. The moment NELF binds may define the position of pausing or NELF could bind earlier during initiation and only lock the tilted conformation upon another signal. Alternatively, the function of a completely different factor may be more appropriate to define as the beginning of promoter-proximal pausing. TFIID is sufficient to cause pausing *in vitro* [37], and it could be that NELF only acts to reinforce the pause afterwards. The fact that TFIID can influence pausing behavior suggests that there may even be cross-talk between initiation and pausing that would further complicate the regulation of these steps [37, 50, 58, 59].

The end of promoter-proximal pausing is similarly challenging to define. Release into elongation and pause release are common terms to describe the transition from pausing to elongation which accurately suggest that another factor actively promotes RNAPII to restart transcription. Typically, P-TEFb kinase activity is described as the signal to transition to elongation. Perhaps phosphorylation of all CDK9 targets in the pause complex is a sufficient definition; however, NELF dissociation is required to release RNAPII from the paused complex and it is unknown how quickly this occurs after phosphorylation. Further complicating the matter, PP2A is known to dephosphorylate CDK9 targets [43, 44], presenting the possibility that multiple rounds of phosphorylation and dephosphorylation of pausing factors could occur before RNAPII is released into elongation.

Finally, it is clear that behavior of RNAPII in early elongation is demonstrably distinct from that during late elongation, yet the boundary between these two phases of transcription is the fuzziest. After release into elongation, RNAPII is much more likely to prematurely terminate transcription within the first 5 kb [55,56], but the features that define the window of early transcription are not well characterized. Integrator is proposed to function as a terminating factor in early elongation and promoter-proximal pausing [43–48, 56, 60]. The past few years have significantly advanced our understanding of Integrator-based premature transcription termination, but there are many outstanding questions about its regulation. For example, what signals Integrator to cleave the nascent transcript, are these signals the same at the pause site and in early elongation, and why would this apparent inefficiency be beneficial? The models for transcriptional quality control ultimately explains premature termination as a resource-saving phenomenon but would require identification of additional

surveillance machinery/function.

Together, early transcriptional steps require an incredibly high temporal and spatial density of functions and factors to accomplish the singular goal of beginning the RNA synthesis process. Early transcription must be tightly controlled to orchestrate the complex pathway, but surprisingly, this process is highly inefficient. Current evidence suggests that only ∼1% of all RNAPII that begin transcription will produce a full-length transcript. The paradox of a tightly-regulated but highly inefficient process will require further work to be fully reconciled. Fortunately, many recent technological advances have the potential to expand our understanding of the fundamentals of early transcription regulation.

1.3 Existing methods and associated challenges to study early transcription

The complex and dynamic nature of early transcriptional events presents a major challenge in studying the precise behavior of RNAPII and associated factors. Early work was limited to studying a single locus in a simple model system or with *in vitro* purified factors [28–30, 61]. While these studies are foundational in the initial characterization of initiation and promoter-proximal pausing, their narrow scope inherently lacks the ability to capture important biological variability that provides insight into regulation at the promoter. In addition, simple model systems may lead to conclusions which do not hold true in higher eukaryotes. For example, TBP is an essential initiation factor in yeast, but accumulating evidence suggests that it is dispensable for RNAPII transcription in higher eukaryotes [62–67].

The explosion of high-throughput techniques in the past decade have brought about an era of genome- and transcriptome-wide studies that have revolutionized the way in which the field approaches transcriptionology. Even more recently, improvements in structural methods provide atomic resolution of massive RNAPII complexes. Nonetheless, there are still many challenges to overcome in studying early transcription. Here, I discuss the advantages and important limitations of modern methods.

1.3.1 Structural works gives insight into transcription machinery

Remarkable insights into early transcription have been gleaned from recent structural work of the preinitiation complex and RNAPII [12, 16, 17, 32, 48, 68–73]. This collection of work made some exceptional breakthroughs with unprecedented resolution of high-order PIC and RNAPII complexes, for example, by revealing the dynamic and flexible nature of the PIC and explaining why a paused RNAPII is physically incapable of elongation. It was shown that TFIID and the PIC adopt multiple conformations depending on the stage of assembly and the promoter DNA sequence, and the NELF tentacles restrict the conformation of the paused RNAPII. In many cases, the structural work complemented existing biochemical data, demonstrating the validity of the approach to visualize transcription.

On the other hand, structural work can suffer from *in vitro* artifacts and resolution limitations. Cryo-EM studies can capture multiple states of the same structure in one experiment, but highly transient and unstable states are difficult to observe and the transition between states cannot necessarily be inferred. Furthermore, the observed structures are constrained to the set of purified factors mixed in the experiment, and it is difficult to know if the set is complete or how many alternative forms of the complex exist *in vivo*. Finally, the structures are not always complete due to the inability to assign density for pieces of the complex. Indicative of this is TAF1, the largest subunit of TFIID at 250 kDa. In all published TFIID/PIC structures to date, no more than 50% of the TAF1 primary sequence is represented. Several explanations could be provided for this result; these regions are too flexible, these are regions not important for initiation and therefore do no need to be ordered in the PIC, or these regions require a more biologically-relevant context, such as chromatin, to be resolved. Nevertheless, these structures have provided invaluable insight which will undoubtedly only be improved upon as more sophisticated structural methods are developed.

1.3.2 Genomics approaches capture chromatin-bound factors

An important question to answer in probing the function of a particular factor in transcription is where it is localized across the genome. Whether studying histone tail modifications

or a TF, genomic studies provide clues as to whether a target is ubiquitous or locus-specific and where in the genome it is functional. With the advent of chromatin immunoprecipitation (ChIP, [74]) and subsequent high-throughput versions ChIP with DNA microarray (ChIP-chip [75]) and ChIP with next generation sequencing (ChIP-seq [76,77]), we can now identify the genome-wide binding pattern of individual proteins. ChIP-seq is heavily favored in the field for high-throughput studies, but it was ChIP-chip that first revealed the ubiquity of promoter-proximal pausing at nearly all genes transcribed by RNAPII [3,78]. Additional treatment of immunoprecipitated DNA with a nuclease [79–81] or transposase [82] provides near single base pair resolution while also improving signal-to-noise and lowering required inputs for protein-DNA interaction-mapping experiments. Performing *in situ* versions of these experiments further minimizes required inputs, even facilitating single-cell profiling [83,84].

While the data produced by these methods are extremely valuable, interpretation of them can be nuanced. It is difficult to know if signal is attributable to direct or indirect binding. Furthermore, signal intensity can vary with binding strength and frequency, where more signal could be due to strong protein-DNA interactions or high interaction frequency. Single-molecule footprinting is capable of measuring the fractional occupancy of promoters genome-wide, but cannot definitively identify the DNA-bound proteins nor infer kinetics of association and dissociation [85]. Importantly, the genomic localization and binding pattern of a specific protein does not directly inform about the regulatory effect on transcription. RNAPII density as measured by ChIP is sometimes used as a proxy for transcriptional activity. This is a reasonable approximation as more RNAPII complexes generally implies more RNA synthesis, but is not absolutely true because RNA synthesis rates also depend on elongation velocity. In addition, these methods are blind to the state of the polymerase and cannot distinguish between a complex which is initiating, paused, or elongating.

1.3.3 Short, capped RNA are observations of paused RNAPII

A strategy to pinpoint the single base pair position of initiation and promoter-proximal pausing is to sequence the short, capped RNA (scRNA) associated with RNAPII at the pause site. Start-seq, developed by the Adelman lab, selects for nuclear, short transcripts of less than 80 nts and depletes uncapped species [57]. The Adelman lab showed that

these transcripts are almost entirely chromatin-associated, and therefore can be used as a one-to-one observation of a paused polymerase [86]. Start-seq generates single-molecule data because 5´ and 3´ ends of each read represent the site of initiation and pausing for a single polymerase, respectively. The relative strength of signal in these data is correlated with the fractional occupancy of the pause site, which is a function of the initiation rate and how quickly RNAPII vacates the pause site either through termination or release into elongation. Alternative versions of the Start-seq protocol which provide similar information have since been published: short, capped RNA-seq (scaRNA-seq) and capped, short RNA-seq (csRNA-seq) [25, 87]. csRNA-seq demonstrated that data similar to Start-seq can be collected without the nuclear isolation step and scaRNA-seq selects for transcripts up to 300 nts to capture information about RNAPII recently released from the pause site.

While powerful as a targeted way to study promoter-proximal pausing, these methods are not designed to probe the behavior of RNAPII in early elongation, with the exception of scaRNA-seq which is limited to the first 300 bp of the gene. Moreover, signal intensity is challenging to interpret. How often the pause site is occupied by an RNAPII molecule depends on the rates of initiation, pause-release, and termination. Without additional information, these rates are impossible to deconvolute. Therefore, changes in signal intensity cannot be assigned to a particular effect. While scaRNA-seq is reported to estimate relative pause-release and initiation rates, the quality of those estimations is not well established and the method does not account for termination at the pause site.

1.3.4 RNAPII-associated nascent RNAs identify transcriptionally engaged RNAPII

Nuclear run-on (NRO) assays probe the location of RNA polymerases transcriptionally engaged with DNA. In these experiments, transcription is halted, nuclei are isolated, nearly all chromatin-associated factors except polymerases are washed away, and transcription is restarted in the presence of nucleotide analogues. RNA labeled with an analogue is then enriched and used as an observation for the position of an engaged polymerase. High throughput versions of nuclear run-on assays, global run-on sequencing (GRO-seq) and precision run-on sequencing (PRO-seq), are effective in profiling the genome-wide distribution

of active RNA polymerase [4, 58]. GRO-seq labels nascent RNA with bromodeoxyuridine (brdU) and allows nascent transcripts to be elongated for the entirety of the run-on time. PRO-seq, on the other hand, labels nascent RNA with biotinylated NTPs which prevent additional NTPs from being added to the transcript. Therefore, PRO-seq provides single base pair resolution as to the positions of the polymerase active site and is generally the preferred NRO-based assay because of this. A major advantage to PRO-seq is that the position of the promoter-proximal pause site can be determined with extremely high precision while also capturing the density of the polymerase over the gene body. Consequently, PRO-seq is commonly used to simultaneously study the regulation of initiation, promoter-proximal pausing, and elongation. Simpler versions of PRO-seq, such as chromatin run-on sequencing (ChRO-seq), have also been developed to facilitate the implementation of the approach [88].

A similar approach is nascent elongation transcript sequencing (NET-seq) which involves the immunoprecipitation of RNAPII and sequencing of the associated RNA [89, 90]. While similar in principle to PRO-seq, NET-seq presents two additional benefits. First, antibody choice in NET-seq allows for selection of specific phosphorylation states of the RNAPII CTD which can reveal information about different stages of RNAPII transcription. Second, NET-seq does not depend on a competent polymerase to incorporate an additional nucleotide and therefore captures nonproductive RNAPII.

While highly useful, both PRO-seq and NET-seq are not without their drawbacks. First, PRO-seq has been criticized for potential creeping of RNAPII during nuclei isolation [91]. If this is the case, it is possible that NET-seq would suffer from the same artifact during chromatin isolation and DNA digestion. The largest impact this could have is on the interpretation of the position of promoter-proximal pausing. This is important because the single base pair position of the pause site is often highly scrutinized and compared to the position of the +1 nucleosome. Secondly, similar to ChIP-seq, both methods are strictly polymerase density assays and not direct measurements of transcriptional activity or RNA synthesis. In order to measure transcriptional activity genome-wide, a different RNA sequencing-based approach is required.

1.3.5 Fluorescence microscopy visualizes transcription dynamics in cells

Single-molecule and bulk fluorescence microscopy experiments have proven to be valuable approaches to visualize the dynamics of TFs and RNAPII in living cells. Residence times of PIC components and other TFs at promoters have been quantified using single-molecule tracking approaches, although much of this work is in budding yeast which lack promoter-proximal pausing [92–95]. Bulk measurements of RNAPII diffusion with fluorescence recovery after photobleaching (FRAP) make population-averaged measurements for rates of initiation, pausing, and elongation [26,27,49]. These FRAP measurements provided some of the first estimates for the steady-state kinetics of early transcription without a drugged perturbation. Finally, single-molecule RNA fluorescence *in-situ* hybridization (smFISH) with probes targeting intronic sequences in combination with RNAPII ChIP unambiguously lead to the conclusions that transcriptional bursting and pause release, and not RNAPII recruitment, are actively regulated upon BET inhibitor treatment [24]. This work demonstrated the power of combinatorial approaches to studying the regulation of early transcription.

A major limitation of fluorescence microscopy approaches is the inherent limit of scope. Current technology restricts experimental setups to one or two labels in live cells. Consequently, published studies have chosen between making bulk population measurements or focusing on a single locus per measurement. In both cases, it is difficult to capture the biological diversity of the measured parameters (eg. PIC residency time or pause duration) in a promoter-specific manner.

1.3.6 Metabolic labeling measures transcription and RNA dynamics

Metabolically labeling RNA with a nucleotide analogue in living cells is a way to gather information about transcription dynamics. A commonly used nucleotide analogue is 4-thiouridine (s^4U) because it is readily incorporated into an RNA transcript by RNAPII in living cells within seconds or minutes [50, 96, 97]. As the polymerase transcribes, it continuously incorporates s^4U into the nascent RNA chain. Therefore, the labeling depends on the position and behavior of polymerase, in contrast to previously described methods that only depend on RNAPII position.

Metabolic labeling followed by nucleotide recoding was developed as an enrichment-free approach to study RNA synthesis and stability (TimeLapse-seq, SLAM-seq, TUC-seq, AMUC-seq [97–102]). Cells are treated with s^4U or 6-thioguanosine (s^6G) for an extended period (2+ h) such that a large proportion of mRNA transcripts are labeled. Upon isolating total RNA, performing nucleotide-recoding chemistry converts the hydrogen bond donor/acceptor pattern of the analogue to the equivalent of another base (s^4U to C or s^6G to A). This allows for computational identification of which reads were newly synthesized during the labeling period. While powerful for studying mRNA synthesis as a complete pathway, these methods are incapable of specifically probing the early transcriptional steps.

A short s^4U pulse (5 min) followed by biotinylation and enrichment of labeled RNA selects for newly synthesized transcripts [96, 97, 103]. Because unstable RNA species (introns, antisense transcripts, etc.) are not substantially degraded within five minutes, sequencing the enriched material provides a readout of the transient transcriptome (TT-seq, TT-TimeLapse-seq). This quantifies the RNA synthesis rate from every position over a transcribed region and provides insight into polymerase dynamics. Application of these methods have been effective in studying the behaviour of RNAPII during early elongation [56, 103, 104]; however, as metabolic labeling depends on the polymerase's enzymatic activity, the accumulation of RNAPII at the promoter-proximal pause site is invisible to these methods.

Recently, we developed Start-TimeLapse-seq (STL-seq, "stall"-seq) as a combination of existing methods, Start-seq and TimeLapse-seq, to directly quantify the kinetics of RNAPII promoter-proximal pausing [50, 57, 97]. STL-seq enriches scRNAs as in Start-seq and combines it with metabolic labeling and nucleotide-recoding chemistry as in TimeLapse-seq. The proportion of labeled scRNA at each TSS is modeled as the fraction of RNAPII initiated during the labeling time, making it possible to estimate the residence time of RNAPII at the pause site. Therefore, STL-seq is the first method to quantify steady-state turnover rates of promoter-proximal paused RNAPII in a genome-wide, TSS-specific, and non-perturbing manner.

Collectively, metabolic labeling-based RNA sequencing methods are the only approaches designed to probe the behavior of transcribing RNAPII and the transcript itself. Unlike

all other approaches which only capture a snapshot of the polymerase or RNA population, labeling of endogenous transcripts over a window of time inherently captures the transcriptome-wide kinetics of transcription, processing, and RNA stability as data points. While no single method on its own is sufficient to follow an RNA from initiation through degradation of the mature form, the toolkit available tiles kinetic information across the entire life cycle of an RNA. To take full advantage of metabolic labeling data, additional computational challenges must be overcome; for example, additional tools are required to accurately identify chemically induced T-to-C or G-to-A mutations and model kinetic parameters [50, 97, 105–107]. As these tools advance and new ones are developed, the utility of metabolic labeling to further our understanding of the regulation of gene expression, particularly in early transcription, will continue to improve.

1.4 Early transcription in human disease

Human genetics is a powerful tool in understanding our biology. A disease-causing mutation in a patient can lead to discoveries of previously unknown proteins or functions. Due to the essential nature of factors globally involved in transcription, complete loss-of-function mutations would generally be fatal. When mutations in PIC components and pausing factors have been identified, they are typically associated with rare disease, and neurological and developmental disorders [108, 109].

When these disorders are found, they are often identified as transcriptomopathies because they are likely to affect transcription at all or a vast majority of genes in the patient. For example, mutations in several Mediator subunits MED12, MED23, and MED25 are linked to disorders associated with intellectual disability [110–114]. Mutations in TFIIH and TFIIE are similarly associated with rare diseases of intellectual disability and neurodegeneration [115, 116]. Interestingly, existing literature establishes TFIID as a hot spot for mutations associated with neurodevelopmental disorders, both in TBP and TAFs [117–129]. The mutations described by this body of work are heterogeneous varying from coding sequence point mutations to a ∼3 kb retrotransposon intronic insertion. Unfortunately, the disease mechanism of nearly all of these disorders are poorly understood, and in many cases

a causal-link is only weakly established.

This family of transcriptomopathies are understudied when considering the potential insight into the complex details of early transcriptional events they can provide. These disorders represent massive unrealized potential to reveal fundamental properties of transcription. In some cases the questions to be answered by focusing on transcriptomopathies are ones the field is already asking, but it is likely that this line of investigation will uncover information we never knew to look for.

1.5 Overview of goals

The overarching goals of my dissertation work were to improve our understanding of the regulation of gene expression at the transcriptional level through application of cutting-edge techniques developed in the Simon lab and development my own method to address the unfilled need of a method to directly observe promoter-proximal pausing behavior. My first major aim was to develop a new metabolic labeling and RNA-seq based technique called Start-TimeLapse-seq (STL-seq) to measure the kinetics of promoter-proximal pausing. The details of the development of STL-seq and what it reveals are described in Chapter 4, and a detailed protocol is provided in Appendix A. STL-seq revealed how release into elongation and premature termination at the pause site are differentially regulated and play distinct roles in gene expression. My second major aim was to understand a rare neurodegenerative disorder called X-Linked Dystonia Parkinsonism (XDP) which is caused by a mutation in the *TAF1* gene. I developed the working hypothesis that the mutation causes a premature cleavage and polyadenylation event for up to 50% of transcribing RNAPII, giving rise to a truncated mRNA transcript and protein. The work investigating the XDP mutation and its functional consequences are described in Chapter 5. A minor aim of my dissertation work, described in Chapter 2, was to optimize and improve TimeLapse chemistry and analysis of RNA sequencing data with nucleotide-recoding chemistry. Finally, my second minor aim was to apply the full suite of metabolic labeling techniques developed in the Simon lab in collaboration with other scientists to reveal as much as possible about regulation of gene expression, transcription, and RNA dynamics. Described in Chapter 3 are four of

my collaborative efforts where TT-TL-seq and/or TimeLapse-seq were applied to study the RNA lifecycle.

Chapter 2

Improving the study of RNA dynamics with advances in RNA-seq with nucleotide-recoding chemistry

This chapter is adapted from:

Zimmer, J.T., Schofield, J.A., Kiefer, L., Vock, I.W., Moon, M.H., Simon, M.D. (*In prep*) Improving the study of RNA dynamics with advances in RNA-seq with nucleotide-recoding chemistry.

2.1 Author contributions

I performed all experiments and data analysis described in this section. All authors contributed to the conception of the work and experimental design.

2.2 Summary

RNA metabolic labeling with nucleotide recoding and RNA sequencing (NR-seq) is a powerful tool to capture the steady-state dynamics of RNA synthesis and decay without the need

for biochemical enrichment. Common to all NR-seq methods is the use of 4-thiouridine (s^4U) as a nucleotide analogue and presence of chemically induced T-to-C mutations in sequencing data. NR chemistry converts s^4U from a uridine analogue to a cytidine analogue, and the apparent T-to-C mutations are then used to identify the population of newly synthesized RNA. Here we show that NR-seq experiments require careful treatment to avoid specific loss of s^4U-labelled RNA during experimental handling and computational processing, an effect referred to as dropout. Experimental dropout is caused by s^4U-containing RNA adhering to plastic surfaces and computational dropout is caused by misalignment of reads containing T-to-C mutations. Importantly, kinetic parameters estimates from all NR-seq methods are equally affected by dropout and all methods are essentially indistinguishable downstream of the s^4U chemical conversion.

2.3 Introduction

RNA-sequencing (RNA-seq) is standard for characterizing the expression profile of cells and quantifying changes in expression upon some treatment. Unfortunately, traditional RNA-seq experiments are not sufficient to capture the difference between upregulation/downregulation in expression via a change in mRNA stability or synthesis rate. Early approaches to distinguish RNA stability from synthesis used transcriptional inhibitors in combination with metabolic labeling and enrichment of labeled RNA [130–133]. This approach requires perturbing steady state conditions with inhibitors, multiple labeling times per condition, and additional handling during enrichment which can inadvertently introduce bias and make analyses more complicated. To address these issues several enrichment-free RNA sequencing-based approaches have been developed to quantify RNA stability while maintaining information about steady state RNA levels [97–102].

These methods (SLAM-seq, TUC-seq, TimeLapse-seq, and AMUC-seq) can generally be classified as a family of nucleotide-recoding RNA-seq (NR-seq) technologies. All four use 4-thiouridine (s^4U) and/or 6-thioguanosine (s^6G) to metabolically label newly synthesized RNA for an extended period of time (typically 1-3 h). Upon purifying RNA, the s^4U is converted from a uridine analogue to a cytidine analogue in terms of its hydrogen bond

donor-acceptor pattern (guanosine to adenosine in the case of s^6G). In TimeLapse, s^4U is oxidized with sodium periodate ($NaIO_4$) to a reactive intermediate and undergoes nucleophilic attack with an amine, 2,2,2-trifluoroethylamine (TFEA). TUC chemistry is very similar but employs osmium tetroxide (OsO_4) as the oxidant and ammonia (NH_3) as the amine. Both TimeLapse and TUC chemistry completely recode the H-bonding pattern of s^4U from that of uridine analogue to that of a cytidine analogue. In SLAM chemistry, s^4U is alkylated with iodoacetamide (IAA) which only partially recodes the H-bonding pattern. When sequenced on an Illumina platform and aligned to the appropriate genome, sites of label incorporation manifest as an apparent mutation. T-to-C or G-to-A mutations are then used to infer the population of newly synthesized RNA and the proportion of newly synthesized RNA can be used to estimate the half-life of the transcript [97, 106].

NR-seq has proven to be a powerful tool in elucidating mechanisms regulating RNA stability and synthesis [134–138]. Unlike biochemical purification approaches to enrich newly synthesized transcripts [96, 103, 132], NR-seq captures valuable information about RNA kinetics without loss of information measured by traditional RNA-seq; however, NR-seq experiments are not widely adopted as a replacement for standard RNA-seq. This may be a result of hesitation to employ a new method which presents a set of challenges that are unknown to or unaddressable by the user without further developmental work. Here, we demonstrate that handling s^4U-containing RNA and processing raw NR-seq data benefit from slightly modified protocols which can be easily implemented by any user of standard RNA-seq, independent of the nucleotide-recoding chemistry. RNA labeled with s^4U can be specifically lost during handling or during alignment to a reference genome, a phenomenon referred to as dropout. Handling dropout occurs because s^4U-containing RNA adheres to surfaces of cell culture dishes and untreated test tubes more than unlabeled RNA. This is addressed by avoiding cell lysis in cell culture dishes and using test tubes designed to minimize nucleotide surface adherence. Computational dropout in NR-seq data occurs because standard aligner softwares assume all mismatches should penalize alignment scores. These penalties force highly mutated reads to drop below filter cutoffs and artificially lower the proportion of mutation-containing reads in processed data. Computational dropout can be addressed using a 3-base alignment strategy and has been demonstrated to be essential

when aligning very short reads containing T-to-C chemically induced mutations [50]. These two solutions increase the yield of label-containing reads and should make challenges of NR-seq data more approachable for users. We also demonstrate that reaction conditions can affect conversion efficiency of s^4U to a cytidine analogue, but overall, all nucleotide-recoding chemistries achieve similar efficiencies and strongly agree with respect to the estimates of RNA degradation rates. Generally, the guidelines presented here for NR-seq data are vital to producing high-quality datasets and taking advantage of the temporal dimension of NR-seq data.

2.4 Results

2.4.1 Additional care must be taken when handling s^4U-labeled RNA to avoid dropout

4-thiouridine (s^4U) is a common reagent to metabolically label newly synthesized RNA and many methods have been developed to take advantage of its chemical properties [50, 96–99, 102, 131, 132]. Despite this, we are not aware of any studies characterizing the behavior of s^4U-containing RNA compared to unlabeled RNA. We found that different RNA handling methods improves the recovery of s^4U-containing transcripts, apparent in sequencing tracks for *DHX9* (Figure 2.1A,B). Adherent cells are commonly lysed directly in cell culture dishes; however, this method is susceptible to specific dropout of s^4U-containing RNA. We found that an improved protocol to handle s^4U-containing RNA involves scraping cells from the cell culture dish and lysing in low nucleotide-binding sample tubes (Figure 2.1A).

We treated adherent cells with s^4U for 2 hours and collected RNA with both protocols. We performed chemistry developed for TimeLapse-seq, SLAM-seq, or TUC-seq on s^4U-labelled RNA from the same samples to demonstrate that dropout is a general effect caused by s^4U and not a specific chemistry. Upon examining the data, we found that NR-seq data correlate well with other data collected with the same handling but not with data collected with the other conditions, both in terms of total reads and T-to-C mutation-containing reads (Figure 2.2A,B). In addition, highly-labeled RNA and nascent transcripts containing introns are visibly depleted from samples collected under dropout conditions

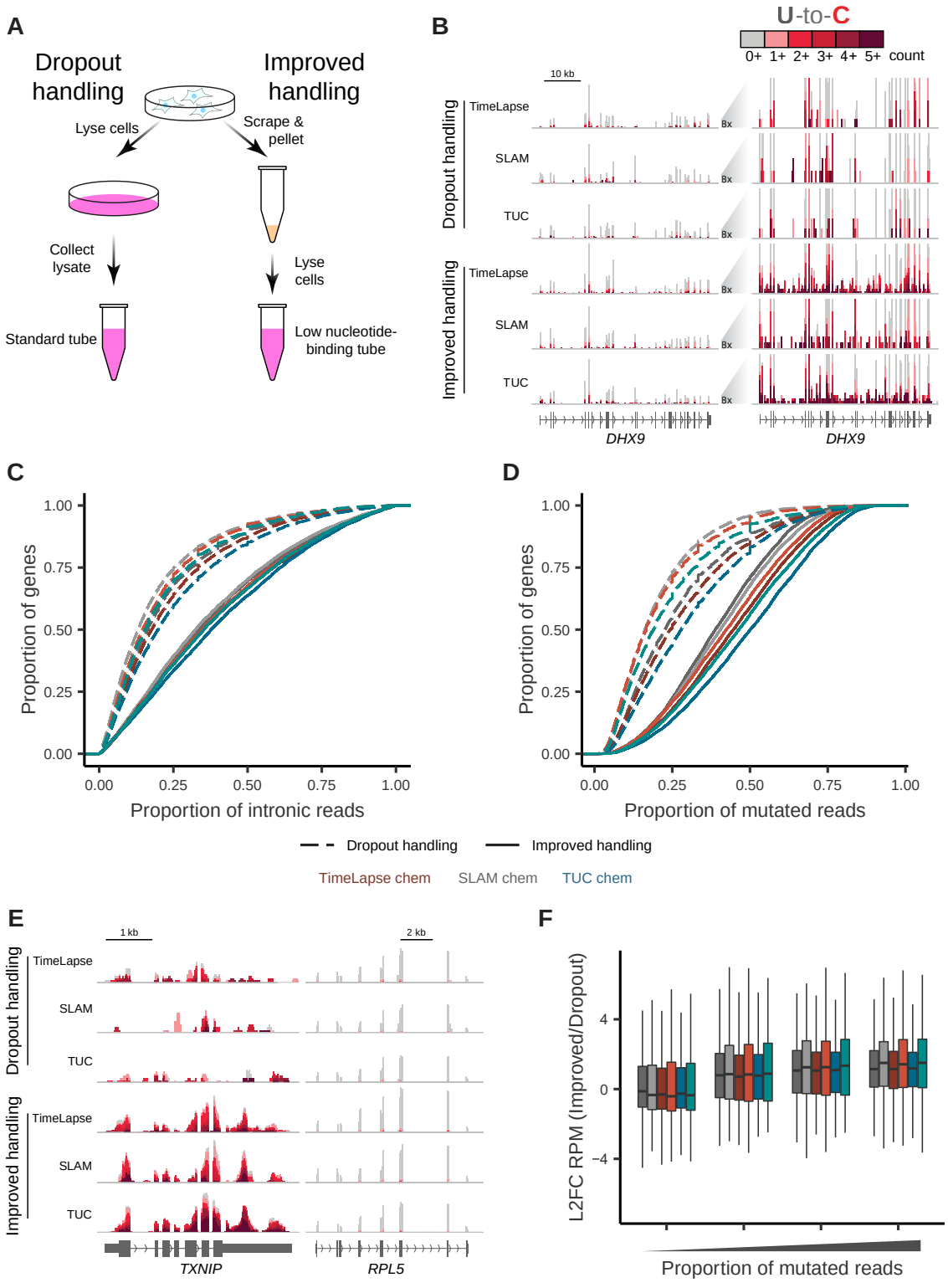

Figure 2.1: Caption on next page.

Figure 2.1: **s^{4}U-labeled RNA is specifically lost during handling** (A) Schematic of the dropout-vulnerable RNA isolation approach to one that improves s^{4}U-labeled RNA yield. (B) Example NR-seq tracks of coverage over a highly intronic gene collected with both handling approaches and treated with TimeLapse, SLAM, or TUC chemistry. (C and D) Cumulative distribution plots of the proportion of reads aligning to an intronic region (C) or proportion of reads containing a T-to-C mutation (D) for all genes in NR-seq data. (E) Example NR-seq tracks of coverage over the gene of a high-turnover (TXNIP) and slow-turnover (RPL5) transcript. (F) Genes were grouped by the proportion of reads containing a mutation in TimeLapse-seq data and compared to the $\log_2$ fold change of non-intronic reads per million in NR-seq data

when compared to data collected with the improved handling (Figure 2.1B). To test the loss of highly-labeled transcripts we calculated the proportion of intronic reads aligning to each gene as we expect these to be essentially 100% labeled due to their short lifetimes relative to the 2 hour labeling period. This analysis demonstrates that nascent RNA is globally underrepresented in the dropout dataset compared to the improved handling dataset independent of the conversion chemistry (Figure 2.1C). To test the effect on all s^{4}U-containing RNA, we calculated the proportion of all reads aligning to each gene which contain a T-to-C mutation (Figure 2.1D). We found that sequencing data from improved handling conditions contain a higher proportion of mutation-containing reads, suggesting that these conditions significantly reduce dropout compared to standard protocols.

Next, we reasoned that s^{4}U-specific dropout should differentially affect high-turnover and slow-turnover transcripts. Fast-turnover transcripts will be highly labeled because most of their population will have been synthesized during the two hour labeling time, whereas slow turnover transcripts should be mostly unlabeled. Exemplified by *TXNIP* (fast-turnover) and *RPL5* (slow-turnover), we found that coverage over a fast-turnover transcripts is strikingly improved with the improved handling protocol and a slow-turnover transcript is relatively unaffected (Figure 2.1E). We examined the change in coverage in reads per million (RPM) of mature, mutation-containing reads aligning to genes grouped by the proportion of reads containing a T-to-C mutation to test if this trend holds true across the entire dataset (Figure 2.1F). This analysis shows that coverage over genes with a low proportion of labeled reads is unaffected by handling, but coverage is increased over genes with a higher proportion of mutation-containing reads. This shows that specific dropout of

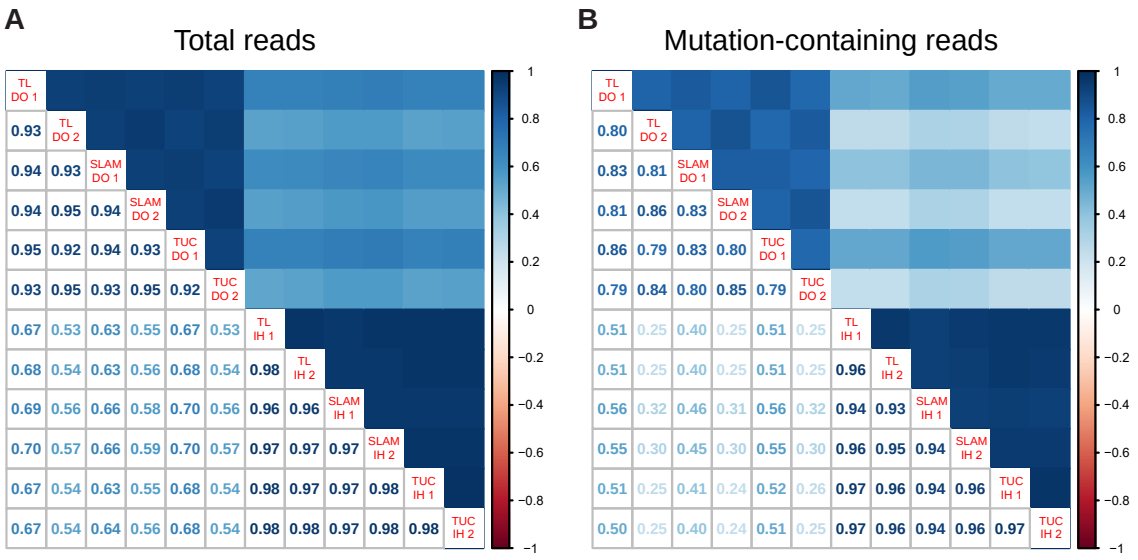

Figure 2.2: **NR-seq data prepared with different handling conditions do not correlate well with each other** (A,B) Correlation plot of the $\log_2$ counts of all (A) or T-to-C mutation containing (B) reads aligning to genes in NR-seq data with both dropout (DO) and improved handling (IH) conditions and aligned with HISAT2.

s^4U-labeled RNA biases NR-seq data, leading to an overrepresentation of previously existing RNA. Ultimately, this will shift estimates of transcript half-lives and turnover kinetics, the major advantage to employing NR-seq approaches. Our improvements to NR-seq protocols reduce bias in estimates for kinetic parameters and show that dropout is an issue with s^4U, not with any NR-seq chemistries.

2.4.2 3-nt alignment minimizes computational dropout in NR-seq data

Most traditional sequencing aligners incur an alignment score penalty when a mismatch is identified within a read. Therefore, NR-induced mutations inadvertently decrease the chance a read will be properly aligned. Most alignment softwares are customizable to allow for mismatch penalty minimization, but this would also reduce the penalty for mismatches not caused by chemical conversion of the nucleotide analogue. Furthermore, we previously demonstrated that a single chemically induced T-to-C mutation in small RNAs is sufficient to cause misalignment [50]. We reasoned that the same effect may be observed in longer sequencing reads of NR-seq data.

Recently, HISAT-3N was developed as a 3-nucleotide (3-nt) version of the splice-aware

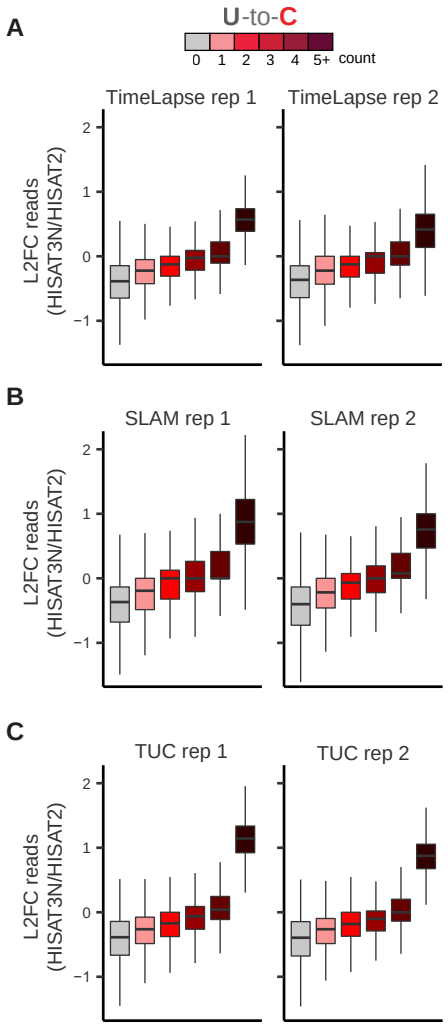

Figure 2.3: **3-base alignment recovers highly mutated reads** (A-C) The distribution of the $\log_2$ fold change in total reads aligned to each gene when comparing HISAT-3N to HISAT2 alignment of NR-seq data when treated with TimeLapse (A), SLAM (B), or TUC (C) chemistry. Reads are binned and colored by the number of T-to-C mutations in the read.

aligner HISAT2 which is commonly used to align RNA-seq data [107, 139]. 3-nt alignment approaches intentionally converts all instances of one nucleotide to another. In the case of NR-seq data using T-to-C chemistry, all T's in the genome and in sequencing data are converted to C's, thereby masking chemically induced T-to-C conversions. HISAT-3N is the first 3-nt aligner specifically developed for NR-seq data and while the authors previously validated its accuracy and efficiency, the recovery of mutation-containing reads in NR-seq data was not tested. We aligned our improved handling NR-seq data using HISAT2 with

slightly relaxed mismatch penalties (--mp 4,2) or HISAT-3N with default settings and found that HISAT-3N does not improve read count disagreement between data collected with each handling condition but does generally agree with HISAT2 alignments (Figure 2.4A-D). To assess HISAT-3N's ability to align mutation-containing reads, we calculated the $\log_2$ fold change in reads aligning to each gene grouped by the number of observed T-to-C mutations (Figure 2.3A-D). HISAT-3N tends to align more reads than HISAT2 as the number of mutations in the read increases. This effect is particularly evident for reads with five or more T-to-C mutations, whose numbers for nearly all genes upon employment of HISAT-3N.

Similarly to handling dropout, the biases introduced by computational dropout will artificially lower the total proportion of mutation-containing reads, and more strongly affect reads derived from high-turnover transcripts. In addition, computational dropout is not specific to any NR-seq chemistry as it only depends on the presence of NR-induced mutations. Therefore, we recommend that all NR-seq data be aligned using a 3-nt approach. HISAT-3N is currently the only validated, splice-aware 3-nt software available, but other software such as Bismark in combination with Bowtie 2 can be used when splicing information is not required [50, 140, 141].

2.4.3 Alternative reaction conditions optimize TimeLapse chemistry with s^4U

Improvements in handling s^4U-containing RNA and NR-seq data analysis should allow us to be more confident that we are observing all chemically converted s^4U incorporation sites. We took advantage of these improvements to determine optimal TimeLapse reaction conditions with biological samples. Standard TimeLapse chemistry is performed under slightly acidic conditions (pH 5.2) to avoid basic conditions which would promote RNA hydrolysis. Sodium periodate ($NaIO_4$) is used as the oxidant and 2,2,2-trifluorethylamine (TFEA) as the nucleophilic amine because our *in vitro* restriction endonuclease assay and NMR experiments suggested this is the most efficient oxidant-amine combination under conditions which do not promote RNA hydrolysis [97]. In addition, $NaIO_4$ is a commonly used oxidant and the low pKa leads to TFEA remaining mostly deprotonated in slightly

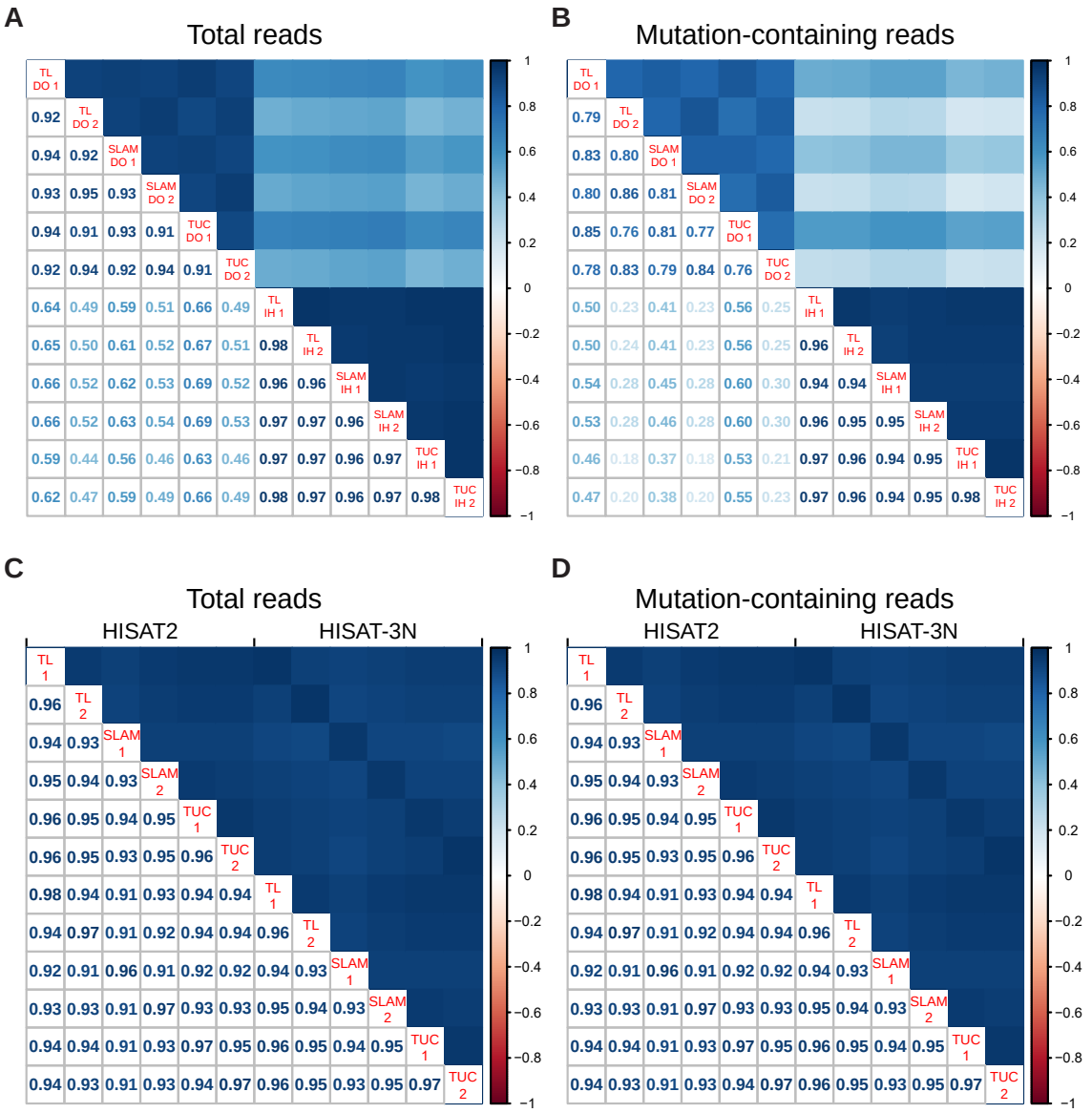

Figure 2.4: **Handling and computational dropout are independent** (A,B) Correlation plot of the $\log_2$ counts of all (A) or T-to-C mutation containing (B) reads aligning to genes in NR-seq data with both dropout (DO) and improved handling (IH) conditions and aligned with HISAT-3N. (C,D) Correlation plot of the $\log_2$ counts of all (A) or T-to-C mutation containing (B) reads aligning to genes in NR-seq data with improved handling conditions and aligned either with HISAT2 or HISAT-3N.

acidic to neutral buffers.

First, we performed TimeLapse-seq in duplicate with labeled RNA purfied using improved handling, only varying the reaction pH to compare standard conditions to a near-neutral buffer in which TFEA should be close to 100% deprotonated. We examined the

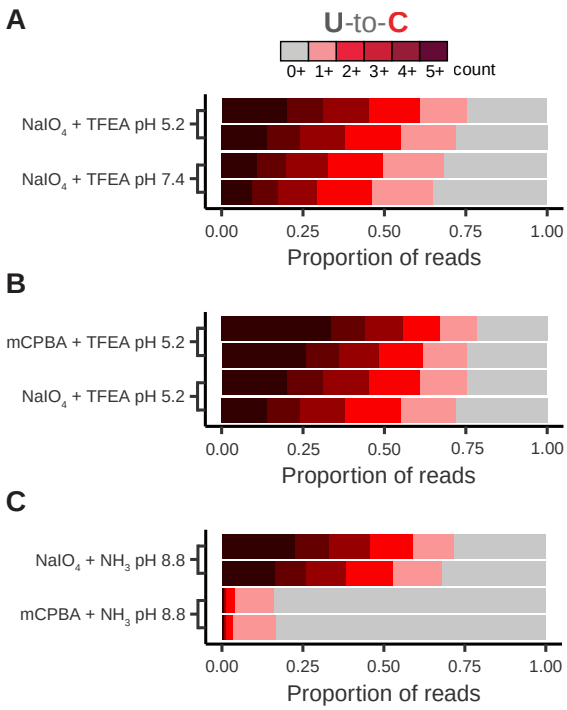

Figure 2.5: **Buffer conditions and reagents affect efficiency of TimeLapse chemistry** (A-C) The proportion of intron-aligning reads which contain T-to-C mutations when comparing buffer pH (A), oxidant (B), or amine (C). Color indicates the number of mutations in the reads.

mutational content of all reads aligning to intronic regions as they are expected to be entirely labeled after a two hour treatment. The total proportion of mutation-containing intronic reads and average mutations per U were higher under more acidic conditions, demonstrating that the reaction is more efficient under slightly acidic conditions despite a higher fraction of deprotonated TFEA (Figures 2.5A & 2.6A).

Next, we sought to directly compare the performance of two oxidants which have both been previously employed for TimeLapse chemistry. $NaIO_4$ is the most common TimeLapse oxidant, but meta-chlorobenzoic acid (mCPBA) was also characterized to efficiently oxidize s^4U under TimeLapse conditions [97]. In addition, $NaIO_4$ oxidizes 3´ diols, requiring the use of mCPBA when a 3´ ligation to RNA is required as part of downstream library prep, as is the case in Start-TimeLapse-seq (STL-seq, [50]). When comparing TimeLapse-seq data generated with both oxidants, we found that the proportion of total intronic reads containing at least one T-to-C mutation is similar with the two oxidants (Figure 2.5B). However, mCPBA leads to an increase in reads with five or more mutations and a notable

increase in the average number of mutations per U (Figures 2.5B & 2.6A). We also tested mCPBA under near-neutral conditions and found that conversion efficiency was not as high as under acidic conditions (Figures 2.5B & 2.6A). Therefore, while mCPBA and $NaIO_4$ both efficiently convert s^4U to a cytidine analogue, mCPBA is slightly more efficient under TimeLapse conditions.

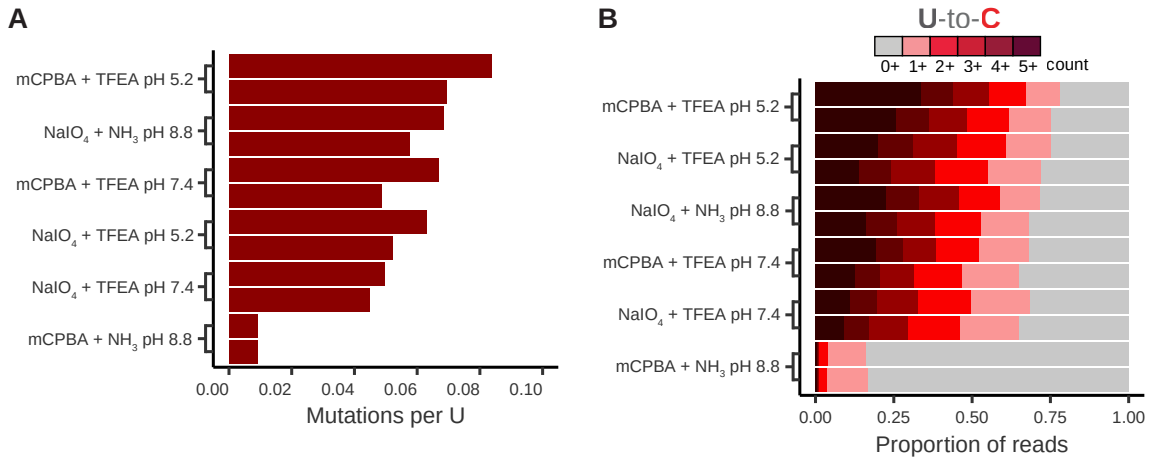

Figure 2.6: **Comparison of mutational content in TimeLapse-seq data using different reaction conditions** (A) The per U mutation rate in all intron-aligning reads with different TimeLapse conditions (B) The proportion of intron-aligning reads which contain T-to-C mutations with different TimeLapse conditions. Color indicates the number of mutations in the reads.

Next, we tested if ammonia could be used as an the amine in TimeLapse-seq as it is expected to convert s^4U directly to a C instead of a C analogue and is used in similar chemistry developed as part of TUC-seq [99]. Again, we performed NR chemistry on the same RNA using ammonia as the amine, both TimeLapse oxidants, and a basic pH due to the high pKa of ammonia. We found that only $NaIO_4$ produced elevated T-to-C mutation rates with ammonia (Figure 2.5C). When holding the oxidant constant as $NaIO_4$, we found that ammonia results in a slightly lower proportion of intronic reads containing a mutation, but the average per U mutation rates are slightly higher than with TFEA as the amine. However, mCPBA with TFEA remains the most efficient combination under these conditions (Figures 2.5B & 2.6A).

2.4.4 Estimates for RNA degradation rate constants agree between all NR-seq methods and are similarly affected by dropout

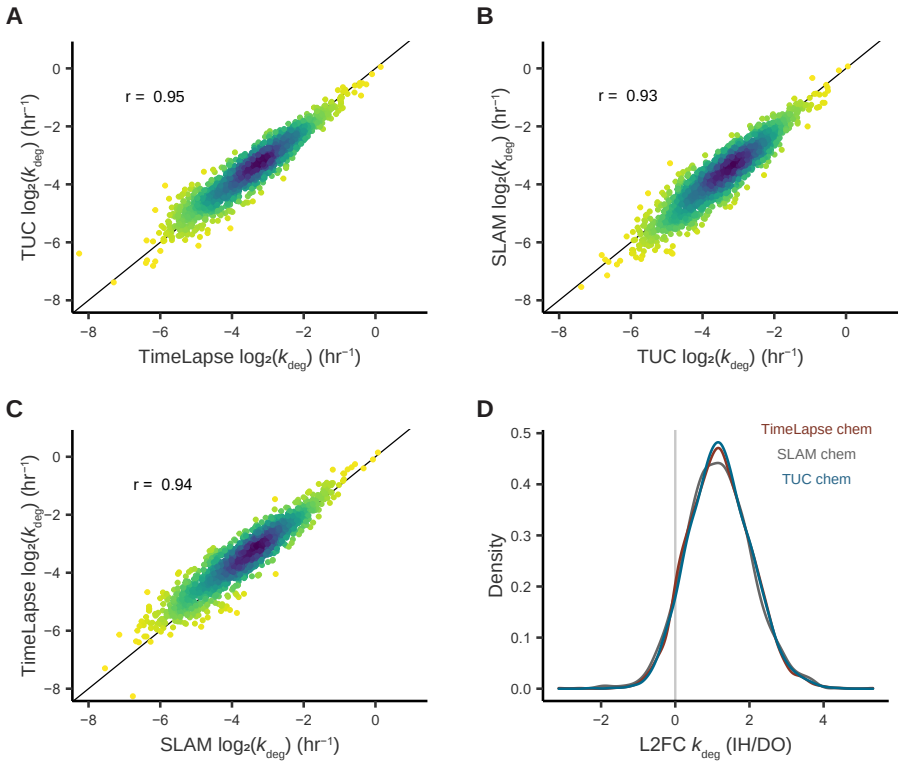

Figure 2.7: **Comparison of RNA degradation rates measured by NR-seq methods and the impact of dropout** (A-C) Scatter plots comparing k_{deg} estimates made with bakR using three NR-seq methods. The Pearson correlation coefficient is shown for each. (D) The distribution of the $\log_2$ fold change in k_{deg} estimates when comparing the improved handling and 3-nt alignment (IH) to dropout-vulnerable handling and 4-nt alignment (DO).

Finally, we asked if any NR chemistry used in TimeLapse-seq (mCPBA + TFEA), SLAM-seq (IAA), and TUC-seq (OsO_4 + NH_3) provides a significant benefit and should be preferred in all NR-seq experiments. Previously, TimeLapse-seq, SLAM-seq, and TUC-seq were determined to perform similarly in estimating mRNA degradation rates, and the authors report to have scraped cells from plates prior to lysing cells but did not employ a 3-nt alignment strategy [142]. We used HISAT-3n-aligned data collected with improved handling conditions to compare transcript-specific degradation rate constant estimates (k_{deg}) obtained with the statistical package bakR between all three NR-seq methods (Vock et al., *in prep*). In agreement with the previously published comparison, k_{deg} estimates produced using data from all three methods strongly agree with each other (Figure 2.7A-C).

We then compared the k_{deg} estimates made with data from improved processing strategies to those made with data collected with dropout conditions and aligned with HISAT2. The distributions of the $\log_2$ fold changes in k_{deg} estimates between the two conditions for each NR-seq method is nearly identical, indicating that experimental and computational dropout equally affects k_{deg} estimates, independent of NR chemistry (Figure 2.7D). On average, k_{deg} estimates are more than 2-fold larger with optimal conditions than with conditions not optimized to minimize dropout of s^4U-labeled RNA, demonstrating the importance in minimizing dropout when making estimates for RNA k_{deg}.

2.5 Discussion

Here we have shown that NR-seq experiments cannot be treated identically to RNA-seq experiments, both in terms of experimental handling and computational processing. Our data demonstrate that RNA labeled with s^4U is specifically lost during RNA isolation, an effect we call handling dropout, most likely due to labeled RNA adhering more strongly to plastic surfaces than unlabeled RNA. Dropout introduces a bias against transcripts with short half-lives. Each fast-turnover RNA molecule is more likely to be labeled with s^4U and therefore more likely to be lost during handling. Ultimately, this leads to an overestimation of transcript half-lives and a dampening of any changes in turnover rates caused by an experimental treatment.

Furthermore, T-to-C mutation-containing reads in sequencing data are more difficult to align due to mismatch penalties applied by standard RNA-seq aligner softwares. We showed that reads with more T-to-C mutations, particularly five or more, are less likely to be aligned with a standard RNA-seq aligner. This is at least partially attributable to standard aligners requiring a seed sequence to perfectly match the reference genome. Typically the seed is twenty base pairs long, but if a perfect twenty base pair long stretch does not exist, the read will fail to be aligned. In addition, each mismatch between the read and reference genome incurs an alignment penalty. If there are a sufficient number of mismatches, the read may fail to align. Each of these parameters can be customized to a certain extent to allow for more mismatches, however this also allows for mismatches of any N-to-N and not just T-to-C. To

solve this issue, we took advantage of a new 3-nt RNA-seq aligner, HISAT-3N, which does not penalize the induced T-to-C mismatches [107]. HISAT-3N follows a similar strategy as aligners developed for bisulfite sequencing: all Ts in the reference genome and sequencing data are converted to C's. The 3-nt alignment prevents T-to-C mismatches from penalizing alignments, and because HISAT-3N stores information about the original sequence, it can be used to identify T-to-C mutations. HISAT-3N was previously demonstrated to be a fast and accurate aligner, and is therefore highly recommended for all NR-seq experiments.

We showed that TimeLapse chemistry efficiently converts s^4U to a cytidine analogue with either $NaIO_4$ or mCPBA and performed best under slightly acidic conditions. mCPBA achieved a modestly higher conversion rate than $NaIO_4$, but, more importantly, preserves 3´ ends of RNAs which can be important for downstream processing steps of a sequencing experiment [50]. Therefore, the difference between $NaIO_4$ and mCPBA is inconsequential unless a 3´ adapter ligation is required. Likewise, TFEA and ammonia can be used as the nucleophilic amine with $NaIO_4$, but ammonia should not be used with mCPBA.

Finally, s^4U conversion chemistries developed as part of TUC-seq, SLAM-seq, and TimeLapse-seq tend to provide comparable estimates of the steady-state kinetics of cellular RNA, unless material availability or safety is a concern. Independent of chemistry selection, additional consideration must be taken when preparing samples and analyzing data NR-seq data. s^4U has long been used as a chemical tool for RNA metabolic labeling, but was never previously characterized as a challenging molecule to handle. On the other hand, NR-seq is a quickly evolving technique and methods to analyze the data are still being developed. With the work presented here, we established new guidelines to address previously unappreciated challenges of performing NR-seq which will improve the power of NR-seq methods as tools to study RNA dynamics.

Chapter 3

Case studies using RNA metabolic labeling to study transcription and RNA dynamics

3.1 Introduction

As part of my dissertation work in the Simon lab, I have sought out opportunities to collaborate with other scientists seeking to better understand trancsriptional behavior and/or RNA dynamics by using the Simon lab's toolkit of transcriptomic techniques. These collaborations have exposed me to dozens of brilliant scientists across more than ten different projects at Yale and other institutions. Each collaboration brought with it new challenges and pushed me to be a better scientist by expanding my knowledge, improving my communication skills, and refining my techniques. I consider my collaborations to be a genuine privilege and some of the most important experiences during my time in the Simon lab. The following are summaries of my contributions to four published studies, presented in chronological order of publication date.

3.2 The role on lncRNA transcription in tumorigenesis

This section is adapted from:

Olivero, C.E., Martínez-Terroba, E., **Zimmer, J.**, Liao, C., Tesfaye, E., Hooshdaran, N., Schofield, J.A., Bendor, J., Fang, D., Simon, M.D., Zamudio, J.R., Dimitrova, N. (2020). p53 Activates the Long Noncoding RNA Pvt1b to Inhibit Myc and Suppress Tumorigenesis. Mol. Cell, 77(4), 761-774. doi: 10.1016/j.molcel.2019.12.014

In collaboration with the Dimotrova lab, we applied metabolic labeling- and TimeLapse-based techniques to determine the function of a novel lncRNA isoform, *Pvt1b*. TT-TL-seq and TimeLapse-seq revealed the induction of the *Pvt1b* isoform and its repressive effect on the Myc gene upon p53 activation.

The tumor suppressor p53 transcriptionally activates target genes to suppress cellular proliferation during stress. p53 has also been implicated in the repression of the proto-oncogene Myc, but the mechanism has remained unclear. Here, we identified *Pvt1b*, a p53-dependent isoform of the long noncoding RNA (lncRNA) *Pvt1*, expressed 50 kb downstream of *Myc*, which becomes induced by DNA damage or oncogenic signaling and accumulates near its site of transcription. We showed that production of the *Pvt1b* RNA is necessary and sufficient to suppress *Myc* transcription in cis without altering the chromatin organization of the locus. Inhibition of *Pvt1b* increased *Myc* levels and transcriptional activity and promotes cellular proliferation. Furthermore, *Pvt1b* loss accelerates tumor growth, but not tumor progression, in an autochthonous mouse model of lung cancer. These findings demonstrated that *Pvt1b* acts at the intersection of the p53 and Myc transcriptional networks to reinforce the anti-proliferative activities of p53.

3.2.1 Activation of a p53-Dependent Pvt1 Isoform, Pvt1b

To characterize the transcripts produced from the Pvt1 locus, we performed TimeLapse-seq. We found evidence for extensive alternative splicing and confirmed that variants containing exon 1b were induced by p53, while exon 1a-containing variants were constitutively expressed (Figure 3.1). Despite the splicing heterogeneity, TimeLapse-seq revealed that stress-induced Pvt1b differed from constitutively expressed Pvt1a solely by the use of exon 1b versus exon 1a and exhibited comparable splicing patterns to downstream exons (Figure 2F). We concluded that p53 activation during genotoxic and oncogenic stress initiated tran-

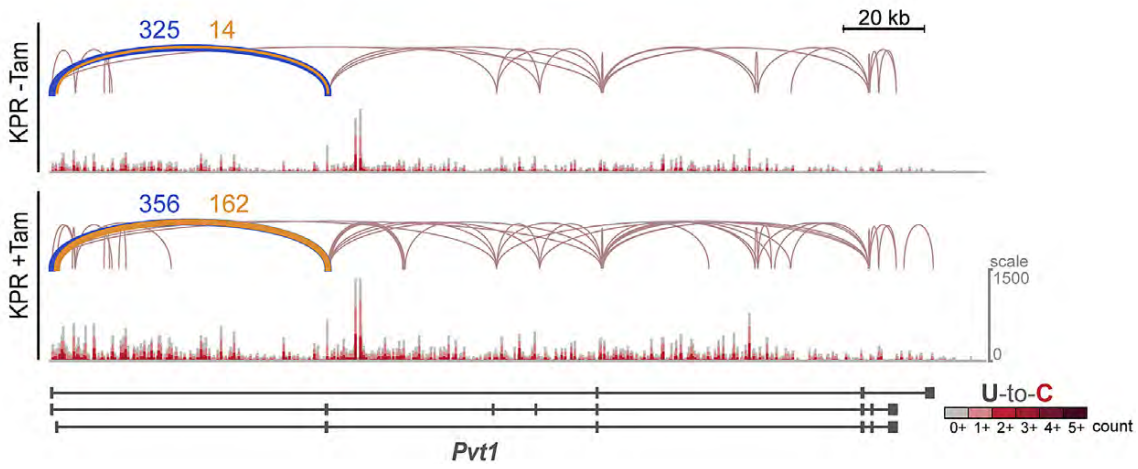

Figure 3.1: **The *Pvt1b* isoform is upregulated by p53 activation** Genome browser tracks and Sashimi plots from TimeLapse-seq data in KPR cells, treated as indicated. Average number of splice junctions from two biological replicates from exon 1a to exon 2 (blue) and from exon 1b to exon 2 (orange) are indicated.

scription in the Pvt1 locus from exon 1b, leading to the production of the p53-dependent isoform *Pvt1b*, whereas *Pvt1a* represented a largely constitutively expressed isoform.

3.2.2 Pvt1b Suppresses Myc Transcriptional Activity *In Vitro*

To test whether *Pvt1b* acted at the transcriptional or post-transcriptional level, we sequenced nascent RNA from untreated and Tam-treated ΔRE and Con *KPR* cells ([97]). We found that nascent *Myc* transcripts were significantly upregulated in ΔRE+Tam compared with Con+Tam *KPR* cells, indicative of transcriptional regulation (Figure 3.2A,B). These data revealed that *Pvt1b* production promotes transcriptional suppression of *Myc*.

Next, we queried how the changes in *Myc* RNA levels affected the Myc transcriptional program by examining the consequence of *Pvt1b* loss on a curated set of 196 Myc target genes (gene set enrichment analysis, HALLMARK_MYC_TARGETS_V1; [143]). We plotted the cumulative frequency distribution of the fold change (FC) of Myc target genes in ΔRE cells relative to Con cells in the presence of stress (log FC [ΔRERE/Con+stress]). Compared with a randomly generated set of control genes expressed at comparable levels, we found a significant increase in the levels of Myc targets in MEFs and *KPR* cells (Figure 3.2C). We concluded that *Myc* derepression by ΔRE mutagenesis led to a small but significant

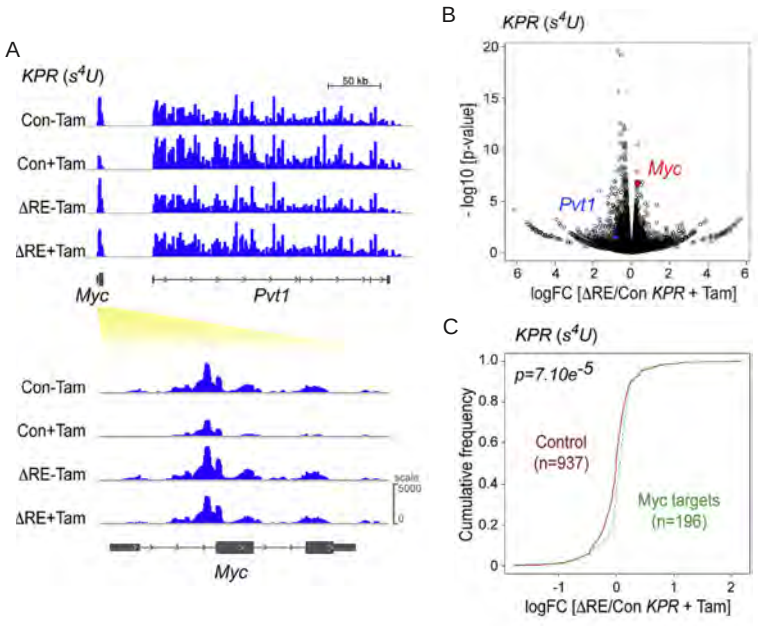

Figure 3.2: ***Myc* and Myc target genes are downregulated following p53 activation** (A) Top: Genome Browser tracks depicting the *Myc-Pvt1* locus from transient-transcriptome (TT) TimeLapse-seq. Bottom: detail of the *Myc* locus. (B) Butterfly plot depicting the fold change (log FC) in gene expression of indicated samples relative to statistical significance (-$\log_{10}$[p value]; KPR, n = 2 biological replicates). Gene expression profiling was performed by TimeLapse-seq of ribosomal cDNA-depleted s^4U-labeled RNA isolated from Con or ΔRE gRNA-expressing KPR cells, untreated or treated with Tam for 16 h. Total Pvt1 (blue) and Myc (red) are labeled. (C) Cumulative frequency distribution plot of differential expression for a set of curated Myc target genes and a matched set of control genes.

increase in the transcriptional activity of Myc.

3.3 Measuring readthrough transcription in the downstream of gene region

This section is adapted from:

Rosa-Mercado, N.A., **Zimmer J.T.**, Apostolidi, M., Rinehart, J., Simon, M.D., and Steitz, J.A. (2021). Hyperosmotic stress alters the RNA polymerase II interactome and induces readthrough transcription despite widespread transcriptional repression. Mol. Cell, 81(3), 502-513.e4. doi: 10.1016/j.molcel.2020.12.002

In collaboration with the Steitz lab, we sought to characterize readthrough transcription in the downstream-of-gene region using transient-transcriptome-TimeLapse-seq to measure transcriptional activity upon hyperosmotic stress and Integrator knockdown.

Stress-induced readthrough transcription results in the synthesis of downstream-of-gene (DoG)-containing transcripts. The mechanisms underlying DoG formation during cellular stress remain unknown. Nascent transcription profiles during DoG induction in human cell lines using TT-TimeLapse sequencing revealed widespread transcriptional repression upon hyperosmotic stress. Yet, DoGs are produced regardless of the transcriptional level of their upstream genes. ChIP sequencing confirmed that stress-induced redistribution of RNA polymerase (Pol) II correlates with the transcriptional output of genes. Stress-induced alterations in the Pol II interactome are observed by mass spectrometry. While certain cleavage and polyadenylation factors remain Pol II associated, Integrator complex subunits dissociate from Pol II under stress leading to a genome-wide loss of Integrator on DNA. Depleting the catalytic subunit of Integrator using siRNAs induces hundreds of readthrough transcripts, whose parental genes partially overlap those of stress-induced DoGs. Our results provide insights into the mechanisms underlying DoG production and how Integrator activity influences DoG transcription.

3.3.1 Hyperosmotic stress causes widespread transcriptional repression

We established the nascent transcriptional profiles accompanying DoG induction by performing TT-TL-seq [97] of untreated HEK293T cells and cells exposed to hyperosmotic stress (Figure 3.3A). Specifically, we exposed cells to 80mM KCl for 60 min but added the nucleoside analog 4-thiouridine (s^4U) during the last 5 min to label RNAs being actively transcribed [96]. After extracting RNA from HEK293T cells, RNA from *Drosophila* S2 cells was added to each sample as a normalization control to ensure accurate differential expression analysis [144, 145]. RNAs containing s^4U were then biotinylated using methanethiosulfonate (MTS) chemistry and enriched on streptavidin beads [132]. Finally, U-to-C mutations were induced using TL chemistry to assess the nascent nature of the enriched RNAs [97]. TT-TL-seq experiments were performed using conditions previously found to induce DoGs [146].

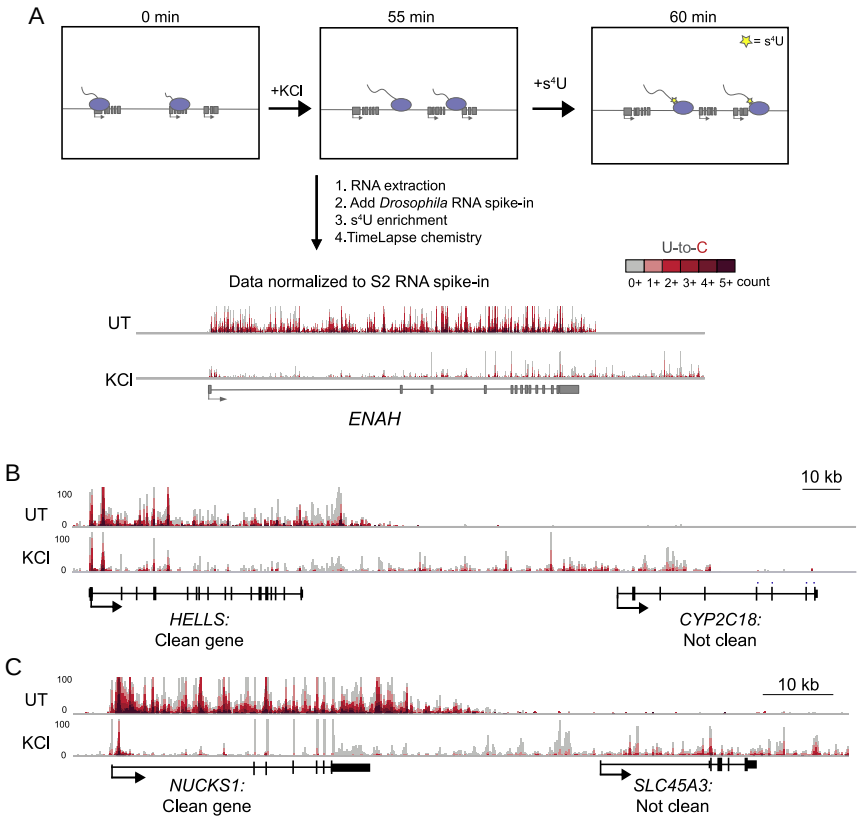

Figure 3.3: **TT-TL-seq reveals transcriptional profiles that accompany DoG induction after hyperosmotic stress** (A) Setup for TT-TimeLapse sequencing (TT-TL-seq) experiments in HEK293T cells. An arrow indicating directionality marks the beginning of each transcription unit. Exons are shown as rectangles, and Pol II molecules are light-purple ovals with attached nascent RNAs. Genome browser views of TT-TL-seq data for *ENAH* provide an example of results from untreated (UT) and KCl treated (KCl) cells after normalization to the spike in control. (B) Browser image of TT-TL-seq data exemplifying a clean gene (*HELLS*) and a gene that does not meet the criteria for a clean gene (*CYP2C18*). The DoG produced from *HELLS* reads into *CYP2C18*, making the latter appear to be transcriptionally activated by hyperosmotic stress ($\log_2$ FC = 5.39). (C) The DoG produced from *NUCKS1* is assigned to *SLC45A3* because of extensive read-in transcription, which also complicates accurate differential expression analysis for *SLC45A3* ($\log_2$ FC = 4.13). In the browser images, ∼4–5 kb upstream of *HELLS* and *NUCKS1* are shown.

It was previously observed that read-in transcription of DoGs into neighboring genes leads to the mis-characterization of overlapping transcripts as being activated by stress ([147–150] Figure 3.3B). Moreover, read-in transcription confounds the assignment of DoGs to the corresponding parent gene (Figure 3.3C). Our analyses suggest that ∼55% of expressed genes experience read-in transcription after hyperosmotic stress. Therefore, to ensure accurate differential expression analyses and DoG characterization, we generated

a sub-list of genes (referred to as "clean genes" throughout the article). The term "clean genes" describes genes that are expressed, do not overlap with readthrough regions that correspond to neighboring genes on either strand, and have higher expression within the gene body than the region 1 kb upstream of the gene's transcription start site (TSS) ([150]; Figure 3.3B and C). We identified 4,584 clean genes in HEK293T cells after hyperosmotic stress and analyzed their transcriptional regulation.

Consistent with previous reports analyzing steady-state RNAs in human cells [151], we identified changes in transcriptional responses after hyperosmotic stress. Our results reveal predominantly decreases in nascent transcript levels after stress (Figure 3.4A–C). Specifically, the number of normalized read counts corresponding to clean genes decreases 3-fold after KCl treatment. Yet, we find that a subset of clean genes bypasses this transcriptional repression (Figure 3.4B,C), including GADD45B (Figure 3.4D), which is known to be induced by hyperosmotic stress [152]. More than 88% of clean genes were repressed after hyperosmotic stress, while only 3% were upregulated (Figure 3.4C).

3.3.2 Stress-induced readthrough transcripts arise independent of gene-transcription levels

Consistent with widespread transcriptional repression, normalized TT-TL-seq read counts within the bodies of DoG-producing clean genes decreased after hyperosmotic stress, while read counts corresponding to DoG regions increased (Figure 3.5A). However, $\log_2$ fold changes in nascent RNAs of DoG-producing clean genes show that DoGs are produced from genes that experience all three types of transcriptional responses (Figure 3.5B,C). Specifically, 2.9% of DoGs arise from activated clean genes, 87.8% arise from repressed clean genes, and 9.3% arise from clean genes that retain comparable expression in stressed and unstressed HEK293T cells (Figure 3.5B). We then asked whether DoGs preferentially arise from genes that are transcriptionally repressed upon hyperosmotic stress. Interestingly, the percentage of DoG-producing genes within each class of transcriptional regulation is consistent, comprising 12%–14% (Figure 3.5C). We conclude that DoGs are produced regardless of the transcriptional level of their upstream genes (Figure 3.5D).

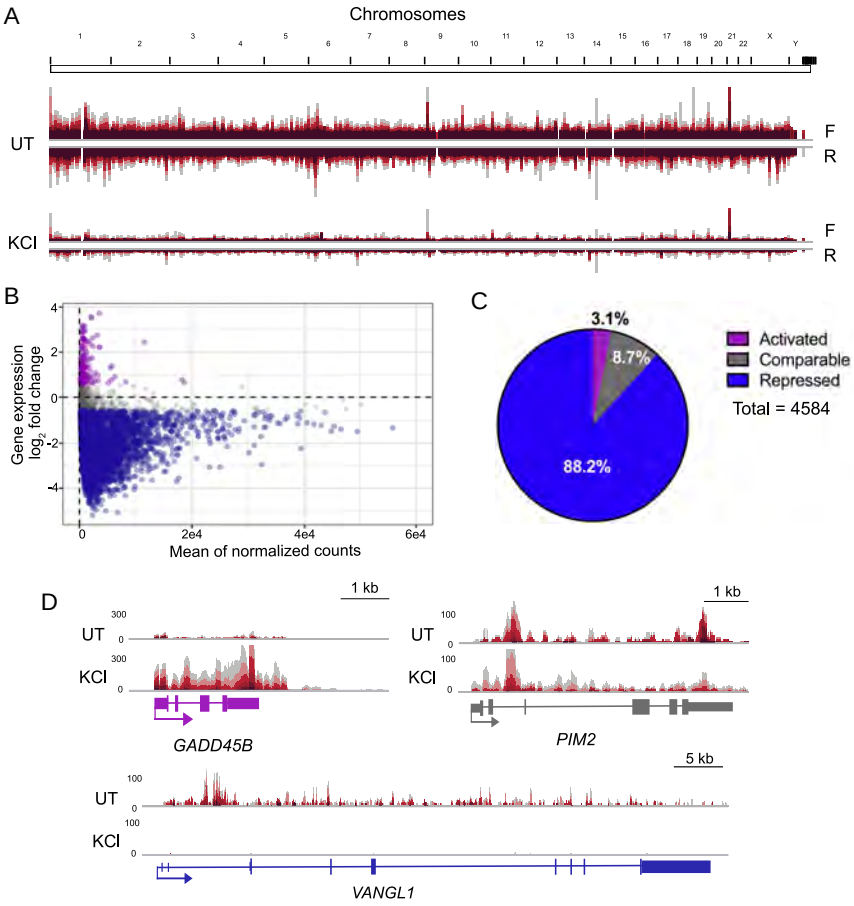

Figure 3.4: **Hyperosmotic stress leads to widespread transcriptional repression** (A) Whole-genome view of TT-TL-seq normalized reads for forward (F) and reverse (R) strands in UT and KCl samples. (B) Minus average plot showing the $\log_2$ fold change for clean genes on the y axis and the mean of normalized counts on the x axis. Activated genes are shown in purple, genes retaining comparable expression are gray, and repressed genes are blue (n = 4584). (C) Pie chart illustrating the percentage of clean genes within each of the 3 categories of transcriptional regulation (activated gene, $\log_2$ FC > 0.58; comparable gene, $\log_2$ FC < 0.58 but > -0.58; repressed gene, $\log_2$ FC < -0.58). (D) Browser shots of TT-TL-seq tracks from HEK293T cells for *VANGL1*, which is transcriptionally repressed by hyperosmotic stress ($\log_2$ FC = -4.97), *PIM2*, which retains comparable expression after KCl treatment ($\log_2$ FC = 0.28), and *GADD45B*, which is activated by hyperosmotic stress ($\log_2$ FC = 3.72).

3.3.3 Clean DoG-producing genes are functionally enriched for transcriptional repression

Previous analyses of DoG-producing genes did not reveal any functional enrichment [146]. We suspected that the challenge of assigning DoGs to the correct gene of origin because of their extension into neighboring genes may have complicated previous efforts (Figure 3.3C).

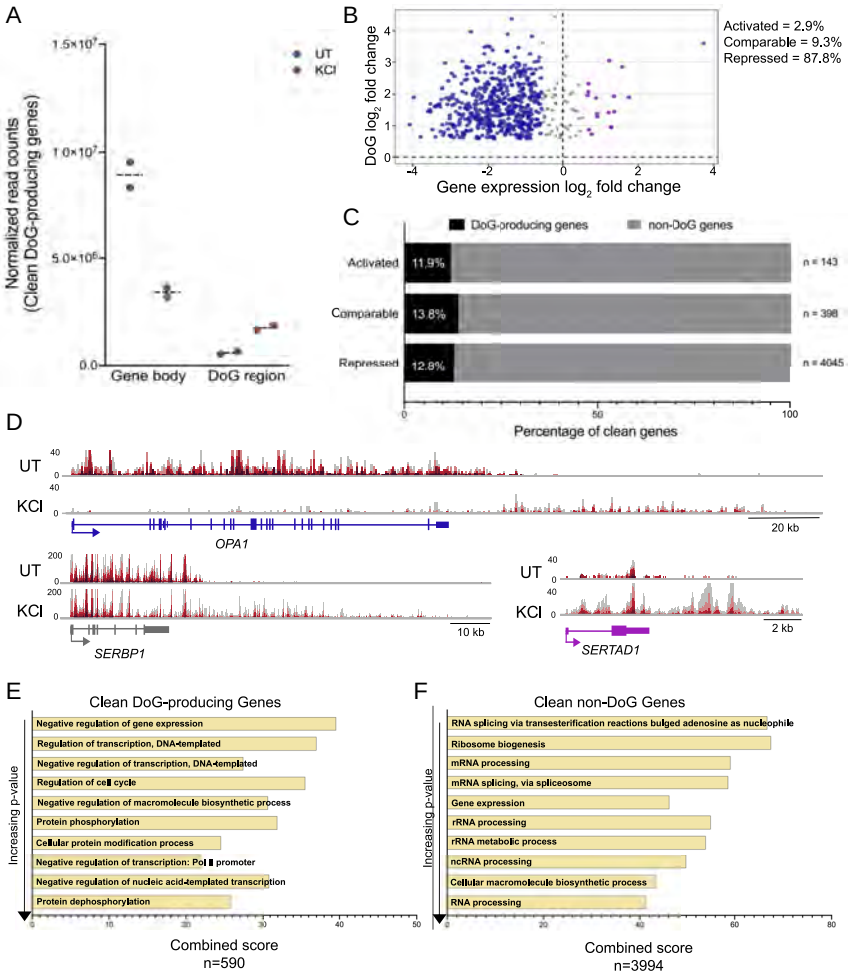

Figure 3.5: **DoGs arise regardless of the transcriptional levels of their upstream genes upon hyperosmotic stress** (A) Interleaved scatterplot showing the sum of normalized TT-TL-seq read counts of DoG-producing clean genes and corresponding DoG regions (n = 590) in untreated and KCl-treated HEK293T cells for two biological replicates. (B) Scatterplot showing clean gene $\log_2$ fold change (FC) for the gene body on the x axis and the $\log_2$ FC for the DoG region on the y axis. DoGproducing genes that are transcriptionally activated upon stress are represented in purple, genes retaining comparable levels of expression are gray, and genes that are repressed are blue. (C) Bar graph showing the percentage of DoG-producing clean genes (black) within each category of transcriptional regulation. (D) Browser image showing UT and KCl TT-TL-seq reads for *OPA1*, a transcriptionally repressed DoG-producing clean gene (gene $\log_2$ FC = -3.22), for *SERBP1*, which retains comparable expression after stress (gene $\log_2$ FC = -0.45), and for a transcriptionally activated DoG-producing clean gene, *SERTAD1* (gene $\log_2$ FC = 1.23). (E and F) Bar graphs show gene ontology combined scores for the 10 most significantly enriched biological processes in order of increasing p value for (E) DoG-producing clean genes and for (F) non-DoG clean genes. Combined scores are the product of the p value and the Z score as calculated by Enrichr [153].

Therefore, we revisited the question of whether DoG-producing genes are enriched for certain biological processes using only clean genes. We performed gene ontology analysis of DoG-producing clean genes and clean genes that fail to generate DoGs (non-DoG genes) using Enrichr [153,154]. Interestingly, 5 out of the 10 enriched terms with the most significant p values for DoG-producing genes are related to transcriptional repression (Figure 3.5E). The remaining 5 terms are related to transcriptional regulation and protein modifications. Non-DoG genes do not show such a striking enrichment for terms related to transcriptional repression compared to other terms (Figure 3.5F). Instead, these genes are strongly enriched for general processes related to RNA processing.

3.3.4 Depletion of Integrator endonuclease leads to DoG production

The Integrator complex regulates transcription termination at many noncoding RNA loci and has been shown to bind the 3´ end of certain protein-coding genes [155, 156]. We investigated whether knocking down the catalytic subunit of the complex, Int11, using siRNAs is sufficient to induce DoGs. We transfected HEK293T cells with an siRNA against Int11 (siInt11) or with a non-targeting siRNA control (siC). HEK293T cells stably expressing FLAG-tagged, siRNA-resistant wild-type (WT), or catalytically inactive Int11 (E203Q) were also transfected with siInt11 [155, 157] for 72 h.

We assessed the extent to which knockdown of endogenous Int11 induces DoGs genome-wide. To increase cell viability, we performed TT-TL-seq on HEK293T cells transfected with an siRNA against Int11 for 48 h. Results obtained from cells lacking functional Int11 reveal hundreds of readthrough sites across the genome that are induced by more than 1.5-fold compared to the siC-transfected sample (Figure 3.6A). According to TT-TLseq data, induction of readthrough transcription after depletion of endogenous Int11 was most evident upon expression of the E203Q mutant (Figure 3.6A and B). Yet, readthrough transcripts observed in siInt11-transfected cells expressing no rescue and in cells expressing the E203Q mutant Int11 were highly correlated (Figure 3.7A). As expected, the most highly induced sites of readthrough transcription corresponded to snRNA genes [155,158]. We also detected readthrough downstream of lncRNA and histone genes as previously described [159, 160]. However, most identified readthrough sites were downstream of protein-coding genes. Of the

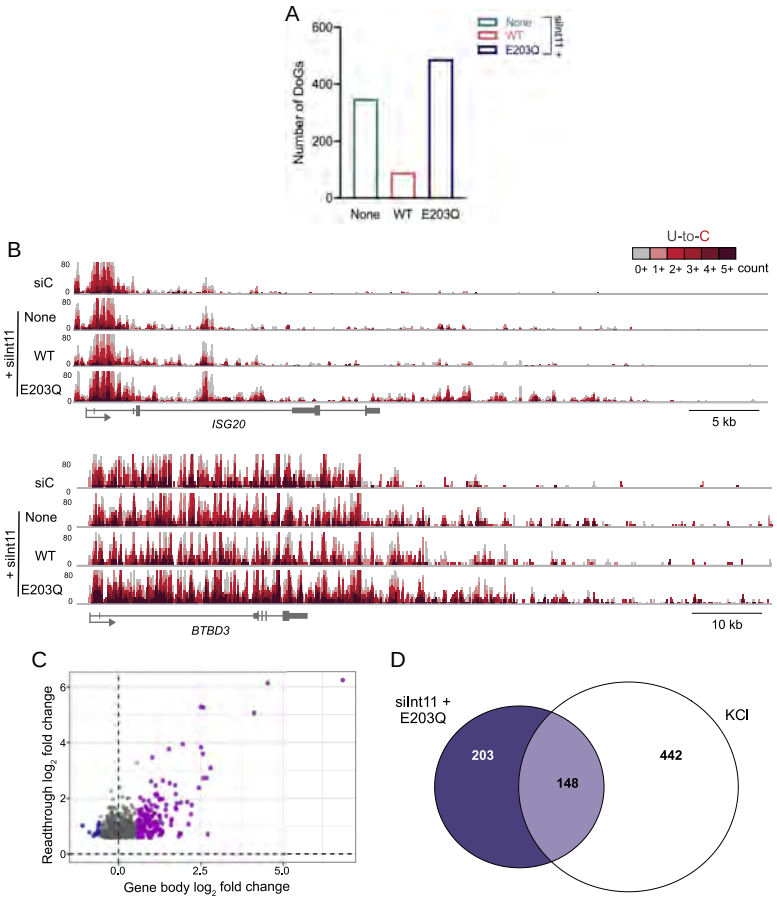

Figure 3.6: **Depletion of Integrator nuclease subunit leads to readthrough transcription** (A) Bar graph showing the number of DoGs induced after siRNA knockdown of endogenous Int11. (B) Browser image of TT-TL-seq data for two genes that produce DoGs upon depletion of functional Int11. (C) Scatterplot showing gene expression $\log_2$ fold change (FC) for siInt11+E203Q HEK293T cells on the x axis and the $\log_2$ FC of the corresponding readthrough transcripts on the y axis. Genes that are activated in siInt11+E203Q cells compared to siC-transfected cells are purple, unaffected genes are gray and repressed genes are blue. All readthrough sites identified in the siInt11+E203Q sample are represented in this plot (n = 840). (D) Venn diagram displaying overlap between the identities of clean genes that produce readthrough transcripts in siInt11+E203Q cells (dark blue) and those of DoG-producing clean genes in KCl-treated samples (white).

840 readthrough transcripts induced by knockdown of functional Int11 (E203Q sample), 489 exhibited greater than 80% read coverage in the region 5 kb downstream of the annotated termination site of the gene of origin and, therefore, met all criteria to be classified as DoG RNAs (Figure 3.6A and B).

Depletion of Integrator subunits has been shown to alter the transcriptional levels of certain genes [45, 46, 60, 156]. Consistently, we found that depletion of functional Int11

in HEK293T cells differentially affects more than a thousand genes (Figure 3.7B). Examination of the expression levels of the parent genes revealed that readthrough transcripts predominantly arise from upregulated genes or from genes that retain comparable expression after knockdown, while very few arise from genes that are transcriptionally repressed (Figure 3.6C).

Given our observation that the interaction between Integrator subunits and Pol II is disrupted by hyperosmotic stress, we asked how the identities of genes producing readthrough transcripts upon depletion of functional Int11 compare to genes that produce DoGs after hyperosmotic stress. We identified 232 clean genes producing readthrough transcripts in siInt11-transfected cells and 351 clean genes producing readthrough transcripts in the siInt11+E203Q mutant sample. Comparison with the 590 DoG-producing clean genes identified in stressed cells showed that up to 25% of KCl-induced DoGs are detected at loci that also produce readthrough transcripts after depletion of functional Int11 (Figures 3.6D and 3.7C). These readthrough transcripts are generally more robustly induced after KCl treatment than upon depletion of functional Int11 (Figure 3.7D), suggesting that, although Int11 knockdown is sufficient to produce readthrough transcription, decreased interactions between Integrator and Pol II are not solely responsible for DoG induction upon hyperosmotic stress.

3.4 Probing the effect of a splicing factor mutant on RNA stability

This section is adapted from:

Biancon, G., Joshi, P., **Zimmer, J.T.**, Hunck, T., Gao, Y., Lessard, M.D., Courchaine, E., Barentine, A.E.S., Machyna, M., Botti, V., Qin, A., Gbyli, R., Patel, A., Song, Y., Kiefer, L., Viero, G., Neuenkirchen, N., Lin, H., Bewersdorf, J., Simon, M.D., Neugebauer, K.M., Tebaldi, T., Halene, S. (2022). Precision analysis of mutant U2AF1 activity reveals deployment of stress granules in myeloid malignancies. Molecular Cell, 82(6), 1107-1122.e7.

In collaboration with the Halene lab, we applied TimeLapse-seq to ask how stress granule

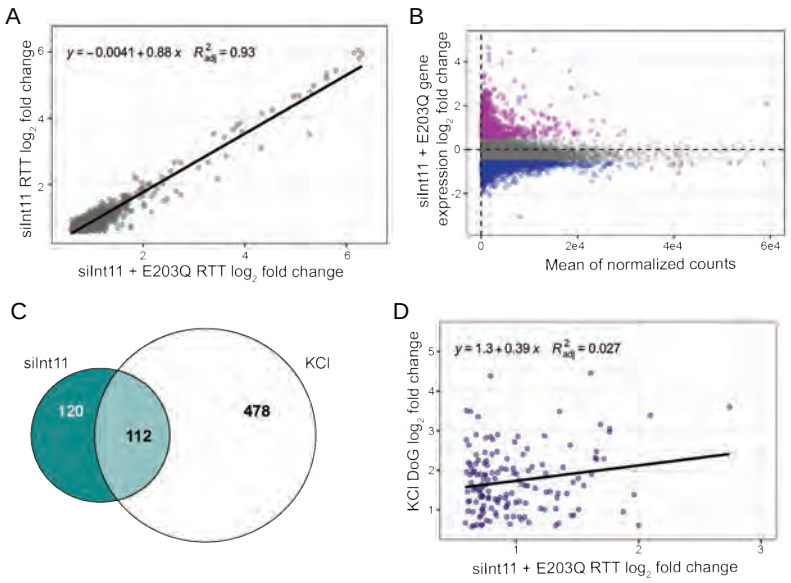

Figure 3.7: **Int11 knockdown leads to readthrough transcription** (A) Scatter plot showing correlation between readthrough transcripts detected in siInt11-transfected cells expressing no rescue and cells expressing the E203Q mutant Int11 (n=576). (B) Mean average plot showing read counts for expressed genes on the x-axis and their $\log_2$ FC after depletion of functional Int11 (E203Q sample) on the y-axis (n=14,640). Activated genes are shown in purple and repressed genes in blue. (C) Venn diagram showing the overlap between the identities of clean genes that produce readthrough transcripts in siInt11-transfected cells (teal) and DoGproducing genes in KCl-treated samples (white). (D) Scatter plot demonstrating a correlation between the $\log_2$ FC of overlapping genes in KCl-treated cells (y-axis) and siInt11+E203Q cells (x-axis) (n=148).

formation in response to two U2AF1 splicing factor mutants affects RNA stability in an cancer-relevant system.

Splicing factor mutations are common among cancers, recently emerging as drivers of myeloid malignancies. U2AF1 carries hotspot mutations in its RNA-binding motifs; however, how they affect splicing and promote cancer remain unclear. The U2AF1/U2AF2 heterodimer is critical for 3´ splice site (3´SS) definition. To specifically unmask changes in U2AF1 function in vivo, we developed a crosslinking and 'precipitation procedure that detects contacts between U2AF1 and the 3´SS AG at single-nucleotide resolution. Our data reveal that the U2AF1 S34F and Q157R mutants establish new 3´SS contacts at -3 and +1 nucleotides, respectively. These effects compromise U2AF2-RNA interactions, resulting predominantly in intron retention and exon exclusion. Integrating RNA binding, splicing, and turnover data, we predicted that U2AF1 mutations directly affect stress granule com-

ponents, which was corroborated by single-cell RNA-seq. Remarkably, U2AF1- mutant cell lines and patient-derived MDS/AML blasts displayed a heightened stress granule response, pointing to a novel role for biomolecular condensates in adaptive oncogenic strategies.

3.4.1 U2AF1 mutations enhance stress granule formation improving cell fitness under stress

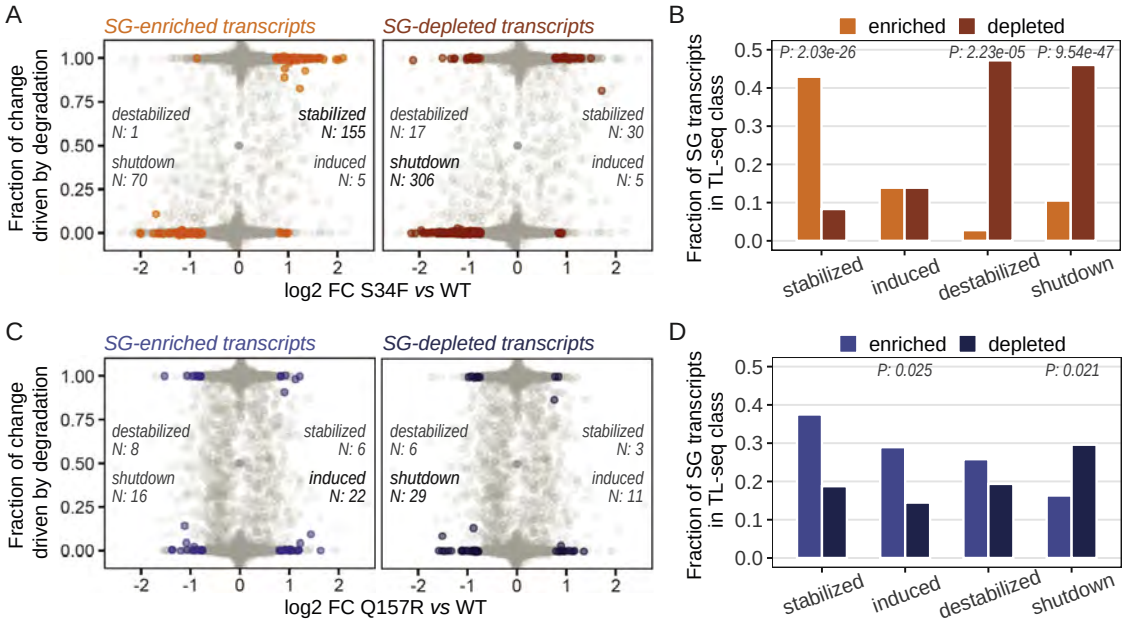

Figure 3.8: **Mutant U2AF1 cells show differential RNA dynamics related to stress granule content** (A and C) Scatter plot of gene expression changes (x axis) and relative stability/degradation contributions (y axis) in S34F (A) or Q157R (C) versus WT, measured by TL-seq (two replicates per condition). Transcripts enriched (left panel) or depleted (right panel) in stress granules are highlighted. N, number of transcripts in each TL-seq class (stabilized, induced, destabilized, and shutdown). (B and D) Fraction of SG-enriched versus depleted transcripts in each TL-seq class in S34F (B) or Q157R (D) versus WT. Differences in fractions within each class were tested with a proportion test.

To understand how mutant U2AF1-induced aberrant binding to pre-mRNA and splicing would result in enhanced SG formation at the RNA level, we analyzed RNA transcript dynamics by TimeLapse-seq (TL-seq) [97]. This technique allows us to disentangle the contributions of RNA synthesis versus stability on total RNA levels. When comparing mutant versus WT U2AF1 cells with TL-seq, transcripts were sorted into four classes: upregulated transcripts with increased stability ("stabilized") versus increased synthesis

rate ("induced"), and downregulated transcripts with reduced stability ("destabilized") versus reduced synthesis rate ("shutdown") (Figure 3.8A). Integration of TL-seq variations with the aforementioned experimental datasets characterizing SG-enriched and SG-depleted transcripts yielded significant over-representation of SG-enriched RNAs among transcripts with increased stability (S34F, Figure 3.8A and B) and synthesis (Q157R, Figure 3.8C and D). Conversely, SG-depleted transcripts were mainly in the destabilized and shutdown classes for both mutants (Figure 3.8A-D).

3.5 Dissecting regulatory function of *lincRNA-p21*

This section is adapted from:
Winkler L., Jimenez M., **Zimmer J.T.**, Williams A., Simon M.D., and Dimitrova N. (2022). Functional elements of the cis-regulatory *lincRNA-p21*. Cell Rep., 39(3). doi: 10.1016/j.celrep.2022.110687

In collaboration with the Dimitrova lab, we applied transient-transcriptome RT-qPCR to a genetic construct to probe whether transcription of the full lncRNA-p21 locus or a portion of it is required for regulation of *p21*.

The p53-induced long noncoding RNA (lncRNA) *lincRNA-p21* is proposed to act in cis to promote p53-dependent expression of the neighboring cell cycle gene, *Cdkn1a/p21*. The molecular mechanism through which the transcribed *lincRNA-p21* regulatory locus activates p21 expression remains poorly understood. To elucidate the functional elements of cis-regulation, we generated a series of genetic models that disrupt DNA regulatory elements, the transcription of *lincRNA-p21*, or the accumulation of mature *lincRNA-p21*. Unexpectedly, we determined that full-length transcription, splicing, and accumulation of *lincRNA-p21* are dispensable for the chromatin organization of the locus and for cis-regulation. Instead, we found that production of *lincRNA-p21* through conserved regions in exon 1 of *lincRNA-p21* promotes cis-activation. These findings demonstrate that the activation of nascent transcription from this lncRNA locus, but not the generation or accumulation of a mature lncRNA transcript, is necessary to enact local gene expression

control.

3.5.1 Development of genetic models to query the role of lincRNA-p21 transcription and transcript accumulation

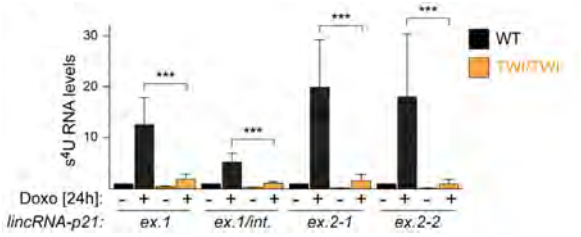

Figure 3.9: **Development of genetic tools to probe the contribution of lincRNA-p21 transcription and accumulation to p21 regulation** Transient transcriptome (TT) qRT-PCR analysis of normalized *lincRNA-p21* levels in s^4U-labeled RNA from Doxo-treated MEFs. Data are represented as mean ± SEM of indicated biological replicates; n.s. not significant, ***$p < 0.001$; paired t test.

To investigate RNA-dependent contributions of the *lincRNA-p21* locus to *p21* regulation, we generated two independent genetic models. To determine whether transcription through the *lincRNA-p21* locus is required for *p21* activation, we used CRISPR-Cas9-mediated genetic engineering to insert the 49-nucleotide synthetic polyadenylation signal (PAS) [161] into exon 1 of the endogenous *lincRNA-p21* locus in murine blastocysts (Figure 2A). In parallel, to determine the contribution of the *lincRNA-p21* RNA to *p21* regulation, we introduced the 74-nucleotide Twister (TWI) self-cleaving ribozyme [162] at the same site in exon 1 of endogenous murine *lincRNA-p21* (Figure 2A). We anticipated that while the PAS element would lead to premature termination, TWI would allow transcription of *lincRNA-p21* but lead to transcript cleavageand degradation. *LincRNA-p21*$^{\mathrm{PAS}}$ and *lincRNA-p21*$^{\mathrm{TWI}}$ founder animals were crossed to C57BL/6J mice to obtain germline transmission (Figures S2A and S2B). Next, heterozygous crosses revealed that mice harboring homozygous PAS and TWI *lincRNA-p21* alleles are viable, born at Mendelian ratios, and do not display any apparent abnormalities.

To distinguish whether the TWI insertion led to post-transcriptional degradation of *lincRNA-p21* or affected *lincRNA-p21* biogenesis, we performed transient transcriptome (TT) analysis of s^4U-labeled RNA. Primers located in exon 1 both upstream and down-

stream of the TWI insertion site revealed that TWI led to the degradation of 60%–80% of newly transcribed lincRNAp21, while primers located in exon 2 approximately 20 kb downstream of the TWI insertion pointed to a 90% reduction in transcription near the transcription termination site (Figure 3.9). These results indicated that the PAS led to efficient transcription termination, while TWI mediated co-transcriptional transcript degradation and only allowed approximately 10% of full-length transcript production.

3.6 Discussion

In principle, metabolic-labeling-based approaches are useful tools to ask targeted questions about transcription, nascent RNA, and RNA stability. In practice, this body of collaborative work exemplifies the utility of TT-TL-seq and TimeLapse-seq to uncover behavior of RNAPII and and RNA which would otherwise be difficult to study. My collaborative work with Christiane Olivero in Nadya Dimitrova's lab demonstrates that TimeLapse-seq can be used in the same manner as RNA-seq to observe alternative splicing events and that TT-TL-seq can be used to detect changes in transcriptional activity at a single locus or across specific genesets [163]. As shown by my collaboration with Lauren Winkler in Nadya Dimitrova's lab, TT-TL-seq can also be used to query transcriptional activity at different positions along the gene body, and not just across the gene as a whole [164]. Extending beyond the gene body, I worked with Nicolle Rosa-Mercado in Joan Stetiz's lab to capture nascent transcripts past the cleavage and polyadenylation site with TT-TL-seq [165]. Finally, working with Giulia Biancon in Stephanie Halene's lab, we were able to integrate TimeLapse-seq-derived kinetic information of RNA transcripts to show that differential RNA stability and synthesis are related to the formation of stress granules as a result of U2AF1 splicing factor mutants [137]. Together, these stories, and others which remain unpublished, are demonstrations that metabolic-labeling based technologies present a unique opportunity to reveal new principles of biology.

Chapter 4

STL-seq reveals pause-release and termination kinetics for promoter-proximal paused RNA polymerase II transcripts

This chapter is adapted from:

Zimmer, J.T., Rosa-Mercado, N.A., Canzio, D., Steitz, J.A., and Simon, M.D. (2021). STL-seq reveals pause-release and termination kinetics for promoter-proximal paused RNA polymerase II transcripts. Mol Cell, 81(21), 4398-4412. doi: 10.1016/j.molcel.2021.08.019.

4.1 Summary

Despite the critical regulatory function of promoter-proximal pausing, the influence of pausing kinetics on transcriptional control remains an active area of investigation. Here, we present Start-TimeLapse-seq (STL-seq), a method that captures the genome-wide kinetics of short, capped RNA turnover and reveals principles of regulation at the pause site. By measuring the rates of release into elongation and premature termination through inhibition of pause release, we determine that pause-release rates are highly variable and most

promoter-proximal paused RNA Polymerase II molecules prematurely terminate (∼80%). The preferred regulatory mechanism upon a hormonal stimulus (20-hydroxyecdysone) is to influence pause-release rather than termination rates. Transcriptional shutdown occurs concurrently with induction of promoter-proximal termination under hyperosmotic stress but paused transcripts from TATA box-containing promoters remain stable, demonstrating an important role for cis-acting DNA elements in pausing. STL-seq dissects the kinetics of pause release and termination, providing an opportunity to identify mechanisms of transcriptional regulation.

4.2 Introduction

Promoter-proximal pausing is a dynamic step in transcription that occurs at most RNA polymerase II (Pol II)-transcribed genes in metazoans and is an important point of regulatory input controlling gene expression [3, 6]. Promoter-proximal pausing is the process by which Pol II stalls 20-60 bp downstream of the transcription start site (TSS), forming a stable complex engaged on chromatin with a short nascent transcript [28–32]. To proceed through pausing and synthesize a full-length transcript, Pol II must be released into elongation, a step promoted by the kinase activity of positive transcription elongation factor b (P-TEFb) [34, 38]. Several studies, however, have demonstrated that not all paused Pol II molecules are released into elongation and some prematurely terminate through eviction from the DNA and rapid degradation of the nascent transcript [27, 43, 45, 46, 49, 60, 86, 166, 167]. Pause release and termination are in kinetic competition and determine the fate of paused Pol II; changing the rate of either can regulate gene expression.

The pause site represents a major node of regulatory input but how pause release and termination respond to regulatory signals genome-wide remains unclear [24,86,167,168]. While nascent RNA sequencing can reveal an increase in the number of Pol II molecules released into elongation, it remains ambiguous whether such observations are due to the distinct biochemical activities of increasing the pause-release rate or of decreasing the premature-termination rate. In addition, measuring premature termination is challenging because

terminated transcripts are rapidly degraded and thus difficult to directly observe. These obstacles limit our understanding of pausing regulation and prompt a more systematic analysis of pausing dynamics that includes determining the rates of release into elongation and premature termination and the regulation of each.

Conventional RNA-seq experiments do not robustly capture the short transcripts associated with paused Pol II and therefore are poorly suited to study pausing. However, an RNA sequencing-based method, Start-seq [57], specifically enriches short, capped RNA transcripts (scRNA) associated with paused Pol II such that each read represents a single engaged Pol II molecule paused at the promoter-proximal site. While Start-seq has provided important insights into steady-state levels of paused RNA Pol II [5, 168], analyzing the dynamics and turnover of these paused transcripts has proven more challenging. Thus, many questions about pausing kinetics remain unanswered, including the fraction of Pol II molecules that are released into elongation from the pause site.

Previous studies estimated paused Pol II half-lives by blocking initiation of new transcripts using triptolide (Trp), an inhibitor of TFIIH helicase activity [59, 85, 86, 169–171]. Yet kinetics upon Trp inhibition may not be reflective of kinetics of the uninhibited state, making these estimates unreliable [167, 170, 172]. Efforts to estimate half-lives of paused Pol II in a Trp-independent manner have been performed by integrating information from multiple nascent RNA-seq methods but require the assumption that premature termination occurs rarely [104, 173]. To our knowledge, these are the only two strategies applied to study paused Pol II behavior in a genome-wide and TSS-specific manner.

We sought to develop an approach that focuses on short nascent transcripts with the specificity of Start-seq and also captures the dynamics of RNA transcripts using RNA metabolic labeling and nucleotide-recoding [97]. Here we present Start-TimeLapse-seq (STL-seq), which measures steady-state kinetics of paused Pol II genome-wide without blocking transcription initiation. We apply STL-seq to fly and human cells and find very similar half-lives of paused Pol II in both systems. We demonstrate that STL-seq, when combined with P-TEFb inhibition, allows deconstruction of Pol II turnover into components of premature termination and release into elongation. We find that Pol II prematurely terminates at a similar rate at nearly all promoter-proximal pause sites. While release into

elongation is infrequent when compared to termination, it is highly variable across the genome and is the primary target of regulation in response to hormonal stimulus by 20-hydroxyecdysone treatment in *Drosophila*. On the other hand, termination is largely unaffected by the stimulus but is induced upon hyperosmotic stress. Our work provides the first direct, global measurements of pausing dynamics using non-perturbing methods and supports a model in which release into elongation regulates expression levels while premature termination functions as a quality control mechanism to ensure competent elongation.

4.3 Results

4.3.1 Short, capped transcripts can be metabolically labeled using s^4U

We sought to develop a method that directly measures the steady-state kinetics of Pol II pausing. We were inspired by Start-seq, which enriches for the short, capped RNA (scRNA) associated with the paused complex [57]. While Start-seq does not inherently capture transcript dynamics, we reasoned that if we could combine it with TimeLapse-seq, an enrichment-free method capable of capturing transcriptional dynamics, we could distinguish newly synthesized and preexisting scRNA through 4-thiouridine (s^4U) metabolic labeling. The fraction of scRNA synthesized during labeling can be revealed by chemically converting s^4U to a cytidine analogue which manifests as an apparent T-to-C mutation in sequencing data [97]. Start-TimeLapse-seq (STL-seq) therefore combines the power of metabolic labeling with the specificity of scRNA enrichment to reveal dynamics of promoter-proximal Pol II pausing.

We treated D. melanogaster S2 cells with s^4U for 5 min (Figure 4.1A), a time well validated for studying transient transcripts using other s^4U-based methods [96,97] and generally in line with previous pause duration estimates [6]. We found that s^4U-treated samples, but not controls, were enriched for TimeLapse-dependent T-to-C mutations (Figure 1B). Use of an alignment strategy that does not penalize T-to-C mismatches improved mapping of STL-seq reads, particularly shorter reads with two or more T-to-C mutations, while maintaining low background mutations (Figure 4.2A-C, Bismark, [140]).

We found that STL-seq reads provide similar profiles from s^4U-labeled and unlabeled

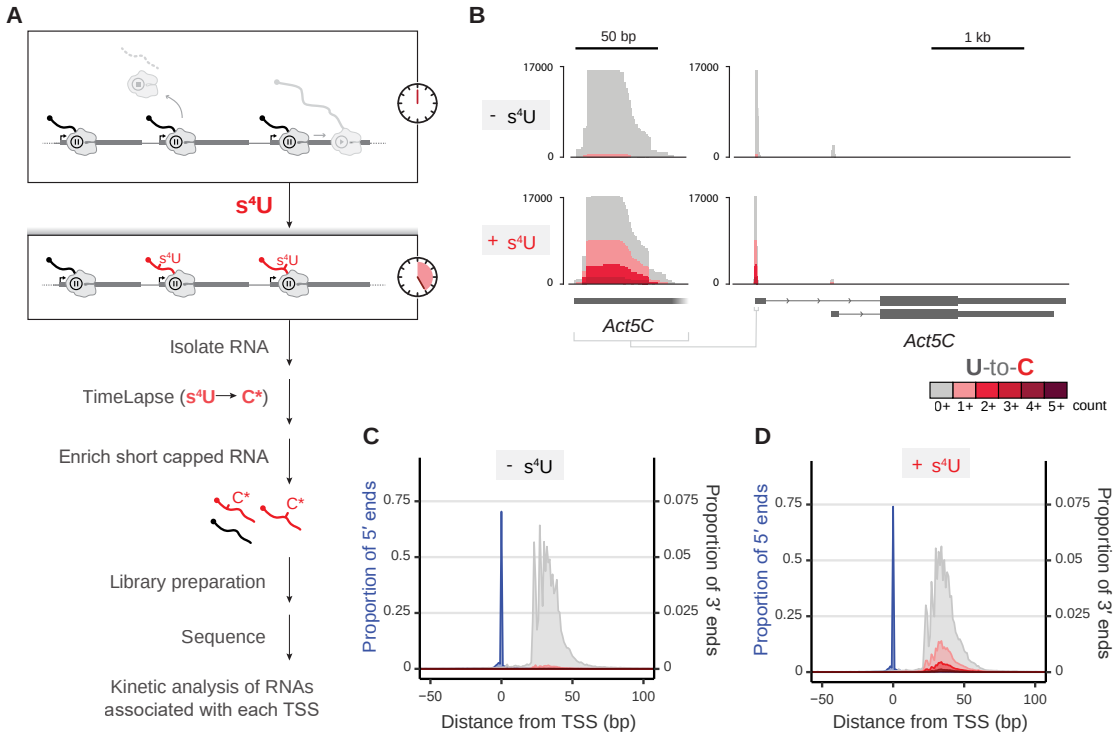

Figure 4.1: **STL-seq captures turnover dynamics of transcripts from promoter-proximal paused polymerase** (A) Scheme of STL-seq. Native RNA is metabolically labeled with s^4U for a short time before isolating RNA. TimeLapse chemistry is performed prior to enriching for short, capped RNA transcripts which are then sequenced. (B) Example STL-seq tracks demonstrating typical Start-seq coverage with elevated T-to-C TimeLapse mutations only in s^4U-labeled samples. The entire Act5C locus is shown (right) with an expanded view of the major TSS (left). (C and D) Metaplots of STL-seq 5´ and 3´ read ends identify the TSS and promoter-proximal pause site relative to the observed TSS location. The single nucleotide location of the TSS (blue, 5´ end of read) and pausing position (grey and red, 3´ end of read) are depicted separately. The 3´ ends are colored by the read's mutational content while the 5´ ends are not. Read ends at each distance from the TSS for the unlabeled (C) and labeled (D) samples are shown as a proportion of the total number of reads. The proportion of 5´ ends corresponds to the left y-axis scale and the proportion of 3´ ends corresponds to the right y-axis scale.

samples, demonstrating that the metabolic labeling and chemical treatment do not interfere with measurements of scRNAs (Figures 4.1C & 4.1D). STL-seq signals at each TSS are highly reproducible, both at the level of total reads (Pearson's r = 0.94, Figure 4.2D) and T-to-C mutation-containing reads in the labeled samples (Pearson's r = 0.91, Figure 4.2E). Correlation between total read counts of labeled and unlabeled samples is also high (Figure 4.2D). Together, these results demonstrate that s^4U can be introduced to label newly synthesized scRNA transcripts without adversely altering the scRNA levels.

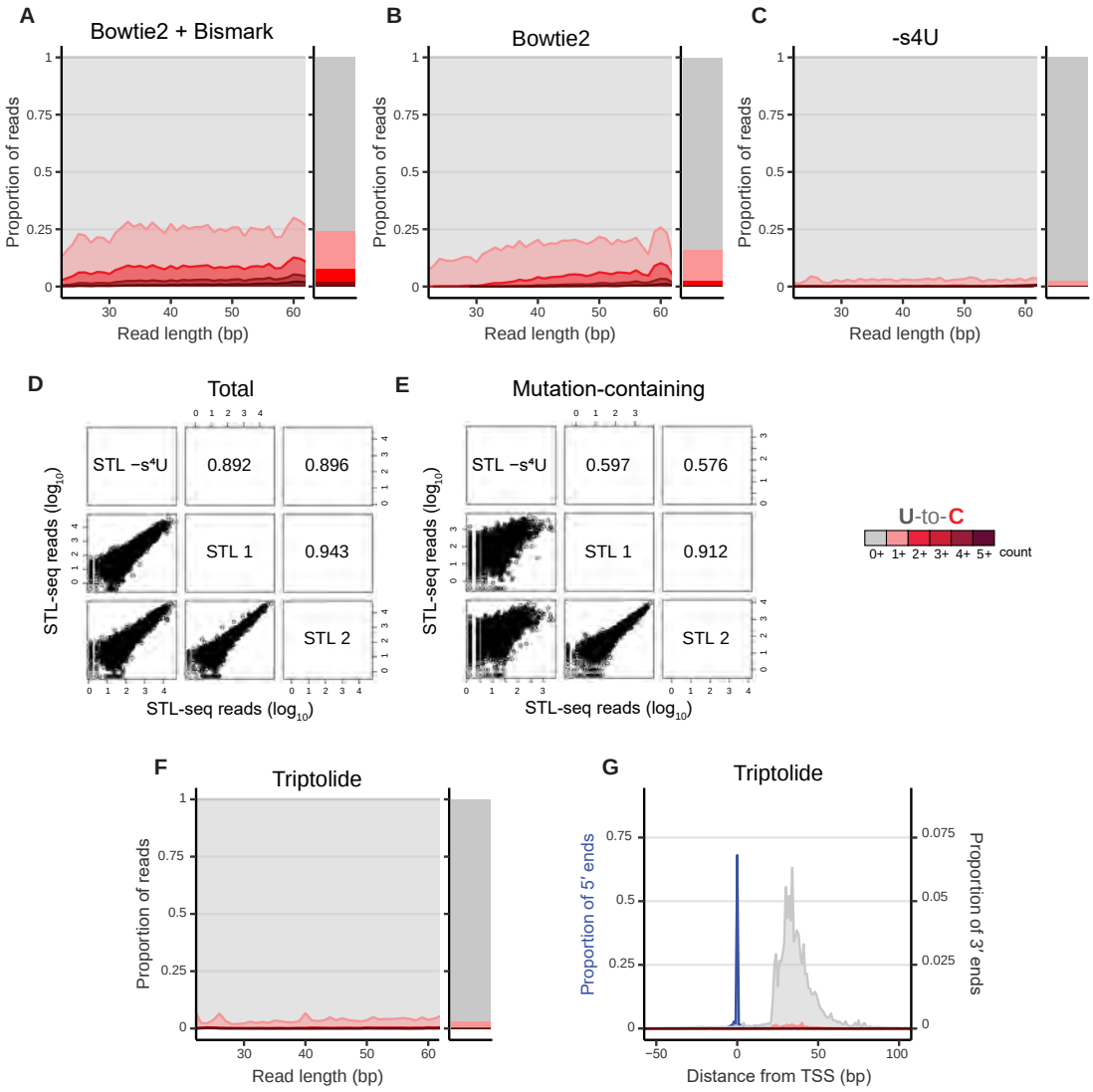

Figure 4.2: Caption on next page.

To verify that mutated reads are synthesized only by newly initiated Pol II, we inhibited initiation by treating cells with Trp prior to metabolic labeling. We did not observe accumulation of STL-seq reads containing T-to-C mutations, indicating that Trp efficiently blocks new initiation and that scRNAs recently released into elongation are not a significant source of signal (Figures 4.2F & 4.2G). Furthermore, previous work has demonstrated scRNA released from chromatin are rare, suggesting that most scRNA are degraded rapidly upon dissociation from chromatin [86]. We conclude that the mutations derived from labeled scRNAs in STL-seq are a result of the transcription of new scRNAs and therefore reflective of newly initiated Pol II which have not yet been released into elongation or terminated.

Figure 4.2: **STL-seq captures mutation information of newly synthesized short, capped transcripts** (A, B, C) The proportion of reads with varying numbers of mutations at each length. Reads are considered by absolute length, and each mutation level includes reads with the same or greater number of mutations. Reads were aligned with Bowtie 2 combined with Bismark (A, C) or Bowtie 2 alone (B) and either labeled (A, B) or unlabeled (C) with s^4U. The total proportion of reads is colored by number of mutations and shown on the right. (D) Pairs plots of total TSS read counts from STL-seq unlabeled and labeled samples. Read counts are plotted on the $\log_{10}$ scale and the Pearson correlation coefficient of each comparison is shown. (E) Pairs plots of mutation-containing TSS read counts from STL-seq unlabeled and labeled samples. Read counts are plotted on the $\log_{10}$ scale and the Pearson correlation coefficient of each comparison is shown. (F) Same as A-C but labeled under triptolide inhibition and aligned with Bismark. (G) Metaplot of STL-seq read ends relative to the TSS location under triptolide inhibition. The single nucleotide location of the TSS (blue, 5´ end of read) and pausing position (grey and red, 3´ end of read) are depicted separately. The 3´ ends are colored by the read's mutational content while the 5´ ends are not. Read ends at each distance from the TSS for the labeled samples are shown as a proportion of the total number of reads. The proportion of 5´ ends corresponds to the left y-axis scale and the proportion of 3´ ends corresponds to the right y-axis scale.

4.3.2 STL-seq data can be used to quantify scRNA turnover accurately and robustly

Data from our single timepoint STL-seq experiment suggested diverse kinetics of scRNA turnover at different TSSs. To further explore scRNA dynamics, we performed an independent STL-seq time series (1.5, 3, 5, 7.5, 10, and 120 min of s^4U labeling) such that nearly all scRNAs across all TSSs were predicted to turn over within the longest labeling period. We observed the expected time-dependent accumulation of T-to-C mutations and found that the rate of accumulation varied at different TSSs, illustrating the capability of STL-seq to reveal a range of pausing kinetics across the genome (Figures 4.3A & 4.4A).

To use the mutational content of STL-seq reads to study scRNA dynamics, we developed statistical methods to robustly quantify turnover. Determining the fraction of newly made transcripts (θ) allows estimation of turnover using a first-order observed rate constant ($\hat{k}_{\text{obs}}$, min^{-1}) for transcripts initiated from each TSS. It is important to use a statistical model to infer the fraction of scRNA that are newly made because some newly made reads will not contain any mutations (Figure 4.4B). The lack of mutations in some newly synthesized reads could lead to a global underestimation of scRNA turnover rates if only mutation-containing reads were considered newly made. Instead we use a binomial mixture model. We define θ

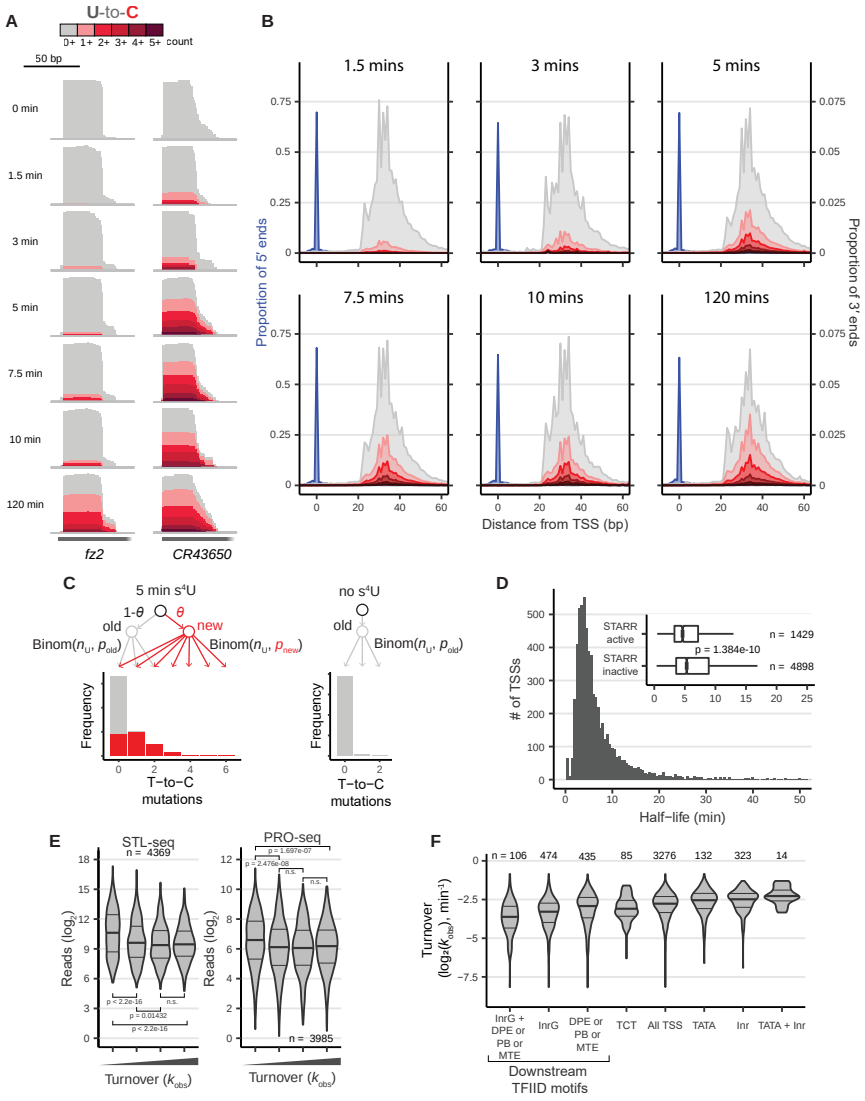

Figure 4.3: Caption on next page.

in relation to $\hat{k}_{\text{obs}}$ with an exponential model such that

$$\theta = 1 - e^{(-\hat{k}_{obs}t)} \tag{4.1}$$

where t is the labeling time. Similar to our previous analyses [97], we estimated the fraction of newly made transcripts using a binomial model of the number of mutations observed (tc) and uridines (n_u) present in each read (Figure 2C), thereby accounting for variable uridine content across TSSs. The model also depends on the TSS-specific background (p_o) mutation rate, which is determined from the unlabeled controls, and the TSS-specific

Figure 4.3: **Estimation of scRNA transcript turnover from STL-seq** (A) STL-seq tracks of the *fz2* and *CR43650* TSSs labeled with s^4U for the indicated times. Tracks are autoscaled to show relative proportion of mutated reads. (B) Metaplots of STL-seq 5´ (blue) and 3´ (grey and red) read ends relative to the TSS labeled with s^4U for the indicated times with similar presentation to Figures 4.1C & 4.1D. (C) The fraction of new scRNA (θ) is estimated with a mixed binomial model. The model estimates the background mutation rate (p_{old}) with the unlabeled control and uses the number of U's in each read (n_{U}) and the distribution of T-to-C mutations in the labeled samples to estimate the TimeLapse-dependent mutation rate (p_{new}). In this simulated example, each read derives from a TSS with a 5 min half-life and average read length of 35 nt with a uridine every 4 nts. Newly synthesized transcripts (red) are synthesized with (p_{new}) = 10% and preexisting reads (grey) are synthesized with (p_{old}) = 0.25%. See STAR methods for more details. (D) Histogram of scRNA half-life estimates made with STL-seq from S2 cells. The inset boxplot separates scRNA half-lives into those aligned to regions with and without STARR-seq enhancer activity. Significance was assessed by a two-sided Wilcoxon rank sum test. (E) The distribution of either STL-seq reads (left) or promoter-proximal PRO-seq reads (right, [45]) grouped into even quartiles by observed scRNA turnover. Significance was assessed by a two-sided Wilcoxon rank sum test. (F) Distribution of the total observed turnover rate constant at promoters grouped by motif content. All motifs are known TFIID binding elements except the polypyrimidine initiator (TCT) and the degenerate initiator (Inr). The pause button (PB), downstream promoter element (DPE), and motif ten element (MTE) were grouped together such that promoters may have one or a combination of these within 50 bp downstream.

TimeLapse mutation rate (p_n). The probability mass function is

$$f\left(tc|n_u, p_n, p_o\right) = \theta \text{BinomialLogit}\left(tc|n_u, p_n\right) + \left(1-\theta\right) \text{BinomialLogit}\left(tc|n_u, p_o\right) \tag{4.2}$$

which is parameterized on the logit scale to avoid hard upper and lower bounds.

We used a Bayesian hierarchical modeling approach to estimate these parameters using RStan software (Version 2.19.3, [174]) that implements no-U-turn Markov Chain Monte Carlo (MCMC) sampling. We defined hierarchical parameters for the global background ($\bar{p}_o$) and TimeLapse ($\bar{p}_n$) mutation rates to account for local variability while allowing for information sharing between TSSs to benefit those with lower coverage. The local mutation rates for the s^{th} TSS were defined with a non-centered parameterization as follows

$$p_{o[s]} = \bar{p}_o + \sigma_o z_{o[s]} \tag{4.3}$$

$$p_{n[s]} = \bar{p}_n + \sigma_n z_{n[s]} \tag{4.4}$$

where σ_n and σ_o are the standard deviations of the global TimeLapse and background mutations rates, respectively, and z_n and z_o are TSS-specific z-scores for the TimeLapse and background mutations rates, respectively. For the complete parameterization and prior definition, see methods. Simulations of scRNA with variable kinetics and uridine content supported the feasibility of using mutational content from STL-seq data to infer scRNA half-lives with this model (Figures 4.4B & 4.4C, see methods).

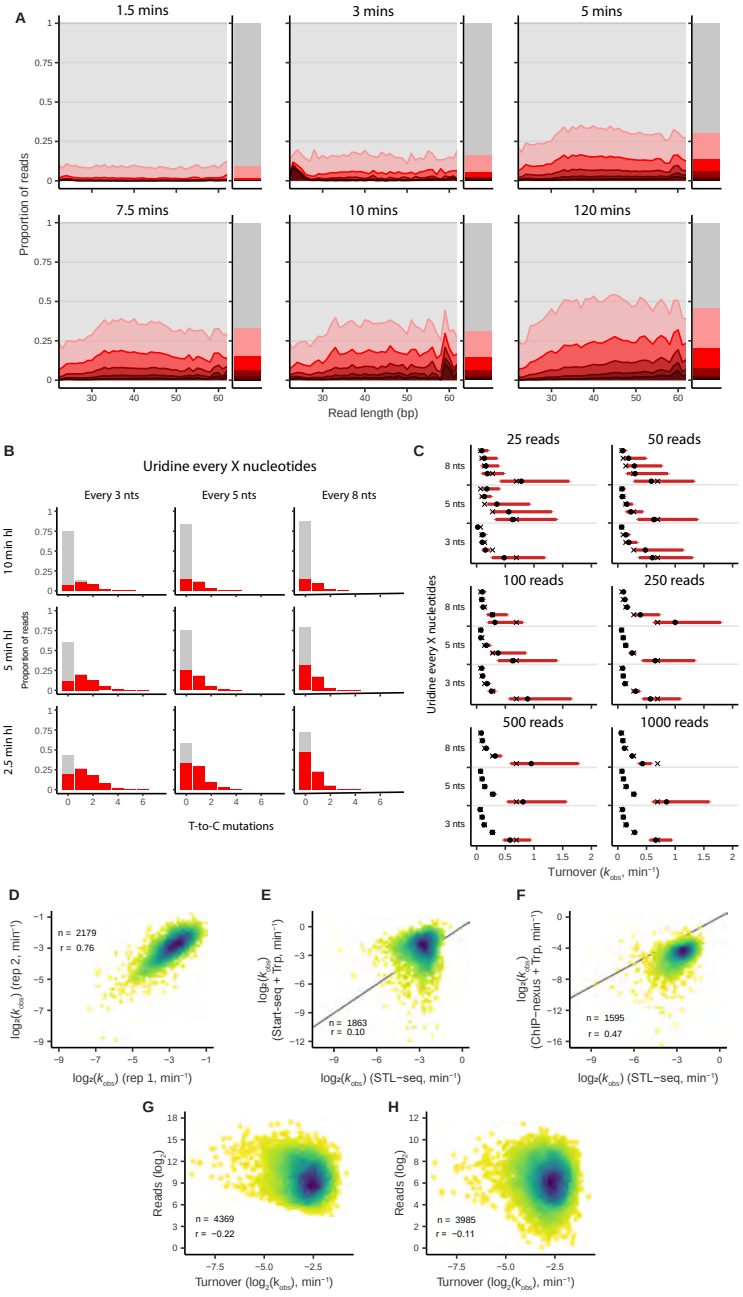

Figure 4.4: Caption on next page.

Figure 4.4: **Binomial modeling of STL-seq mutation data** (A) The proportion of reads with varying numbers of mutations at each length when labeled for varying times with s^4U. Reads are considered by absolute length and each mutation level includes reads with the same or greater number of mutations. The total proportion of reads is colored by number of mutations and shown on the right of the plot for each label time. (B) Simulated STL-seq mutational content of TSSs with various pausing half-lives (hl) and uridine content. Reads are colored by whether they are considered new (red) and synthesized during the label time or old (grey) and synthesized prior to labeling. See methods for details of simulated data. (C) Using the model to estimate the rate constant of simulated data generated as in B in addition to varying degrees of coverage. The true rate constant is indicated with a cross. The median value of the posterior estimate is indicated with a solid circle and the 80% credible interval is indicated by red bars. See methods for details of simulated data. (D) Correlation plot comparing estimates from single STL-seq replicates. Plotted points represent the median value of the posterior estimate. The density of plotted points is indicated by color (blue, high; yellow, low). The Pearson correlation coefficient is shown. (E) Correlation plot comparing $\hat{k}_{\text{obs}}$ estimates made with STL-seq and previously published Start-seq data under Trp inhibition. The median value of the STL-seq posterior estimate is plotted and the 1:1 line is shown. The density of plotted points is indicated as in D. The Pearson correlation coefficient is shown. (F) Correlation plot comparing $\hat{k}_{\text{obs}}$ estimates made with STL-seq and previously published ChIP-nexus data under Trp inhibition. The median value of the STL-seq posterior estimate is plotted and the 1:1 line is shown. The density of plotted points is indicated as in D. The Pearson correlation coefficient is shown. (G) Correlation plot comparing STL-seq $\hat{k}_{\text{obs}}$ estimates and STL-seq read counts at each TSS. The median value of the STL-seq posterior estimate is plotted. The density of plotted points is indicated as in D. The Pearson correlation coefficient is shown. (H) As in G but comparing STL-seq $\hat{k}_{\text{obs}}$ estimates and PRO-seq read counts in the promoter-proximal region of each TSS.

4.3.3 STL-seq reveals high turnover of scRNAs at most TSSs

We applied the binomial mixture model to our genome-wide STL-seq data and found that median $\hat{k}_{\text{obs}}$ estimates of high confidence TSSs (low uncertainty in parameter estimates, see STAR Methods) agree well between replicates (Figure S2D). By combining both replicates to estimate a single $\hat{k}_{\text{obs}}$ for each TSS, we find the median half-life of scRNA to be about five minutes with half-lives spanning from minutes to tens of minutes (inner 90% range spanning 2.1 to 24 min, Figure 2D). In agreement with previous findings [5], scRNA initiated from regions with enhancer activity as measured by STARR-seq turn over with half-lives faster than those initiated from regions without enhancer activity. However, we do not find evidence of scRNA with extremely long average half-lives (one hour or longer) that were observed in previous Trp inhibition experiments [5,59,85,175]. More generally, STL-seq $\hat{k}_{\text{obs}}$

estimates show moderate agreement with $\hat{k}_{\text{obs}}$ estimates made with previously published Trp inhibition data (Figures S2E & S2F); however, the slower estimates made with Trp inhibition data [59] buttress previous concerns that Trp may stabilize paused Pol II. These results demonstrate that the overall rate of paused scRNA turnover is fast regardless of the TSS type and led us to investigate what TSS and promoter features associate with variability in scRNA turnover.

We asked if the level of Pol II occupancy at the pause site influences pausing kinetics. As Pol II spends little time loaded on the promoter in the preinitiation complex (PIC), promoter-proximal pausing is a major rate-limiting step during early transcription [6, 176]. Accordingly, pause sites should always be close to fully occupied so long as the promoter is in an active state. We used STL-seq read counts from high confidence TSSs (see methods) as an indicator of Pol II occupancy and found that slow turnover is not strongly correlated with higher occupancy (Figures 4.3E & 4.4G). To further probe this relationship, we re-analyzed available PRO-seq data [45] and counted reads in the promoter-proximal region. This analysis showed a similar relationship where slow turnover is weakly associated with higher read counts (Figures 4.3E & 4.4H). Thus, STL-seq data provide further evidence that pausing is a principal rate-limiting step prior to elongation.

TFIID, a bridge-like PIC component, is sufficient to induce pausing *in vitro* [37]. Cis-acting DNA elements, especially those related to TFIID binding, have been shown to influence Pol II pausing [59, 177]. TFIID contacts the TATA box through its TATA-binding protein (TBP) subunit and makes additional DNA contacts downstream of the promoter at the initiator motif (InrG), downstream promoter element (DPE), motif ten element (MTE), and pause button (PB). Presence of these downstream motifs tends to extend pausing half-lives while the TATA box tends to shorten them [59,177]. Our data recapitulate these results at high confidence TSSs and demonstrate the destabilizing effect of the degenerate, G-less initiator motif (Inr, [178]) (Figure 4.3F). The polypyrimidine initiator (TCT) motif, which is similar to InrG but does not bind TFIID, appears to be associated with similar kinetics as InrG. Our robust and reproducible measurements of $\hat{k}_{\text{obs}}$ support previous observations and provide the foundation to further examine the principles underlying promoter-proximal pausing.

4.3.4 Termination is generally faster but less variable than release into elongation

Next, we sought to determine the proportion of paused Pol II molecules that are prematurely terminated at each TSS prior to entering productive elongation. Previous work established that premature termination is an important fate of the paused complex, but the relative contributions of pause release and premature termination were not determined for TSSs genome-wide [27, 45, 49, 60, 86, 166, 167, 169]. We used flavopiridol (FP) treatment (prior to s^4U labeling) to inhibit release into elongation and allow for measurement of premature termination (Figure 4.5A). FP increases STL-seq reads at the majority of TSSs, except those with the most scRNA reads, perhaps because they are already fully saturated with paused Pol II (Figure 4.6A). This increase in STL-seq reads indicates a stabilization of the paused complex due to inhibition of release into elongation by FP.

We further developed the model described above to assume that the observed turnover rate constant ($\hat{k}_{\text{obs}}$) at steady state is the sum of termination and pause-release rate constants (Figure 4.5A, see methods). Previous studies demonstrated that FP does not perturb premature termination [49, 85]. Therefore, under FP inhibition, the observed turnover is attributed only to premature termination ($\hat{k}_{\text{term}}$). We calculate the pause-release rate constant ($\hat{k}_{\text{rel}}$) as the difference between $\hat{k}_{\text{obs}}$ and $\hat{k}_{\text{term}}$. We find that pause-release constants on average are slow but vary widely (median 0.027 min^{-1}; inner 90% range 0.0015-0.31 min^{-1}), while termination constants are fast and more tightly distributed (median 0.11 min^{-1}; inner 90% range 0.027-0.23 min^{-1}) (Figure 4.5B).

Because polymerases must be released from the pause site to transcribe the rest of the gene body, we expected that transcriptional activity in the gene body should be a function of how quickly scRNA are released into elongation. Transient-Transcriptome-TimeLapse-seq (TT-TL-seq) enriches for nascent RNA and is therefore a good measure of transcriptional activity. We performed TT-TL-seq and compared coverage to STL-seq $\hat{k}$ estimates at TSSs where we could make high confidence estimates of $\hat{k}_{\text{rel}}$ (n = 2865; see methods). As expected, we found that $\hat{k}_{\text{rel}}$ is the best predictor of TT-TL-seq signal when compared to $\hat{k}_{\text{term}}$ and $\hat{k}_{\text{obs}}$ (Figures 4.5C & 4.6B-E). We also assessed this relationship using an orthogonal measure

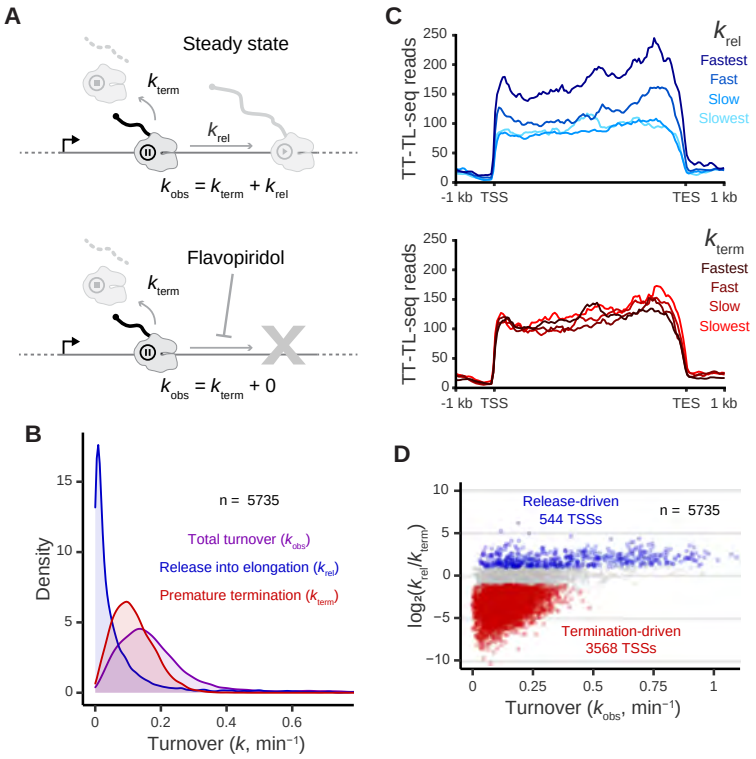

Figure 4.5: **Termination is fast while release into elongation explains variability of Pol II turnover at pause sites** (A) Representation of pausing kinetics under steady-state and flavopiridol-inhibited conditions. The steady-state observed turnover ($\hat{k}_{\text{obs}}$) is the sum of the rates of release into elongation ($\hat{k}_{\text{rel}}$) and premature termination ($\hat{k}_{\text{term}}$). Upon flavopiridol inhibition, observed turnover is caused only by premature termination. (B) The distribution of the first-order rate constants for total turnover, release into elongation, and premature termination. (C) Metaplots of TT-TL-seq signal grouped into even quartiles by release and termination of the respective high confidence TSS (n = 2422 genes). Coverage is determined over 50 nt bins. (D) Total observed rate constant plotted versus the $\log_2$ ratio of the release rate and termination rate. Points are colored if the 80% credible interval of the $\log_2$ ratio does not overlap zero and the median value is greater than 1 (blue) or less than -1 (red).

of elongating Pol II activity (gene-body PRO-seq reads; [45]; Figure 4.6F), which further supported our conclusion that STL-seq pause-release rates are more tightly linked to gene body transcription.

To provide additional validation of our approach to estimate $\hat{k}_{\text{rel}}$ and $\hat{k}_{\text{term}}$ we reasoned that genes with significant levels of paused polymerase at the TSS but very low transcriptional activity in their gene body must have low $\hat{k}_{\text{rel}}$. Therefore, we expect $\hat{k}_{\text{term}} \approx \hat{k}_{\text{obs}}$ at these TSSs and those rate constants should not be perturbed by FP. We used TT-TL-seq data to identify genes with the lowest 10% of transcriptional activity where we expect

$\hat{k}_{\text{term}} \approx \hat{k}_{\text{obs}}$. We then identified confident TSSs at these genes and found $\hat{k}_{\text{term}}$ and $\hat{k}_{\text{obs}}$ using a model designed to estimate differences (see methods). We found that turnover under FP inhibition and in the uninhibited state were not substantially different at these TSSs, further supporting our assumption that FP does not alter $\hat{k}_{\text{term}}$ (Figure 4.6G).

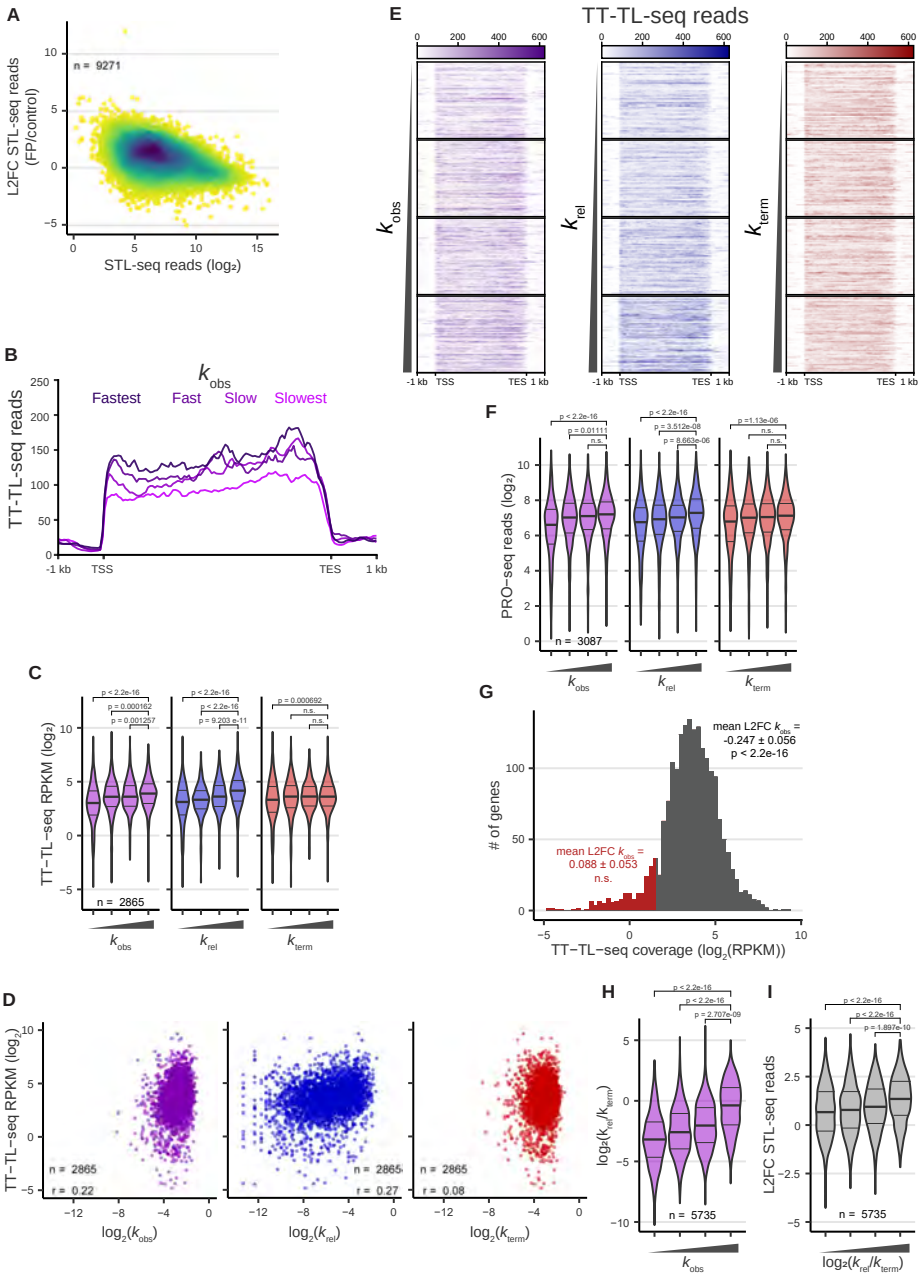

Figure 4.6: Caption on next page.

To examine the relative amount of Pol II terminated or released into elongation at each TSS, we took the $\log_2$-transformed ratio of the rates of pause release versus termination. We

Figure 4.6: **Release from the pause site predicts downstream transcriptional activity** (A) The change in normalized STL-seq reads at all TSS upon flavopiridol inhibition plotted versus the average normalized STL-seq read count in untreated controls. The density of plotted points is indicated by color (blue, high; yellow, low). (B) Metaplots of TT-TL-seq signal grouped into even quartiles by total observed turnover of the respective TSS. Coverage is determined over 50 nt bins. (C) The distribution of TT-TL-seq coverage over the entire gene body by RPKM. Genes are grouped into even quartiles by total observed turnover, release, or termination of the respective TSS. Significance was assessed by a two-sided Wilcoxon rank sum test. (D) Correlation plot comparing $\hat{k}_{\text{obs}}$, $\hat{k}_{\text{rel}}$, or $\hat{k}_{\text{term}}$ estimates made with STL-seq and TT-TL-seq coverage in RPKM. The median value of the STL-seq posterior estimate is plotted. The Pearson correlation coefficient is shown. (E) Heat maps of TT-TL-seq data grouped as in C and ordered by the indicated rate constant. (F) The distribution of PRO-seq ([45] reads over the region from 0.5 kb to 1.5 kb downstream of the TSS. Genes are grouped into even quartiles by total observed turnover, release, or termination of the respective TSS. Significance was assessed by a two-sided Wilcoxon rank sum test. (G) Histogram of RPKM values for genes in TT-TL-seq data. The red highlighted region represents the 10% of genes with the least coverage. The mean $\log_2$ fold change of $\hat{k}_{\text{obs}}$ upon FP treatment and standard error are shown for TSSs of the bottom 10% of genes (red) or the entire genome (black). Statistical difference from zero was assessed with a one-sample Wilcoxon signed rank test. (H) The distribution of the $\log_2$ ratio of the release rate to termination rate at all TSSs grouped into even quartiles by the total observed turnover. Significance was assessed by a two-sided Wilcoxon rank sum test. (I) The distribution of the $\log_2$ fold change in normalized STL-seq read counts upon flavopiridol inhibition at all TSSs grouped into even quartiles by the $\log_2$ ratio of the release rate to termination rate. Significance was assessed by a two-sided Wilcoxon rank sum test.

found that termination is faster than pause release at most TSSs (62%) while the converse is uncommon (9%) (Figures 4.5D & 4.6H). However, TSSs with the fastest total turnover ($\hat{k}_{\text{obs}} > 0.5\text{min}^{-1}$) are more likely to release scRNA into productive elongation than terminate the transcript. We compared the change in scRNA read counts upon FP treatment to the ratio of $\hat{k}_{\text{rel}}$ to $\hat{k}_{\text{term}}$ and found that more frequent pause release is associated with the accumulation of more reads, as would be expected (Figure 4.6I). Taken together, these results reveal that on average, termination is about four times faster than pause release and therefore ∼80% of total turnover, while pause release is typically slower but has a larger dynamic range.

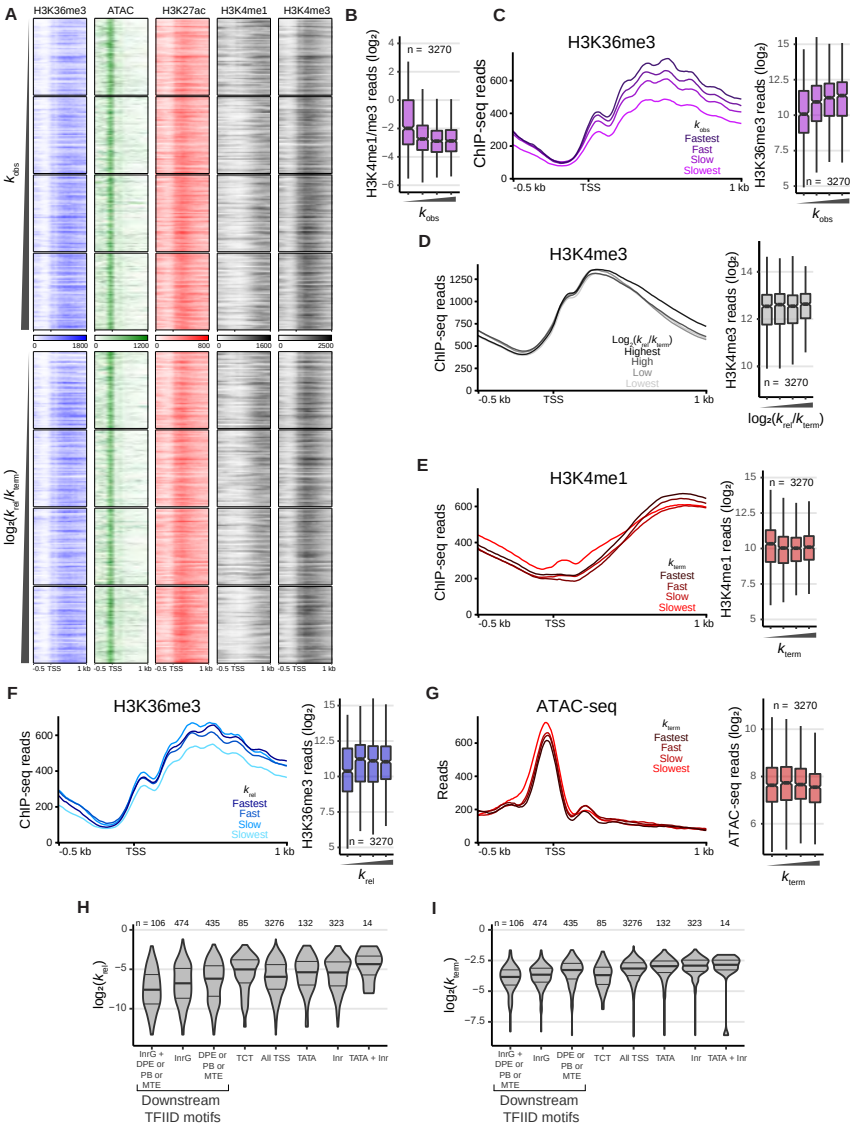

Figure 4.7: Caption on next page.

4.3.5 Certain histone tail modifications are associated with less permissive pausing dynamics

The local chromatin environment around promoters is important for the regulation and maintenance of transcriptional activity. To examine how the local chromatin landscape around pause sites is related to scRNA dynamics revealed by STL-seq, we focused our analysis on high confidence promoter TSSs. We found that pause sites with the least stable scRNA are modestly enriched for chromatin marks typically associated with active promoters. For example, low monomethylation and high trimethylation of histone 3 lysine

Figure 4.7: **Chromatin structure defines unique profiles of pausing kinetics** (A) Heat maps of ATAC-seq [45] and ChIP-seq H3K27ac, H3K4me1, H3K4me3 [5], and H3K36me3 [179] around promoters grouped into even quartiles and ordered by total turnover or $\log_2$ ratio of the release rate to the termination rate of the respective TSS. Heatmaps are centered on the STL-seq TSS with a window of 0.5 kb upstream and 1 kb downstream. (B) The $\log_2$ ratio of H3K4me1 to H3K4me3 ChIP-seq reads in the window 0.5 kb upstream and 1 kb downstream of the promoter TSS, grouped into even quartiles by total turnover (n = 3270 promoters). Significance was assessed by a two-sided Wilcoxon rank sum test. (C) Metaplot (left) and read count distribution (right) of H3K36me3 ChIP-seq data around promoters with a window of 0.5 kb upstream and 1 kb downstream of the TSS, grouped into even quartiles by total turnover. (D) Metaplot (left) and read count distribution (right) of H3K4me3 ChIP-seq data around promoters with a window of 0.5 kb upstream and 1 kb downstream of the TSS, grouped into even quartiles by the $\log_2$ ratio of the release rate to the termination rate. (E) Metaplot (left) and read count distribution (right) of H3K4me1 ChIP-seq data around promoters with a window of 0.5 kb upstream and 1 kb downstream of the TSS, grouped into even quartiles by termination. (F) Same as C but grouped into even quartiles by pause release. (G) Metaplot (left) and read count distribution (right) of ATAC-seq data around promoters with a window of 0.5 kb upstream and 1 kb downstream of the TSS grouped into even quartiles by termination. (H and I) The distribution of the release rate (H) or termination rate constants (I) at promoters grouped by motif content. The pause button (PB), downstream promoter element (DPE), and motif ten element (MTE) were grouped together such that promoters may have one or a combination of these in the downstream region.

4 (H3K4me1/me3) indicate high promoter activity [180]. We find that slower turnover of scRNA is related to a larger ratio of mono-to-trimethylation at H3K4 (Figures 4.7A & 4.7B). Additionally, trimethylated histone 3 lysine 36 (H3K36me3), whose deposition is signaled by active elongation (reviewed in [9]), is enriched immediately downstream of fast turnover sites (Figures 4.7A & 4.7C).

We found distinct H3K4 methylation profiles to be associated with promoters depending on their relative rates of pause release and termination (Figures 4.8A & 4.7A). H3K4me3 promotes PIC assembly and transcription initiation [22]. We observe that relative rates of pause release and termination are not significantly related to H3K4me3 levels (Figure 4.7D), supporting the notion that H3K4me3 only behaves as a signal to activate promoters. On the other hand, H3K4me1 is depleted at promoters that are the most likely to release Pol II into elongation (Figure 4.8B). By examining the relationship of pause release and termination with H3K4me1 separately, we find that the modification is more strongly related to pause release (Figures 4.8C & 4.7E). This negative correlation suggests H3K4me1 could play a

role in suppressing the release of Pol II into elongation, an idea consistent with the fact that H3K4me1 is enriched at enhancers where productive elongation is rare (reviewed in [181]). In further agreement, scRNA initiated from promoters with low H3K4me1/me3 ratios are more likely to be released into elongation and behave the least similarly to scRNA initiated from enhancer TSSs (eTSSs, Figure 4.8D).

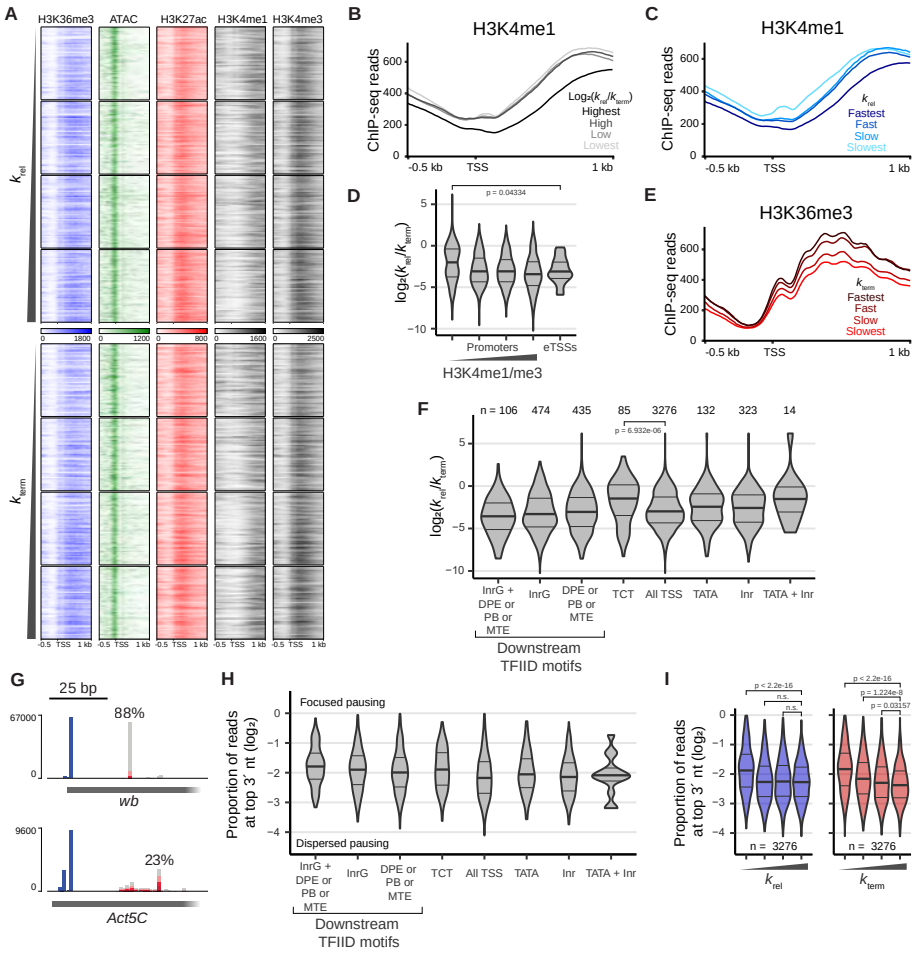

Figure 4.8: Caption on next page.

We examined the relationship between scRNA dynamics and H3K36me3 levels and found the mark to exhibit a significant positive relationship with premature termination (Figure 4.8E). We also find that H3K36me3 is depleted from promoter-proximal regions at the slowest releasing sites but has similar levels at the remaining promoters (Figure 4.7F). As H3K36me3 suppresses cryptic initiation from weak downstream promoters [182], our data raise the intriguing possibility that this suppression is supported by promoting premature

Figure 4.8: Weak promoter architecture leads to rapid termination of paused Pol II (A) Heat maps of ATAC-seq [45] and H3K27ac, H3K4me1, H3K4me3 [5], and H3K36me3 [179] ChIP-seq around promoters grouped into even quartiles and order by pause release and termination at the respective TSS. Heatmaps are centered on the STL-seq TSS with a window of 0.5 kb upstream and 1 kb downstream. (B) Metaplot of H3K4me1 ChIP-seq data around promoters with a window of 0.5 kb upstream and 1 kb downstream of the TSS grouped into even quartiles by the $\log_2$ ratio of the pause-release rate to the termination rate (n = 3270 promoters). (C) Metaplot of H3K4me1 ChIP-seq data around promoters with a window of 0.5 kb upstream and 1 kb downstream of the TSS grouped into even quartiles by pause release (n = 3270 promoters). (D) Distribution of the $\log_2$ ratio of the release rate to the termination rate at promoters grouped by whether TSSs are of high confidence promoters or enhancers (eTSSs, n = 21). Promoters are further grouped into even quartiles by H3K4me1 enrichment determined by ChIP-seq. Significance was assessed by a two-sided Wilcoxon rank sum test. (E) Metaplot of H3K36me3 ChIP-seq data around promoters with a window of 0.5 kb upstream and 1 kb downstream of the TSS grouped into even quartiles by termination (n = 3270 promoters). (F) Distribution of the $\log_2$ ratio of the release rate to the termination rate at promoters grouped by motif content. The pause button (PB), downstream promoter element (DPE), and motif ten element (MTE) were grouped together such that promoters may have one or a combination of these in the downstream region. Significance was assessed by a two-sided Wilcoxon rank sum test. (G) Example STL-seq tracks where the single nucleotide location of the TSS (blue, 5´ end of read) and pausing position (grey and red, 3´ end of read) are depicted separately. The 3´ ends are colored by the read's mutational content while the 5´ ends are not. The maximum percent of reads with the same 3´ end is shown above the read position. (H) Distribution of the proportion of reads with 3´ ends located at the most frequent pause position. At each promoter, the most common position of the 3´ read end was identified, and the proportion of reads at this position was determined. Promoters are separated by promoter motif as in F. (I) Distribution of the proportion of reads with 3´ ends located at the most frequent pause position as in H but promoters are grouped into even quartiles by pause release (left) or termination (right). Significance was assessed by a two-sided Wilcoxon rank sum test.

termination. Fast termination is also found at less accessible promoters as measured by ATAC-seq, bolstering the association between premature termination and weak promoters (Figure 4.7G).

In summary, H3K4me1 and H3K36me3 are enriched at pause sites that are less likely to release Pol II into elongation. It is possible that histone tail modifications locally repress gene expression by influencing dynamics at promoter-proximal pause sites. Our data support a model in which H3K4me1 blocks release into elongation, while H3K36me3 recruits factors that promote premature termination.

4.3.6 Promoter and pause-site architecture are associated with stability of the paused complex

Together with previous work [37,59,177], our findings demonstrate the importance of TFIID binding elements in pausing and led us to examine how promoter motif content relates to both pause-release and termination kinetics. We find no evidence that downstream TFIID binding motifs substantially alter the proportion of Pol II which is terminated or released into elongation (Figure 4.8F). As would therefore be expected from our $\hat{k}_{\text{obs}}$ estimates, downstream TFIID motifs are associated with slow rates for both $\hat{k}_{\text{rel}}$ and $\hat{k}_{\text{term}}$ (Figures 4.7H & 4.7I). The TCT motif marks TSS with a high proportion of Pol II that is released into elongation. The TCT motif is primarily found at promoters of ribosomal proteins which are typically among the most highly expressed genes [183]. Therefore, it is unsurprising to observe elevated pause-release rates at TSSs with the TCT motif (Figure 4.7H). Further investigation of these TSSs will likely provide a deeper understanding of how cis-acting DNA can promote release into elongation.

Strong downstream TFIID binding stabilizes the paused Pol II complex and coordinates the focused pausing of Pol II molecules [58,59]. We reasoned that if the total scRNA turnover and pause site dispersion is influenced by the organization of the PIC, relative rates of pause release and termination could also be differentially affected. As a measure of focused versus dispersed pausing, we determined the proportion of scRNA with the identical and most common 3´ position at each TSS (Figure 4.8G). Consistent with previous findings [58], focused pause sites are associated with downstream TFIID binding motifs (Figure 4.8H). When comparing pause site dispersion with scRNA kinetics, the fastest terminating pause sites are associated with less focused pausing profiles while dispersion had little bearing on the pause-release rate (Figure 4.8I). These data demonstrate that cis-regulatory DNA elements in the promoter-proximal region mark TSSs with distinct kinetic and physical pausing profiles. In addition, premature termination is more likely to occur at TSSs that do not reproducibly pause Pol II in the same position.

4.3.7 Enhanced release into elongation is the major response to hormone stimulus

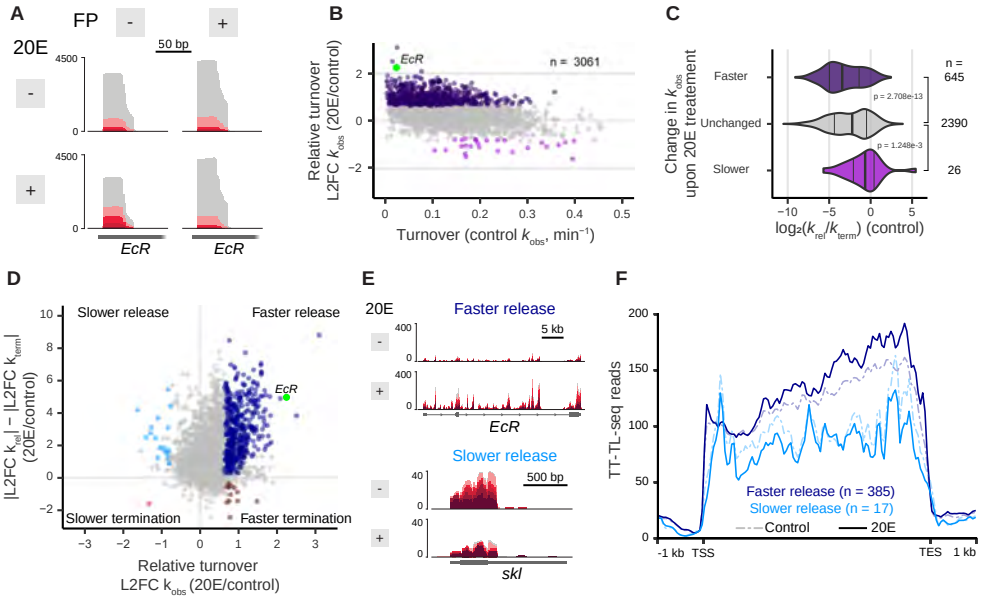

Figure 4.9: **Hormonal stimulus by 20E preferably regulates release into elongation** (A) STL-seq tracks of the EcR gene TSS when treated with or without 20E and with or without flavopiridol inhibition. (B) Total observed rate constants of high confidence promoters plotted versus the $\log_2$ fold change when S2 cells are stimulated with 20E for thirty minutes. Points are colored if the 80% credible interval is greater than $\log_2(1.5)$ (dark purple) or less than -$\log_2(1.5)$ (light purple). The EcR TSS shown in A is highlighted in green. (C) Distribution of the $\log_2$ ratio of the pause-release rate versus termination rate at promoters grouped by the change in the total turnover as determined in B. Significance was assessed by a two-sided Wilcoxon rank sum test. (D) The $\log_2$ fold change of promoters plotted versus the difference of the magnitudes of the $\log_2$ fold change of the release and termination constants upon 20E stimulus. Points are colored if the 80% credible interval of L2FC $\hat{k}_{\text{obs}}$ is entirely greater than $\log_2(1.5)$ or less than -$\log_2(1.5)$ and if the 80% credible interval of the difference in magnitudes does not overlap zero. The EcR TSS shown in A is highlighted in green. (E) TT-TL-seq tracks of EcR and skl +/- 20E stimulus as examples of genes where 20E-induced changes in scRNA turnover are driven by faster or slower pause release, respectively. (F) Metaplots of TT-TL-seq signal without (dashed) and with (solid) 20E stimulus separated by whether release from the TSS was faster (dark blue) or slower (light blue) upon 20E stimulus as determined in D. Coverage is determined over 50 nt bins.

A major outstanding question which has not been broadly addressed is whether release into elongation, premature termination, or both are targets for the regulation of gene expression. STL-seq presents an opportunity to quantify changes in pause release and termination in response to a regulatory stimulus. By treating cells with 20-hydroxyecdysone

(20E), a hormone known to both induce and repress expression of target genes [184, 185], we can determine the preferred mechanism of regulation at the pause site. If altered initiation rates were solely responsible for the transcriptional response, we would expect to see correlation between changes in STL-seq reads and TT-TL-seq reads, but this was not the case (Figure 4.10A). To dissect the relative changes in pause-release and termination rates, we pretreated 20E-stimulated cells with FP and performed STL-seq. In uninhibited samples, 20E stimulus markedly increased the proportion of mutation-containing scRNA from TSSs of genes well-characterized as 20E targets (e.g., Figure 4.9A). To quantify these changes genome-wide, we used the same model as described above to estimate termination and pause-release rates. At high confidence TSSs, we find that 20E stimulus generally increases the total observed turnover of scRNA at many TSSs (Figure 4.9B).

Because 20E-stimulated TSSs tend to release Pol II into elongation slowly under normal conditions, we expected upregulation of $\hat{k}_{\text{rel}}$ to be the more likely response (Figure 4.9C). Indeed, the inflation of pause-release rates is more dramatic than the diminution of termination both in effect size and in the number of TSSs (Figures 4.10B & 4.10C). At each TSS, we examined the difference in magnitude of the $\log_2$ fold change of $\hat{k}_{\text{rel}}$ and $\hat{k}_{\text{term}}$ in the context of the change in $\hat{k}_{\text{obs}}$ (Figures4.95D & 4.105D). The four possible regulatory options of induction or suppression of either pause release or termination are separated into the four quadrants. Independent of whether a TSS is repressed or induced, our data reveal that most are regulated primarily at the level of release into elongation. Even at the fastest terminating TSSs where suppression of premature termination has the most potential to lead to increased expression, we see that pause release is preferably regulated (Figures 4.10E & 4.10F). To confirm these findings, we performed TT-TL-seq with the same RNA as collected for STL-seq (Figure 4.9E). We binned genes by the quadrant in which their TSSs are found in Figure 4.9D. In support of the predicted transcriptional changes detected by STL-seq, we observed that 20E treatment leads to increased transcriptional activity over genes with induced pause release, as well as loss of transcription over genes with repressed pause release (Figure 4.9F).

In summary, we observed that hormone treatment broadly elevates scRNA turnover at many TSSs. STL-seq demonstrates that 20E-induced changes in the total turnover rate

constants in most cases are driven by regulation of release into elongation while premature termination contributes only minor regulatory effects.

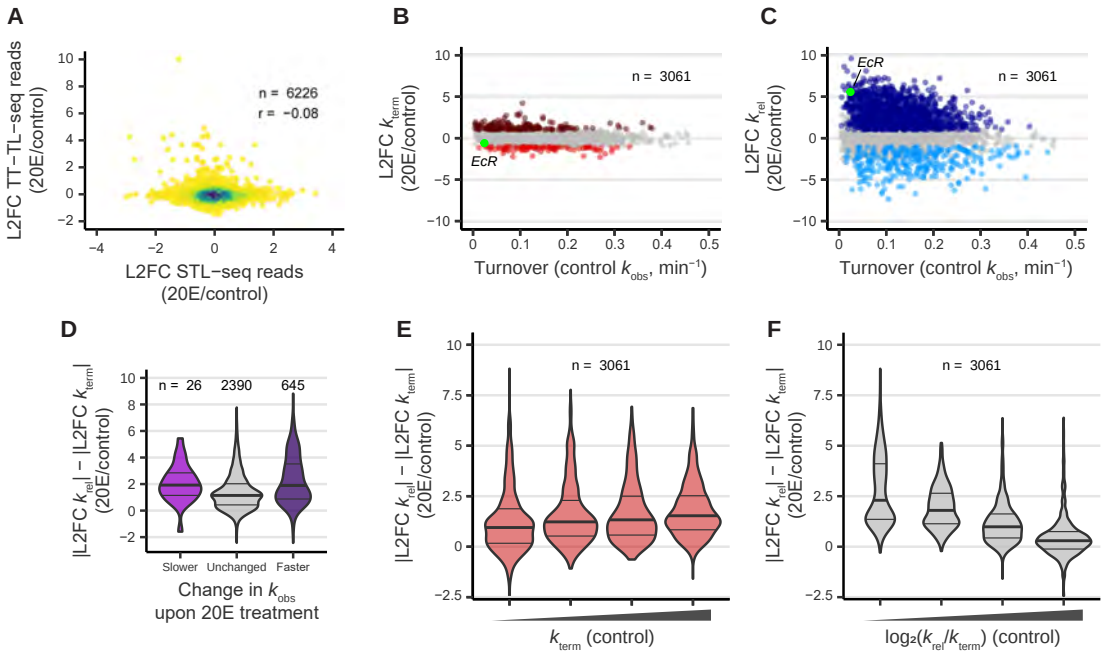

Figure 4.10: **Termination is rarely regulated in response to hormone stimulus** (A) The $\log_2$ fold change in normalized STL-seq reads at all TSS upon 20E treatment versus the $\log_2$ fold change in normalized TT-TL-seq read counts of the associated gene upon hormonal treatment. The density of plotted points is indicated by color (blue, high; yellow, low). The Pearson correlation coefficient is shown. (B and C) The total observed turnover rate constants plotted versus the $\log_2$ fold change in termination (B) or release (C) upon 20E stimulus. Points are colored if the 80% credible interval of the $\log_2$ fold change does not overlap zero and the median value is greater than 1 (dark red/blue) or less than -1 (light red/blue). (D) The difference between the magnitudes of the $\log_2$ fold change of release and termination grouped by the change in the total turnover as determined in Figure 4.9B. (E) The difference between the magnitudes of the $\log_2$ fold change of release and termination grouped by the change into even quartiles by termination. (F) The difference between the magnitudes of the $\log_2$ fold change of release and termination grouped by the change into even quartiles by the $\log_2$ ratio of the release rate to the termination rate.

4.3.8 Hyperosmotic stress induces premature termination

While hormone treatment primarily regulates $\hat{k}_{\text{rel}}$, we wondered if other stimuli might influence $\hat{k}_{\text{term}}$. Hyperosmotic stress alters the transcriptional landscape of human cells by inducing readthrough transcription as well as widespread transcriptional repression [146,165]. Previous Pol II ChIP-seq experiments under the same conditions revealed loss of Pol II over

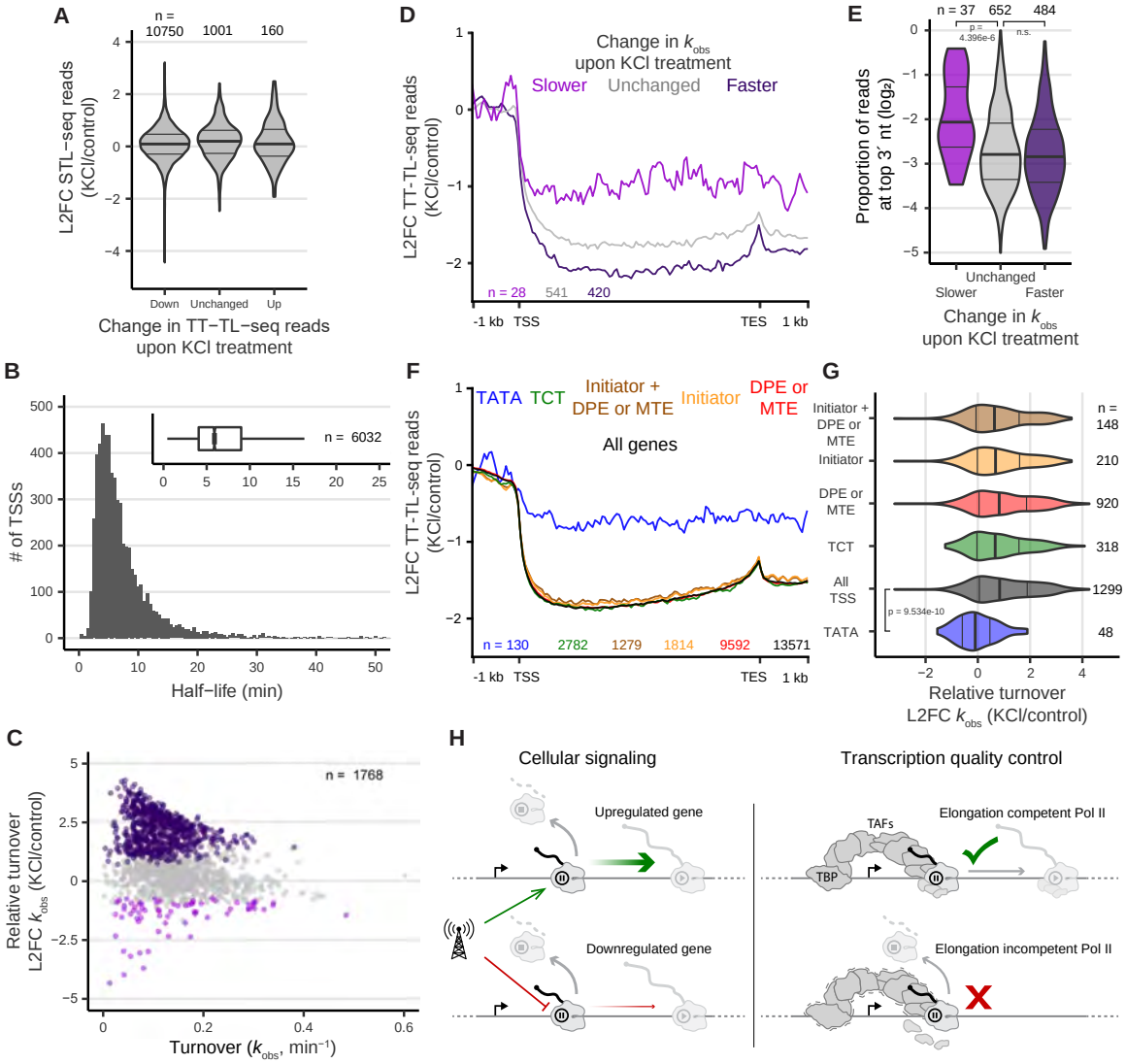

Figure 4.11: Caption on next page.

the body of repressed genes. This observation suggests that salt stress-induced transcriptional repression is at least partially accomplished at or prior to promoter-proximal pausing. Therefore, STL-seq is uniquely suited to provide insight into the mechanism accounting for this transcriptional repression.

We treated human 293T cells with 80 mM KCl for one hour and performed STL-seq to assess changes in initiation, termination, and pause release. We found that STL-seq signals were highly reproducible for both total and mutation-containing reads (Figures 4.12A & 4.12B). If reduced initiation were solely responsible for transcriptional downregulation, we would expect to see substantial loss of STL-seq reads at the promoters of downregulated

Figure 4.11: **Hyperosmotic stress induces premature termination at TATA-less promoters** (A) Change STL-seq reads at promoter TSSs grouped by the change in TT-TL-seq signal over the gene body. (B) Histogram of scRNA half-life estimates made with STL-seq from human 293T cells. The inset boxplot depicts the distribution of all TSSs. (C) Total observed rate constants of TSSs plotted versus the $\log_2$ fold change when cells are exposed to hyperosmotic stress for one hour. Points are colored if the 80% credible interval is entirely greater than $\log_2(1.5)$ (dark purple) or less than -$\log_2(1.5)$ (light purple). (D) Metaplots of the $\log_2$ fold change of TT-TL-seq signal upon hyperosmotic stress grouped by the change in turnover of scRNA at the gene's TSS as determined by STL-seq. Coverage was determined over 50 nt bins before calculating the fold change of each bin. (E) Distribution of the proportion of reads at promoters with 3´ ends located at the most frequent pause position grouped by the change in scRNA turnover and colored as in D. At each TSS the most common position of the 3´ read end was identified, and the proportion of reads at this position was determined. (F) Metaplots of the $\log_2$ fold change of TT-TL-seq signal upon hyperosmotic stress grouped by the motif content of the associated STL-seq TSS. Coverage was determined over 50 nt bins before calculating the fold change of each bin. (G) The distribution of the $\log_2$ fold change of scRNA turnover rate constants at promoters upon hyperosmotic stress. TSSs are grouped by motif content and colored as in F. Significance was assessed by a two-sided Wilcoxon rank sum test. (H) Proposed model for the distinct roles of release into elongation and premature termination at the promoter-proximal pause site. To alter gene expression, cells signal for either an increase or decrease in release into elongation (left). Premature termination does not contribute greatly to the response to cellular signaling but acts to evict paused Pol II whose elongation factors do not assemble properly (right). Coordinated binding of TFIID subunits, TBP and TAFs, is important for maturation of an elongation-competent Pol II and significantly stabilizes the mature complex.

genes, but we did not (Figure 4.11A). Thus, hyperosmotic stress must induce transcriptional repression via a reduction in the pause-release rate or an increase in the termination rate.

We applied the same model as described above to estimate $\hat{k}_{\text{obs}}$ of scRNA genome wide. Notably, the distribution of steady-state scRNA half-life estimates in human 293T cells is very similar to that of fly S2 cells (Figure 4.11B). When comparing stressed and unstressed cells, we found scRNA transcripts from many high confidence TSSs in untreated conditions to be much less stable upon hyperosmotic stress (Figure 4.11C). This observation and the loss of gene body transcription suggest that induction of Pol II premature termination at the pause site is a major response to hyperosmotic stress. We then compared changes in turnover to previously published TT-TL-seq data [165]. Supporting our model that hyperosmotic stress induces premature termination, active transcription is more repressed over genes with destabilized scRNA (Figures 4.11D & 4.12C). Promoters with decreased turnover produce more focused pause sites than those with unchanged or induced turnover

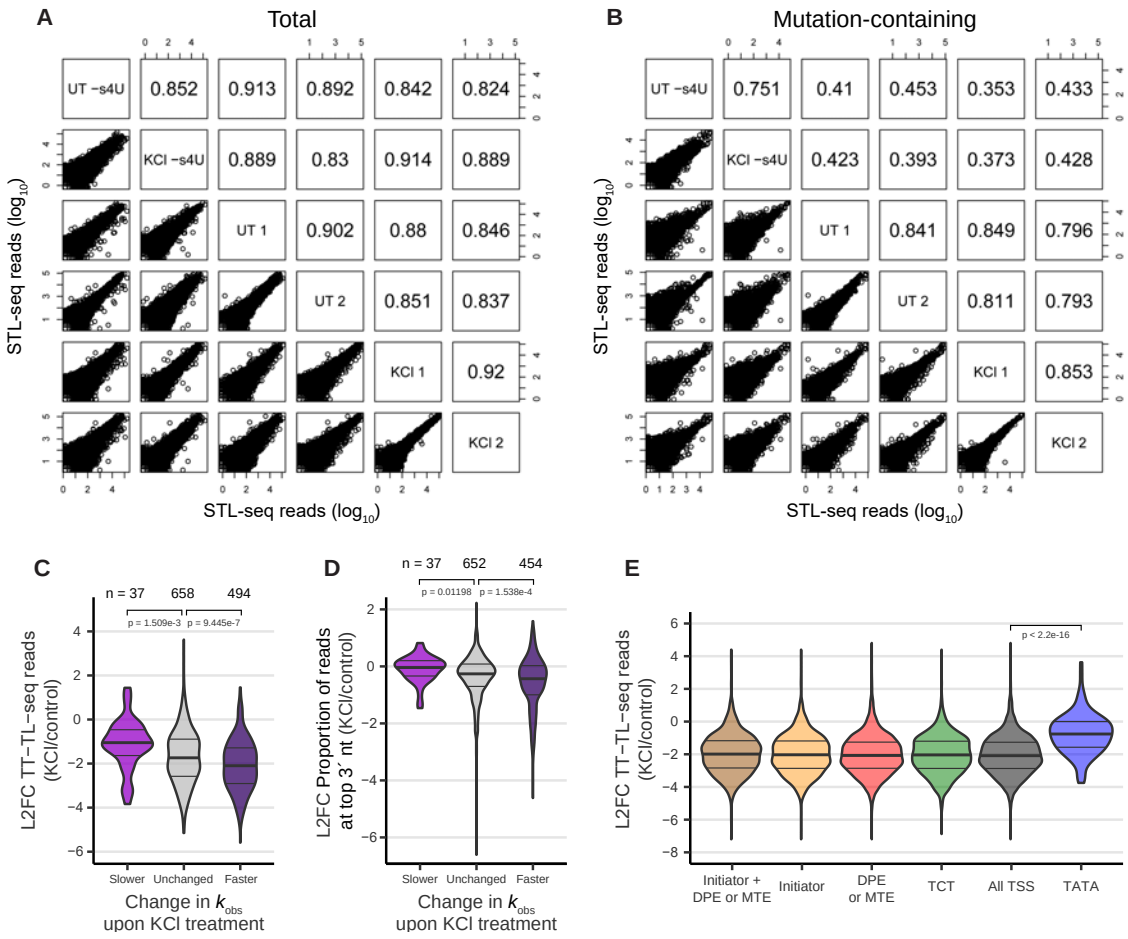

Figure 4.12: **Increased termination at the pause site causes widespread transcriptional repression upon hyperosmotic stress in human cells** (A) Pairs plots of total TSS read counts from STL-seq unlabeled and labeled samples under normal and hyperosmotic stress conditions. Read counts are plotted on the $\log_{10}$ scale and the Pearson correlation coefficient of each comparison is shown. (B) Pairs plots of mutation-containing TSS read counts from STL-seq unlabeled and labeled samples under normal and hyperosmotic stress conditions. Read counts are plotted on the $\log_{10}$ scale and the Pearson correlation coefficient of each comparison is shown. (C) Distribution of the $\log_2$ fold change of TT-TL-seq signal upon hyperosmotic stress grouped by the change in turnover of the TSS called to each gene. Significance was assessed by a two-sided Wilcoxon rank sum test. (D) Distribution of the $\log_2$ fold change of the proportion of reads with 3´ ends located at the most frequent pause position grouped by the change in turnover of the scRNA. At each promoter, the most common position of the 3´ read end is identified, and all reads are considered by their 3´ end relative to the most common. The change in the proportion of reads at the most common position upon hyperosmotic stress is plotted. Significance was assessed by a two-sided Wilcoxon rank sum test. (E) Distribution of the $\log_2$ fold change of TT-TL-seq signal upon hyperosmotic stress grouped by the motif content of the associated STL-seq TSS. Significance was assessed by a two-sided Wilcoxon rank sum test.

(Figure 4.11E). Upon hyperosmotic stress, the induced TSSs become even less focused (Figure 4.12D). Together, these results suggest that the promoters of genes downregulated upon KCl treatment are prone to stress-induced termination at the pause site.

We hypothesized that TSSs prone to stress-induced termination may lack cis-acting DNA elements that recruit components of the PIC or other pausing factors. We again binned promoters by their motif content (using the consensus TATA box described by [186], see methods). Strikingly, genes with promoters containing a TATA box were protected from transcriptional repression (Figures 4.11F & 4.12E). The TSSs of these genes were also protected from stress-induced termination at the pause site (Figure 4.11G). However, none of the downstream TFIID motifs generally protected genes from transcriptional repression or premature termination despite their ability to stabilize the paused Pol II complex. Taken together, these results indicate that termination at the pause site is an important regulatory process that is associated with cis-acting DNA elements at the promoter.

4.4 Discussion

STL-seq provides genome-wide insight into the dynamics of promoter-proximal pausing by combining the time resolution of metabolic labeling [97] with the TSS specificity of Start-seq [57]. Our results demonstrate that STL-seq reliably captures kinetic information of scRNA, allowing inference of the kinetic behavior of the Pol II paused complex.

We found total observed turnover are similar between human and fly (Figures 2D and 6B), suggesting that pausing dynamics and regulation may be conserved across metazoans. To better understand the complex behavior of Pol II at the pause site, we dissected total observed turnover into rates of release into elongation and premature termination. This revealed that while only a small fraction of paused Pol II enters productive elongation, pause release is highly dynamic and responds to 20E stimulus in *Drosophila*. On the other hand, premature termination does not determine gene expression and is insensitive to the same hormonal stimulus. These findings provide detailed kinetic support for the concept that active regulation at the pause site occurs by altering the rate of release into elongation (Figure 6H).

We also sought to identify a function for premature termination. Similar to Beckedorff et al. (2020), we favor a model in which termination at the promoter-proximal pause site occurs as a quality check mechanism to ensure that members of a mature elongation complex (EC) correctly and completely assemble on Pol II (Figure 6H). In support of this model, we found relatively fast termination rates at TSSs with features that we view as hallmarks of inefficient EC assembly. These features include high H3K36me3, the lack of downstream TFIID binding motifs, and less focused pause sites. H3K36me3 functions to repress cryptic initiation and leads to the erasure of other activating histone tail modifications [20, 22, 182]. The absence of downstream TFIID binding motifs leads to less focused pausing, which we suspect is a symptom of poorly assembled ECs.

Previous work [58] demonstrated that weaker contacts between the PIC and the paused complex lead to less focused pause sites. Our data supports an extension of this model in which weaker interactions between the PIC and the paused complex lead to faster premature termination. We hypothesized that stressing cells in a manner that disrupts the transcriptional machinery may lead to increased premature termination at the pause site. Hyperosmotic stress alters the Pol II interactome and leads to transcriptional silencing genome-wide [165]. We showed that induction of premature termination is at least partly responsible for the response to hyperosmotic stress that results in genome-wide transcriptional repression. Therefore, we speculate that hyperosmotic stress disrupts elongation factor assembly and results in a larger proportion of incompetent ECs reaching the pause site. These incompetent ECs are then signaled for premature termination before they can enter productive elongation. However, leaky pause release of ECs which lack critical processing machinery may lead to production of downstream-of-gene containing transcripts (DoGs), which are a product of readthrough transcription [146], and/or misspliced transcripts [187]. In this manner, premature termination at the pause site may provide a form of kinetic proofreading.

Recently, the Integrator complex has been the focus of studies examining premature termination at the pause site [43, 45, 46, 60], and one study proposes that Integrator acts in early elongation [56]. Interestingly, hyperosmotic stress causes dissociation of Integrator from Pol II [165], suggesting that Integrator is not responsible for the induction of pre-

mature termination observed under hyperosmotic stress. More generally, future STL-seq experiments may help clarify the roles of Integrator at the promoter-proximal pause site.

Here, TFIID has emerged as a critical factor that acts beyond initiation to establish and maintain proper kinetics of promoter-proximal pausing. Strikingly, we found that the presence of TATA box prevents stress-induced premature termination, highlighting the vital role of TFIID through initiation and pause release. Our findings illustrate the unique capabilities of STL-seq to reveal the dynamics and regulation of promoter-proximal pausing as well as to identify essential pausing factors. This method will enable future studies investigating the mechanism of promoter-proximal pausing which were not previously possible.

4.5 Limitations of the study

STL-seq is a powerful tool to study pausing kinetics and provide mechanistic insight into promoter-proximal pausing at most TSSs. Accurate estimation of kinetic parameters using STL-seq is limited by the read depth and mutational content of reads mapping to each TSS. In practice, we have found that rate estimates at a TSS are less reliable when the mutational content at the TSS is low. In this case, there may not be enough information to use our Bayesian models to confidently determine rate constants, leading to large credible intervals for our parameter estimates. Low numbers of observed mutations could result from low read coverage, few uridines in a scRNA, or from a low s^4U incorporation rate. We developed criteria for identification and removal of unreliable TSSs from our analyses (see Methods), allowing us to restrict analyses to the thousands of TSSs where kinetic parameters can be confidently estimated.

To dissect the steady-state observed turnover of scRNA at the pause site, we took advantage of the rapid inhibition of P-TEFb activity by flavopiridol (FP) to block pause release. We provide evidence that FP does not perturb premature-termination rate constants, but the accuracy of our $\hat{k}_{\text{rel}}$ and $\hat{k}_{\text{term}}$ estimates would be reduced at some TSSs if P-TEFb inhibition directly or indirectly influences premature-termination rates constants. Nonetheless, our hyperosmotic stress experiments demonstrate that additional information about transcription over the gene body (e.g., from TT-TimeLapse-seq data) is sufficient to

identify changes in $\hat{k}_{\text{rel}}$ and $\hat{k}_{\text{term}}$ without the need for P-TEFb inhibition.

Chapter 5

Probing the functional consequences of an SVA insertion in the *TAF1* gene in XDP

5.1 Author contributions

Cris Bragg, Shivangi Shah, Christine Vaine, Sherman Weissman, and Anna Szekely provided patient-derived cell lines and performed the s^4U treatments and cell lysis for all TimeLapse-seq and TT-TL-seq experiments using these cell lines. Michelle Moon helped perform TimeLapse-seq for some of the experiments with patient-derived cell lines. Giselle Fisher performed the *in vitro* pull-down experiments. I performed all other experiments and data analyses described in this section. Matthew Simon, Jeremy Schofield, and I contributed the the conception of the work and experimental design.

5.2 Summary

X-Linked Dystonia Parkinsonism is a rare, but severe, neurodegenerative genetic disorder whose disease mechanism is not known. Here we provide evidence that the causally-linked SINE-VNTR-Alu (SVA) retrotransposon insertion found in intron 32 of *TAF1* causes a premature transcription termination event of some actively transcribing RNA polymerase

II (RNAPII) complexes. The resulting transcript (*xTAF1*) would be translated into a truncated isoform which is predicted to differ from canonical TAF1 (cTAF1) in the binding pocket of the second bromodomain. In cells expressing xTAF1, we find that this truncation gives rise to changes in gene expression which are related to the binding pattern of TAF1 and the local histone modification profile at the promoter. Expressing xTAF1 recapitulates the same pausing phenotype observed in XDP patient-derived cells and gives credence for our model that xTAF1 is pathogenic. Furthermore, this work provides previously unappreciated biological relevance for the activity of the TAF1 bromodomain 2.

5.3 Introduction

X-Linked Dystonia Parkinsonism (XDP) is a rare neurodegenerative disorder endemic to the island of Panay in the Philippines. XDP was first described in 1976 as a torsion dystonia exclusively displaying in males with a relatively late average age at onset [188]. Since then, XDP has been diagnosed in more than 500 patients and was recently causally linked to a SINE-VNTR-Alu (SVA) retrotransposon insertion in intron 32 of the gene encoding TATA-box binding protein (TBP) associated factor 1 (TAF1) on the X chromosome [124–126, 189]. The SVA is inserted antisense relative to the direction of transcription of *TAF1*, and is composed of five major sequence elements. Listed by their order in the direction of transcription of *TAF1*, these elements are a poly T sequence, a short interspersed nuclear element (SINE), a variable number tandem repeats (VNTR), an Alu element, and a $(CCCTCT)_n$ repeat sequence. The causal link derives from a hexanucleotide repeat sequence where the number of repeats is negatively associated with the age at onset of XDP symptoms [124, 189]. This is the only direct connection between the XDP phenotype and the SVA insertion, and there is no prevailing model for the disease mechanism. Unfortunately, challenges associated with understanding XDP pathogenesis compound in the inability to develop a therapy to treat XDP patients. Therefore, understanding the underlying molecular phenotypes are an important step in improving patient quality of life.

The revelation of a causal link between the SVA and XDP lead to a concerted effort to characterize transcripts originating from the *TAF1* locus in patient-derived cells [125].

TAF1 is about one megabase long with 38 exons in the canonical isoform (*cTAF1*) and an additional six-base-pair microexon, which is spliced in as exon 34´in the neuron-specific *TAF1* isoform (nTAF1). nTAF1 was previously suggested to be repressed in XDP and possibly important to XDP pathogenesis, but more recent work refutes these claims [125, 190–192]. *TAF1* expression is moderately downregulated in patient samples and patient-derived cells, and interestingly seems to be stronger for exons downstream of the SVA than exons upstream of the SVA [125, 189, 191, 193]. In addition, long-read sequencing identified a rare XDP-specific and SVA-dependent isoform which lacks exons 33-38 and retains a portion of intron 32 upstream of the SVA as the novel last exon (TAF1-32i) [125, 194]. The last notable XDP-specific change in the *TAF1* transcriptome is a second SVA-dependent intron retention event of nearly half of intron 32. PolyA-selected short-read RNA-seq from patient-derived cells showed that the first half of intron 32 upstream of the SVA is retained across multiple stages of development [125].

Several other mutations in *TAF1* have already been linked to neurodevelopmental and intellectual disability disorders, but unlike XDP, these disorders are associated with changes in the *TAF1* coding sequence (CDS) and typically display early in life or at birth [121, 128, 195–198]. This suggests that TAF1 function is critical for proper neuronal development; however, relatively little is known about TAF1 outside of its context as a TFIID subunit and there is no defined, direct role for TAF1 or nTAF1 specific in neuronal health. TAF1 is the largest subunit of the general transcription factor (GTF) TFIID and is an essential factor responsible for promoter recognition during assembly of the preinitiation complex (PIC) [68, 199–203]. TAF1 contains tandem bromodomains which are known to bind acetylated histone H4 tails, is suggested to form a heterodimer with TAF7 that recognizes repressive histone tail modifications, and has been shown to have histone acetyl transferase (HAT) and kinase activity [19, 204–206]. Evidence for these activities is primarily from *in vitro* data and the biological relevance is yet to be thoroughly characterized. Since the SVA insertion does not directly change the CDS of *TAF1*, it is possible that XDP is a TAF1 insufficiency disorder associated with one or more of the functions described above. Another possibility is that alternative *TAF1* transcripts caused by the SVA insertion encode a truncated TAF1 protein isoform with a deleterious gain of function.

Here we provide evidence for an SVA-dependent premature transcription termination event near the site of insertion which gives rise to a stable XDP-specific *TAF1* isoform (*xTAF1*). *xTAF1* is predicted to contain exons 1-32 and the entire first half of intron 32 up to the site of the SVA insertion. The *xTAF1* transcript has potential to be translated into an xTAF1 protein which would replace the 288 C-terminal residues of cTAF1 with 17 intronically-encoded residues. The xTAF1 protein product completely lacks the predicted C-terminal kinase domain and replaces the last helix of the second bromodomain (BD2) with the intronically encoded residues. AlphaFold predicts that the binding pocket of xTAF1 BD2 is intact but enlarged relative to cTAF1 BD2.

We profiled the transcriptional landscape in XDP patient-derived cells and the TAF1-expressing cells with our suite of nucleotide-recoding RNA-seq methods (TimeLapse-seq [97], TT-TL-seq [97, 132], and STL-seq [50]). We found that RNA synthesis is perturbed in XDP and xTAF1-expressing cells, and promoter-proximal paused RNA polymerase II (RNAPII) is similarly redistributed in both cell models relative to controls. To assess the consequences of an altered BD2 pocket on TAF1 promoter-binding activity, we performed ChIP-seq in cTAF1- and xTAF1-expressing human cells and found that xTAF1 binds promoters in a manner similar to but distinct from cTAF1. Our data suggest that BD2 activity is inhibitory to TAF1 binding at some promoters. The truncation mutation causes xTAF1 to associate with promoters more strongly, and this could be the gain of function responsible for XDP. This work proposes a model for XDP pathogenisis, and in doing so, provides biological relevance for TAF1 BD2.

5.4 Results

5.4.1 Early steps in RNA synthesis are perturbed in XDP cells

Previously, patient derived cell lines were developed as a model system to study X-Linked Dystonia Parkinsonism [125]. This work revealed that XDP patient-derived cells have hundreds of differentially expressed genes when compared to control and the sets of differentially expressed genes vary widely across development stages. This suggests that XDP may have very cell-type specific affects that cannot be dissected with traditional RNA-seq experi-

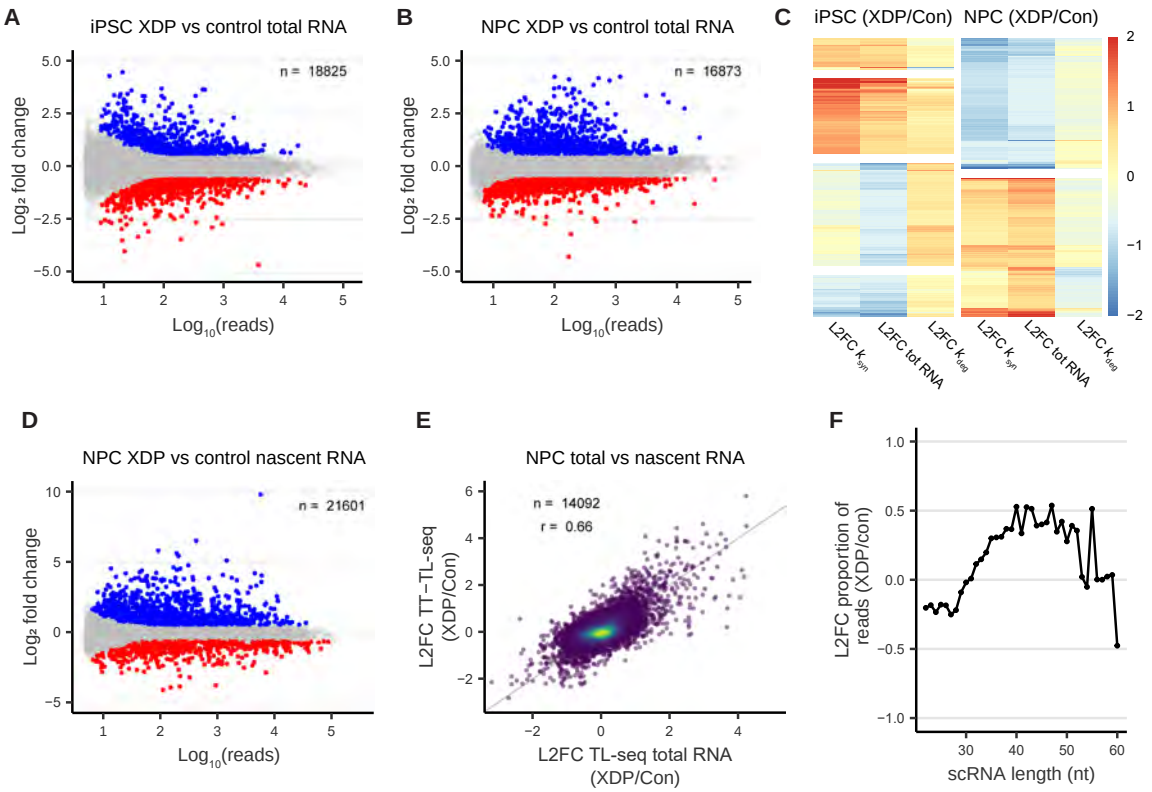

Figure 5.1: **RNA synthesis is perturbed in XDP patient-derived cells** (A & B) MA plots for total RNA TL-seq data comparing XDP patient-derived cells versus controls at the iPSC (A) or NPC (B) stage. Transcripts are highlighted if the DESeq2 $p_{adj} < 0.05$. (C) Heatmaps of the $\log_2$ fold change in k_{syn}, total RNA, or k_{deg} of significantly differentially expressed transcripts in total RNA in iPSCs (left) or NPCs (right). Genes are highlighted if the DESeq2 $p_{adj} < 0.05$. (D) MA plots for TT-TL-seq data comparing XDP patient-derived cells versus controls at the NPC stage. (E) Scatter plot comparing the $\log_2$ fold change of read counts in TL-seq and TT-TL-seq at the NPC stage. The color represents the density of plotted points where yellow and purple correspond to high and low density, respectively. The y=x line is plotted and the Pearson correlation coefficient is shown. (F) The $\log_2$ fold change of the proportion of total reads in STL-seq data comparing XDP patient-derived cells versus controls at the NPC stage. scRNA reads were grouped by their absolute length.

ments. Furthermore, it is unclear if the affects on expression are due to altered stability of the mature transcripts or perturbations to transcription. To probe mRNA degradation and synthesis, we performed TimeLapse-seq in patient-derived induced pluripotent stem cells (iPSCs) and neural progenitor cells (NPCs). Consistent with the previous report, by total read counts, we observed hundreds of significantly upregulated and downregulated transcripts in both data sets when comparing XDP to control cell lines, and there was no strong correlation between cell types (Figures 5.1A,B and 5.2A). We took advantage of the

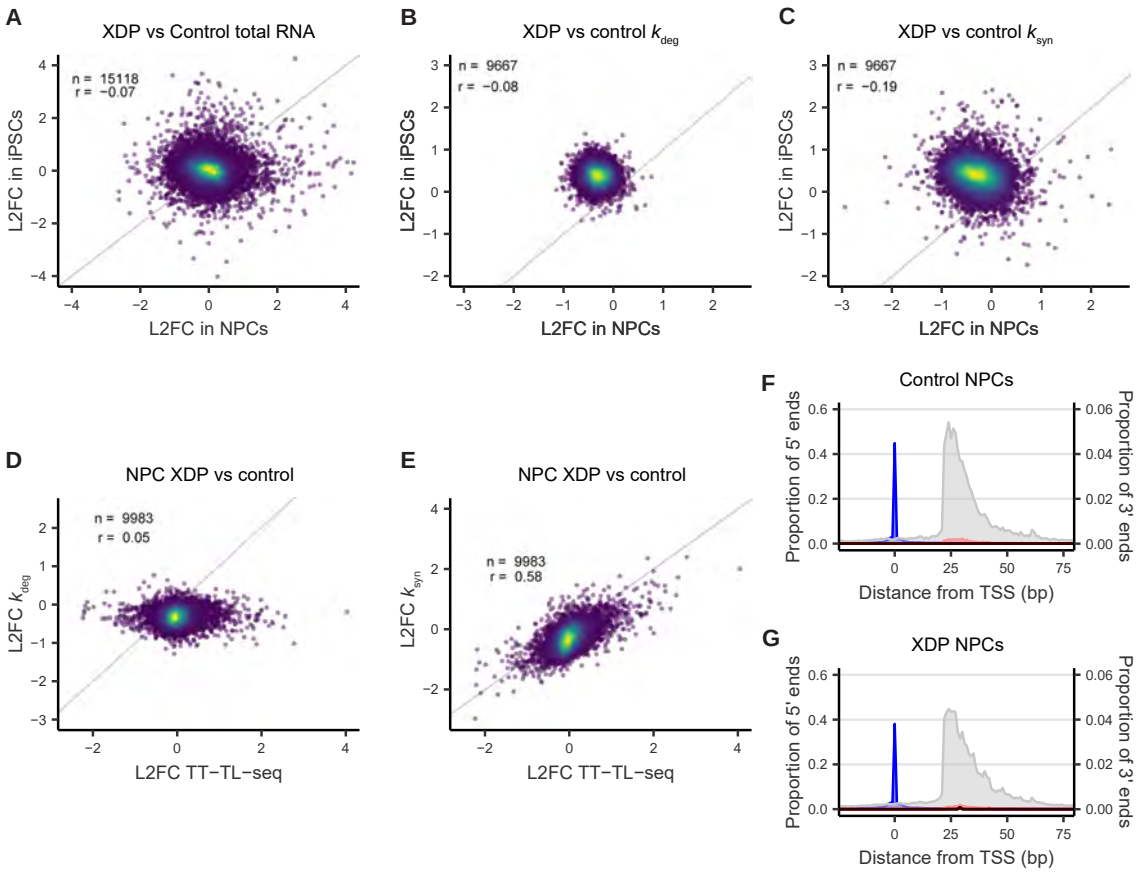

Figure 5.2: **RNA synthesis and not RNA degradation drives changes in gene expression in XDP cells** (A-C) Scatter plots comparing the $\log_2$ fold change of TL-seq read counts (A), k_{deg} (B), or k_{syn} (C) between patient-derived cells at the iPSC and NPC stages. The color represents the density of plotted points where yellow and purple correspond to high and low density, respectively. The y=x line is plotted and the Pearson correlation coefficient is shown. (D & E) Scatter plots comparing the $\log_2$ fold change of TT-TL-seq read counts against $\log_2$ fold change in TL-seq k_{deg} (D) or k_{syn} (E). The color represents the density of plotted points where yellow and purple correspond to high and low density, respectively. The y=x line is plotted and the Pearson correlation coefficient is shown. (F & G) Metaplots of STL-seq 5′ and 3′ read ends. The single-nucleotide location of the TSS (blue, 5′ end of read) and pausing position (grey and red, 3′ end of read) are depicted separately. The 3′ ends are colored by the read's mutational content. Read ends at each distance from the TSS for control (F) and XDP (G) NPCs are shown as a proportion of the total number of reads. The proportion of 5′ ends corresponds to the left y-axis scale and the proportion of 3′ ends corresponds to the right y-axis scale.

temporal information captured by TimeLapse-seq to determine if changes in total RNA are driven by RNA synthesis or degradation. Using the hybrid model implementation of bakR (Vock et al., *in prep*). We examined the changes in RNA synthesis (k_{syn}) and degradation (k_{deg}) rate constants and found that expression changes in XDP patient-derived cells

are primarily due to changes in k_{syn} across developmental stages (Figure 5.1C). Similar to changes in total RNA, the effects on k_{deg} or k_{syn} do not strongly correlate between cell types. This suggests that XDP is related to some perturbation during RNA synthesis that is highly cell-type specific.

To further dissect the effects on RNA synthesis, we performed Transient-Transcriptome-TimeLapse-seq (TT-TL-seq) to probe the activity of RNAPII across the genome in patient-derived NPCs [97,132]. We found many genes to be differentially transcribed, in agreement with TimeLapse-seq which revealed that RNA synthesis is perturbed in SVA-containing cell lines (Figure 5.1D). In addition, changes in nascent RNA correlate well with changes in total RNA and $k_{\texttt{syn}}$ but not $k_{\texttt{deg}}$, suggesting that a perturbation in an early step of transcription is the major cause for differential expression at the total RNA level (Figures 5.1E and 5.2D,E).

Finally, we performed Start-TimeLapse-seq (STL-seq) on patient cell lines at the NPC stage to directly observe the behavior of RNAPII during initiation and promoter-proximal pausing [50]. While the TimeLapse T-to-C mutation rate is not high enough to estimate the kinetics of promoter-proximal pausing, we used the data as a single-molecule approach to determine the positions of initiation and pausing. Each read in STL-seq data captures the single-nucleotide position of the site of initiation and pausing (Figure 5.2F,G). When we compared the proportion of read lengths in STL-seq data from SVA-containing patient cell lines to that of controls we found that SVA cell lines lose reads shorter than 30 bp and gain reads of lengths between 30 and 60 bp (Figure 5.1F). This observation could be explained in three ways; RNAPII transcribes farther on average before pausing, RNAPII paused 20-30 bp downstream of the TSS is destabilized, or RNAPII paused beyond 30 bp downstream of the TSS is stabilized in XDP patient cells relative to control cells. In any case, the insight provided by our RNA metabolic labeling approaches and nucleotide-recoding chemistry point towards an XDP-specific perturbation of early steps in transcription, and in particular promoter-proximal pausing genome wide may be directly affected.

5.4.2 The SVA insertion gives rise to a stable, intron-retained XDP-specific *TAF1* transcript

TAF1 is known to play an important role in promoter recognition and DNA binding as a TFIID subunit, and DNA motifs known to be bound by TAF1 have been shown to be related to the duration of RNAPII promoter-proximal pausing [50, 68, 178, 199–203]. Several mutations in *TAF1* have been linked to intellectual disability and developmental delays [121, 128, 195–197]. While this body of work establishes *TAF1* as an important factor in human neurodevelopment, we lack an understanding for the role of TAF1 in the pathogenisis of these disorders. XDP is unique in that the causal mutation of the SVA insertion does not directly change the coding sequence, but has been shown to give rise to at least two novel *TAF1* transcripts through partial retention of intron 32 or inclusion of a novel exon in intron 32 (TAF1-32i) [125]. We thought to further characterize the affect of the SVA insertion on the *TAF1* locus with a targeted analysis of our TimeLapse-seq and TT-TL-seq data.

We began by examining *TAF1* in our TimeLapse-seq data from XDP and control patient-derived cells at the iPSC and NPC stages, as well as XDP patient-derived NPCs from which the SVA has been deleted with CRISPR/cas9 excision (Figure 5.3A,B). As part of our differential expression and kinetic analyses, we defined three additional *TAF1* features: the region of *TAF1* intron 32 upstream of the SVA insertion, exons upstream of intron 32 (exons 1-32), and exons downstream of intron 32 (exons 33-38). Consistent with previous results, *TAF1* is modestly downregulated in XDP cells relative to controls, and the intron retention event is stronger at earlier developmental stages (Figure 5.3C) [125]. We also found that exons 33-38 are more affected by the presence of the SVA insertion than exons 1-32, and deletion of the SVA insertion rescues the downregulation of exons 33-38 in NPCs. Next, we took advantage of TimeLapse data to demonstrate that the effect on the downstream exons is primarily attributable to a decrease in synthesis, suggesting that cells containing the SVA insertion do not produce transcripts containing exons 33-38 as frequently (Figure 5.3D). This effect is reversed in NPCs by deleting the SVA insertion, showing that this effect on exon synthesis is SVA-dependent. Finally, we found intron 32

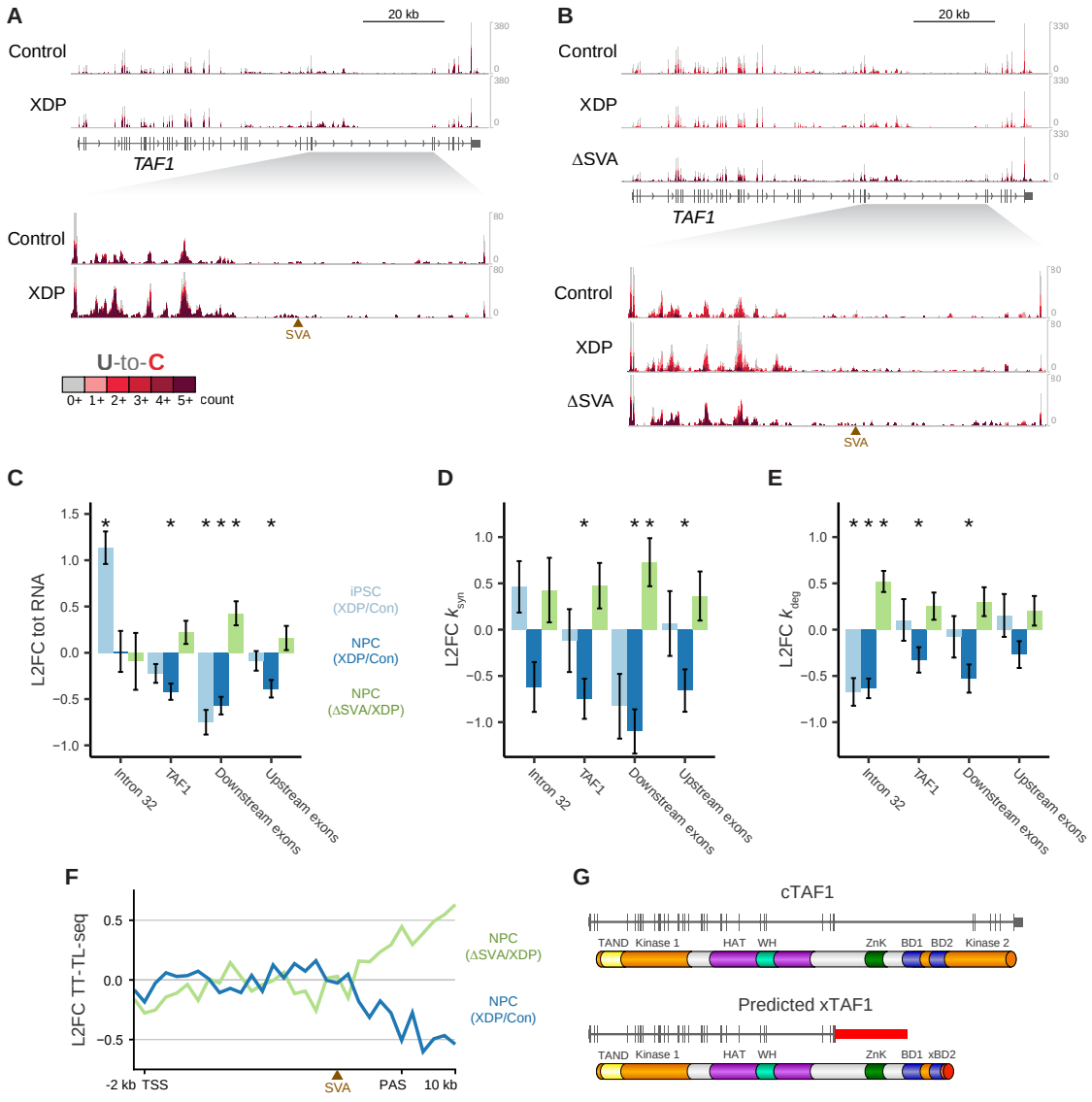

Figure 5.3: **The SVA insertion perturbs transcription of the *TAF1* locus** (A & B) Representative of TL-seq tracks of *TAF1* from patient-derived iPSCs (A) or NPCs (B). (C-E) The $\log_2$ fold change in total RNA (C), degradation rates (k_{deg}, D), or synthesis rates (k_{syn}, E) from TL-seq data between XDP and control patient cells in iPSCs and NPCs and between ΔSVA and XDP cells in NPCs. Values are separately determined for the region of *TAF1* intron 32 upstream from the SVA insertion, full-length *TAF1*, and exons upstream or downstream of intron 32. Significance was assessed by DESeq2 or bakR and the standard error is shown. (F) The $\log_2$ fold change in TT-TL-seq coverage over the *TAF1* gene from the transcription start site (TSS) to the cleave and polyadenylation site (PAS). Comparisons from NPCs between XDP and control cells and between ΔSVA and XDP cells are depicted. (G) Model for the predicted xTAF1 transcript and protein that would arise from a cleavage and polyadenylation event introduced near the site of the SVA insertion. The predicted xTAF1 truncation replaces the last six exons with the first half of intron 32 and substitutes 288 C-terminal amino acids with 17 different residues.

upstream of the SVA insertion is stabilized in XDP cells relative to controls, and deletion of the SVA destablizes the same region (Figure 5.3E). These data are consistent with the previously observed intron retention event, and suggest that the intron-containing transcript is stable and does not contain exons 33-38. In addition, published polyA-enriched data from XDP neural stem cells (NSCs) suggests that intron 32 upstream of the SVA insertion is included in a polyadenylated transcript [125].

Next, we compared coverage over *TAF1* in our TT-TL-seq data from NPCs to determine the effect of the SVA insertion on transcription over the gene. When comparing XDP NPCs to controls, coverage begins to drop off near the site of the SVA insertion (Figure 5.3F), consistent with a premature transcription termination event. Comparing coverage in TT-TL-seq data between ΔSVA and XDP NPCs, we find that the deletion rescues the apparent loss of nascent RNA downstream of the SVA insertion (Figure 5.3F). Therefore, the transcriptional effect is also SVA-dependent and is consistent with a premature transcription termination event caused by the SVA insertion. When considering the previous polyA-enriched RNA-seq data, we propose that the SVA insertion introduces or activates a cryptic cleavage and polyadenylation site (PAS) in intron 32. This cryptic PAS does not cause termination of all polymerases, but would give rise to a truncated XDP-specific *TAF1* transcript (*xTAF1*). As *xTAF1* would have the same 5´ untranslated region as cTAF1, it would have a high translation potential. The predicted truncation of TAF1 substitutes 288 C-terminal amino acids of cTAF1 with 17 intronically-encoded residues before the first in-frame stop codon (Figure 5.3G). This truncation deletes the second annotated TAF1 kinase domain and replaces a small piece of the second TAF1 bromodomain (BD2).

5.4.3 The XDP truncation of TAF1 affects the structure and function of BD2

The cTAF1 tandem bromodomains are known to bind acetylated histone H4 tails and are thought to aid in TAF1/TFIID recruitment to promoters [19]. Interestingly, a point mutation which causes a serine-to-glycine missense mutation in the penultimate helix of BD2 has already been implicated in a rare intellectual disability syndrome [197]. In this case, the mutation does not directly change the residues which shape the binding pocket

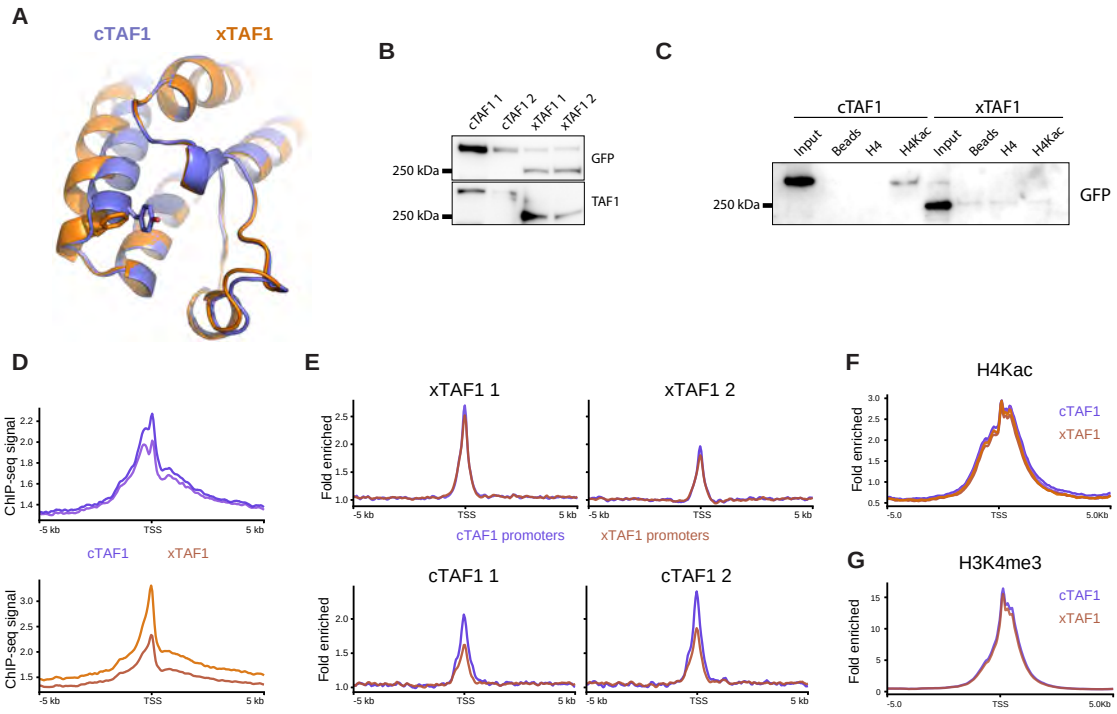

Figure 5.4: **xTAF1 binds chromatin in a similar but distinct pattern as cTAF1** (A) AlphaFold predictions of cTAF1 BD2 and xTAF1 BD2 overlayed to show major differences between the structures. The most important residues related to changes in the binding pocket surface are shown: a tyrosine in cTAF1 BD2 and a phenylalanine in xTAF1 BD2. (B) Western blots using whole-cell lysates from cells expressing GFP-cTAF1 or GFP-xTAF1 using an antibody raised against GFP (top) or TAF1 (bottom). A total of two clones were generated for each TAF1 isoform. (C) Western blot of H4 tail peptide pull-downs using an antibody raised against GFP. Pull-downs were performed in the absence of H4 peptides or with unmodified H4 peptide tails (H4) or H4 peptide tails acetylated at lyseines 5 and 12 (H4Kac).The TAF1 isoform expressed in cells from which lysates were prepared are indicated. (D) GFP ChIP-seq metaplots from cells expressing GFP-cTAF1 (top) or GFP-xTAF1 (bottom) centered on the annotated TSS of all expressed genes. (E) Metaplots of the fold enrichment in GFP ChIP-seq data from GFP-cTAF1 and GFP-xTAF1 expressing cells centered on the TSS of promoters identified to be bound by cTAF1 or xTAF1. (F & G) Metaplots of H4Kac (F) and H3K4me3 (G) CUT&RUN fold enrichment data from cells expressing GFP-cTAF1 or GFP-xTAF1 centered on the annotated TSS of all expressed genes.

but was speculated to affect the flexibility and binding preferences of BD2. Because of this precedence for the importance of BD2 in neurodevelopment, we decided to determine how the XDP-specific truncation of TAF1 affects the domain and its function. We used AlphaFold to predict the structure of cTAF1 BD2 (BD2) and xTAF1 BD2 (xBD2) [207]. The overall structure predictions of BD2 does not differ significantly between the two TAF1

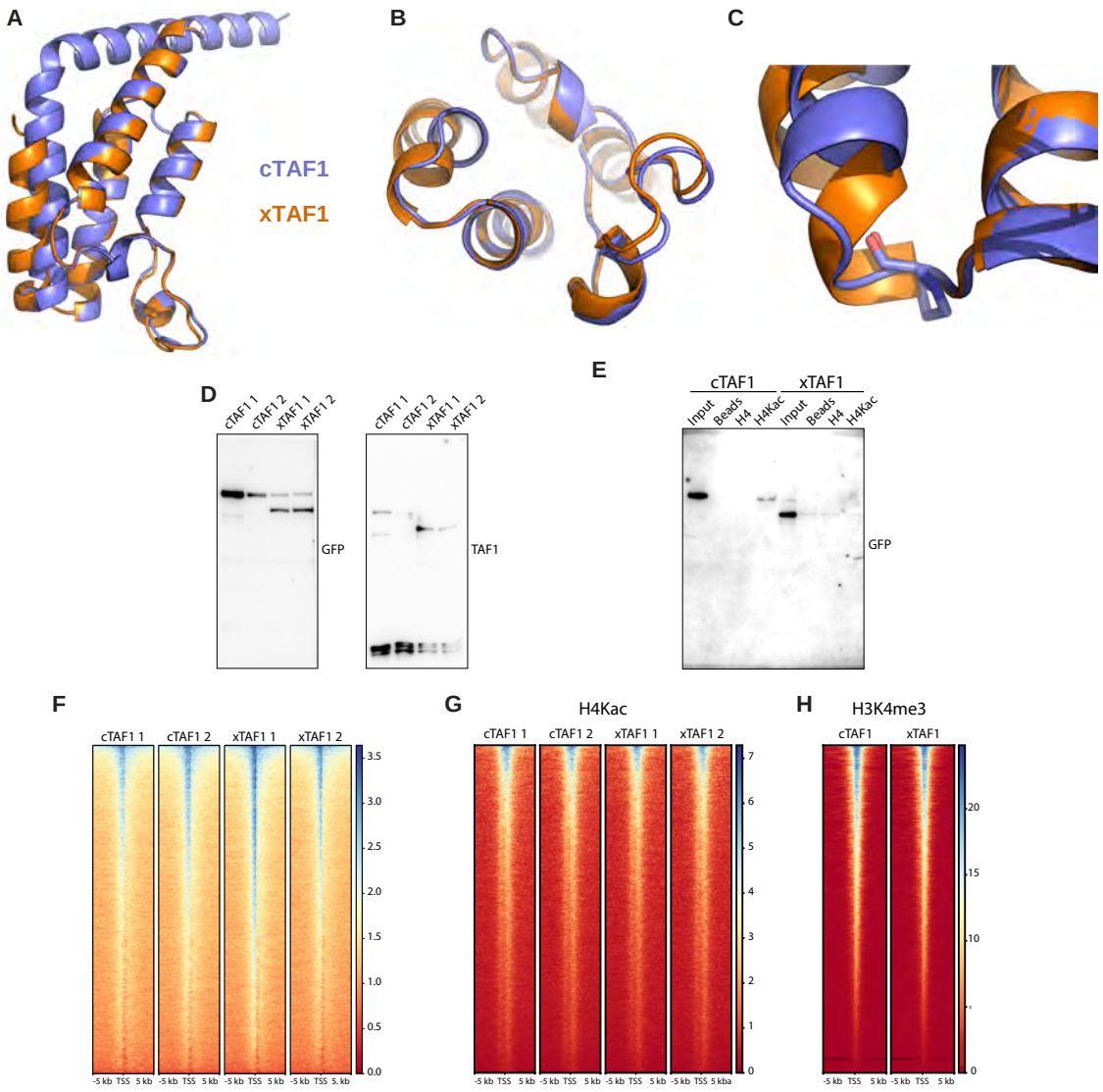

Figure 5.5: **Expressing xTAF1 does not globally perturb the chromatin landscape** (A) AlphaFold predictions of the complete BD2 of cTAF1 and xTAF1 overlayed to show major differences between the structures. (B) Overlayed AlphaFold predictions of cTAF1 BD2 and xTAF1 BD2 looking up into the binding pocket to illustrate the expansion of the binding pocket and shift of the substituted helix. (C) Overlayed AlphaFold predictions of cTAF1 BD2 and xTAF1 BD2 with an expanded view of the P>S mutation which causes a conformational shift in the predicted xTAF1 structure. (D & E) The full Western blots shown in 5.4B (D) and 5.4C (E). (F) GFP ChIP-seq heatmaps from cells expressing GFP-cTAF1 or GFP-xTAF1 centered on the annotated TSS of all expressed genes and ordered by decreasing average signal. (G & H) Heatmaps of H4Kac (G) and H3K4me3 (H) CUT&RUN fold enrichment data from cells expressing GFP-cTAF1 or GFP-xTAF1 centered on the annotated TSS of all expressed genes and ordered by decreasing average signal.

isoforms (RMSD=0.374, Figure 5.5A). Interestingly, the xBD2 differs from BD2 in sequence and structure only in the last helix of the bromodomain. A proline between the last two

helices of BD2 introduces a loop which delays the formation of a helix by three additional positions, allowing the last helix to wrap around and form the final side of the bromodomain binding pocket (Figure 5.5B,C). In xBD2, this proline is replaced with a glycine, which AlphaFold predicts to allow the helix to form prematurely and considerably enlarge the xBD2 binding pocket (Figure 5.4A). The tyrosine gatekeeping residue of BD2 is replaced with an equally bulky phenylalanine, but this is not predicted to be sufficient to close off this side of the xBD2 pocket.

We developed 293T cells stably expressing GFP fused to the N-terminus of cTAF1 or xTAF1 to facilitate studies of the functional consequences of expressing xTAF1 *in vivo* (Figures 5.4B and 5.5D). We synthesized biotinylated H4 peptides which were either unmodified or acetylated at lysines 5 and 12 (H4K5ac/12ac) and performed an *in vitro* pull-down with whole-cell lysate. As expected GFP-cTAF1 associated with H4K5ac/12ac peptides and not unmodified peptides, but GFP-xTAF1 signal was not detected above background for either peptide (Figures 5.4C and 5.5E). This confirms that the binding preference of xBD2 differs from BD2 and motivated us to check how the genome-wide distribution may differ between cTAF1 and xTAF1.

We performed ChIP-seq with an antibody raised against GFP to characterize the chromatin binding pattern of cTAF1 and xTAF1. We found that xTAF1 is distributed similarly across the genome and is most frequently observed in promoter regions, demonstrating that xTAF1 has the capability to bind promoter DNA and association to acetylated H4 tails is not required for TAF1 recruitment (Figure 5.5F). Interestingly, we observed a minor upstream peak in cTAF1 ChIP-seq data that was not present in xTAF1 data (Figure 5.4D). The TAF1 bromodomains are generally thought to associate with the +1 nucleosome, but our results suggest that TAF1 bromodomains associate with acetylated histones upstream from the promoter. Next, we used MACS2 to call peaks over promoters bound by cTAF1 or xTAF1 and found that xTAF1 ChIP-seq signal is similar over both promoter sets. On the other hand, cTAF1 ChIP-seq signal is weaker than over promoters bound by xTAF1 than over promoters bound by cTAF1 (Figure 5.4E). Next, we performed CUT&RUN for histone H3 lysine 4 trimethylation (H3K4me3) and histone H4 lysine 5, 8, or 12 acetylation (H4K5ac/8ac/12ac) but found no major differences, suggesting that changes in histone

modification patterns are not responsible for the observed changes in TAF1 binding (Figures 5.4F,G and 5.5G,H). Together, these data suggest that xTAF1 binds a subset of promoters more strongly than cTAF1, and this could be facilitated through a relief of an inhibitory effect enforced by TAF1 BD2 binding activity.

5.4.4 xTAF1 perturbs early steps of RNA synthesis

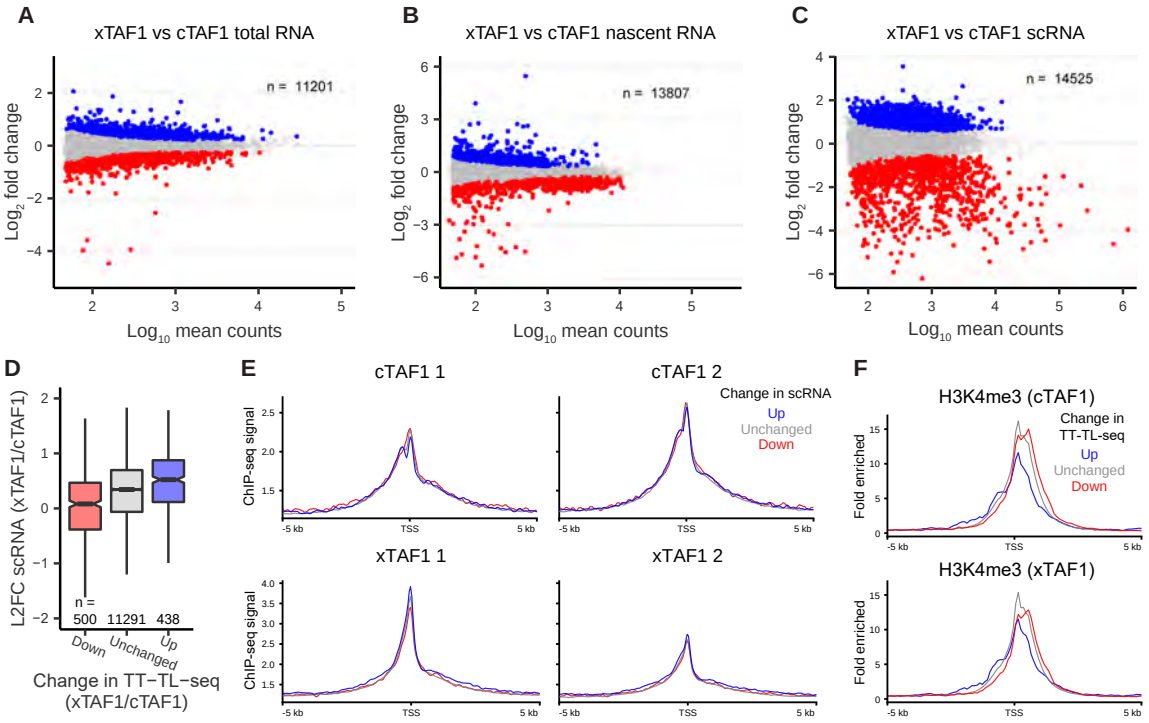

Figure 5.6: **Perturbations in early transcription caused by xTAF1 expression are associated with distinct chromatin states** (A-C) MA plots comparing expression of mature transcripts (RNA-seq, A), nascent RNA (TT-TL-seq, B), and scRNA at promoters (STL-seq, C) in xTAF1-expressing cells against cTAF1-expressing cells. Transcripts, genes, or promoters are highlighted if the DESeq2 $p_{adj} < 0.05$. (D) Genes were grouped by their change in TT-TL-seq data as in B. The $\log_2$ fold change of in scRNA reads from STL-seq data of all TSSs associated with the gene are plotted. (E) Metaplots of cTAF1 or xTAF1 ChIP-seq data centered on the TSS of genes determined by STL-seq data and separated by whether the read counts at the TSS are upregulated, not significantly changed, or downregulated in STL-seq data. (F) Metaplots of H3K4me3 CUT&RUN data centered on the the TSS of genes determined by STL-seq data and separated by whether the gene is upregulated, not significantly changed, or downregulated in TT-TL-seq data.

To better understand the consequences of mutating the TAF1 BD2 binding pocket on the function of TAF1 as a TFIID subunit, we performed a comprehensive characterization

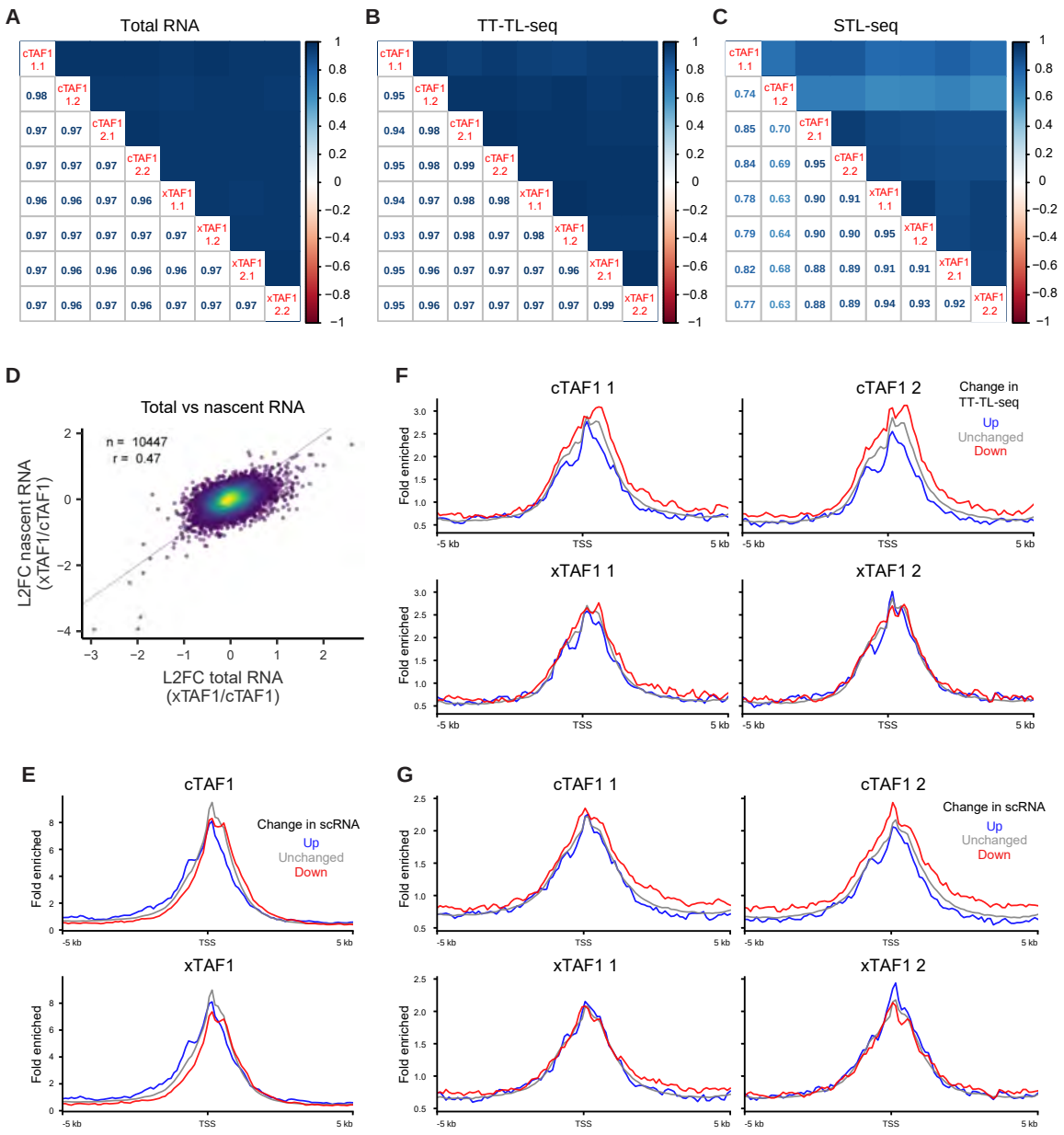

Figure 5.7: **Characterizing the transcriptome of cells expressing cTAF1 and xTAF1** (A-C) Correlation plots of total read counts in total RNA-seq (A), TT-TL-seq (B), and STL-seq (C) data. (D) Scatter plot comparing the $\log_2$ fold change in read counts from total RNA-seq and TT-TL-seq data. (E) Metaplots of H3K4me3 CUT&RUN data centered on the TSS determined by STL-seq data of genes and separated by whether scRNA read counts are upregulated, not significantly changed, or downregulated over the promoter in STL-seq data (F & G) Metaplots of H4Kac CUT&RUN data centered on the TSS of genes determined by STL-seq data and separated by whether read counts are upregulated, not significantly changed, or downregulated over the gene body in TT-TL-seq data (F) or the promoter in STL-seq data (G).

of RNA synthesis in cTAF1- and xTAF1-expressing cells with RNA-seq, TT-TL-seq, and STL-seq. We found good correlation across samples in all experiments (Figure 5.7A-C). At the level of total RNA, hundreds of transcripts are differentially expressed between cells expressing xTAF1 and cTAF1 (Figure 5.6A). xTAF1 expression leads to the same degree of differential transcriptional activity as measured by TT-TL-seq (Figure 5.6B), and we found agreement between changes in total and nascent RNA (Figure 5.7D). Similar to the patient-derived cell lines, this suggests that changes in early transcription causally lead to altered expression in total RNA. Considering read counts from STL-seq data to probe how promoter-proximal paused RNAPII is distributed across the genome, we found thousands of alternatively used TSSs (Figure 5.6C). While the strongest affects resulted in a loss of scRNA, most promoters gained scRNA and this affect was strongest at the TSSs of transcriptionally induced genes (Figure 5.6D). This indicates that the frequency of transcription initiation in cells expressing xTAF1 is slightly larger than in cells expressing only cTAF1 and the increased initiation frequency leads to more RNAPII entering elongation.

Next, we explored how the local chromatin environment is related to the changes observed in RNAPII behavior at promoters. First, we asked if the distinct binding profiles of cTAF1 and xTAF1 could lead to changes in early transcription. Binning by change in STL-seq read counts, promoters which gain scRNA upon xTAF1 expression show the most prominent upstream peak in cTAF1 ChIP-seq data while the profiles are identical in xTAF1 data (Figure 5.6E). Promoters which lose the upstream peak are therefore associated with an increased frequency of transcription initiation upon xTAF1 expression. Next, we reasoned that if TAF1 mediating TFIID recruitment to acetylated H4 histones is an important PIC assembly mechanism, ablation of this recruitment mechanism may alter the transcriptional program depending on the landscape of H4Kac or other histone modifications. We binned promoters by the change in their transcriptional output in TT-TL-seq data and generated metaplots of H3K4me3 CUT&RUN data (Figure 5.6F). Interestingly, the H3K4me3 profiles of each group differed in a similar manner independent of the expressed TAF1 isoform. A strong peak associated with the +1 nucleosome is ubiquitously present, but transcriptional downregulation upon xTAF1 expression is associated with a stronger H3K4me3 peak immediately downstream the +1 nucleosome peak. Similar H3K4me3 profiles were also observed

when performing the same analysis with STL-seq read counts (Figure 5.7E). The same trends were observed to a lesser degree in H4Kac CUT&RUN data from cells, and we also found that upregulated promoters/genes present with a minor peak upstream of the TSS (Figure 5.7F,G). Given the *in vitro* pull-down and cTAF1 ChIP-seq results, it is tempting to speculate that cTAF1 BD2 associates with this population of acetylated H4 histones.

Taken together, these results suggest that the cTAF1 binding profile and local chromatin environment impact how genes are differentially regulated when xTAF1 is expressed. This provides preliminary evidence that TAF1 BD2 binding activity has an inhibitory effect on gene expression at the level of initiation. While we observe a general increase in initiation, we did not observe a global bias in one direction for transcription or expression. Therefore, xTAF1 must have a separate effect on the kinetics of promoter-proximal pausing which further complicates the transcriptional response.

5.4.5 xTAF1 destabilizes and redistributes the position of promoter-proximal paused RNAPII

Finally, we performed a more in-depth analysis of our STL-seq data in xTAF1-expressing cells to determine the effects of the TAF1 BD2 mutant on promoter-proximal pausing. bakR analysis of more nearly 3,500 promoters revealed a general upregulation in the observed first order rate constant (k_{STL}) of promoter-proximal paused RNAPII turnover (Figure 5.8A). Promoters which lost scRNA were also much more likely to be associated with the increase in k_{STL} (Figure 5.8B). As promoters with upregulated initiation rates were associated with an increase in transcriptional activity, it is unlikely that this increase in turnover can be attributed to an increase in release into elongation. Therefore, the most likely explanation is that this increase in turnover is attributed to an increase in premature termination of RNAPII.

Next, we sought to determine if xTAF1 expression could recapitulate the redistribution of RNAPII pausing position that was found in XDP patient-derived cells. We found that xTAF1 expression produces a similar pausing phenotype relative to cTAF1-expressing cells in which a relative loss of scRNA shorter than 30 nts is observed (Figure 5.8C). Furthermore, normalization of these data with *Drosophila* RNA spike ins allows us to quantitatively de-

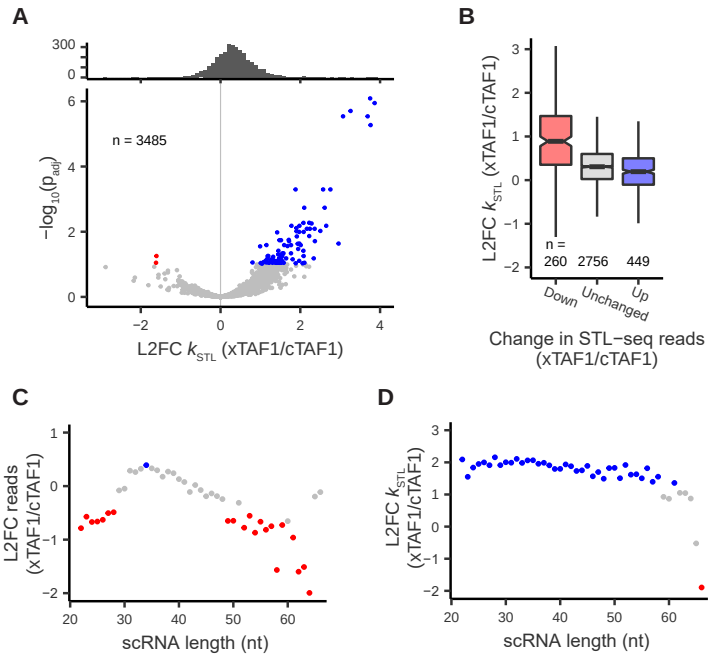

Figure 5.8: **xTAF1 perturbs RNAPII promoter-proximal pausing** (A) Volcano plot the $\log_2$ fold change in the observed first order rate constant (k_{STL}) for promoters in STL-seq data as measured by bakR. Points are highlighted if $p_{adj} < 0.1$. A histogram for the $\log_2$ fold change values is shown above. (B) Promoters were grouped by their change in STL-seq read count data as in 5.6C. The $\log_2$ fold change of in k_{STL} from STL-seq data are plotted. (C) The $\log_2$ fold change of absolute read counts in STL-seq comparing xTAF1-expressing cells to cTAF1-expressing cells. scRNA reads were grouped by their absolute length. (D) The $\log_2$ fold change of k_{STL} from STL-seq comparing xTAF1-expressing cells to cTAF1-expressing cells. scRNA reads were grouped by their absolute length.

termine that the relative shift is attributable to an absolute loss of the shortest transcripts. This phenomenon could be caused by a downstream shift in the pause position or destabilization of RNAPII paused less than 30 bp from the TSS. We used bakR to analyze the pause site turnover dependent on the length of the read rather than individual promoters. This analysis revealed that scRNAs up to 60 nts in length are similarly destabilized in xTAF1-expressing cells relative to cTAF1-expressing cells (Figure 5.8D). Therefore, xTAF1 expression leads to RNAPII transcribing farther on average before undergoing promoter-proximal pausing.

5.5 Discussion

Here we provide evidence that the X-Linked Dystonia Parkinsonism (XDP) causal SVA retrotransposon insertion in intron 32 of *TAF1* gives rise to a truncated *TAF1* transcript (*xTAF1*) with translation potential. It is clear that canonical, full-length *TAF1* isoform (*cTAF1*) is repressed in XDP patient-derived cell lines, but is downregulated only up to 50% depending on the patient and cell-type. Importantly, our data indicate that RNAPII is not differentially initiated from the *TAF1* promoter, suggesting that the downregulation of *TAF1* can be attributed to loss of coverage over exons 33-38. It is possible that cTAF1 insufficiency may produce a similar phenotype; however, our data suggest that total copy numbers of *TAF1* transcripts may be unchanged in XDP patient-derived cells. To date, a consensus molecular model for the disease mechanism of XDP does not exist beyond the SVA insertion. We propose that XDP is a transcriptomopathy caused by an insufficiency of TAF1 BD2 activity. This model provides direction to improve our understanding of XDP pathogenicity and reveals a new fundamental role for TAF1/TFIID function.

Many challenges in studying XDP are derived from the lack of a model system. Establishing XDP-patient derived cell lines was an important step for XDP researchers and facilitated the work presented here, but did not immediately reveal how the SVA insertion gives rise to XDP [125]. In combination with polyA+ sequencing data from Aneichyk et al. (2018), our TimeLapse-seq and TT-TL-seq data provide evidence for a cryptic cleavage and polyadenylation site (PAS) that is either activated or introduced by the SVA insertion. Interestingly, the cryptic PAS does not induce transcription termination for the majority of RNAPII elongation complexes (ECs) that transcribe over the region. It is known that a longer SVA hexanucleotide repeat sequence is associated with younger age at onset [124, 189]. It may be that the hexanucleotide repeats slow RNAPII elongation and increase the opportunity for a cleavage event to occur. If this is the case, more repeats would also be associated with a larger proportion of the total TAF1 population that is truncated, and this could lead to manifestation of XDP symptoms at a younger age. It is unclear why cleavage would be induced for a subset of transcriptional events but this would be an interesting line of investigation for future work.

We developed a cell model coexpressing xTAF1 and the endogenous cTAF1. Excitingly, the xTAF1-expressing cells recapitulated the promoter-proximal paused RNAPII redistribution observed in XDP-patient derived cells. If found to reproduce across other stages of development, and if the effect is more severe at early stages of development where the SVA has a larger impact on synthesis of the *cTAF1* transcript, this would be the first XDP-specific molecular phenotype. Utilizing this phenotype as a readout of other XDP model systems may prove to be a valuable tool. Recapitulating the pausing phenotype with xTAF1 expression in human cells also gives credence to our proposal that the xTAF1 protein is expressed in XDP patients. AlphaFold predicts that the xBD2 binding pocket is significantly enlarged relative to that of wild type BD2. This presents an opportunity to leverage the unique binding pocket as a therapeutic target. We predict that the lack of xTAF1 BD2 affinity for histone modification bound by cTAF1 BD2 relieves an inhibitory effect on transcription initiation (Figure 5.9). Therefore, in order to target the pathogenic protein as a therapy, inhibition is not sufficient and degradation or deletion would be required. Designing a Protein Targeting Chimera (PROTAC) small molecule that specifically recognizes the xBD2 pocket would be an efficient strategy to prevent the protein from causing a dysregulation of pausing and transcription in patients [208].

xTAF1 provides insight into the biological relevance for cTAF1 BD2 histone binding activity (Figure 5.9). It has been previously shown, and we confirmed, that the cTAF1 tandem bromodomains bind acetylated lysines on histone H4 tails (H4Kac), but the relevance for this activity has never been pursued further [19]. The field assumes that TAF1 association with histone acetylation is a mechanism of recruitment and promoter recognition for the TFIID complex. However, counter to this conventional line of thinking, TAF1 binding to promoters is stronger and initiation rates globally increase when TAF1 H4Kac binding activity is destroyed. Furthermore, we found evidence for the association of cTAF1 BD2 with histones upstream of the TSS, another result contrary to the traditionally assumed behavior of TAF1 [209]. Strikingly, we found that the strongest changes in TAF1 binding profiles at upregulated promoters, raising the possibility that BD2 plays a role in fine-tuning the kinetics associated with early steps of transcription. We found a unique chromatin signature depending on the promoter response, suggesting that the xTAF1-associated phenotypes

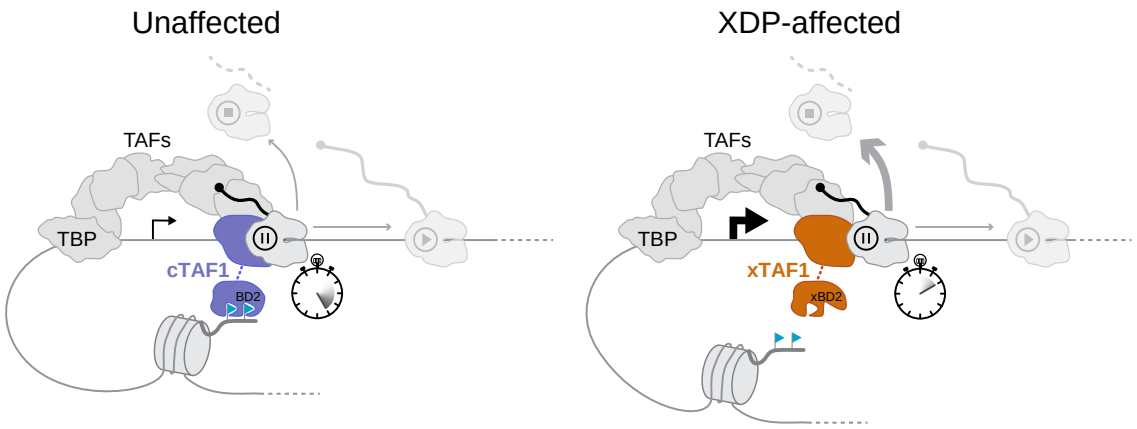

Figure 5.9: **Model for XDP pathogenesis and TAF1 BD2 function.** In XDP-affected cells, we propose that a novel TAF1 isoform (xTAF1) is expressed as a result of the SVA insertion in intron 32 of *TAF1* which encodes the canonical TAF1 isoform (cTAF1). TAF1 contains two tandem bromodomains (BD) near the C-terminus, and in xTAF1 the binding pocket of BD2 is mutated. This mutation ablates TAF1 binding activity towards acetylated histone H4 tails (blue flags) upstream of the promoter and perturbs the normal behavior of RNAPII in early transcription. Loss of BD2 binding activity results in an increased initiation rate, faster turnover at the pause site attributed to an increase in premature termination, and a lengthening of the average distance RNAPII transcribes before undergoing promoter-proximal pausing.

are related to specific TAF1-chromatin interactions. As the chromatin landscape is highly cell type specific, this may explain why it has been difficult to detect correlations in XDP-induced expression changes across stages of development in patient-derived cell lines. Lastly, the redistribution of promoter-proximal paused RNAPII in xTAF1-expressing cells suggests that BD2 association with histones coordinates the pause site. Therefore, this leads us to conclude that a matrix of information sharing is facilitated by a network of contacts between DNA upstream of the TSS, nucleosomes, TAF1/TFIID, RNAPII, and DNA downstream of the TSS to determine exactly where RNAPII will pause following transcription initiation. It has never been shown if the single base pair position of the pause site is an important determinant of the RNA synthesis pathway, but our results suggest it is tightly controlled.

Finally, our work on XDP exemplifies the value in investing in studies of rare human genetic disorders. By the nature of essential factors such as members of the preinitiation complex (PIC), mutations in these genes are not associated with common medical conditions. Unfortunately, this results in small patient populations with untreatable conditions

and represents missed opportunities to reveal new, fundamental biology. More than a dozen mutations in TAF1 alone give rise to neurodevelopmental phenotypes, most commonly intellectual disability [121,128,195–198]. Many of these mutations are not within the tandem bromodomains or the DNA-binding domains of TAF1 suggesting that other domains play important roles in cells. Considering the shear number of individual proteins in the PIC, the transcription field has likely overlooked many other essential functions. Moving forward, addressing the needs of patient populations afflicted with rare genetic disorders while deeply investigating the molecular phenotypes would be an impactful approach to advance biology and medicine simultaneously.

Chapter 6

Methods and data analysis

6.1 Methods

6.1.1 Cell lines and s^4U metabolic labeling

Metabolic labeling of cells was performed as described previously [97]. For STL-seq in *Drosophila*, S2 cells were grown to approximately 3-4 million cells/mL and spiked with s^4U (1 mM). Cells were incubated at 27°C for the appropriate labeling time, fully resuspended and transferred to 4-5 volumes of ice-cold PBS in an ice bath. Cells were pelleted by centrifuging at 500Xg for 5 min. PBS was removed, cells were resuspended in 1 mL TRIzol, and frozen at -80°C. For STL-seq and TT-TL-seq, HEK293T cells were grown to approximately 70% confluency when the media was spiked with s^4U (1 mM). For all NR-seq experiments, HEK293T cells were grown to approximately 70% confluency when the media was spiked with s^4U (100 µM). For all STL-seq and TT-TL-seq experiments and NR-seq experiments with improved handling conditions, plates were immediately placed on ice and washed with ice-cold PBS. Cells were scraped from plates, transferred to low nucleotide-binding tubes, and pelleted by centrifuging at 500Xg for 5 min. PBS was removed and cells were resuspended in 1 mL TRIzol and frozen at -80°C. For all NR-seq experiments with dropout handling conditions, ice cold TRIzol was added directly to plates, pipetted up and down to spread across the plate, and left on ice for 5 min to fully lyse the cells. Cellular lysate was then transferred to a separate tube and frozen at -80°C.

When spike ins were performed for STL-seq and TT-TL-seq, HEK293T cells were spiked into S2 TRIzol samples at 5% by cell count. Total *Drosophila* RNA was spiked into total human RNA at 5% by mass.

6.1.2 Generation of cell lines expressing GFP-TAF1

The GFP-TAF1 plasmid was a gift from Kyle Miller (Addgene plasmid # 65395). The GFP-TAF1 plasmid was modified to remove the C-terminal V5 tag and then introduce the xTAF1 truncation. Wild type HEK293T cells were transfected with plasmid encoding either GFP-cTAF1 or GFP-xTAF1 using the Lipofectamine 3000 reagent. Cells were grown at 37°C for 48 h when blasticidin was introduced at a final concentration of 2 ng/μL. Cells were allowed to further expand until stably expressing colonies appeared. These colonies were picked and allow to expand in separate culture dishes and sorted by FACS.

6.1.3 Immunoblots

Cells were lysed in RIPA buffer. 1 µL of benzonase was added to the lysate and incubated at 4°C for 1 h with rotation. The lysate was cleared by spinning at max for 10 min at 4°C. The supernatant was transferred to a fresh 1.5 mL tube. Lysate was electrophoresed on an SDS-PAGE gel for 45 min at 200 V and transferred to a PVDF membrane. The membranes were blocked with 5% milk for 1 h with shaking, incubated with primary for 1 h at room temperature or overnight at 4°C, incubated with secondary antibody for 1 h at room temperature or overnight at 4°C. The membranes were washed 3X with TBST buffer and 1X with PBST. The Western blot was developed with ECL reagent and imaged with chemiluminescence.

6.1.4 *In vitro* histone tail pull-down assays

Cells were lysed in RIPA buffer. 1 µL of benzonase was added to the lysate and incubated at 4°C for 1 h with rotation. The lysate was cleared by spinning at max for 10 min at 4°C. Per reaction, 55 µL MyOne C1 streptavidin resin was loaded with 25 µg H4 tail peptide in PBS and incubated at room temperature for 30 min with rotation. Beads were washed 4X with 200 µL PBS and resuspended in 55 µL PBS per reaction. 50 µL of beads were

aliquoted into separate PCR tubes and washed once with 200 µL bead wash buffer (20 mM HEPES, pH 8.4, 5% glycerol, 0.2 mM EDTA, 0.1% Triton X-100, 150 mM NaCl, protease inhibitors). The beads were captured on a magnet, the supernatant was removed, 50 µg of protein extract was loaded onto each sample, and reactions were incubated overnight at 4°C with rotation. Beads were washed 5X with 1 mL wash buffer (20 mM HEPES, pH 8.4, 5% glycerol, 0.2 mM EDTA, 0.1% Triton X-100, 350 mM NaCl, protease inhibitors) and eluted in two round by adding 25 µL SDS loading buffer and heating at 95°C for 5 min.

6.1.5 Drug and KCl treatments

For STL-seq and TT-TL-seq, *D. melanogaster* S2 cells were treated with 42 µM 20-hydroxyecdysone or DMSO for 30 min. S2 were treated with either 10 µM Triptolide for 10 min, 500 nM Flavopiridol for 40 min, or DMSO for the same time as a control. For combined Flavopiridol and 20-hydroxyecdysone treatments, S2 cells were pretreated with 500 nM Flavopiridol for 10 min before adding 20-hydroxyecdysone directly to cell media. Labeling times were always the last 5 min of any treatment.

To induce hyperosmotic stress, HEK293T cells were treated with 80 mM KCl for a total of 1 h as described in [146]. Metabolic labeling was performed during the last 5 min of stress.

6.1.6 NR-seq (TimeLapse-seq, SLAM-seq, TUC-seq)

Genomic DNA was depleted by treating with TURBO DNase and total RNA was extracted with one equivalent volume of Agencourt RNAClean XP beads according to manufacturer's instructions. 5 µg of total RNA was subjected to TimeLapse, SLAM, or TUC chemistry as previously described with some modifications [97–99].

For TimeLapse-seq, RNA was mixed with 600 mM TFEA or NH_3, 1 mM EDTA and 100 mM sodium acetate pH 5.2 or 100 mM Tris pH 7.4. Then, $NaIO_4$ or mCPBA was added to 10 mM final and the reaction was incubated at 45°C for 1 h. RNA was purified with one volume of Agencourt RNAClean XP beads and eluted with nuclease-free water. RNA was mixed with 10 mM DTT, 10 mM Tris pH 7.4, 5 mM EDTA, and 50 mM NaCl and incubated at 37°C for 30 min. RNA was purified with one volume of Agencourt RNAClean

XP beads and eluted with nuclease-free water.

For SLAM-seq, RNA was mixed with 50% DMSO, 50 mM sodium phosphate buffer pH 8.0 and 10 mM IAA and incubated at 50°C for 15 min. The reaction was stopped by adding excess DTT. RNA was purified with one volume of Agencourt RNAClean XP beads and eluted with nuclease-free water.

For TUC-seq, RNA was mixed with 180 mM NH_4Cl and 450 µM OsO_4 and incubated at 25°C for 3 h. RNA was purified with one volume of Agencourt RNAClean XP beads and eluted with nuclease-free water.

For each sample, 10 ng of RNA input was used to prepare sequencing libraries from the Clontech SMARTer Stranded Total RNA-Seq kit (Pico Input) with ribosomal cDNA depletion. Libraries were sequenced on a NovaSeq 6000 2X100bp.

6.1.7 STL-seq

Total RNA from S2 and 293T cells suspended in TRIzol was purified as described previously with minor changes [97]. Following TRIzol extraction, RNA was precipitated with one volume of isopropanol supplemented with 1 mM DTT. Extracted RNA was immediately subjected to TimeLapse chemistry as previously described with minor modifications. The oxidant used was meta-chloroperoxybenzoic acid (mCPBA) to avoid modifying the 3´ ends of RNA and interfering with downstream ligations. All purifications with Agencourt RNAClean XP beads were performed with 2 volumes of beads and supplemented with isopropanol to improve recovery of short RNA. Start-seq was performed on total RNA essentially as previously described with minor modifications [57]. Briefly, total RNA was electrophoresed on a 15% denaturing Urea-PAGE gel for 1 h at 200 V. RNA between the sizes of ∼20 and ∼80 nt was excised, extracted from the gel with a crush-soak method, and ethanol precipitated. The short RNA was then treated successively with RNA 5´ polyphosphatase (VWR), Terminator 5´-phosphate-dependent exonuclease (Lucigen), and ligated to a custom, pre-adenylated DNA adapter with T4 RNA ligase 2 truncated (NEB). Short, capped RNA was then electrophoresed on a 15% denaturing Urea-PAGE gel for 1 h at 200 V. RNA between the sizes ∼40 and ∼100 nt was excised, extracted, and ethanol precipitated. Ligated RNA was treated successively with calf intestinal alkaline phosphatase

(NEB), RNA 5´ Pyrophosphohydrolase with ThermoPol buffer (NEB), and T4 RNA ligase 1 (NEB) to ligate a custom RNA oligo. RNA was reverse transcribed with SuperScript RT III and finally amplified with Phusion polymerase. Amplified libraries were purified by electrophoresis on a 6% native TBE PAGE gel, extraction, and ethanol precipitation. Libraries were sequenced either on a NovaSeq 6000 2X100bp or HiSeq 4000 2X150bp.

6.1.8 TT-TL-seq

Where paired STL-seq and TT-TL-seq data exists, RNA previously collected for STL-seq was used for TT-TL-seq. Genomic DNA was depleted by treating with TURBO DNase and total RNA was extracted with one equivalent volume of Agencourt RNAClean XP beads according to manufacturer's instructions. 50 µg of total RNA was subjected to MTS chemistry and biotinylation followed by streptavidin enrichment essentially as previously described [97]. TimeLapse chemistry was performed as described above. For each sample, 10 ng of RNA input was used to prepare sequencing libraries from the Clontech SMARTer Stranded Total RNA-Seq kit (Pico Input) with ribosomal cDNA depletion. Libraries were sequenced on a NovaSeq 6000 2X100bp.

6.1.9 ChIP-seq

30 million HEK293T cells grown to ∼75% confluency were used to perform ChIP-seq for every sample. Formaldehyde was added to media to 1% final and cells were crosslinked for 10 min. Formaldehyde was quenched with 125 mM glycine for 5 min, followed by 2 rinses with cold PBS. Cells were scraped from plates in 5 mL cold PBS and pelleted by spinning at 500Xg for 5 min at 4°C. Supernatant was removed and pellets were frozen at -80°C. Pellets were thawed and cell were resuspended in 4 mL lysis buffer (50 mM HEPES pH 7.9, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100), incubated on ice for 10 min, and pelleted at 4°C for 5 min at 4,000 RPM. The upernatant was removed and the pellet was washed twice with 3 mL wash buffer (10 mM Tris-Cl pH 8.1, 200 mM NaCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0). Samples were spun at 4,000 RPM at 4°C after each wash. The pellet was gently washed without resuspension twice with 1.5 mL shearing buffer (0.1% SDS, 1 mM EDTA, 10 mM pH 8.1). The pellet was then resuspended

in 1 mL shearing buffer and sheared in a 1 mL Covaris tube at 140 W, 5% duty cycle, and 200 burst/cycle for 12 min. 115 µL 10% Triton X-100 and 34.5 µL 5 M NaCl and spun in 1.5 mL Eppendorf tubes at max speed for 10 min at 4°C. 5 µg GFP antibody was added to each sample. Samples were incubated at 4°C overnight, then added to 30 ul Protein G DynaBeads that had been pre-equilibrated with 0.5 ml ChIP Shearing Buffer. Then, samples were incubated at 4°C for 1.5 h. The beads were washed with 1 mL for 5 min for each of the following steps: 2x ChIP low salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Hepes-KOH pH 7.9, 150 mM NaCl), 2x high salt buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Hepes-KOH pH 7.9, 500 mM NaCl), 1x LiCl buffer (100 mM Tris-HCl pH 7.5, 0.5 M LiCl, 1% NP-40, 1% Sodium deoxycholate), 1x TE buffer. Beads were resuspended in 100 µL PK buffer (20 mM HEPES pH 7.9, 1 mM EDTA, 0.5% SDS) and 2 µL proteinase K was added. Samples were incubated at 50°C for 30 min with shaking. Beads were captured with a magnet, the supernatant was collected, and 3 µL 5M NaCl and 0.5 µL RNase A were added. Samples were incubated overnight at 65°C. 1.5 µL proteinase K was added and samples were incubated for 1 h at 50°C with shaking. DNA was purified using a Qiaquick PCR purification kit and eluted in 100 µL water.

Eluted DNA was mixed with 13 µL NEB T4 ligase buffer (M0202S), 5 µL 10 mM dNTPs, 5 µL NEB T4 polynucleotide kinase (M0201L), 5 µL NEB T4 DNA polymerase (M0210S), 1 µL NEB DNA Polymerase I, Large Klenow fragment (M0203S). The sample was mixed and incubated at 20°C for 30 min. DNA was purified using a Qiaquick PCR purification kit and eluted in two rounds of 16 µL water. Next, purified DNA was mixed with 5 µL NEB buffer #2, 10 µL 1 mM dATP (N0440S), 3 µL NEB Klenow fragment (M0212S). Samples were incubated at 37°C for 25 min. 1 µL of RNase (DNase free, Roche 11119915001) and incubated at 37°C for 5 min. DNA was then purified with 90 µL SPRI beads and eluted in 8 µL water. Purified DNA was mixed with 12.5 µL NEB blunt/TA ligase mastermix (M0367L), 2 µL 1 µM of each Illumina adapter, and 2.5 µL water and incubate at 25°C for 15 min. DNA was then purified with 40 µL SPRI beads and eluted in 15 µL water. Finally, 5 µL of each sample was mixed with 25 µL 2X Phusion mastermix (M0531S), 2 µL 10 µM Illumina universal primer, 2 µL 10 µM Illumina index primer, and 16 µL and samples were PCR amplified for a total of 12 cycles. The final libraries were purified with 75 µL SPRI

beads and eluted in 15 µL water. Libraries were sequenced on a NovaSeq 6000 2X100bp.

6.1.10 CUT&RUN

CUT&RUN was performed according to manufacturer's instructions provided in the CUTANA™ ChIC/CUT&RUN Kit (EpiCypher, 14-1048). Briefly, 11 µL of ConA beads were activated per reaction washing beads twice in 100 µL cold bead activation buffer per reaction and resuspending in 11 µL cold bead activation buffer per reaction. Per reaction, 500,000 cells grown to 75% confluency were washed with 5 mL room temperature PBS, scraped,, pelleted by spinning at 600Xg for 3 min at room temperature, transferred to a fresh 1.5 mL tube in 1 mL PBS, pelleted again, and the supernatant was removed. Cells were washed twice with 100 µL wash buffer per reaction and resuspended in 100 µL wash buffer per reaction. 10 µL of ConA beads were added to cells per reaction and incubated at room temperature for 10 min. The celll/bead slurry was captured on magnets, the supernatant was removed, 50 µL antibody buffer was quickly added per reaction, and samples were gently vortexed. Reactions were split into separate PCR tubes, 0.5 µg of primary antibody was added to each reaction, and samples were incubated at 4°C overnight with nutation. The cell/bead slurry was captured on a magnet, the supernatant was removed, and washed twice with 200 µL cold digitonin buffer. The slurry was resuspended in 50 µL cold digitonin buffer by gently vortexing. Next, 2.5 µg pAG-MNase was added to each reaction, the samples were gently vortexed, and incubated at room temperature for 10 min. The slurry was captured on magnets, washed twice with 200 µL cold digitonin buffer, resuspended in 50 µL cold digitonin buffer, and gently vortexed to resuspend. Next, 1 µL of 100 mM $CaCl_2$ was added to each reaction, samples were gently vortexed, and incubated at 4°C 2 h with nutation. To stop the reaction, 33 µL stop buffer was added to each reaction, samples were vortexed, and reactions were incubated at 37°C for 10 min. Samples were spun briefly, captured on magnets, and the supernatant was transferred to fresh PCR tubes. DNA was purified with provided cleanup columns and eluted in 12 µL water.

Sequencing libraries were prepared from CUT&RUN DNA were prepared according to manufacturer's instructions prepared in CUTANA™ CUT&RUN Library Prep Kit (EpiCypher, 14-1001). Briefly, end repair was performed by mixing 5 ng of CUT&RUN DNA with 0.1X

TE buffer up to 25 µL, 4.2 µL end prep buffer, and 1.8 µL end prep and incubating at 20°C for 20 min followed by 65°C for 30 min. Adapter ligation and U-excision was performed by adding 1.25 µL adapter, 16.5 µL ligation mix, and 0.55 µL ligation enhancer to each sample and incubating at 20°C for 15 min. Then, 1 µL U-excision enzyme was mixed with each sample and incubated at 37°C for 15 min. DNA was purified with 47.75 µL SPRI beads and eluted in 12 µL TE buffer. 10.5 µL of each elution was transferred to a fresh tube and to each 1 µL if i7 primer, 1 µL of i5 primer, and 12.5 µL high fidelity 2X PCR master mix was added to each sample. Libraries were amplified for a total of 14 cycles, DNA was purified with 25 µL SPRI beads, and eluted in 12 µL water. Libraries were sequenced on a NovaSeq 6000 2X100bp.

6.2 Data analysis

6.2.1 NR-seq and TT-TL-seq alignment and mutation calling

Filtering and alignment to the human GRCh38 genome version 26 (Ensembl 88) or the *Drosophila* dm6 genome (or a combined genome when using spike ins for normalization) were performed as described above for STL-seq with some differences. Reads were trimmed of adaptor sequences with Cutadapt v1.16 [210] and aligned to GRCh38 or dm6 using HISAT-3N [107] with default parameters and --base-change T,C (or HISAT2 [211] with --mp 4,2 where mentioned). Reads aligning to transcripts were quantified with HTSeq [212] htseq-count. SAMtools v1.5 [213] was used to collect only read pairs with a mapping quality greater than 2 and concordant alignment (sam FLAG = 99/147 or 83/163). Mutation calling was performed essentially as described previously [97]. Briefly, T-to-C mutations were only considered if they met several conditions. Mutations must have a base quality score greater than 40 and be more than 3 nucleotides from the read's end. Sites of likely single-nucleotide polymorphisms (SNPs) and alignment artifacts were identified with bcftools or from sites of high mutation levels in the non-s^4U treated controls (binomial likelihood of observation p ¡ 0.05). These sites were not considered in mutation calling. Browser tracks were made using STAR v2.5.3a [214]. Normalization scale factors were calculated with edgeR [215] using read counts from the spike-in species (calcNormFactors using method = 'upperquartile'). If

using spike ins for normalization, only reads aligning to the genome of the spike in species were used for normalization with edgeR.

6.2.2 Alignment of new and previously published sequencing data

PRO-seq sequencing data was filtered and aligned in the same manner as TT-TL-seq data except to use HISAT2 [211] with the default mismatch penalties, and mutation calling was not performed.

All ChIP-seq, CUT&RUN, ATAC-seq, and STARR-seq data were treated identically. Reads were filtered to remove duplicate sequences with FastUniq [216], trimmed of adaptor sequences with Cutadapt v1.16 and aligned to the *Drosophila* dm6 genome using the Bowtie 2 v2.2.9 [141]. SAMtools v1.5 was used to collect reads with a mapping quality greater than 2 and concordant alignment (sam FLAG = 0/16 for single-end data and 147/99 or 83/163 for paired-end data). MACS2 was used to call peaks [217].

Previously published Start-seq data was processed identically to STL-seq data to align reads and normalize counts.

6.2.3 STL-seq alignment, mutational analysis, and TSS calling

For STL-seq, filtering and alignment to the human GRCh38 genome version 26 (Ensembl 88) or the *Drosophila* dm6 genome were performed as described previously with some modifications [97]. Paired-end sequencing formats with 100 bp reads or longer caused low quality score for most second reads in each pair. Consequently, data was treated as single-end data by using only the first read in each pair. Reads were trimmed of adaptor sequences with Cutadapt v1.16 [210] and aligned to GRCh38 or dm6 using the Bowtie 2 option of Bismark v0.22.2 [140] with default parameters except --local. Bismark was used in concert with Bowtie 2 v2.2.9 [141]. Bismark alignment was a critical step as standard alignment software does not efficiently align short reads with one or more T-to-C mutations. Reads aligning to transcripts were quantified with HTSeq [212] htseq-count. SAMtools v1.5 [213] was used to collect only read pairs with a mapping quality greater than 2 and concordant alignment (sam FLAG = 0/16). Mutation calling was performed essentially as described previously [97]. Briefly, T-to-C mutations were only considered if they met several con-

ditions. Mutations must have a base quality score greater than 40 and be more than 3 nucleotides from the read's end. Sites of likely single-nucleotide polymorphisms (SNPs) and alignment artifacts were identified with bcftools or from sites of high mutation levels in the non-s^4U treated controls (binomial likelihood of observation p ¡ 0.05). These sites were not considered in mutation calling. Browser tracks were made using STAR v2.5.3a [214]. Reads which did not align in the initial alignment step were aligned to either the dm6 or GRCh38 genome (according to the spike-in species) in the same manner as above. Normalization scale factors were calculated with edgeR [215] using read counts from the spike-in species (calcNormFactors using method = 'upperquartile').

TSS calling was performed with TSScall to identify annotated (obsTSS) and unannotated (uTSS) transcription start sites [57]. Aligned sequencing reads from all samples of one species were pooled and analyzed with the TSScall pipeline. For *Drosophila* data, default settings were used except --annotation_search_window 500, --annotation_join_distance 100, and --call_method global. For human data, default settings were used except --annotation_search_window 1000, --annotation_join_distance 200, and --call_method big_winner. BEDTools [218] was used to assign reads to the nearest TSS within a 200bp window upstream and downstream of the read's ends. TSSs are considered to be promoters if they are classified as an obsTSS. See STARR-seq methods below for eTSS calling.

6.2.4 Estimation of RNA decay and synthesis kinetics

For all kinetic analyses of NR-seq data, the bakR R package was used to estimate RNA degradation rates and the change in these rates upon treatment (Vock et al., *in prep*). Default settings were used with the MCMC model (StanFit = TRUE). RNA synthesis rates and changes in synthesis rates were determined as outlined in the bakR manual using DESeq2 to estimate changes in total RNA.

6.2.5 Estimation of Pol II turnover with previous data under triptolide inhibition

To estimate scRNA half-lives from previously published Start-seq data under Trp inhibition, TSSs with low read counts in the uninhibited control samples were removed and all samples

were normalized to the control. The data were transformed to the log scale and each TSS was fit with a linear model. We found that normalized Start-seq signal increases at many TSSs upon Trp treatment, suggesting that Trp affects the kinetics of Pol II at the promoter-proximal pause site and making it an unreliable approach. This artifact produced negative $\hat{k_{\text{obs}}}$ estimates which are biologically impossible. TSSs demonstrating this behavior and were calculated to have a negative rate constant were removed from our analysis.

For previously published ChIP-nexus data under Trp inhibition, pausing half-lives were published with the associated transcript isoform.

6.2.6 Estimation of the new fraction of scRNA and kinetic parameters of scRNA

For data collected from xTAF1- and cTAF1-expressing cells, kinetic parameters from STL-seq were estimated using bakR with settings Ucut = 0.25, AvgU = 3, totcut = 100, and HybridFit = TRUE (Vock et al., *in prep*). For all other STL-seq samples treated with 5-min s^4U feeds were modeled with the same Binomial model. For each treatment, the number of uridines (n_u) and T-to-C mutations (TC_i) in each read is determined and reads are grouped by the TSS to which they map. The s^4U-untreated samples were used as unlabeled controls (c) to determine the background mutation rate attributed to reverse transcription mistakes, sequencing error, or other sources. The new fraction of scRNA (θ) and mutation rate were modeled as a mixture of two binomial distributions of either true TimeLapse or background mutations parametrized on the logistic scale. The probability mass function of the model is:

$$f\left(tc|n_u, p_n, p_o\right) = \theta \text{BinomialLogit}\left(tc|n_u, p_n\right) + \left(1-\theta\right)\text{BinomialLogit}\left(tc|n_u, p_o\right) \tag{6.1}$$

where p_n is the TimeLapse mutation rate in new transcripts and p_o is the background mutation rate. Under normal steady-state conditions, we assume an exponential model relating the new fraction of transcripts at the sth TSS and the observed turnover rate

constant for scRNA ($\hat{k}_{\text{obs}[s]}$) such that

$$\theta_{ss[s]} = 1 - e^{(-\hat{k}_{\text{obs}[s]}t)} \tag{6.2}$$

where t is the s^4U labeling time of the experiment. To estimate termination and release, the termination rate constant at each TSS ($\hat{k}_{\text{term}_{[}s]}$) was defined with an upper boundary of the total observed rate constant such that

$$\hat{k}_{\text{term}[s]} = \frac{\hat{k}_{\text{obs}[s]}}{e^{a_{[s]}}} \tag{6.3}$$

where a is a real value with lower limit of 0. The new fraction of transcripts under FP inhibition was related to $\hat{k}_{\text{term}}$ in the same manner as $\hat{k}_{\text{obs}}$.

$$\theta_{FP[s]} = 1 - e^{(-\hat{k}_{\text{term}[s]}t)} \tag{6.4}$$

The TSS specific pause release rate constant ($\hat{k}_{\text{rel}_{[}s]}$) was calculated as the difference between $\hat{k}_{\text{obs}_{[}s]}$ and $\hat{k}_{\text{term}_{[}s]}$. This parameterization of $\hat{k}_{\text{term}}$ and $\hat{k}_{\text{rel}}$ avoided cases where release is very slow, and the tail of the posterior distribution may extend into the negative range due to an unrestricted model.

To estimate these parameters, we used a Bayesian hierarchical modeling approach using RStan software (Version 2.19.3, [174]) that implements no-U-turn Markov Chain Monte Carlo (MCMC) sampling. We designed non-centered hierarchical models to estimate global TimeLapse mutation rate ($\bar{p}_{n[j]}$) for the j^{th} treatment condition while also allowing for variability by estimating TSS-specific mutation probabilities ($p_{n[j,s]}$). For the background mutation rate, we estimated a single global parameter ($\bar{p}_o$) while allowing for local variation among TSSs by estimating TSS-specific mutation probabilities ($p_{o[s]}$). We used weakly informative priors for global mutation rates on the logistic scale which covered the range of previously observed mutation rates that could be reasonably expected. The TSS-specific mutation rates were found by estimating a standard deviation (σ) for each global parameter and a TSS-specific z-score (z). Finally, s^4U-labeled and unlabeled control samples are indicated by I where if sample c is labeled with s^4U $I = 1$ and zero if the sample is

unlabeled.

Global parameter priors:

$$\bar{p}_o \sim \text{Normal}(-6, 0.5) \tag{6.5}$$

$$\bar{p}_{n[j]} \sim \text{Normal}(-2.5, 0.5) \tag{6.6}$$

$$\sigma_o \sim \text{HalfCauchy}(0, 1) \tag{6.7}$$

$$\sigma_{n[j]} \sim \text{HalfCauchy}(0, 1) \tag{6.8}$$

$$I_{[c]} = \begin{cases} 0, & \text{if } c \in \text{controls} \\ 1, & \text{otherwise} \end{cases} \tag{6.9}$$

$$s \in \{1, 2, \ldots, n_{\text{TSS}}\} \tag{6.10}$$

$$j \in \{1, 2, \ldots, n_{\text{treatment}}\} \tag{6.11}$$

$$\tag{6.12}$$

Local parameter priors:

$$z_{o[s]} \sim \text{Normal}(0, 1) \tag{6.13}$$

$$z_{n[j,s]} \sim \text{Normal}(0, 1) \tag{6.14}$$

$$p_{o[s]} = \bar{p}_o + \sigma_o z_{o[s]} \tag{6.15}$$

$$p_{n[j,s]} = \bar{p}_{n[j]} + \sigma_{n[j]} z_{n[j,s]} \tag{6.16}$$

$$\hat{k}_{\text{obs}[j,s]} \sim \text{Gamma}(0.5, 1.75) \tag{6.17}$$

$$a_{[j,s]} \sim \text{HalfNormal}(0, 2) \tag{6.18}$$

For reads $i \in \{1, 2, \ldots, n_{[s]}\}$:

$$f\left(tc_{[i]} \middle| \theta_{[j,s]}, n_{u[i]}, p_{n[j,s]}, p_{o[s]}\right) =$$

$$\prod_{i\,=\,1}^{n_{[s]}} \left(I_{[c]}\theta_{[j,s]}\text{BinomialLogit}\left(y_{[i]}|n_{u[i]}, p_{n[j,s]}\right) \; + \left(1 \; - I_{[c]}\theta_{[j,s]}\right)\text{BinomialLogit}\left(y_{[i]}\middle|n_{u[i]}, p_{o[s]}\right)\right) \tag{6.19}$$

The definition of θ in terms of $\hat{k}_{\text{obs}}$ and $\hat{k}_{\text{term}}$ is included in the model which allows retrieval of posterior distributions of all parameters. Similarly, $\hat{k}_{\text{rel}}$ is calculated within the model as a generated quantity thereby generating a posterior distribution of estimates. Fits from these models converged well at all TSSs when run on the complete dataset with a minimum average read cutoff in the s^{4}U-untreated controls (50 reads from fly cells and 100 reads from human cells). We limited our analysis to TSSs with an 80% CI size for $\hat{k}_{\text{obs}}$ was smaller than 1 on the natural log scale to avoid TSSs where we could not make a precise estimate. To identify the high confidence TSSs, we further limited our analysis to TSSs with an 80% CI that was smaller than 0.5 on the natural log scale. We only consider $\hat{k}_{\text{rel}}$ estimates to be high confidence if both $\hat{k}_{\text{obs}}$ and $\hat{k}_{\text{term}}$ for the TSS qualified as high confidence. In all cases, we report estimates of the parameters using the median value of the posterior distribution.

6.2.7 Estimation of the global effect of flavopiridol on premature termination

To assess if flavopiridol influences $\hat{k}_{\text{term}}$, we developed a model designed to test for flavopiridol-induced changes in turnover. This model is similar to the model described above. We defined a TSS-specific effect parameter ($f_{[s]}$) such that the turnover at a TSS under FP inhibition depends on $\hat{k}_{\text{obs}[s]}$ and $f_{[s]}$ as defined below

$$\hat{k}_{\text{term}[s]} = e^{f_{[s]}} \hat{k}_{\text{obs}[s]} \tag{6.20}$$

where f is unrestricted and the scaled value of $\hat{k}_{\text{obs}[s]}]$ is guaranteed to be greater than zero. Therefore, the definition of θ_{FP} is

$$\theta_{FP[s]} = 1 - e^{\left(e^{f_{[s]}} \hat{k}_{\text{obs}[s]} t\right)} \tag{6.21}$$

While the definition of θ_{SS} is unchanged in this model. In addition, a hierarchical parameter for the effect of FP (f_g) was defined so that the prior for the local effect at each TSS depends on the global effect of FP and the standard deviation of the global effect (σ_f).

Global parameter priors:

$$f_g \sim \text{Normal}(0, 1) \tag{6.22}$$

$$\sigma_f \sim \text{HalfNormal}(0, 1) \tag{6.23}$$

Local parameter prior:

$$f_{[s]} \sim \text{Normal}(f_g, \sigma_f) \tag{6.24}$$

Parameterizations and definitions for the other parameters described above remain unchanged. Estimates for all kinetic parameters were made within the same model and $f_{[s]}$ was transformed into the $\log_2$ fold change within the generated quantities of the model. This model converged well when run on the complete dataset with a minimum average read cutoff of 50 reads in the s^4U-untreated controls. We performed the same filtering as described above to determine high confidence TSSs under uninhibited and inhibited conditions. We identified TSSs where $\hat{k}_{\text{rel}}$ should be very close to zero by those whose gene-body coverage in TT-TL-seq was in the bottom 10% of all genes. As previously, the median value for the FP-induced $\log_2$ fold change at each TSS was used as a point estimate for the true value.

6.2.8 Simulation of STL-seq data

Local TSS-specific mutation rates were randomly chosen from a normal distribution centered on 0.1 and 0.0025 on the logistic scale for true TimeLapse (p_n) and background (p_o) mutations, respectively. The mean values were chosen based on observed mutation rates observed in previous TimeLapse data. For nreads as defined in the text, we simulated

scRNA reads from a TSS with half-life hl using the following model:

$$i \in \{1, 2, \ldots, n_{reads}\} \tag{6.25}$$

$$l_{[i]} \sim \text{ceiling}\big(\text{Normal}(35, 6)\big) \tag{6.26}$$

$$n_{u[i]} \sim \text{ceiling}\left(\frac{l_{[i]}}{nt}\right) \tag{6.27}$$

$$\theta = 1 - e^{\frac{-\log(2)}{hl}t} \tag{6.28}$$

$$X_{[i]} \sim \text{Bernoulli}(\theta) \tag{6.29}$$

$$TC_{[i]} \sim \begin{cases} \text{Binomial}(n_{u[i]}, p_n), & \text{if } X_{[i]} = 1 \\ \text{Binomial}(n_{u[i]}, p_o), & \text{otherwise} \end{cases} \tag{6.30}$$

where the i^{th} read contains $n_{u[i]}$ uridines which are evenly spaced along the read every nt nucleotides. The uridine frequency was chosen this way because scRNA initiated from the same TSS will contain identical sequences which only vary by the distance transcribed ($l_{[i]}$). The mean length and standard deviation were selected to closely reflect the true distribution of read lengths across all scRNA reads. The new fraction of reads (θ) depends on the half-life (hl) of scRNA and the s^4U labeling time (t) of the experiment. Whether a read is new was randomly assigned with a Bernoulli distribution with probability θ. If a read is new, the mutation rate and the number of mutations observed in a read ($TC_{[i]}$) is determined according to a binomial distribution with $n_{u[i]}$ trials and probability p_n. If a read is old, the number of mutations is determined similarly but with probability p_o. Five TSSs with half-lives 1, 2.5, 5, 7.5, and 10 min were simulated together with varying degrees of coverage (25 to 1000 reads) and treated as data from a STL-seq experiment. These data were modeled with the same binomial model described above to estimate the scRNA half-life. The estimated turnover rates were compared to the true values used as input to the simulation.

6.2.9 TT-TL-seq data analysis

RPKM was calculated with the total length of each transcript isoform. TSScall identified transcript isoforms associated with each called TSS. If a single isoform cannot be unam-

biguously assigned to a TSS, the longest isoform was chosen. Transcripts were grouped into equal quartiles by kinetic parameters of their TSS or subsets as defined in the text. Metaplots and heatmaps were produced with deepTools2 [219].

6.2.10 PRO-seq data analysis

As a measure of promoter-proximal pausing, PRO-seq reads were counted within the first 250bp downstream of every TSS called from STL-seq. To determine transcriptional activity over the gene body, PRO-seq reads were counted in the range of 251-1250bp downstream of every TSS called from STL-seq.

6.2.11 ChIP-seq and ATAC-seq data analysis

Aligned ChIP-seq reads for all datasets were counted within the first 500bp downstream of all TSSs identified in STL-seq data. Aligned ATAC-seq reads within the window of -200 to +100 around each TSS were counted. TSSs were grouped into equal quartiles by kinetic parameters and deepTools2 was used to generate metaplots and heatmaps.

6.2.12 STARR-seq data analysis and eTSS identification

To identify STARR-seq peaks from previously published data, aligned bam files of STARR-seq biological replicates from either a developmental core promoter (dCP) or housekeeping core promoter (hkCP) were merged and analyzed with the STARRpeaker tool using default parameters (except --mincov 1). Resulting peak calls for dCP and hkCP were merged and TSSs were assigned as a STARR-seq active TSS if they were within 500 bp of a peak. TSSs were considered to be enhancer TSSs (eTSS) if they are classified as a uTSS by TSScall and have STARR-seq enhancer activity.

6.2.13 Identification of Promoter motifs

PWMTools was used to search for consensus sequences of each motif within the specified window around annotated promoter TSSs identified in STL-seq data. Matches to the consensus sequence were not allowed to contain any mismatches.

Drosophila motif	Consensus motif	TSS search window
TATA box	STATAWAWR [220]	-110 to +1
Initiator +G (InrG)	TCAGTY [177, 220]	-5 to -1
Initiator -G (Inr)	TCAHTY [177, 178, 220]	-5 to -1
TCT motif	YYCTTTYY [183]	-5 to -1
Downstream promoter element (DPE)	KCGGTTSK [220]	+1 to +50
Motif ten element (MTE)	CSARCSSA [221]	+1 to +50
Pause Button (PB)	KCGRWCG [177]	+1 to +50

Table 6.1: *Drosophila* promoter motif sequences and the window searched around the TSS for the motif

Human motif	Consensus motif	TSS search window
TATA box	TATAWAAR [186, 222, 223]	-110 to +1
Initiator	YYANWYY [224]	-5 to -1
TCT motif	YCTYTYY [183]	-5 to -1
Downstream promoter element (DPE)	RGWYV [225]	+1 to +50
Motif ten element (MTE)	CSARCSSA [221]	+1 to +50

Table 6.2: Human promoter motif sequences and the window searched around the TSS for the motif

Appendix A

Start-TimeLapse-seq (STL-seq) protocol

A.1 Important notes to be aware of before beginning

1. DNase treatment is not necessary as genomic DNA is selected against in the size-selection step.

2. It is recommended to start with >10 million cells.

3. **CRITICAL!** Use **mCPBA** as oxidant during TimeLapse chemistry. $NaIO_4$ modifies RNA such that 3´ ligations cannot be performed.

A.2 s^4U treatment and cell harvesting

1. Plate and grow cells to ∼70% confluence

2. Supplement media with 1 mM s^4U and incubate cells for a time determined based on desired application (typically 5 min for STL-seq but can vary 1.5-10 min).

   Note: s^4U is photosensitive, keep solutions wrapped in foil and minimize exposure of samples to light.

3. After incubation period, immediately place cell culture plates on ice. Aspirate media from plate, gently rinse plate once with ice cold PBS and aspirate again.

4. Add 1 mL ice cold PBS to cells. Gently scrape cells from plate using a cell scraper, and transfer cell suspension to a 1.5 mL loBind epi tube.

5. Pellet cells in a pre-chilled (4°C) centrifuge at 700 x g for 3 min. Carefully aspirate PBS from cell pellet.

6. Resuspend pellet in 1 mL Trizol by gently pipetting up and down ∼10 times.

   Note: Trizol is toxic, use with care and in well ventilated area.

7. Trizol samples can be stored overnight at -80°C or kept on ice for RNA isolation.

A.3 RNA isolation

1. Thaw Trizol samples at room temperature. Once completely thawed, keep samples at room temperature for 5 min in the dark.

2. Add 200 µL chloroform to the 1 mL Trizol samples. Shake the tubes vigorously for 15 sec and let sit for 2 min in the dark (drawer is fine).

3. Optional – For easier separation of phases, transfer sample to pre-spun heavy phase-lock tubes.

4. Centrifuge the tubes for 5 min at 12,000 × g, 4°C. Transfer aqueous phase (∼500 µL) to new DNA loBind tubes with 1 µL RNase-free glycogen (20 µg).

5. To each aqueous phase from step 2, add 500 µL (1 eq) of 100% isopropanol with 1 mM DTT final concentration (*make 10 mL isopropanol + 10 µL of 1 M DTT master mix using freshly dissolved DTT). Invert tube ∼10 times or until thoroughly mixed. Incubate samples at room temperature for 10 min.

   Note: The s^4U -RNA is light sensitive and prone to oxidation. While these steps can be performed under standard laboratory lighting, try to minimize the time of

light exposure. The DTT is included to help minimize oxidation of the s^4U and reduce disulfides. DTT will oxidize over time. To be safe either make a fresh 1M stock from solid DTT or aliquot a 1M stock and only freeze thaw each aliquot 1-2 times.

6. Centrifuge samples 20 min at 20,000 × g, 4°C. Carefully remove the supernatant from the RNA/glycogen pellet.

7. Add 1 mL of room temperature freshly prepared 75% ethanol to the pellet, vortex quickly and centrifuge 3 min at 12,000 × g, 4°C.

8. Remove the ethanol completely from the RNA/glycogen pellet. First remove most of the ethanol with a 1000-µL (P1000) pipet tip, then spin the tubes again on a countertop microcentrifuge. Use a gel-loading/10 µL tip to remove the remaining ethanol. Let the pellet air-dry for 2 min. Be careful not to overdry, which will result in loss of RNA.

9. Resuspend each pellet in 16 µL (or desired volume) of nuclease-free (i.e., DEPC-treated water). Measure the RNA concentrations using a Nanodrop spectrophotometer.

A.4 TimeLapse chemistry (25 µL volume, scalable if needed)

1. Prepare RNA into 15 µL nuclease-free water.

2. Prepare TimeLapse master mix in the order shown (8.7 µL per sample).

   0.84 µL 3 M sodium acetate pH 5.2

   0.2 µL 500 m/molar EDTA

   6.36 µL nuclease-free water

   1.3 µL Trifluoroethylamine (TFEA)

   Note: TFEA is volatile, use care when pipetting to ensure adding proper volume. Pipetting TFEA up and down a few times will equilibrate the vapor pressure.

3. To each sample, add 8.7 µL TimeLapse master mix. Flick tubes to combine well and briefly spin to collect sample at bottom of tube.

4. Add freshly prepared 1.3 µL of 200 mM mCPBA.

   *Note: mCPBA **must** be used as oxidant (mCPBA is soluble in ethanol)*

5. Close PCR tubes and flick tubes to mix well and spin down.

6. Incubate in PCR cycler at 45°C for 1h.

7. Add 2 µL of 1 M DTT, mix well, and spin down.

8. Incubate in PCR cycler at 37°C for 30 min.

9. Bring total volume to 200 µL with nuclease free H_2O.

10. Add 20 µL 3 M NaOAc, 1 µL glycoblue. Mix. Add 250 µL isopropanol (or 550 µL EtOH). Mix by inversion and freeze at -80°C for 2 h or overnight.

11. Spin at max, 4°C for 30 min. Wash with 1 mL 75% EtOH. Air dry pellet for at least 2 min or until no EtOH is visible.

12. Resuspend pellet in 10-20 µL 1X RNA loading dye.

A.5 Size selection

1. Prepare 15% 10-well Urea-PAGE gel (1 gel per 4 or 5 samples).

2. Prerun gel for ∼30 min at 200V prior to running samples.

3. Prepare RNA samples for running.

   (a) Add 10 µL 2X RNA loading dye to each sample.

   (b) Mix 1 µL low range ssRNA ladder with 4 µL H_2O and 5 µL loading dye.

   (c) Heat all samples at 70°C for 5 min and ice immediately prior to loading.

4. Clean wells of gel and load samples with empty well between each sample. Run gel at 200V for 60-75 min.

5. With a hot 22G needle poke a hole in the bottom of a 0.5 mL tube for each sample. Place them in a 1.5 mL tube.

6. Carefully remove and wash gel with buffer and water. Incubate gel in 1X gelgreen for 10-15 min in dark with rotation. Image with typhoon on SYBR green setting.

7. Cut smallest possible piece of gel to extract RNA from ∼20 nt to ∼80 nt (about halfway between dyes and just below smallest prominent band, respectively). Place gel slices in 0.5 mL tubes with hole.

8. Spin 0.5 mL tubes in 1.5 mL tubes at max for 2 min. If some gel remains in tube, use clean forceps to crush and transfer to 1.5 mL tube.

9. Add 400 µL H_2O and 40 µL 3 M NaOAc. Vortex briefly to mix. Shake @ RT, 1000 RPM for 2.5 h or 4°C with rotation O/N.

10. Vortex for 30 sec at medium intensity. Cut off tip of 1 mL pipette tip and transfer slurry to spin filter. Spin at 1000 Xg for 2 min.

11. Add 1 µL glycoblue and 500 µL isopropanol (1150 µL EtOH is acceptable but requires -80°C storage). Mix by inversion.

12. Pellet by spinning at max, 4°C for 30 min. Remove supernatant and wash with 1 mL 75% EtOH. Air dry pellet for at least 2 min or until no EtOH is visible.

A.6 Cap selection

1. Resuspend pellet in 17.5 µL H_2O and add 2 µL 10X 5´ polyphosphatase buffer and 1 µL 5´ polyphosphatase (*removes γ and β phosphates from RNA*). Mix by pipetting. Incubate for 30 min at 37°C.

2. To stop reaction, add 180 µL H_2O, transfer to heavy phase lock tubes, add 200 µL phenol-chloroform, shake vigorously for 15 sec, and spin at max for 2 min.

3. Transfer aqueous phase to new 1.5 mL tube and add 20 µL 3M NaOAc, 1 µL glycoblue. Mix. Add 250 µL isopropanol (550 µL EtOH is acceptable but requires -80°C storage.

Better if leaving overnight). Mix. Spin at max, 4°C for 30 min. Wash with 1 mL 75% EtOH. Air dry.

4. Resuspend in 17.5 µL H_2O and add 2 µL 10X 5´ Terminator buffer A and 1 µL 5´ Terminator exonuclease (*degrades uncapped RNA species*). Mix pipetting. Incubate at 30°C for 1 h.

5. Phenol-chloroform extract as previously. Isopropanol or ethanol precipitate as previously.

A.7 3´ ligation

1. Resuspend in 4 µL H_2O and 1 µL pre-adenylated 5´ adapter (∼15 µM). In parallel run a positive control ligation with an RNA of known length.

2. Incubate at 70°C for 2 min. Ice 2 min.

3. Add 1 µL T4 10X RNA ligase buffer, 1 µL RNase inhibitor, 2 µL 50% PEG, 1 µL T4 RNA ligase 2 truncated (*specifically adds pre-adenylated adapter to 5´ end of RNA*). Mix by pipetting. Spin briefly. Incubate at 28°C for 60 min or longer.

4. Add 10 µL of 2X RNA loading dye to stop the reaction.

5. Separate samples on a 15% 10-well Urea-PAGE gel as previously. Image and extract RNA (∼40 nt to ∼150) as previously.

A.8 5´ ligation

1. Resuspend in 17.5 µL H_2O. Add 2 µL 10X Cutsmart buffer and 1 µL CIP alkaline phosphatase (*removes all phosphates from RNA ends. Removes phosphates from leftover 5´ adapters*). Mix by pipetting. Incubate at 37°C for 10 min.

2. Phenol-chloroform extract and precipitate as previously. Resuspend in 5.7 µL H_2O, 0.5 µL RNase inhibitor, and 0.8 µL 10X ThermoPol buffer and transfer to fresh PCR tubes.

3. Add 1 µL RppH (*5´ pyrophosphohydrolase. Removes 5´ cap from RNA and leaves a monophosphate*). Mix by pipetting. Incubate at 37°C for 1 h.

4. Add 1.2 µL 10X T4 RNA ligase buffer. Mix by pipetting and keep on ice.

5. Incubate 1 µL 5´ adapter (20 µM) per sample at 70°C for 2 min. Ice 2 min. Per sample, add 1 µL 10 mM ATP. Mix. Per sample add 1 µL T4 RNA ligase 1 (*ligates 5´ monophosphorylated RNA to 3´ end of RNA adapter*). Mix (include 10% extra).

6. Add 3 µL of adapter/ATP/enzyme mix to RNA/buffer/RNase mix. Mix by pipetting. Spin briefly. Incubate at 28°C for 60 min or longer (or 16°C overnight).

A.9 RT and library amplification

1. Add 1 µL RT primer (20 µM). Mix by pipetting. Incubate at 70°C 2 min. Ice 2 min.

2. In a separate tube, mix the following volumes per sample: 4 µL 5X FS buffer, 1 µL 0.1 M DTT, 1 µL SSRT III, 0.5 µL 12.5 mM dNTP, 0.5 µL RNase inhibitor (Include 10% extra). Mix by pipetting.

3. Add 7 µL of RT mix to each sample. Mix by pipetting. Incubate at 55°C for 60 min.

4. Use 2 µL cDNA to perform qPCR to determine number of cycles to run in PCR. Prep PCR mix containing the following volumes per sample: 8 µL H_2O, 12.5 µL 2X Phusion mix, 0.5 µL 50X SYBR Green (include 10% extra). Add 1 µL of each primer.

5. Prep PCR mix containing the following volumes per sample: 13 µL H_2O, 25 µL 2X Phusion mix (include 10% extra).

6. Take 8 µL cDNA and add 2 µL of each indexing primer and 38 µL of Phusion mix. Mix by pipetting. PCR amplify for determined cycles (14 is default if step 4 not performed).

    1 cycle

        98°C – 30 sec

X cycles

98°C – 10 sec

65°C – 30 sec

72°C – 15 sec

1 cycle

72°C – 10 min

7. Perform SPRI purification with 90 µL beads. Elute in 20 µL H_2O and 4 µL 6X DNA loading dye. Run each sample in 2 lanes of a 6% TBE gel with 1 µL Ultra Low Range DNA ladder.

8. Run gel for 45 min at 200V.

9. Image and extract as previously except to use EtOH instead of IPA. Adapter-adapter ligation is expected at ∼140 bp. Cut just above adapter-adapter product up to ∼250. See example gel 3.

10. Run a bioanalyzer to assess concentration and adapter-adapter ligation product contamination.

Bibliography

[1] F. de Santa, I. Barozzi, F. Mietton, S. Ghisletti, S. Polletti, B. K. Tusi, H. Muller, J. Ragoussis, C. L. Wei, and G. Natoli. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS Biol*, 8(5):e1000384, 2010.

[2] T. K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman, and M. E. Greenberg. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–7, 2010.

[3] M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88, 2007.

[4] L. J. Core, J. J. Waterfall, and J. T. Lis. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science*, 322(5909):1845, 2008.

[5] T. Henriques, B. S. Scruggs, M. O. Inouye, G. W. Muse, L. H. Williams, A. B. Burkholder, C. A. Lavender, D. C. Fargo, and K. Adelman. Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev*, 32(1):26–41, 2018.

[6] L. Core and K. Adelman. Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes Dev*, 33(15-16):960–982, 2019.

[7] S. Feng, S. J. Cokus, X. Zhang, P.-Y. Chen, M. Bostick, M. G. Goll, J. Hetzel, J. Jain, S. H. Strauss, M. E. Halpern, C. Ukomadu, K. C. Sadler, S. Pradhan, M. Pellegrini, and S. E. Jacobsen. Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences of the United States of America*, 107(19):8689–8694, 2010.

[8] A. Zemach, I. E. McDaniel, P. Silva, and D. Zilberman. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science (New York, N.Y.)*, 328(5980):916–919, 2010.

[9] L. A. Gates, C. E. Foulds, and B. W. O'Malley. Histone Marks in the 'Driver's Seat': Functional Roles in Steering the Transcription Cycle. *Trends Biochem Sci*, 42(12):977–989, 2017.

[10] V. Haberle and A. Stark. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol*, 19(10):621–637, 2018.

[11] L. Tora. A unified nomenclature for TATA box binding protein (TBP)-associated factors (TAFs) involved in RNA polymerase II transcription. *Genes Dev*, 16(6):673–5, 2002.

[12] A. B. Patel, B. J. Greber, and E. Nogales. Recent insights into the structure of TFIID, its assembly, and its binding to core promoter. *Curr Opin Struct Biol*, 61:17–24, 2020.

[13] G. Orphanides, T. Lagrange, and D. Reinberg. The general transcription factors of RNA polymerase II. *Genes Dev*, 10(21):2657–83, 1996.

[14] S. Sainsbury, C. Bernecky, and P. Cramer. Structural basis of transcription initiation by RNA polymerase II. *Nat Rev Mol Cell Biol*, 16(3):129–43, 2015.

[15] J. Soutourina. Transcription regulation by the Mediator complex. *Nature Reviews Molecular Cell Biology*, 19(4):262–274, 2018.

[16] R. Abdella, A. Talyzina, S. Chen, C. J. Inouye, R. Tjian, and Y. He. Structure of the human Mediator-bound transcription preinitiation complex. *Science*, 372(6537):52–56, 2021.

[17] X. Chen, X. Yin, J. Li, Z. Wu, Y. Qi, X. Wang, W. Liu, and Y. Xu. Structures of the human Mediator and Mediator-bound preinitiation complex. *Science*, 372(6546):eabg0635, 2021.

[18] S. Rengachari, S. Schilbach, S. Aibara, C. Dienemann, and P. Cramer. Structure of the human Mediator-RNA polymerase II pre-initiation complex. *Nature*, 594(7861):129–133, 2021.

[19] R. H. Jacobson, A. G. Ladurner, D. S. King, and R. Tjian. Structure and Function of a Human TAFII250 Double Bromodomain Module. *Science*, 288(5470):1422–1425, 2000.

[20] M. Vermeulen, K. W. Mulder, S. Denissov, W. W. M. P. Pijnappel, F. M. A. van Schaik, R. A. Varier, M. P. A. Baltissen, H. G. Stunnenberg, M. Mann, and H. T. M. Timmers. Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. *Cell*, 131(1):58–69, 2007.

[21] H. van Ingen, F. M. van Schaik, H. Wienk, J. Ballering, H. Rehmann, A. C. Dechesne, J. A. Kruijzer, R. M. Liskamp, H. M. Timmers, and R. Boelens. Structural Insight into the Recognition of the H3K4me3 Mark by the TFIID Subunit TAF3. *Structure*, 16(8):1245–1256, 2008.

[22] S. M. Lauberth, T. Nakayama, X. Wu, A. L. Ferris, Z. Tang, S. H. Hughes, and R. G. Roeder. H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell*, 152(5):1021–1036, 2013.

[23] A. G. Li, L. G. Piluso, X. Cai, B. J. Gadd, A. G. Ladurner, and X. Liu. An Acetylation Switch in p53 Mediates Holo-TFIID Recruitment. *Molecular Cell*, 28(3):408–421, 2007.

[24] C. R. Bartman, N. Hamagami, C. A. Keller, B. Giardine, R. C. Hardison, G. A. Blobel, and A. Raj. Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation. *Mol Cell*, 73(3):519–532 e4, 2019.

[25] M. S. C. Larke, R. Schwessinger, T. Nojima, J. Telenius, R. A. Beagrie, D. J. Downes, A. M. Oudelaar, J. Truch, B. Graham, M. A. Bender, N. J. Proudfoot, D. R. Higgs, and J. R. Hughes. Enhancers predominantly regulate gene expression during differentiation via transcription initiation. *Mol Cell*, 81(5):983–997.e7, 2021.

[26] X. Darzacq, Y. Shav-Tal, V. de Turris, Y. Brody, S. M. Shenoy, R. D. Phair, and R. H. Singer. In vivo dynamics of RNA polymerase II transcription. *Nature Structural & Molecular Biology*, 14(9):796–806, 2007.

[27] B. Steurer, R. C. Janssens, B. Geverts, M. E. Geijer, F. Wienholz, A. F. Theil, J. Chang, S. Dealy, J. Pothof, W. A. van Cappellen, A. B. Houtsmuller, and J. A. Marteijn. Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. *PNAS*, 115(19):E4368–E4376, 2018.

[28] N. W. Fraser, P. B. Sehgal, and J. E. Darnell. DRB-induced premature termination of late adenovirus transcription. *Nature*, 272:590–593, 1978.

[29] P. Gariglio, M. Bellard, and P. Chambon. Clustering of RNA polymerse B molecules in the 5' moiety of the adult -globin gene of hen erythocytes. *Nucleic Acids Res*, 9(11):2589–2598, 1981.

[30] D. S. Gilmour and J. T. Lis. RNA Polymerase II Interacts with the Promoter Region of the Noninduced hsp70 Gene in Drosophila melanogaster Cells. *Mol Cell Biol*, 6(11):3984–3989, 1986.

[31] K. Adelman and J. T. Lis. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet*, 13(10):720–31, 2012.

[32] S. M. Vos, L. Farnung, H. Urlaub, and P. Cramer. Structure of paused transcription complex Pol II-DSIF-NELF. *Nature*, 2018.

[33] T. Wada, T. Takagi, Y. Yamaguchi, A. Ferdous, T. Imai, S. Hirose, S. Sugimoto, K. Yano, G. A. Hartzog, F. Winston, S. Buratowski, and H. Handa. DSIF, a novel

transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes & development*, 12(3):343–356, 1998.

[34] T. Wada, T. Takagi, Y. Yamaguchi, D. Watanabe, and H. Handa. Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro. *The EMBO journal*, 17(24):7395–7403, 1998.

[35] Y. Yamaguchi, T. Takagi, T. Wada, K. Yano, A. Furuya, S. Sugimoto, J. Hasegawa, and H. Handa. NELF, a Multisubunit Complex Containing RD, Cooperates with DSIF to Repress RNA Polymerase II Elongationd. *Cell*, 97:41–51, 1999.

[36] T. Narita, Y. Yamaguchi, K. Yano, S. Sugimoto, S. Chanarat, T. Wada, D.-k. Kim, J. Hasegawa, M. Omori, N. Inukai, M. Endoh, T. Yamada, and H. Handa. Human transcription elongation factor NELF: identification of novel subunits and reconstitution of the functionally active complex. *Molecular and cellular biology*, 23(6):1863–1873, 2003.

[37] C. B. Fant, C. B. Levandowski, K. Gupta, Z. L. Maas, J. Moir, J. D. Rubin, A. Sawyer, M. N. Esbin, J. K. Rimel, O. Luyties, M. T. Marr, I. Berger, R. D. Dowell, and D. J. Taatjes. TFIID Enables RNA Polymerase II Promoter-Proximal Pausing. *Mol Cell*, 2020.

[38] N. F. Marshall and D. H. Price. Purification of P-TEFb, a Transcription Factor Required for the Transition into Productive Elongation. *Journal of Biological Chemistry*, 270(21):12335–12338, 1995.

[39] N. F. Marshall, J. Peng, Z. Xie, and D. H. Price. Control of RNA Polymerase II Elongation Potential by a Novel Carboxyl-terminal Domain Kinase*. *Journal of Biological Chemistry*, 271(43):27176–27183, 1996.

[40] P. Wei, M. E. Garber, S.-M. Fang, W. H. Fischer, and K. A. Jones. A Novel CDK9-Associated C-Type Cyclin Interacts Directly with HIV-1 Tat and Mediates Its High-Affinity, Loop-Specific Binding to TAR RNA. *Cell*, 92(4):451–462, 1998.

[41] K. Fujinaga, D. Irwin, Y. Huang, R. Taube, T. Kurosu, and B. M. Peterlin. Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. *Molecular and cellular biology*, 24(2):787–795, 2004.

[42] T. Yamada, Y. Yamaguchi, N. Inukai, S. Okamoto, T. Mura, and H. Handa. P-TEFb-Mediated Phosphorylation of hSpt5 C-Terminal Repeats Is Critical for Processive Transcription Elongation. *Molecular Cell*, 21(2):227–237, 2006.

[43] K.-L. Huang, D. Jee, C. B. Stein, N. D. Elrod, T. Henriques, L. G. Mascibroda, D. Baillat, W. K. Russell, K. Adelman, and E. J. Wagner. Integrator Recruits Protein Phosphatase 2A to Prevent Pause Release and Facilitate Transcription Termination. *Mol Cell*, 80(2):345–358.e9, 2020.

[44] S. J. Vervoort, S. A. Welsh, J. R. Devlin, E. Barbieri, D. A. Knight, S. Offley, S. Bjelosevic, M. Costacurta, I. Todorovski, C. J. Kearney, J. J. Sandow, Z. Fan, B. Blyth, V. McLeod, J. H. A. Vissers, K. Pavic, B. P. Martin, G. Gregory, E. Demosthenous, M. Zethoven, I. Y. Kong, E. D. Hawkins, S. J. Hogg, M. J. Kelly, A. Newbold, K. J. Simpson, O. Kauko, K. F. Harvey, M. Ohlmeyer, J. Westermarck, N. Gray, A. Gardini, and R. W. Johnstone. The PP2A-Integrator-CDK9 axis fine-tunes transcription and can be targeted therapeutically in cancer. *Cell*, 184(12):3143–3162.e32, 2021.

[45] N. D. Elrod, T. Henriques, K. L. Huang, D. C. Tatomer, J. E. Wilusz, E. J. Wagner, and K. Adelman. The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes. *Mol Cell*, 76(5):738–752 e7, 2019.

[46] F. Beckedorff, E. Blumenthal, L. F. daSilva, Y. Aoi, P. R. Cingaram, J. Yue, A. Zhang, S. Dokaneheifard, M. G. Valencia, G. Gaidosh, A. Shilatifard, and R. Shiekhattar. The Human Integrator Complex Facilitates Transcriptional Elongation by Endonucleolytic Cleavage of Nascent Transcripts. *Cell Rep*, 32(3):107917, 2020.

[47] H. Zheng, Y. Qi, S. Hu, X. Cao, C. Xu, Z. Yin, X. Chen, Y. Li, W. Liu, J. Li, J. Wang, G. Wei, K. Liang, F. X. Chen, and Y. Xu. Identification of Integrator-PP2A complex (INTAC), an RNA polymerase II phosphatase. *Science*, 370(6520), 2020.

[48] I. Fianu, Y. Chen, C. Dienemann, O. Dybkov, A. Linden, H. Urlaub, and P. Cramer. Structural basis of Integrator-mediated transcription regulation. *Science*, 374(6569):883–887, 2021.

[49] M. S. Buckley, H. Kwak, W. R. Zipfel, and J. T. Lis. Kinetics of promoter Pol II on Hsp70 reveal stable pausing and key insights into its regulation. *Genes Dev*, 28(1):14–9, 2014.

[50] J. T. Zimmer, N. A. Rosa-Mercado, D. Canzio, J. A. Steitz, and M. D. Simon. STL-seq reveals pause-release and termination kinetics for promoter-proximal paused RNA polymerase II transcripts. *Mol Cell*, 81(21):4398–4412, 2021.

[51] D. A. Gilchrist, G. Dos Santos, D. C. Fargo, B. Xie, Y. Gao, L. Li, and K. Adelman. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell*, 143(4):540–51, 2010.

[52] B. Bartkowiak, P. Liu, H. P. Phatnani, N. J. Fuda, J. J. Cooper, D. H. Price, K. Adelman, J. T. Lis, and A. L. Greenleaf. CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes & development*, 24(20):2303–2316, 2010.

[53] I. Jonkers and J. T. Lis. Getting up to speed with transcription elongation by RNA polymerase II. *Nature Reviews Molecular cell biology*, 16(3):167–177, 2015.

[54] R. M. Sheridan, N. Fong, A. D'Alessandro, and D. L. Bentley. Widespread Backtracking by RNA Pol II Is a Major Effector of Gene Activation, 5' Pause Release, Termination, and Transcription Elongation Rate. *Mol Cell*, 73(1):107–118 e4, 2019.

[55] R. Sousa-Luís, G. Dujardin, I. Zukher, H. Kimura, C. Weldon, M. Carmo-Fonseca, N. J. Proudfoot, and T. Nojima. POINT technology illuminates the processing of polymerase-associated intact nascent transcripts. *Molecular Cell*, 81(9):1935–1950.e6, 2021.

[56] S. Lykke-Andersen, K. Žumer, E. Molska, J. O. Rouvière, G. Wu, C. Demel, B. Schwalb, M. Schmid, P. Cramer, and T. H. Jensen. Integrator is a genome-wide attenuator of non-productive transcription. *Molecular Cell*, 81(3):514–529.e6, 2021.

[57] S. Nechaev, D. C. Fargo, G. d. Santos, L. Liu, Y. Gao, and K. Adelman. Global Analysis of Short RNAs Reveals Widespread Promoter-Proximal Stalling and Arrest of Pol II in Drosophila. *Science*, 327:335–338, 2010.

[58] H. Kwak, N. J. Fuda, L. J. Core, and J. T. Lis. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science*, 339, 2013.

[59] W. Shao and J. Zeitlinger. Paused RNA polymerase II inhibits new transcriptional initiation. *Nat Genet*, 49:1045–1051, 2017.

[60] D. C. Tatomer, N. D. Elrod, D. Liang, M. S. Xiao, J. Z. Jiang, M. Jonathan, K. L. Huang, E. J. Wagner, S. Cherry, and J. E. Wilusz. The Integrator complex cleaves nascent mRNAs to attenuate transcription. *Genes Dev*, 33(21-22):1525–1538, 2019.

[61] S. Buratowski, S. Hahn, L. Guarente, and P. A. Sharp. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell*, 56(4):549–61, 1989.

[62] E. Wieczorek, M. Brand, X. Jacq, and L. Tora. Function of TAFII-containing complex without TBP in transcription by RNA polymerase II. *Nature*, 393(6681):187–191, 1998.

[63] J.-C. Dantonel, S. Quintin, L. Lakatos, M. Labouesse, and L. Tora. TBP-like Factor Is Required for Embryonic RNA Polymerase II Transcription in C. elegans. *Molecular Cell*, 6(3):715–722, 2000.

[64] L. Kaltenbach, M. A. Horner, J. H. Rothman, and S. E. Mango. The TBP-like Factor CeTLF Is Required to Activate RNA Polymerase II Transcription during C. elegans Embryogenesis. *Molecular Cell*, 6(3):705–713, 2000.

[65] F. Müller, L. Lakatos, J.-C. Dantonel, U. Strähle, and L. Tora. TBP is not universally required for zygotic RNA polymerase II transcription in zebrafish. *Current Biology*, 11(4):282–287, 2001.

[66] I. Martianov, S. Viville, and I. Davidson. RNA Polymerase II Transcription in Murine Cells Lacking the TATA Binding Protein. *Science*, 298(5595):1036–1039, 2002.

[67] E. Gazdag, U. G. Jacobi, I. van Kruijsbergen, D. L. Weeks, and G. J. Veenstra. Activation of a T-box-Otx2-Gsc gene network independent of TBP and TBP-related factors. *Development*, 143(8):1340–50, 2016.

[68] R. K. Louder, Y. He, J. R. López-Blanco, J. Fang, P. Chacón, and E. Nogales. Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. *Nature*, 531(7596):604–609, 2016.

[69] S. M. Vos, L. Farnung, M. Boehning, C. Wigge, A. Linden, H. Urlaub, and P. Cramer. Structure of activated transcription complex Pol II-DSIF-PAF-SPT6. *Nature*, 2018.

[70] X. Chen, Y. Qi, Z. Wu, X. Wang, J. Li, D. Zhao, H. Hou, Y. Li, Z. Yu, W. Liu, M. Wang, Y. Ren, Z. Li, H. Yang, and Y. Xu. Structural insights into preinitiation complex assembly on core promoters. *Science*, 372(6541):eaba8490, 2021.

[71] S. Schilbach, S. Aibara, C. Dienemann, F. Grabbe, and P. Cramer. Structure of RNA polymerase II pre-initiation complex at 2.9 Å defines initial DNA opening. *Cell*, 2021.

[72] S. Aibara, S. Schilbach, and P. Cramer. Structures of mammalian RNA polymerase II pre-initiation complexes. *Nature*, 594(7861):124–128, 2021.

[73] A. C. Schier and D. J. Taatjes. Everything at once: cryo-EM yields remarkable insights into human RNA polymerase II transcription. *Nature Structural & Molecular Biology*, 2021.

[74] M. J. Solomon and A. Varshavsky. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proceedings of the National Academy of Sciences of the United States of America*, 82(19):6470–6474, 1985.

[75] Ren Bing, Robert François, Wyrick John J., Aparicio Oscar, Jennings Ezra G., Simon Itamar, Zeitlinger Julia, Schreiber Jörg, Hannett Nancy, Kanin Elenita, Volkert

Thomas L., Wilson Christopher J., Bell Stephen P., and Young Richard A. Genome-Wide Location and Function of DNA Binding Proteins. *Science*, 290(5500):2306–2309, 2000.

[76] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837, 2007.

[77] Johnson David S., Mortazavi Ali, Myers Richard M., and Wold Barbara. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, 316(5830):1497–1502, 2007.

[78] G. W. Muse, D. A. Gilchrist, S. Nechaev, R. Shah, J. Parker, S. Grissom, J. Zeitlinger, and K. Adelman. RNA polymerase is poised for activation across the genome. *Nat Genet*, 39(12):1507–1511, 2007.

[79] H. Rhee and B. Pugh. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell*, 147(6):1408–1419, 2011.

[80] P. J. Skene and S. Henikoff. A simple method for generating high-resolution maps of genome-wide protein binding. *eLife*, 4:e09225–e09225, 2015.

[81] Q. He, J. Johnston, and J. Zeitlinger. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature Biotechnology*, 33(4):395–401, 2015.

[82] C. Schmidl, A. F. Rendeiro, N. C. Sheffield, and C. Bock. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nature Methods*, 12(10):963–965, 2015.

[83] P. J. Skene and S. Henikoff. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife*, 6:e21856, 2017.

[84] H. S. Kaya-Okur, S. J. Wu, C. A. Codomo, E. S. Pledger, T. D. Bryson, J. G. Henikoff, K. Ahmad, and S. Henikoff. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature communications*, 10(1):1930–1930, 2019.

[85] A. R. Krebs, D. Imanci, L. Hoerner, D. Gaidatzis, L. Burger, and D. Schubeler. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol Cell*, 67(3):411–422 e4, 2017.

[86] T. Henriques, D. A. Gilchrist, S. Nechaev, M. Bern, G. W. Muse, A. Burkholder, D. C. Fargo, and K. Adelman. Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. *Mol Cell*, 52(4):517–28, 2013.

[87] S. H. Duttke, M. W. Chang, S. Heinz, and C. Benner. Identification and dynamic quantification of regulatory elements using total RNA. *Genome research*, 29(11):1836–1846, 2019.

[88] T. Chu, E. J. Rice, G. T. Booth, H. H. Salamanca, Z. Wang, L. J. Core, S. L. Longo, R. J. Corona, L. S. Chin, J. T. Lis, H. Kwak, and C. G. Danko. Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. *Nature genetics*, 50(11):1553–1564, 2018.

[89] L. S. Churchman and J. S. Weissman. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469(7330):368–373, 2011.

[90] T. Nojima, T. Gomes, A. R. F. Grosso, H. Kimura, M. J. Dye, S. Dhir, M. Carmo-Fonseca, and N. J. Proudfoot. Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell*, 161(3):526–540, 2015.

[91] R. Dollinger and D. S. Gilmour. Regulation of Promoter Proximal Pausing of RNA Polymerase II in Metazoans. *Journal of Molecular Biology*, 433(14):166897, 2021.

[92] D. Mazza, A. Abernathy, N. Golob, T. Morisaki, and J. G. McNally. A benchmark for chromatin binding measurements in live cells. *Nucleic Acids Research*, 40(15):e119–e119, 2012.

[93] V. Q. Nguyen, A. Ranjan, S. Liu, X. Tang, Y. H. Ling, J. Wisniewski, G. Mizuguchi, K. Y. Li, V. Jou, Q. Zheng, L. D. Lavis, T. Lionnet, and C. Wu. Spatiotemporal coordination of transcription preinitiation complex assembly in live cells. *Mol Cell*, 2021.

[94] I. Baek, L. J. Friedman, J. Gelles, and S. Buratowski. Single-molecule studies reveal branched pathways for activator-dependent assembly of RNA polymerase II preinitiation complexes. *Mol Cell*, 2021.

[95] T. Lionnet and C. Wu. Single-molecule tracking of transcription protein dynamics in living cells: seeing is believing, but what are we seeing? *Current Opinion in Genetics & Development*, 67:94–102, 2021.

[96] B. Schwalb, M. Michel, B. Zacher, K. Frühauf, C. Demel, A. Tresch, J. Gagneur, and P. Cramer. TT-seq maps the human transient transcriptome. *Science*, 352(6290):1225–1228, 2016.

[97] J. A. Schofield, E. E. Duffy, L. Kiefer, M. C. Sullivan, and M. D. Simon. TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat Methods*, 15(3):221–225, 2018.

[98] V. A. Herzog, B. Reichholf, T. Neumann, P. Rescheneder, P. Bhat, T. R. Burkard, W. Wlotzka, A. von Haeseler, J. Zuber, and S. L. Ameres. Thiol-linked alkylation of RNA to assess expression dynamics. *Nature Methods*, 14(12):1198–1204, 2017.

[99] C. Riml, T. Amort, D. Rieder, C. Gasser, A. Lusser, and R. Micura. Osmium-Mediated Transformation of 4-Thiouridine to Cytidine as Key To Study RNA Dynamics by Sequencing. *Angew Chem Int Ed Engl*, 56(43):13479–13483, 2017.

[100] L. Kiefer, J. A. Schofield, and M. D. Simon. Expanding the Nucleoside Recoding Toolkit: Revealing RNA Population Dynamics with 6-Thioguanosine. *J Am Chem Soc*, 140(44):14567–14570, 2018.

[101] C. Gasser, I. Delazer, E. Neuner, K. Pascher, K. Brillet, S. Klotz, L. Trixl, M. Himmelstoß, E. Ennifar, D. Rieder, A. Lusser, and R. Micura. Thioguanosine Conversion Enables mRNA-Lifetime Evaluation by RNA Sequencing Using Double Metabolic Labeling (TUC-seq DUAL). *Angew Chem Int Ed Engl*, 59(17):6881–6886, 2020.

[102] Y. Chen, F. Wu, Z. Chen, Z. He, Q. Wei, W. Zeng, K. Chen, F. Xiao, Y. Yuan, X. Weng, Y. Zhou, and X. Zhou. Acrylonitrile-Mediated Nascent RNA Sequenc-

ing for Transcriptome-Wide Profiling of Cellular RNA Dynamics. *Advanced Science*, 7(8):1900997, 2020.

[103] E. E. Duffy, D. Canzio, T. Maniatis, and M. D. Simon. Solid phase chemistry to covalently and reversibly capture thiolated RNA. *Nucleic Acids Res*, 46(14):6996–7005, 2018.

[104] S. Gressel, B. Schwalb, T. M. Decker, W. Qin, H. Leonhardt, D. Eick, and P. Cramer. CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife*, 6, 2017.

[105] A. Uvarovskii and C. Dieterich. pulseR: Versatile computational analysis of RNA turnover from metabolic labeling experiments. *Bioinformatics*, 33(20):3305–3307, 2017.

[106] C. Jürges, L. Dölken, and F. Erhard. Dissecting newly transcribed and old RNA using GRAND-SLAM. *Bioinformatics*, 34(13):i218–i226, 2018.

[107] Y. Zhang, C. Park, C. Bennett, M. Thornton, and D. Kim. Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N. *Genome Res*, 2021.

[108] T. I. Lee and R. A. Young. Transcriptional Regulation and its Misregulation in Disease. *Cell*, 152(6):1237–1251, 2013.

[109] K. Izumi. Disorders of Transcriptional Regulation: An Emerging Category of Multiple Malformation Syndromes. *Molecular Syndromology*, 7(5):262–273, 2016.

[110] H. Risheg, J. M. Graham, R. D. Clark, R. C. Rogers, J. M. Opitz, J. B. Moeschler, A. P. Peiffer, M. May, S. M. Joseph, J. R. Jones, R. E. Stevenson, C. E. Schwartz, and M. J. Friez. A recurrent mutation in MED12 leading to R961W causes Opitz-Kaveggia syndrome. *Nature Genetics*, 39(4):451–453, 2007.

[111] S. Hashimoto, S. Boissel, M. Zarhrate, M. Rio, A. Munnich, J.-M. Egly, and L. Colleaux. MED23 mutation links intellectual disability to dysregulation of immediate early gene expression. *Science (New York, N.Y.)*, 333(6046):1161–1163, 2011.

[112] A. T. Vulto-van Silfhout, B. B. A. de Vries, B. W. M. van Bon, A. Hoischen, M. Ruiterkamp-Versteeg, C. Gilissen, F. Gao, M. van Zwam, C. L. Harteveld, A. J. van Essen, B. C. J. Hamel, T. Kleefstra, M. A. A. P. Willemsen, H. G. Yntema, H. van Bokhoven, H. G. Brunner, T. G. Boyer, and A. P. M. de Brouwer. Mutations in MED12 cause X-linked Ohdo syndrome. *American Journal of Human Genetics*, 92(3):401–406, 2013.

[113] L. Basel-Vanagaite, P. Smirin-Yosef, J. L. Essakow, S. Tzur, I. Lagovsky, I. Maya, M. Pasmanik-Chor, A. Yeheskel, O. Konen, N. Orenstein, M. Weisz Hubshman, V. Drasinover, N. Magal, G. Peretz Amit, Y. Zalzstein, A. Zeharia, M. Shohat, R. Straussberg, D. Monté, M. Salmon-Divon, and D. M. Behar. Homozygous MED25 mutation implicated in eye-intellectual disability syndrome. *Human Genetics*, 134(6):577–587, 2015.

[114] T. Figueiredo, U. S. Melo, A. L. S. Pessoa, P. R. Nobrega, J. P. Kitajima, I. Correa, M. Zatz, F. Kok, and S. Santos. Homozygous missense mutation in MED25 segregates with syndromic intellectual disability in a large consanguineous family. *Journal of Medical Genetics*, 52(2):123–127, 2015.

[115] V. Viprakasit, R. J. Gibbons, B. C. Broughton, J. L. Tolmie, D. Brown, P. Lunt, R. M. Winter, S. Marinoni, M. Stefanini, L. Brueton, A. R. Lehmann, and D. R. Higgs. Mutations in the general transcription factor TFIIH result in beta-thalassaemia in individuals with trichothiodystrophy. *Human Molecular Genetics*, 10(24):2797–2802, 2001.

[116] C. Kuschal, E. Botta, D. Orioli, J. Digiovanna, S. Seneca, K. Keymolen, D. Tamura, E. Heller, S. Khan, G. Caligiuri, M. Lanzafame, T. Nardo, R. Ricotti, F. Peverali, R. Stephens, Y. Zhao, A. Lehmann, L. Baranello, D. Levens, K. Kraemer, and M. Stefanini. GTF2E2 Mutations Destabilize the General Transcription Factor Complex TFIIE in Individuals with DNA Repair-Proficient Trichothiodystrophy. *American Journal of Human Genetics*, 98(4):627–642, 2016.

[117] W. M. C. van Roon-Mom, S. J. Reid, R. L. M. Faull, and R. G. Snell. TATA-binding protein in neurodegenerative disease. *Neuroscience*, 133(4):863–872, 2005.

[118] L. Rooms, E. Reyniers, S. Scheers, R. van Luijk, J. Wauters, L. Van Aerschot, Z. Callaerts-Vegh, R. D'Hooge, G. Mengus, I. Davidson, W. Courtens, and R. F. Kooy. TBP as a candidate gene for mental retardation in patients with subtelomeric 6q deletions. *European journal of human genetics: EJHG*, 14(10):1090–1096, 2006.

[119] M. J. Friedman, A. G. Shah, Z.-H. Fang, E. G. Ward, S. T. Warren, S. Li, and X.-J. Li. Polyglutamine domain modulates the TBP-TFIIB interaction: implications for its normal function and neurodegeneration. *Nature Neuroscience*, 10(12):1519–1528, 2007.

[120] S. Hellman-Aharony, P. Smirin-Yosef, A. Halevy, M. Pasmanik-Chor, A. Yeheskel, A. Har-Zahav, I. Maya, R. Straussberg, D. Dahary, A. Haviv, M. Shohat, and L. Basel-Vanagaite. Microcephaly thin corpus callosum intellectual disability syndrome caused by mutated TAF2. *Pediatric Neurology*, 49(6):411–416.e1, 2013.

[121] J. O'Rawe, Y. Wu, M. Dörfel, A. Rope, P. Au, J. Parboosingh, S. Moon, M. Kousi, K. Kosma, C. Smith, M. Tzetis, J. Schuette, R. Hufnagel, C. Prada, F. Martinez, C. Orellana, J. Crain, A. Caro-Llopis, S. Oltra, S. Monfort, L. Jiménez-Barrón, J. Swensen, S. Ellingwood, R. Smith, H. Fang, S. Ospina, S. Stegmann, N. Den Hollander, D. Mittelman, G. Highnam, R. Robison, E. Yang, L. Faivre, A. Roubertie, J.-B. Rivière, K. Monaghan, K. Wang, E. Davis, N. Katsanis, V. Kalscheuer, E. Wang, K. Metcalfe, T. Kleefstra, A. Innes, S. Kitsiou-Tzeli, M. Rosello, C. Keegan, and G. Lyon. TAF1 Variants Are Associated with Dysmorphic Features, Intellectual Disability, and Neurological Manifestations. *American Journal of Human Genetics*, 97(6):922–932, 2015.

[122] B. Yuan, D. Pehlivan, E. Karaca, N. Patel, W.-L. Charng, T. Gambin, C. Gonzaga-Jauregui, V. R. Sutton, G. Yesil, S. T. Bozdogan, T. Tos, A. Koparir, E. Koparir, C. R. Beck, S. Gu, H. Aslan, O. O. Yuregir, K. Al Rubeaan, D. Alnaqeb, M. J. Alshammari, Y. Bayram, M. M. Atik, H. Aydin, B. B. Geckinli, M. Seven, H. Ulucan,

E. Fenercioglu, M. Ozen, S. Jhangiani, D. M. Muzny, E. Boerwinkle, B. Tuysuz, F. S. Alkuraya, R. A. Gibbs, and J. R. Lupski. Global transcriptional disturbances underlie Cornelia de Lange syndrome and related phenotypes. *The Journal of Clinical Investigation*, 125(2):636–651, 2015.

[123] H. Hu, S. A. Haas, J. Chelly, H. Van Esch, M. Raynaud, A. P. M. de Brouwer, S. Weinert, G. Froyen, S. G. M. Frints, F. Laumonnier, T. Zemojtel, M. I. Love, H. Richard, A.-K. Emde, M. Bienek, C. Jensen, M. Hambrock, U. Fischer, C. Langnick, M. Feldkamp, W. Wissink-Lindhout, N. Lebrun, L. Castelnau, J. Rucci, R. Montjean, O. Dorseuil, P. Billuart, T. Stuhlmann, M. Shaw, M. A. Corbett, A. Gardner, S. Willis-Owen, C. Tan, K. L. Friend, S. Belet, K. E. P. van Roozendaal, M. Jimenez-Pocquet, M.-P. Moizard, N. Ronce, R. Sun, S. O'Keeffe, R. Chenna, A. van Bömmel, J. Göke, A. Hackett, M. Field, L. Christie, J. Boyle, E. Haan, J. Nelson, G. Turner, G. Baynam, G. Gillessen-Kaesbach, U. Müller, D. Steinberger, B. Budny, M. Badura-Stronka, A. Latos-Bieleńska, L. B. Ousager, P. Wieacker, G. Rodríguez Criado, M.-L. Bondeson, G. Annerén, A. Dufke, M. Cohen, L. Van Maldergem, C. Vincent-Delorme, B. Echenne, B. Simon-Bouy, T. Kleefstra, M. Willemsen, J.-P. Fryns, K. Devriendt, R. Ullmann, M. Vingron, K. Wrogemann, T. F. Wienker, A. Tzschach, H. van Bokhoven, J. Gecz, T. J. Jentsch, W. Chen, H.-H. Ropers, and V. M. Kalscheuer. X-exome sequencing of 405 unresolved families identifies seven novel intellectual disability genes. *Molecular Psychiatry*, 21(1):133–148, 2016.

[124] D. C. Bragg, K. Mangkalaphiban, C. A. Vaine, N. J. Kulkarni, D. Shin, R. Yadav, J. Dhakal, M.-L. Ton, A. Cheng, C. T. Russo, M. Ang, P. Acuña, C. Go, T. N. Franceour, T. Multhaupt-Buell, N. Ito, U. Müller, W. T. Hendriks, X. O. Breakefield, N. Sharma, and L. J. Ozelius. Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. *PNAS*, 114(51):E11020–E11028, 2017.

[125] T. Aneichyk, W. T. Hendriks, R. Yadav, D. Shin, D. Gao, C. A. Vaine, R. L. Collins, A. Domingo, B. Currall, A. Stortchevoi, T. Multhaupt-Buell, E. B. Penney, L. Cruz, J. Dhakal, H. Brand, C. Hanscom, C. Antolik, M. Dy, A. Ragavendran, J. Underwood,

S. Cantsilieris, K. M. Munson, E. E. Eichler, P. Acuna, C. Go, R. D. G. Jamora, R. L. Rosales, D. M. Church, S. R. Williams, S. Garcia, C. Klein, U. Muller, K. C. Wilhelmsen, H. T. M. Timmers, Y. Sapir, B. J. Wainger, D. Henderson, N. Ito, N. Weisenfeld, D. Jaffe, N. Sharma, X. O. Breakefield, L. J. Ozelius, D. C. Bragg, and M. E. Talkowski. Dissecting the Causal Mechanism of X-Linked Dystonia-Parkinsonism by Integrating Genome and Transcriptome Assembly. *Cell*, 172(5):897–909 e21, 2018.

[126] D. C. Bragg, N. Sharma, and L. J. Ozelius. X-Linked Dystonia-Parkinsonism: recent advances. *Current opinion in neurology*, 32(4):604–609, 2019.

[127] F. El-Saafin, C. Curry, T. Ye, J.-M. Garnier, I. Kolb-Cheynel, M. Stierle, N. L. Downer, M. P. Dixon, L. Negroni, I. Berger, T. Thomas, A. K. Voss, W. Dobyns, D. Devys, and L. Tora. Homozygous TAF8 mutation in a patient with intellectual disability results in undetectable TAF8 protein, but preserved RNA polymerase II transcription. *Human Molecular Genetics*, 27(12):2171–2186, 2018.

[128] N. Okamoto, H. Arai, T. Onishi, T. Mizuguchi, and N. Matsumoto. Intellectual disability and dysmorphic features in male siblings arising from a novel TAF1 mutation. *Congenit Anom (Kyoto)*, 60(1):40–41, 2020.

[129] P. García-Gutiérrez and M. García-Domínguez. BETting on a Transcriptional Deficit as the Main Cause for Cornelia de Lange Syndrome. *Frontiers in Molecular Biosciences*, 8:709232, 2021.

[130] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, 2011.

[131] G. Fuchs, Y. Voichek, S. Benjamin, S. Gilad, I. Amit, and M. Oren. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biology*, 15(5):R69, 2014.

[132] E. E. Duffy, M. Rutenberg-Schoenberg, C. D. Stark, R. R. Kitchen, M. B. Gerstein, and M. D. Simon. Tracking distinct RNA populations using efficient and reversible covalent chemistry. *Mol Cell*, 59(5):858–866, 2015.

[133] H.-B. Li, J. Tong, S. Zhu, P. J. Batista, E. E. Duffy, J. Zhao, W. Bailis, G. Cao, L. Kroehling, Y. Chen, G. Wang, J. P. Broughton, Y. G. Chen, Y. Kluger, M. D. Simon, H. Y. Chang, Z. Yin, and R. A. Flavell. m6A mRNA methylation controls T cell homeostasis by targeting IL-7/STAT5/SOCS pathway. *Nature*, 548(7667):338–342, 2017.

[134] M. Muhar, A. Ebert, T. Neumann, C. Umkehrer, J. Jude, C. Wieshofer, P. Rescheneder, J. J. Lipp, V. A. Herzog, B. Reichholf, D. A. Cisneros, T. Hoffmann, M. F. Schlapansky, P. Bhat, A. von Haeseler, T. Köcher, A. C. Obenauf, J. Popow, S. L. Ameres, and J. Zuber. SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis. *Science*, 360(6390):800–805, 2018.

[135] Z. Na, Y. Luo, J. A. Schofield, S. Smelyansky, A. Khitun, S. Muthukumar, E. Valkov, M. D. Simon, and S. A. Slavoff. The NBDY Microprotein Regulates Cellular RNA Decapping. *Biochemistry*, 59(42):4131–4142, 2020.

[136] Y. Luo, J. A. Schofield, Z. Na, T. Hann, M. D. Simon, and S. A. Slavoff. Discovery of cellular substrates of human RNA-decapping enzyme DCP2 using a stapled bicyclic peptide inhibitor. *Cell Chem Biol*, 28(4):463–474.e7, 2021.

[137] G. Biancon, P. Joshi, J. T. Zimmer, T. Hunck, Y. Gao, M. D. Lessard, E. Courchaine, A. E. S. Barentine, M. Machyna, V. Botti, A. Qin, R. Gbyli, A. Patel, Y. Song, L. Kiefer, G. Viero, N. Neuenkirchen, H. Lin, J. Bewersdorf, M. D. Simon, K. M. Neugebauer, T. Tebaldi, and S. Halene. Precision analysis of mutant U2AF1 activity reveals deployment of stress granules in myeloid malignancies. *Molecular Cell*, 82(6):1107–1122.e7, 2022.

[138] H. Alalam, J. A. Zepeda-Martínez, and P. Sunnerhagen. Global SLAM-seq for accurate mRNA decay determination and identification of NMD targets. *RNA*, 28(6):905–915, 2022.

[139] D. Kim, B. Langmead, and S. L. Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*, 12(4):357–60, 2015.

[140] F. Krueger and S. R. Andrews. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, 27(11):1571–1572, 2011.

[141] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–9, 2012.

[142] E. Boileau, J. Altmüller, I. S. Naarmann-de Vries, and C. Dieterich. A comparison of metabolic labeling and statistical methods to infer genome-wide dynamics of RNA turnover. *Briefings in Bioinformatics*, 22(6):bbab219, 2021.

[143] A. Liberzon, C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 1(6):417–425, 2015.

[144] J. Lovén, D. A. Orlando, A. A. Sigova, C. Y. Lin, P. B. Rahl, C. B. Burge, D. L. Levens, T. I. Lee, and R. A. Young. Revisiting Global Gene Expression Analysis. *Cell*, 151(3):476–482, 2012.

[145] K. Chen, Z. Hu, Z. Xia, D. Zhao, W. Li, and J. K. Tyler. The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. *Molecular and Cellular Biology*, 36(5):662–667, 2016.

[146] A. Vilborg, M. Passarelli, T. Yario, K. Tycowski, and J. Steitz. Widespread Inducible Transcription Downstream of Human Genes. *Mol Cell*, 59(3):449–461, 2015.

[147] A. Rutkowski, F. Erhard, A. L'Hernault, T. Bonfert, M. Schilhabel, C. Crump, P. Rosenstiel, S. Efstathiou, R. Zimmer, C. Friedel, and L. Dölken. Widespread disruption of host transcription termination in HSV-1 infection. *Nature Communications*, 6, 2015.

[148] T. Hennig, M. Michalski, A. J. Rutkowski, L. Djakovic, A. W. Whisnant, M.-S. Friedl, B. A. Jha, M. A. P. Baptista, A. L'Hernault, F. Erhard, L. Dölken, and C. C. Friedel.

HSV-1-induced disruption of transcription termination resembles a cellular stress response but selectively increases chromatin accessibility downstream of genes. *PLOS Pathogens*, 14(3):e1006954, 2018.

[149] J. F. Cardiello, J. A. Goodrich, and J. F. Kugel. Heat Shock Causes a Reversible Increase in RNA Polymerase II Occupancy Downstream of mRNA Genes, Consistent with a Global Loss in Transcriptional Termination. *Molecular and Cellular Biology*, 38(18):e00181–18, 2018.

[150] S. Roth, S. Heinz, and C. Benner. ARTDeco: Automatic readthrough transcription detection. *BMC Bioinformatics*, 21(1), 2020.

[151] R. Amat, R. Böttcher, F. L. Dily, E. Vidal, J. Quilez, Y. Cuartero, M. Beato, E. d. Nadal, and F. Posas. Rapid reversible changes in compartments and local chromatin organization revealed by hyperosmotic shock. *Genome Research*, 29(1):18–28, 2019.

[152] S. K. Mak and D. Kültz. Gadd45 Proteins Induce G2/M Arrest and Modulate Apoptosis in Kidney Cells Exposed to Hyperosmotic Stress*. *Journal of Biological Chemistry*, 279(37):39075–39084, 2004.

[153] E. Chen, C. Tan, Y. Kou, Q. Duan, Z. Wang, G. Meirelles, N. Clark, and A. Ma'ayan. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14, 2013.

[154] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, 2016.

[155] D. Baillat, M.-A. Hakimi, A. M. Näär, A. Shilatifard, N. Cooch, and R. Shiekhattar. Integrator, a Multiprotein Mediator of Small Nuclear RNA Processing, Associates with the C-Terminal Repeat of RNA Polymerase II. *Cell*, 123(2):265–276, 2005.

[156] A. Gardini, D. Baillat, M. Cesaroni, D. Hu, J. M. Marinis, E. J. Wagner, M. A. Lazar, A. Shilatifard, and R. Shiekhattar. Integrator regulates transcriptional initiation and pause release following activation. *Mol Cell*, 56(1):128–139, 2014.

[157] M. Xie, W. Zhang, M.-D. Shu, A. Xu, D. A. Lenis, D. DiMaio, and J. A. Steitz. The host Integrator complex acts in transcription-independent maturation of herpesvirus microRNA 3 ends. *Genes & Development*, 29(14):1552–1564, 2015.

[158] T. Albrecht and E. Wagner. snRNA 3 end formation requires heterodimeric association of integrator subunits. *Molecular and Cellular Biology*, 32(6):1112–1123, 2012.

[159] T. Nojima, M. Tellier, J. Foxwell, C. Ribeiro de Almeida, S. Tan-Wong, S. Dhir, G. Dujardin, A. Dhir, S. Murphy, and N. Proudfoot. Deregulated Expression of Mammalian lncRNA through Loss of SPT6 Induces R-Loop Formation, Replication Stress, and Cellular Senescence. *Molecular Cell*, 72(6):970–984.e7, 2018.

[160] J. R. Skaar, A. L. Ferris, X. Wu, A. Saraf, K. K. Khanna, L. Florens, M. P. Washburn, S. H. Hughes, and M. Pagano. The Integrator complex controls the termination of transcription at diverse classes of gene targets. *Cell Res*, 25(3):288–305, 2015.

[161] N. Levitt, D. Briggs, A. Gil, and N. J. Proudfoot. Definition of an efficient synthetic poly(A) site. *Genes & Development*, 3(7):1019–1025, 1989.

[162] A. Roth, Z. Weinberg, A. G. Y. Chen, P. B. Kim, T. D. Ames, and R. R. Breaker. A widespread self-cleaving ribozyme class is revealed by bioinformatics. *Nature Chemical Biology*, 10(1):56–60, 2014.

[163] C. E. Olivero, E. Martínez-Terroba, J. Zimmer, C. Liao, E. Tesfaye, N. Hooshdaran, J. A. Schofield, J. Bendor, D. Fang, M. D. Simon, J. R. Zamudio, and N. Dimitrova. p53 Activates the Long Noncoding RNA Pvt1b to Inhibit Myc and Suppress Tumorigenesis. *Mol Cell*, 77(4):761–774, 2020.

[164] L. Winkler, M. Jimenez, J. T. Zimmer, A. Williams, M. D. Simon, and N. Dimitrova. Functional elements of the cis-regulatory lincRNA-p21. *Cell Reports*, 39(3), 2022.

[165] N. A. Rosa-Mercado, J. T. Zimmer, M. Apostolidi, J. Rinehart, M. D. Simon, and J. A. Steitz. Hyperosmotic stress alters the RNA polymerase II interactome and induces readthrough transcription despite widespread transcriptional repression. *Mol Cell*, 81(3):502–513.e4, 2021.

[166] K. Brannan, H. Kim, B. Erickson, K. Glover-Cutter, S. Kim, N. Fong, L. Kiemele, K. Hansen, R. Davis, J. Lykke-Andersen, and D. L. Bentley. mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. *Mol Cell*, 46(3):311–24, 2012.

[167] K. A. Nilson, C. K. Lawson, N. J. Mullen, C. B. Ball, B. M. Spector, J. L. Meier, and D. H. Price. Oxidative stress rapidly stabilizes promoter-proximal paused Pol II across the human genome. *Nucleic Acids Res*, 45(19):11088–11105, 2017.

[168] L. H. Williams, G. Fromm, N. G. Gokey, T. Henriques, G. W. Muse, A. Burkholder, D. C. Fargo, G. Hu, and K. Adelman. Pausing of RNA polymerase II regulates mammalian developmental potential through control of signaling networks. *Mol Cell*, 58(2):311–322, 2015.

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

[170] B. Erickson, R. M. Sheridan, M. Cortazar, and D. L. Bentley. Dynamic turnover of paused Pol II complexes at human promoters. *Genes Dev*, 32(17-18):1215–1225, 2018.

[171] T. T. Tettey, X. Gao, W. Shao, H. Li, B. A. Story, A. D. Chitsazan, R. L. Glaser, Z. H. Goode, C. W. Seidel, R. C. Conaway, J. Zeitlinger, M. Blanchette, and J. W. Conaway. A Role for FACT in RNA Polymerase II Promoter-Proximal Pausing. *Cell Rep*, 27(13):3770–3779 e7, 2019.

[172] C. Dienemann, B. Schwalb, S. Schilbach, and P. Cramer. Promoter Distortion and Opening in the RNAPolymerase II Cleft. *Mol Cell*, 73:97–106, 2019.

[173] M. G. Jaeger, B. Schwalb, S. D. Mackowiak, T. Velychko, A. Hanzl, H. Imrichova, M. Brand, B. Agerer, S. Chorn, B. Nabet, F. M. Ferguson, A. C. Muller, A. Bergthaler, N. S. Gray, J. E. Bradner, C. Bock, D. Hnisz, P. Cramer, and G. E. Winter. Selective Mediator dependence of cell-type-specifying transcription. *Nat Genet*, 2020.

[174] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A Probabilistic Programming Language. *J Stat Softw*, 76(1):32, 2017.

[175] F. Chen, X. Gao, and A. Shilatifard. Stably paused genes revealed through inhibition of transcription initiation by the TFIIH inhibitor triptolide. *Genes Dev*, 29(1):39–47, 2015.

[176] A. Krumm, L. B. Hickey, and M. Groudine. Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. *Genes Dev*, 9:559–572, 1995.

[177] D. A. Hendrix, J.-W. Hong, J. Zeitlinger, D. S. Rokhsar, and M. S. Levine. Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. *PNAS*, 105(22):7762–7767, 2008.

[178] W. Shao, S. G. M. Alcantara, and J. Zeitlinger. Reporter-ChIP-nexus reveals strong contribution of the Drosophila initiator sequence to RNA polymerase pausing. *eLife*, 8:e41461, 2019.

[179] Y. Chen, N. Negre, Q. Li, J. O. Mieczkowska, M. Slattery, T. Liu, Y. Zhang, T.-K. Kim, H. H. He, J. Zieba, Y. Ruan, P. J. Bickel, R. M. Myers, B. J. Wold, K. P. White, J. D. Lieb, and X. S. Liu. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature Methods*, 9(6):609–614, 2012.

[180] N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, and B. Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet*, 39(3):311–8, 2007.

[181] E. Calo and J. Wysocka. Modification of Enhancer Chromatin: What, How, and Why? *Mol Cell*, 49(5):825–837, 2013.

[182] L. Xie, C. Pelz, W. Wang, A. Bashar, O. Varlamova, S. Shadle, and S. Impey. KDM5B regulates embryonic stem cell self-renewal and represses cryptic intragenic transcription. *EMBO J*, 30(8):1473–84, 2011.

[183] T. J. Parry, J. W. Theisen, J. Y. Hsu, Y. L. Wang, D. L. Corcoran, M. Eustice, U. Ohler, and J. T. Kadonaga. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev*, 24(18):2013–8, 2010.

[184] N. Yamanaka, K. F. Rewitz, and M. B. O'Connor. Ecdysone control of developmental transitions: lessons from Drosophila research. *Annu Rev Entomol*, 58:497–516, 2013.

[185] C. M. Uyehara and D. J. McKay. Direct and widespread role for the nuclear receptor EcR in mediating the response to ecdysone in Drosophila. *PNAS*, 116(20):9893–9902, 2019.

[186] T. Juven-Gershon, J. Y. Hsu, J. W. Theisen, and J. T. Kadonaga. The RNA polymerase II core promoter - the gateway to transcription. *Curr Opin Cell Biol*, 20(3):253–9, 2008.

[187] K. A. Reimer, C. A. Mimoso, K. Adelman, and K. M. Neugebauer. Co-transcriptional splicing regulates 3' end cleavage during mammalian erythropoiesis. *Mol Cell*, 81(5):998–1012.e7, 2021.

[188] L. V. Lee, F. M. Pascasio, F. D. Fuentes, and G. H. Viterbo. Torsion dystonia in Panay, Philippines. *Adv Neurol*, 14:137–51, 1976.

[189] A. Westenberger, C. J. Reyes, G. Saranza, V. Dobricic, H. Hanssen, A. Domingo, B.-H. Laabs, S. Schaake, J. Pozojevic, A. Rakovic, K. Grütz, K. Begemann, U. Walter, D. Dressler, P. Bauer, A. Rolfs, A. Münchau, F. J. Kaiser, L. J. Ozelius, R. D. Jamora, R. L. Rosales, C. C. E. Diesta, K. Lohmann, I. R. König, N. Brüggemann,

and C. Klein. A hexanucleotide repeat modifies expressivity of X-linked dystonia parkinsonism. *Annals of Neurol*, 85(6):812–822, 2019.

[190] S. Makino, R. Kaji, S. Ando, M. Tomizawa, K. Yasuno, S. Goto, S. Matsumoto, M. D. Tabuena, E. Maranon, M. Dantes, L. V. Lee, K. Ogasawara, I. Tooyama, H. Akatsu, M. Nishimura, and G. Tamiya. Reduced neuron-specific expression of the TAF1 gene is associated with X-linked dystonia-parkinsonism. *Am J Hum Genet*, 80(3):393–406, 2007.

[191] N. Ito, W. T. Hendriks, J. Dhakal, C. A. Vaine, C. Liu, D. Shin, K. Shin, N. Wakabayashi-Ito, M. Dy, T. Multhaupt-Buell, N. Sharma, X. O. Breakefield, and D. C. Bragg. Decreased N-TAF1 expression in X-linked dystonia-parkinsonism patient-specific neural stem cells. *Dis Model Mech*, 9(4):451–62, 2016.

[192] S. Capponi, N. Stöffler, E. B. Penney, K. Grütz, S. Nizamuddin, M. W. Vermunt, B. Castelijns, C. Fernandez-Cerado, G. P. Legarda, M. S. Velasco-Andrada, E. L. Muñoz, M. A. Ang, C. C. E. Diesta, M. P. Creyghton, C. Klein, D. C. Bragg, P. De Rijk, and H. T. M. Timmers. Dissection of TAF1 neuronal splicing and implications for neurodegeneration in X-linked dystonia-parkinsonism. *Brain communications*, 3(4):fcab253–fcab253, 2021.

[193] A. Domingo, D. Amar, K. Grütz, L. V. Lee, R. Rosales, N. Brüggemann, R. D. Jamora, E. Cutiongco-Dela Paz, A. Rolfs, D. Dressler, U. Walter, D. Krainc, K. Lohmann, R. Shamir, C. Klein, and A. Westenberger. Evidence of TAF1 dysfunction in peripheral models of X-linked dystonia-parkinsonism. *Cellular and molecular life sciences: CMLS*, 73(16):3205–3215, 2016.

[194] J. Pozojevic, S. M. Algodon, J. N. Cruz, J. Trinh, N. Bruggemann, J. Lass, K. Grutz, S. Schaake, R. Tse, V. Yumiceba, N. Kruse, K. Schulz, V. K. A. Sreenivasan, R. L. Rosales, R. D. G. Jamora, C. C. E. Diesta, J. Matschke, M. Glatzel, P. Seibler, K. Handler, A. Rakovic, H. Kirchner, M. Spielmann, F. J. Kaiser, C. Klein, and A. Westenberger. Transcriptional Alterations in X-Linked Dystonia-Parkinsonism Caused by the SVA Retrotransposon. *Int J Mol Sci*, 23(4), 2022.

[195] T. S. Niranjan, C. Skinner, M. May, T. Turner, R. Rose, R. Stevenson, C. E. Schwartz, and T. Wang. Affected kindred analysis of human X chromosome exomes to identify novel X-linked intellectual disability genes. *PloS one*, 10(2):e0116454–e0116454, 2015.

[196] J. A. Kosmicki, K. E. Samocha, D. P. Howrigan, S. J. Sanders, K. Slowikowski, M. Lek, K. J. Karczewski, D. J. Cutler, B. Devlin, K. Roeder, J. D. Buxbaum, B. M. Neale, D. G. MacArthur, D. P. Wall, E. B. Robinson, and M. J. Daly. Refining the role of de novo protein-truncating variants in neurodevelopmental disorders by using population reference samples. *Nat Genet*, 49(4):504–510, 2017.

[197] S. E. Hurst, E. Liktor-Busa, A. Moutal, S. Parker, S. Rice, S. Szelinger, G. Senner, M. F. Hammer, L. Johnstone, K. Ramsey, V. Narayanan, S. Perez-Miller, M. Khanna, H. Dahlin, K. Lewis, D. Craig, E. H. Wang, R. Khanna, and M. A. Nelson. A novel variant in TAF1 affects gene expression and is associated with X-linked TAF1 intellectual disability syndrome. *Neuronal signaling*, 2(3):NS20180141–NS20180141, 2018.

[198] Y. Xu, N. Man, D. Karl, C. Martinez, F. Liu, J. Sun, C. J. Martinez, G. M. Martin, F. Beckedorff, F. Lai, J. Yue, A. Roisman, S. Greenblatt, S. Duffort, L. Wang, X. Sun, M. Figueroa, R. Shiekhattar, and S. Nimer. TAF1 plays a critical role in AML1-ETO driven leukemogenesis. *Nature Communications*, 10(1):4925, 2019.

[199] C. E. Metcalf and D. A. Wassarman. DNA binding properties of TAF1 isoforms with two AT-hooks. *J Biol Chem*, 281(40):30015–23, 2006.

[200] M. Anandapadamanaban, C. Andresen, S. Helander, Y. Ohyama, M. I. Siponen, P. Lundström, T. Kokubo, M. Ikura, M. Moche, and M. Sunnerhagen. High-resolution structure of TBP with TAF1 reveals anchoring patterns in transcriptional regulation. *Nature Structural & Molecular Biology*, 20(8):1008–1014, 2013.

[201] H. Wang, E. C. Curran, T. R. Hinds, E. H. Wang, and N. Zheng. Crystal structure of a TAF1-TAF7 complex in human transcription factor IID reveals a promoter binding module. *Cell Research*, 24(12):1433–1444, 2014.

[202] E. C. Curran, H. Wang, T. R. Hinds, N. Zheng, and E. H. Wang. Zinc knuckle of TAF1 is a DNA binding module critical for TFIID promoter occupancy. *Scientific Reports*, 8(1):4630, 2018.

[203] T. Bhuiyan and H. T. M. Timmers. Promoter Recognition: Putting TFIID on the Spot. *Trends in Cell Biology*, 29(9):752–763, 2019.

[204] R. Dikstein, S. Ruppert, and R. Tjian. TAFII250 Is a Bipartite Protein Kinase That Phosphorylates the Basal Transcription Factor RAP74. *Cell*, 84(5):781–790, 1996.

[205] C. A. Mizzen, X.-J. Yang, T. Kokubo, J. E. Brownell, A. J. Bannister, T. Owen-Hughes, J. Workman, L. Wang, S. L. Berger, T. Kouzarides, Y. Nakatani, and C. D. Allis. The TAFII250 Subunit of TFIID Has Histone Acetyltransferase Activity. *Cell*, 87(7):1261–1270, 1996.

[206] S. Bhattacharya, X. Lou, P. Hwang, K. R. Rajashankar, X. Wang, J.- Gustafsson, R. J. Fletterick, R. H. Jacobson, and P. Webb. Structural and functional insight into TAF1–TAF7, a subcomplex of transcription factor II D. *Proceedings of the National Academy of Sciences*, 111(25):9103–9108, 2014.

[207] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[208] K. M. Sakamoto, K. B. Kim, A. Kumagai, F. Mercurio, C. M. Crews, and R. J. Deshaies. Protacs: chimeric molecules that target proteins to the Skp1-Cullin-F box complex for ubiquitination and degradation. *Proceedings of the National Academy of Sciences of the United States of America*, 98(15):8554–8559, 2001.

[209] A. B. Patel, R. K. Louder, B. J. Greber, S. Grünberg, J. Luo, J. Fang, Y. Liu, J. Ranish, S. Hahn, and E. Nogales. Structure of human TFIID and mechanism of TBP loading onto promoter DNA. *Science*, 362(6421):eaau8872, 2018.

[210] M. Martin. Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. *EMBnet journal*, 17:10–12, 2011.

[211] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8):907–915, 2019.

[212] S. Anders, P. T. Pyl, and W. Huber. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–9, 2015.

[213] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and S. Genome Project Data Processing. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–9, 2009.

[214] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

[215] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, 2010.

[216] H. Xu, X. Luo, J. Qian, X. Pang, J. Song, G. Qian, J. Chen, and S. Chen. FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One*, 7(12):e52249, 2012.

[217] E. G. Wilbanks and M. T. Facciotti. Evaluation of Algorithm Performance in ChIP-Seq Peak Detection. *PLOS ONE*, 5(7):e11471, 2010.

[218] A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–2, 2010.

[219] F. Ramirez, D. P. Ryan, B. Gruning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dundar, and T. Manke. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*, 44(W1):W160–5, 2016.

[220] U. Ohler, G.-c. Liao, H. Niemann, and G. M. Rubin. Computational analysis of core promoters in the Drosophila genome. *Genome Biol*, 3(12), 2002.

[221] C. Y. Lim, B. Santoso, T. Boulay, E. Dong, U. Ohler, and J. T. Kadonaga. The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev*, 18(13):1606–17, 2004.

[222] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engström, M. C. Frith, A. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6):626–35, 2006.

[223] J. Ponjavic, B. Lenhard, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, and A. Sandelin. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol*, 7(8):R78, 2006.

[224] R. Javahery, A. Khachi, K. Lo, B. Zenzie-Gregory, and S. T. Smale. DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol Cell Biochem*, 14(1):116–127, 1994.

[225] T. W. Burke and J. T. Kadonaga. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev*, 10:711–724, 1996.