

**FUNCTIONAL STUDIES OF TRANSCRIPTION, FROM RNA-PROTEIN
INTERACTIONS, TO PROMOTER PROXIMAL PAUSING, TO THE
FUNDAMENTAL UNITS OF TRANSCRIPTION INITIATION**

A Dissertation

Presented to the Faculty of the Graduate School
of Cornell University

In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Jacob Michael Tome

May 2018

© 2018, Jacob Michael Tome

FUNCTIONAL STUDIES OF TRANSCRIPTION, FROM RNA-PROTEIN INTERACTIONS, TO PROMOTER PROXIMAL PAUSING, TO THE FUNDAMENTAL UNITS OF TRANSCRIPTION INITIATION

Jacob Michael Tome, Ph. D.

Cornell University 2018

To understand gene regulation, we need both targeted approaches to probe individual regulatory components, and systems level approaches to understand the functional state of cells. Presented here are several studies at different points on this functional spectrum, with a focus on the crucial regulatory step of promoter proximal pausing.

RNA-protein interactions have critical roles in gene regulation. We adapted an Illumina GAIIx sequencer to make several millions of these measurements with a High-Throughput Sequencing – RNA Affinity Profilng (HiTS-RAP) assay. Millions of cDNAs are sequenced, bound by the *E. coli* replication terminator protein Tus, and transcribed *in situ*, whereupon Tus halts transcription leaving RNA stably attached to its template DNA. Binding of fluorescently-labeled protein is then quantified in the sequencer. By measuring the affinity of mutagenized libraries of an RNA aptamer to NELF-E, an RNA binding subunit of the pausing factor NELF, we show that this interaction is due to a short RNA motif, but the three dimensional structure of the aptamer is critical for its high affinity. We used this aptamer as an *in vivo* inhibitor of the interaction between NELF-E and nascent RNA in *Drosophila* S2 cells. Pausing was globally reduced, but promoters with the transcription factor GAF were

unchanged. Thus, the interaction between NELF-E and nascent RNA is not critical for pausing when GAF aids NELF recruitment, but is a strong component of NELF recruitment elsewhere.

In higher eukaryotes, the timing and level of transcription at gene promoters by RNA Polymerase II (Pol II) is specified largely by the sum of information from the promoter itself and from distal enhancers. Transcription widely occurs at enhancers, suggesting Pol II may be a ubiquitous nexus of regulatory signaling. To explore this, we sequenced nascent RNAs at single-molecule resolution to identify Pol II initiation, capping, and pause sites. Our analyses reveal distinct sequence-specified pause classes associated with differences in RNA capping dynamics. Initiation typically occurs within large clusters, especially at gene promoters. Integrated analysis of nearby chromatin and transcription factors suggests a model of gene regulation in which Pol II initiation provides a biophysical scaffold to create and maintain regulatory domains.

BIOGRAPHICAL SKETCH

I was born in 1988 in Denver to Sue and Tracy Tome. When I was two, we moved back to their hometown of Meadville, PA. Though the move took us away from the mountains, it brought us back to a place where all of my grandparents, aunts, uncles, and cousins were. Because of this, I got to grow up with my extended family close by. Though we may not be wealthy, I'm extremely lucky in another way. I come from a wonderful, supportive family. My mom stayed home with us, and is a constant source of support. We always spent our time together, even when my dad worked 60+ hours per week in a toolshop.

Meadville is very much a rust belt town. Through growing up here, I feel that I have a good, emotional understanding of our current political climate. This also instilled in me a desire to be a part of our society's future, where we work together to unravel the secrets of the universe, and use this to improve lives. When I was in middle school, my dad, a toolmaker, went to college to become a teacher. Seeing him work overtime while raising a family and going to college absolved any doubts I had about the value of education, and taking advantage of it as early as possible. Though I may come from a small town, I got an excellent education here: when I went to college, I was somewhat surprised to find that I could hold my own with the best: I tested into the highest possible Spanish language class, ahead of kids who went to private schools that cost what my parents make in a year.

When I went to Bucknell University as an undergraduate, I studied Biochemistry. Ever since I can remember, I've always known that I wanted to study the science of life (though maybe not with any concrete end goal, I've found as I decide what I want to do after my Ph.D). Cornell for my Ph.D was also an easy choice. As soon as I started the interview here, it just felt right. And it still feels right.

To everyone who's helped me get here

ACKNOWLEDGMENTS

First and foremost, I'd like to thank John Lis. If I could match even 10% of your enthusiasm for science, then I'd be set. It was truly a pleasure to work with a mentor who is so excited about the work we do, and who takes interest even in the small stuff. Because of this, I'm leaving your lab with better training than I could have ever hoped to have received. Your boundless interest in transcription means that there are no bounds to what we can learn with you. I'd also like to thank you for creating such a wonderful environment in the lab. I've worked with some pretty great people in the Lis lab.

I'd also like to thank Abdullah Ozer. I'm absolutely serious when I say that the best grad students in our lab, and the people I trust most, rotated with you. Most of what I know about biochemistry, I learned through you. I owe a lot of my success to you giving me a great project as a rotation student (HiTS-RAP), and being a great mentor and collaborator every step of the way.

Nate Tippens has also been a great collaborator. I think our skill sets and ways of operating complement each other pretty well. I could not have gotten off of the ground with coPRO analysis without all of your hard work, and I've become a better programmer with your help. I'm also much better at arguing.

John Pagano was a huge help in getting started with NELF aptamer work. He generously let provided me with everything I needed to use NELFapt for HiTS-RAP, and to start using it *in vivo*.

I'd like to thank Fabiana Duarte for always being there. I'm so lucky that my best friend was also my baymate.

Janis Werner has always looked after us in the lab. A lot of what we do would be much harder without you.

Everyone at the Cornell genomics core has been extremely helpful. Sequencing runs incredibly smoothly here. I'm very thankful that I've had Peter to consult with when I design a new library. Ann, Tom, and Jen were a extremely helpful as I learned to run my old dinosaur of a GAIIX.

HiTS-RAP would not have been possible without the help of Illumina. Gary Schroth got us valuable support from them. Dan Gheba taught me most of what I know about the GAIIX, and was always there when I had questions (or problems).

I'd like to thank everyone else that I've worked with. I feel so grateful to too many of you to mention, and look forward to being part of the same science family for decades to come.

Thanks to all of my friends in Ithaca. You've been there for the good times, and are there for the bad.

I'd like to thank my family. Everything that I've done is only because of the love and support that I've gotten from you. Even though we'll be further apart after I move on, I'm sure we'll stay just as close. I've had endless love and support from my parents throughout my life. Sam, I'm so thankful that you've been at Cornell through most of my time here. Our breakfasts, bike rides, and backpacking trips together are the highlights of my time in Ithaca. Eli, I'm looking forward to seeing you more from Seattle. Our trip out West was a turning point in my

life: I went from someone who wants to get out and explore, to someone who takes every possible opportunity to go and do something rad everyday.

I'd like to thank Ithaca for being the best place I can imagine to live. I never wanted to leave, in the end. I think we have the best of everything here: natural beauty, great food, and awesome people. I never ran out of places to explore here. In my empty moments throughout the day, I'll often have a flashback to some beautiful place I biked or ran past in the last few years.

Table of Contents

BIOGRAPHICAL SKETCH	iv
ACKNOWLEDGMENTS	vi
Table of Contents	ix
List of Figures	xiii
List of Tables	xvi
List of Abbreviations	xvii
Chapter 1: Introduction	1
Transcription	1
The Transcription Cycle	1
Modes of Gene Regulation	5
The Core Promoter.....	6
Transcription Factors	7
Chromatin Environment.....	8
Three-Dimensional Organization and Enhancers	9
Steps in the Transcription Cycle Important in This Dissertation.....	13
Initiation	13
Capping.....	14
Promoter Proximal Pausing	16
The Role of DSIF.....	21
The Role of NELF	23
P-TEFb Mediated Pause Escape	25
Strategies for Studying Gene Regulation.....	26
Observe (and Re-observe).....	26
High-throughput Sequencing	26
PRO-seq	28
Perturb.....	30
Chapter 2: Comprehensive Analysis of RNA-Protein Interactions by High Throughput Sequencing-RNA Affinity Profiling	32
Introduction.....	32
Results.....	34

Tus Stably Halts Transcription	34
Transcription Halting with Tus is Functional on the GAIIX Sequencing Instrument	39
Measuring Equilibrium Binding Constants for the GFP Aptamer and Mutants with HiTS-RAP.....	43
Comprehensive mutagenesis of the GFP Aptamer	45
High-Throughput Affinity Profiling of Drosophila NELF-E	50
Discussion.....	55
Methods.....	58
Purification of proteins	58
Library preparation	59
EMSA of halted transcription complex	60
Transcription halting on 454 beads.....	61
Transcription halting on glutathione beads.....	61
HiTS-RAP.....	62
Sequencing and data extraction	64
Loss of signal correction.....	65
K _d calculation.....	67
EMSA of GFPapt and NELFapt and mutants.....	69
RNA secondary structure predictions	69
Notes on HiTS-RAP	71
Efficiency of halting	71
Strategies for solving K _d s.....	72
HiTS-RAP as a Tool for Identification, Characterization, and Optimization of Aptamers	73
Overview of HiTS-RAP.....	75
RNA encoding DNA Library.....	75
Writing an .xml recipe for a GAIIX run	77
Cluster generation and sequencing.	85
dsDNA regeneration and transcription halting.	86
Protein Binding.	87
Data Analysis.....	88

Anticipated Results	91
Chapter 3: In Vivo Disruption of NELF-E’s Interaction with Nascent RNA by NELFapt Expression.....	96
Introduction.....	96
Results.....	99
Overview of NELFapt Inhibition Strategy	99
Characterization of the NELFapt Construct	102
NELFapt Inhibition in the First Experiment Caused a Downstream Shift in Pausing	109
NELF Aptamer Inhibition in a Second Experiment Confirmed GAF’s Role in Recruiting NELF.....	120
Conclusions and Discussion	129
NELF-E’s Interaction with Nascent RNA Is More Important at Moderately Paused Genes	129
Difficulties in Determining the Effect of NELFapt from These Data	132
Future Directions	135
Materials and Methods.....	138
Transfection and Stable Cell Line Generation.....	138
Northern Blot	138
EMSA	139
PRO-seq	139
Chapter 4: Human transcription characterized by single molecule coupling of Pol II initiation and active site.....	140
Introduction: Nascent RNA Sequencing for Dissecting Gene Regulatory Mechanisms	140
Results.....	142
CoPRO Facilitates Identification of Transcription Initiation Sites.....	142
Two Types of Promoter Proximal Pausing.....	147
Sequence Largely Determines Pause Position.....	147
Different Capping Dynamics in Early and Late Pause Classes	150
Early Pausing Is Linked to a Poised State of Chromatin	151
Comparison of Pause Sites Used by Nearby TSNs	154
Transcription Initiation Occurs in Clusters.....	156

Arrangement of TSSes within TIDs	156
The Strongest TSSes Determine Nucleosome Phasing within TIDs	158
Gene Promoters Are Found in Large TIDs.....	165
TID Structure and Chromatin Environment Are Tightly Linked	165
Transcription Factors' Distribution within TIDs Reflects Their Function	172
Possible Functional Roles of TIDs in Gene Regulation	172
Methods.....	174
coPRO Experiments.....	174
Cap State Spike-ins	179
Sequence alignment	180
Read summarization and normalization.....	181
Defining transcription start nucleotides (TSNs), start sites (TSSs), and initiation domains (TIDs)	182
Pause classification	183
Metaplots and heatmaps.....	184
Chapter 5: General Conclusions and Future Directions.....	185
General Conclusions	185
Future Directions	186
HiTS-RAP.....	186
Aptamers as Inhibitors	190
Termination, Capping and Transcription Initiation Domains.....	192
Bibliography	197

List of Figures

Chapter 1

Figure 1. 1: The Transcription Cycle	3
Figure 1. 2: Steps leading to promoter proximal pausing	16
Figure 1. 3: Status of promoter proximally paused polymerase	20

Chapter 2

Figure 2. 1: T7 RNA polymerase halting with Tus gives stable complexes containing DNA and functional RNA	35
Figure 2. 2: Transcription halts or terminates at Tus-bound ter sites in the non-permissive orientation	36
Figure 2. 3: EGFP only binds to transcription complexes with full-length GFP aptamer RNA ..	38
Figure 2. 4: Schematic of HiTS-RAP templates	39
Figure 2. 5: RNA-protein interactions can be assayed by HiTS-RAP on an Illumina GAIIx instrument	40
Figure 2. 6: Corrections applied to HiTS-RAP data	42
Figure 2. 7: Confirmation of HiTS-RAP measured affinities by EMSA	44
Figure 2. 8: Analysis of GFPapt by HiTS-RAP	46
Figure 2. 9: Secondary structure predictions of GFP aptamer single point mutants with higher or lower affinities	47
Figure 2. 10: Secondary structure predictions of GFP aptamer double point mutants with significantly higher or lower affinities than predicted by the single point mutants	49
Figure 2. 11: EMSA confirmation of HiTS-RAP affinities for NELFapt	51
Figure 2. 12: Analysis of NELFapt by HiTS-RAP	52
Figure 2. 13: Secondary structure predictions of NELF-E aptamer double point mutants with ..	54
Figure 2. 14: HiTS-RAP data analysis workflow	88
Figure 2. 15: Example binding curves.	93

Chapter 3

Figure 3. 1: NELFapt PRO-seq Experimental Design	99
Figure 3. 2: NELFapt <i>in vivo</i> expression strategy	100
Figure 3. 3: <i>In vitro</i> characterization of the NELFapt transgene	103
Figure 3. 4: Quantification of <i>in vivo</i> NELFapt expression by qPCR	105
Figure 3. 5: Quantification of <i>in vivo</i> NELFapt expression by Northern Blot	106
Figure 3. 6: Correlation between replicates in the first PRO-seq experiment	108
Figure 3. 7: Effects of copper treatment in S2 cells	109
Figure 3. 8: Differences between WT and NELFapt are greater than either upon induction	110
Figure 3. 9: Pausing at NELF Bound genes	111
Figure 3. 10: Effect of NELFapt inhibition in pause regions with NBEs	113
Figure 3. 11: CG5854 shows a downstream shift in pausing in the first PRO-seq	114

Figure 3. 12: Quantifying shifts in pause distributions with CDF area.	115
Figure 3. 13: Calling genes with shifts in pause distribution in the first PRO-seq.....	116
Figure 3. 14: A shift in pause distributions with NELFapt the first PRO-seq experiment.....	117
Figure 3. 15: Genes with shifts in pause distribution are depleted for GAF, but have high nucleosome occupancy and positioning.....	118
Figure 3. 16: Shifted pause distributions are a result of reduction of proximal pausing with no change in pausing at the first nucleosome.	119
Figure 3. 17: Correlation between replicates in the second PRO-seq experiment.....	121
Figure 3. 18: <i>CG5854</i> 's downstream shift is not reproduced in the second PRO-seq.....	121
Figure 3. 19: The shift in pause distributions with NELFapt was not reproduced in the second PRO-seq experiment	122
Figure 3. 20: NELFapt and scrambled control have very similar effects	124
Figure 3. 21: NELFapt and scrambled NELFapt cause a global reduction in pausing at most promoters that is amplified upon induction.	126
Figure 3. 22: Pause regions affected in the second PRO-seq are depleted for GAF, and have high nucleosome positioning and occupancy.....	128
Figure 3. 23: Aptamer inhibition of NELF-E results in a reduction of pausing at genes where NELF-E to nascent RNA interactions are critical.....	130
Figure 3. 24: NELF-E's interaction with nascent RNA has a greater contribution to NELF recruitment in the absence of GAF.	131

Chapter 4

Figure 4. 1: coPRO simultaneously measures initiation site and active site of engaged RNA polymerase II genome-wide.....	142
Figure 4. 2: Dispersion of initiation within TSSes	144
Figure 4. 3: <i>ACTB</i> , an extremely focused promoter.	146
Figure 4. 4: Two distinct modes of promoter proximal pausing exhibit different capping dynamics.	149
Figure 4. 5: DNA sequence around and pause sites.....	150
Figure 4. 6: Transcription factors, chromatin environment, and Pol II phosphorylation at Early and Late pause TSNs.	152
Figure 4. 7: Expression Level and Pausing at Early and Late TSNs	153
Figure 4. 8: Dispersion of Pausing.....	153
Figure 4. 9: Many TSNs Pause at Shared Sites in the <i>MAPK1</i> Promoter.....	155
Figure 4. 10: A global view of transcription initiation shows rules for divergent pairing and reveals widespread complex organization.	156
Figure 4. 11: The first nucleosome is positioned by paused polymerase	158
Figure 4. 12: The maxTSS is usually part of a divergent pair	160
Figure 4. 13: Nucleosome phasing in complex TIDs	162

Figure 4. 14: Minor TSSes are restricted to the gaps between nucleosomes phased by stronger TSSes.	163
Figure 4. 15: Large TIDs Are Strong Outliers	164
Figure 4. 16: ARID1B, an Exceptionally Large TID.....	164
Figure 4. 17: TID organization is linked to chromatin environment	166
Figure 4. 18: Features of TIDs Irrespective of Transcript Stability.....	167
Figure 4. 19: Distinct binding modes within TIDs	168
Figure 4. 20: Native MNase ChIP for histone modifications at TIDs reveals tight covariation with transcription initiation that is supported by multiple datasets	169
Figure 4. 21: Histone modifications show strong patterns around TIDs.	171
Figure 4. 22: Schematic of coPRO	175
Figure 4. 23: Enzymatic steps for cap state selection in coPRO	176
Figure 4. 24: coPRO library design schematic	178

Chapter 5

Figure 5. 1: Observing Pol II termination with coPRO Uncapped.	194
---	-----

List of Tables

Table 3. 1: Oligonucleotides Used in HiTS-RAP..... 63

Table 4. 1: Oligonucleotides used for coPRO 179

List of Abbreviations

ADP	Adenosine Diphosphate
ATP	Adenosine Triphosphate
BEAF	Boundary Element Associated Factor
bp	Basepairs
Brd4	bromodomain containing factor 4
BrU	BromoUridine
CAGE	Capped Analysis of Gene Expression
Cdk	Cyclin-Dependent Kinase __
ChIP	Chromatin ImmunoPrecipitation
coPRO	Coordinated 5' and 3' Precision Run-On
CPA	Cleavage and PolyAdenylation (site)
CTD	Carboxy-Terminal Domain
CTP	Cytosine Triphosphate
DNA	DeoxyriboNucleic Acid
DSIF	5,6-Dichloro-1- β -D-ribofuranosylbenzimidazole (DRB) Sensitivity Inducing Factor
EMSA	ElectroMobility Shift Assay
eRNA	Enhancer RNA
GAF	GAGA Factor
GFP	Green Fluorescent Protein
GRO-seq	Global Run-on Sequencing
GTP	Guanidine Triphosphate
H3K27ac	Histone H3 Lysine 4 acetyl
H3K4me1	Histone H3 Lysine 4 monomethyl
H3K4me3	Histone H3 Lysine 4 trimethyl
HiTS-RAP	High Throughput Sequencing, RNA Affinity Profiling

kb	KiloBases (1000 basepairs)
lncRNA	Long Noncoding RNA
M1BP	Motif 1 Binding Protein
mM	MilliMolar
MNase	Micrococcal Nuclease
mRNA	Messenger RNA
NELF	Negative ELongation Factor
NELFapt	NELF-E aptamer
nM	NanoMolar
nt	Nucleotides
PARP-1	Poly(ADP - ribose) polymerase 1
PIC	Pre-Initiation Complex
Pol I	RNA Polymerase I
Pol II	RNA Polymerase II
Pol III	RNA Polymerase III
pppRNA	5' triphosphate RNA
pRNA	5' monophosphat RNA
PRO-seq	Precision Run-On Sequencing
P-TEFb	Positive Transcription Elongation Factor b
RNA	RiboNucleic Acid
SEC	Super Elongation Complex
TBP	TATA Binding Protein
TFIIx	Transcription initiation Factor ____
TID	Transcription Initiation Domain
TSN	Transcription Start Nucleotide
TSS	Transcription Start Site
UTR	UnTranslated Region

Chapter 1: Introduction

Transcription

Transcription is the process by which genetic information in deoxyribonucleic acid, or DNA is transcribed into ribonucleic acid, or RNA, which then can serve as the template for protein synthesis. Since this basic framework of gene regulation was developed (Jacob and Monod, 1961), our understanding of how transcription is regulated has evolved constantly. Transcription in metazoans is carried out by three RNA polymerase enzymes, Pol I, Pol II, and Pol III, named after the ammonium sulfate fractions in which they were originally purified (Roeder and Rutter, 1969). Pol II is responsible for transcription of most protein coding genes, and for transcription of many non-coding transcripts. Mammalian genomes are pervasively transcribed, giving rise to a plethora of both protein coding and noncoding transcripts (Djebali *et al.*, 2012). Varying transcriptional programs are the means by which a genome that is essentially identical in every cell within an organism gives rise to different cell types, and aberrations in these programs are the cause of many diseases (Lee and Young, 2013). Therefore, understanding transcription and its regulation is paramount for our understanding of development, normal homeostasis, and disease.

The Transcription Cycle

Transcription by Pol II must proceed through several steps to produce a full length transcript (Fuda, Ardehali and Lis, 2009), each of which is potentially rate limiting in certain cases. A schematic of this process is presented in Figure 1. 1. For transcription factors and the general transcription machinery to bind, chromatin must first be made accessible. Next, transcription factors recruit the general transcription machinery to promoters. The general

transcription factors then recruit RNA Polymerase II (Pol II) and facilitate initiation (Murakami *et al.*, 2013). To accomplish this, general transcription factors and activators first assemble a pre-initiation complex (PIC) at a promoter. These protein factors recruit Pol II to the promoter. There is then an ATP dependent remodeling step in which the DNA is unwound at the transcription start site, allowing Pol II to begin transcribing (Sainsbury, Bernecky and Cramer, 2015). Now, the initiated polymerase must break its contacts with the promoter to become a transcribing polymerase (Saunders, Core and Lis, 2006; Fuda, Ardehali and Lis, 2009). During this process, general transcription factors are evicted from the polymerase, contacts with the promoter and associated factors are broken, the transcription bubble collapses, TFIIF phosphorylates Ser5 of the CTD, and Pol II proceeds quickly to the site of the promoter proximal pause (Saunders, Core and Lis, 2006; Fuda, Ardehali and Lis, 2009; Adelman and Lis, 2012; Sainsbury, Bernecky and Cramer, 2015). As Pol II transitions from initiation to promoter proximal pausing, capping takes place (Rasmussen and Lis, 1993). Pol II remains in a promoter proximal paused state for varying amounts of time, providing an opportunity for regulatory input and assembly of an elongation complex. Pause release and entry into productive elongation is marked by P-TEFb mediated phosphorylation of Pol II. Unstable transcripts, like upstream divergent and eRNA transcripts will terminate shortly after pause release. For mRNA and lncRNA genes, Pol II transcribes through a transcription unit that could be up to many hundreds of kilobases. This phase is marked by cotranscriptional splicing, and gradual maturation and acceleration of the transcribing Pol II (Jonkers and Lis, 2015). At the end of the transcription unit, cleavage and polyadenylation takes place after Pol II transcribes through the cleavage and polyadenylation (CPA) site. Pol II continues transcribing beyond the CPA site, at a slower rate, as termination is carried out (Proudfoot, 2016; Schwalb *et al.*, 2016). The final step is recycling

of the Pol II enzyme and associated factors, so that they can be reused for subsequent rounds of transcription.

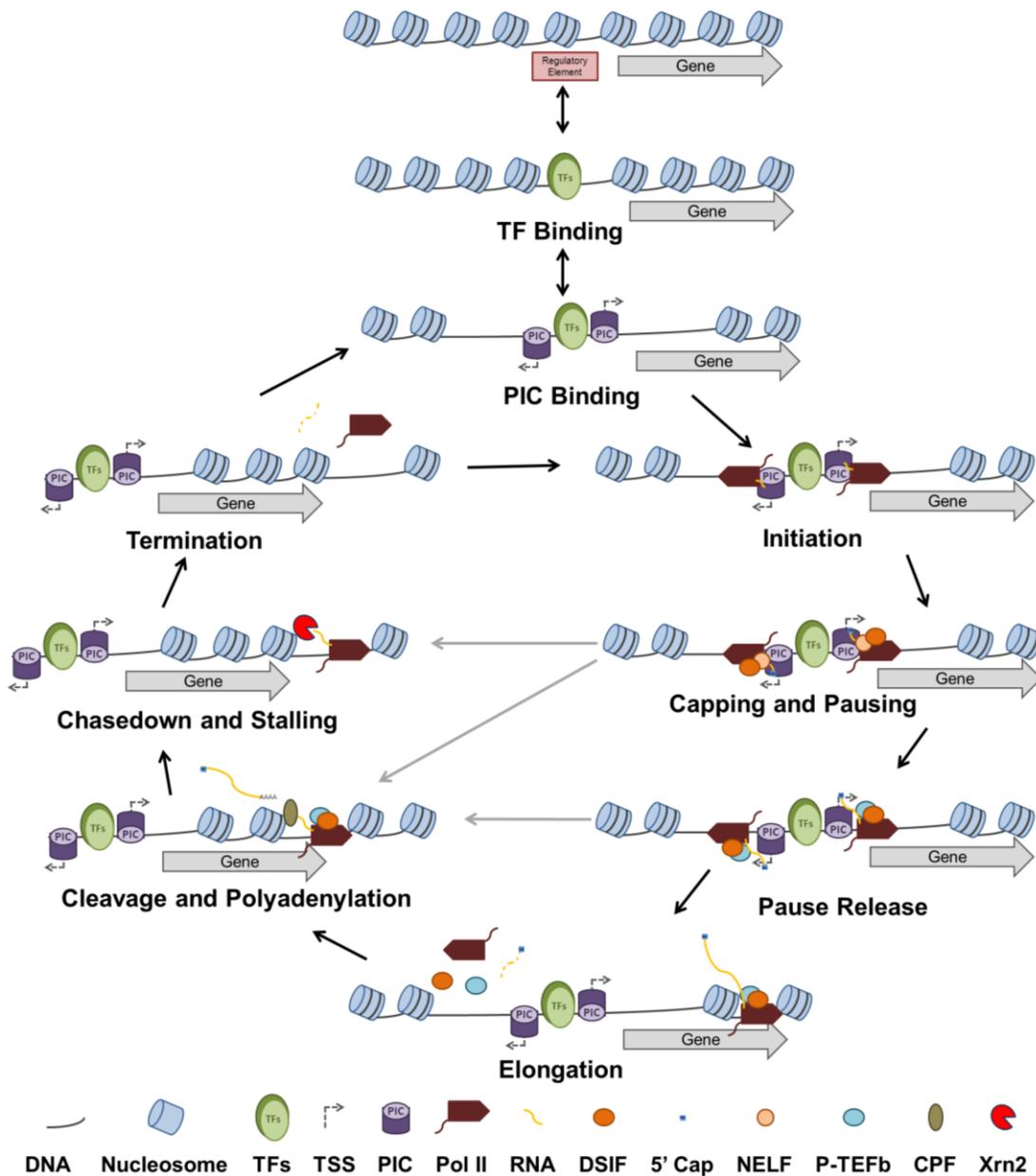


Figure 1. 1: The Transcription Cycle

Schematic of transcription adapted from (Fuda, Ardehali and Lis, 2009). The path of the normal cycle is indicated with black arrows. Alternative paths, such as termination before capping, and early termination after pause release, are indicated with gray arrows. Reversible steps are indicated by two headed arrows. Bidirectional mammalian promoters are presented here: after pause release, the upstream divergent polymerase terminates, and the polyadenylated transcript is degraded. Enhancer transcription would be most similar to all upstream divergent transcription, with termination soon after pause release.

Though the process of transcription is presented here as a cycle, in reality transcription at any given site in the genome moves through this process as more of a web, indicated with gray arrows in Figure 1. 1. For example, chromatin does not need to be made accessible for each round: accessible sites often remain as such in differentiated cells(Neph *et al.*, 2012; Thurman *et al.*, 2012). Many transcription units (such as eRNAs) terminate shortly after pause release, and thus do not go through an elongation phase(Almada *et al.*, 2013). At some rate, paused Pol II terminates before undergoing release(Buckley *et al.*, 2014; Krebs *et al.*, 2017). Even from an mRNA promoter, Pol II could terminate after pause release with some frequency(Brannan *et al.*, 2012). At any given locus, multiple transcripts could be taking any number of paths through this web. For example, at an mRNA promoter, the mRNA transcript will undergo pause release and enter productive elongation while the upstream divergent transcript terminates shortly after release(Leighton J Core *et al.*, 2014). Transcription can serve many purposes, of which production of a functional RNA product is only one. For example, promoter proximal pausing has a role in maintaining nucleosome free promoters and modulating chromatin environment(Adelman and Lis, 2012), and enhancers produce transcripts as they carry out their role of regulating transcription at a distance(Leighton J Core *et al.*, 2014). These extra functions can involve any step in the transcription cycle: for example, splicing at 'lncRNA' enhancers has been shown to promote transcriptional activity at nearby targets(Engreitz *et al.*, 2016). Thus, these many paths through the transcription cycle all serve the purpose of regulating gene expression.

Throughout the transcription cycle, the status of the carboxyl-terminal domain (CTD) of the largest subunit of Pol II (Rpb1) is a bellwether of these events, and plays a crucial role in recruitment of critical regulatory factors at different stages of the cycle(Saunders, Core and Lis,

2006). This disordered domain consists of many repeats of the seven amino acid consensus sequence YSPTSPS. The CTD is important in regulation of the activity of the polymerase and of RNA processing events through the transcription cycle. It serves as a site of assembly for processing factors including capping proteins, the spliceosome, and termination factors, as well as a site of binding for regulatory proteins (Prelich, 2002). Many such proteins function in part by directly modifying the CTD. Notably, serines 2, 5, and 7 (Ser2, Ser5, and Ser7) and threonine 4 (Thr4) are phosphorylated depending on the step of the transcription cycle. Generally, Ser5 and Ser7 phosphorylation peak near the beginning of the gene, while Ser2 phosphorylation peaks in the gene body (Saunders, Core and Lis, 2006), and Thr4P peaks at the end of the gene. These modifications are critical, as a modified CTD plays a role in recruiting different regulatory factors to polymerase through the transcription cycle (Harlen and Churchman, 2017). A recent study by the Churchman lab used mass spectrometry to identify factors that associate with Pol II with various CTD modifications in yeast (Harlen *et al.*, 2016). They find that Ser2P is most associated with elongation, and that termination factors are preferentially associated with Pol II with a Thr4P modified CTD. Beyond this, modifications in different regions of the CTD can have different functions, as different heptad repeats within the CTD have different proclivities to modification (Schuller *et al.*, 2016).

Modes of Gene Regulation

Any step of the transcription cycle can be modulated to tune the levels of mRNA produced for a given gene. Much of this dissertation will focus on promoter proximal pausing as a regulated step in gene expression. Outlined below are some of the basic mechanisms of regulation. A basic theme of this section is that regulatory potential expands exponentially as

larger functional units of gene expression are considered. Thereby, information poor core promoters are subjected to extremely tight control.

The Core Promoter

Transcription initiation in eukaryotes is somewhat enigmatic. Though many core promoter elements are specified by well-defined DNA sequences that are recognized by the general transcription factors (Vo Ngoc *et al.*, 2017), many promoters have very poor or no matches to these sequence elements (Kadonaga, 2012; Lenhard, Sandelin and Carninci, 2012; Siebert and Söding, 2014). Furthermore, promoters that do contain some of these sequence elements often do not contain the others (Ohler *et al.*, 2002), so a ‘complete’ core promoter is exceedingly rare. In fact, the simple rational design of a ‘super core promoter’ resulted in a promoter that drives transcription at a level higher even than many viral promoters (Juven-Gershon, Cheng and Kadonaga, 2006). Thus, a perfect promoter would in fact be undesirable at an endogenous gene as it would drive transcription at a level that is too high and thus affords little opportunity for fine tuning. This means that a well-organized core promoter is absolutely not required for transcription initiation, especially in mammals, and that the core promoters of genes are in reality relatively information poor, though some core promoter elements are associated with certain types of transcription units, such as the TCT motif which is present at most ribosomal protein genes (Kadonaga, 2012; Lenhard, Sandelin and Carninci, 2012). Despite this, transcription is regulated with exquisite precision, through components of the transcription cycle beyond the core promoter and initiation. Many of these modes of regulation conspire to adjust the accessibility of different sequence elements to transcription factors, and to tune the availability of the transcription machinery at a locus, so that initiation occurs at the best match to a core promoter at the site where these other factors have driven the basal transcription

machinery. Steps beyond initiation, such as promoter proximal pausing and entry into productive elongation then provide additional layers of control.

Transcription Factors

Regulation of transcription is carried out at many levels beyond the core promoter elements. Foremost is the availability of transcription factors (TFs). The human genome encodes approximately 1,400 sequence specific transcription factors that can fine-tune the expression of promoters by binding their sequence elements near the promoter and then recruiting Pol II and the general transcription machinery, pause factors, factors involved in altering chromatin environment, or carry out any number of other roles(Vaquerizas *et al.*, 2009). Transcription factors are some of the most potent marks of cell identity and drivers of cellular responses(Lee and Young, 2013). As such, the genes encoding key transcription factors are among those with the most elaborate systems in place for their regulation, from super-enhancers(Hnisz *et al.*, 2013), to extremely complex enhancer organization(Fulco *et al.*, 2016) and functional relationships(Diao *et al.*, 2016), to vast chromatin domains around their promoters(Chen *et al.*, 2015). With the vast number of transcription factors expressed in any give cell type, the possible combinatorics for promoters are huge. TFs tend to bind in clusters, which are often anchored by a few stable binders(Wei *et al.*, 2017). Thus, transcription factor binding underlies most of the other mechanisms of gene regulation that will be discussed here.

Understanding the DNA binding specificities of transcription factors is critical for assessing their function. Several methods exist to quantitatively characterize transcription binding specificities. SELEX-based method such as PB-seq(Guertin *et al.*, 2012) or Bind-N-Seq(Zykovich, Korf and Segal, 2009) have been used to quantitatively determine affinities of

transcription factors to naked genomic DNA or random oligonucleotides. The HiTS-FLIP method(Nutiu *et al.*, 2011), which serves as the foundation of my work with RNA binding measurements that is described in Chapter 2, images the binding of fluorescently labeled transcription factors to random oligonucleotides at different protein concentrations on an Illumina sequencing instrument to quantitatively determine affinity to every possible n-mer (up to 12-mers). In recent years, quantitative characterizations of transcription factor binding have been scaled up immensely by the Taipale group. They used SELEX to determine the binding specificities of hundreds of TFs(Jolma *et al.*, 2013), to characterize the specificity of pairs of TFs which synergize to bind in tandem in ways that would not be predicted from each individual factors' specificity(Jolma *et al.*, 2015), and to characterize affinity in the context of the a nucleosome(Zhu *et al.*, 2017). The hope of these types of characterizations is that they could be used to model the regulatory state of a cell if enough parameters are known, such as the availability of transcription factors, their affinity to different DNA elements, and the barrier that different chromatin environments present.

Chromatin Environment

Chromatin environment is another mechanism by which gene regulation occurs. Many transcription factors are not able to bind their target sequence elements when they are buried in closed chromatin(Guertin and Lis, 2010). The first step in activating transcription in a region of closed chromatin is often the binding of a transcription factor that is able to recognize its target within closed chromatin (pioneer factor)(Zaret and Carroll, 2011). After this initial recruitment, there is constant cross-talk between chromatin environment and transcription. The presence of the transcription machinery keeps promoters accessible(Adelman and Lis, 2012). Furthermore, the histones of nearby nucleosomes are post translationally modified(Berger, 2002), and these

modifications themselves serve to modulate the structure of chromatin and recruit coactivators that recognize them(Bannister and Kouzarides, 2011). Chromatin state is often used to infer the activity at loci across the genome. Generally, the histone modifications H3K4me3 and H3K27ac are used as marks of active promoters, and H3K4me1 and H3K27ac for active enhancers(Heintzman *et al.*, 2007). The different histone modifications can thus be used to predict activity with varying success: for example, H3K9ac and H3K4me2 have been found to be most predictive of Pol II occupancy(Chen *et al.*, 2011). Integrated analysis of many histone marks has been used to develop an HMM based classifier to characterize chromatin state genome-wide(Ernst and Kellis, 2012): these powerful predictions are widely used to characterize loci as active/inactive and promoter/enhancer, etc. Thus, transcription factor binding, chromatin structure, histone post-translational modifications, and the presence of many co-activators all serve to expand the information content at promoters to tune gene expression. A recent study by the Wysocka and Adelman groups highlights these features and how interconnected and multifunctional each component is. They find that the catalytic activity of the H3K4 methyltransferase MLL3/4 is dispensable for its role in activation of transcription through enhancers, but that the presence of MLL is essential(Dorigi *et al.*, 2017). This indicates that these factors have a plethora of functions, and that the environment around promoters is extremely complex.

Three-Dimensional Organization and Enhancers

Finally, three dimensional genome organization is another important avenue of regulation. If the possible space for regulatory logic is expanded by considering different levels from the core promoter to TF binding to chromatin environment, then another massive gain in regulatory potential is made by considering large scale genome organization and the close spatial

proximity between loci separated by a long genomic distance that it creates. Each locus that joins a regulatory hub is subject to all of the other regulatory levels discussed up to this point. Enhancers, or genomic loci that regulate promoters from a distance, have been known to be critical in the regulation of gene expression for decades (Serfling, Jasin and Schaffner, 1985). Enhancers often contain critical regulatory information for genes, and are thus frequently misregulated in disease (Smith and Shilatifard, 2014). Thus, recent efforts have focused on mapping the three dimensional organization of the genome to understand where functional distal interactions occur. In order to pack a 2 meter long DNA genome into a cell's nucleus, it undergoes a great deal of packaging. The genome is folded into large Topologically Associated Domains (TADs) (Dixon *et al.*, 2012), which constrain potential distal interactions. The logic for formation of three dimensional interactions is complex. CTCF is a DNA-binding master regulator of folding (Tang *et al.*, 2015), and the cohesion complex is a critical cofactor which forms a ring which physically tethers distant loci. Pol II could serve as the motor for extruding DNA through cohesion rings to mediate distal looping interactions (Busslinger *et al.*, 2017). In addition to activating interactions, transcriptional repression is often carried out through high-order structures as well. For example, a repressive chromatin environment can spread across entire TADs (Lieberman-aiden *et al.*, 2009). In summary, three dimensional genome organization can expand all of the regulatory principles discussed up to this point to many loci that act at a distance to affect transcription from a gene promoter, thus constituting a major fraction of the total regulatory potential at promoters.

The nature of enhancers is somewhat difficult to capture with a simple definition. Classically, they are defined as a DNA element that can activate transcription at a distance independent of their orientation. They are subject to many of the same levels of regulation as

promoters, and thus can be used to tune the environment around their target sites. They are sites of transcription factor binding(Whyte *et al.*, 2013), carry histone modifications(Calo and Wysocka, 2013), are often transcribed(Kim *et al.*, 2010), and loop to promoters in three-dimensional space(Zabidi and Stark, 2016). Much of the field uses histone modifications to identify enhancers(Heintzman *et al.*, 2007; Ernst and Kellis, 2012). In the Lis lab, many of our efforts to characterize enhancers focus on the fact that they are transcribed. PRO-cap, a method for mapping transcription initiation sites genome-wide by sequencing the 5' ends of capped nascent RNAs(Kruesi *et al.*, 2013), is an extremely sensitive way of identifying sites of transcription initiation. Thus, it provides a way of calling regulatory elements, whether they are enhancers or promoters, in a sensitive, comprehensive fashion from a single assay. Putative enhancers are sites of initiation that do not produce a stable transcript, when assessed by to Capped Analysis of Gene Expression, or CAGE(Carninci *et al.*, 2006), a method for mapping the start sites of accumulated mRNAs. Fundamentally, transcription at enhancers is very similar to transcription a gene promoters, and in fact, the histone modifications normally used to distinguish promoters from enhancers are strongly correlated with transcriptional activity, and are therefore not likely a reliable mark of enhancer activity(Andersson *et al.*, 2014; Leighton J. Core *et al.*, 2014). Therefore, gene promoters, upstream divergent promoters, and eRNA promoters share a common architecture, so it is not surprising that any site of transcription initiation can potentially act as an enhancer: even gene and lncRNA promoters can have enhancer activity(Diao *et al.*, 2017). However, not all enhancers are transcribed: some enhancers that are activated during heat shock and seem to be functional carry histone H4 acetylation, but no detectible transcription(Vihervaara *et al.*, 2017). In fact, different features of enhancers may be critical for their activity in different contexts. In an elegant study, the Lander group found

through a series of deletion mutants that elongation and splicing from lncRNA genes that act as enhancers is critical for their enhancer activity in some cases, while in others just initiation and promoter proximal pausing there is sufficient(Engreitz *et al.*, 2016). Similarly, the Shilatifard group has shown that Paf1 at enhancers can help stimulate pause release at their target promoters(Chen *et al.*, 2017).

Some of the most exciting work in the field of gene regulation in the past several years seeks to catalogue enhancers comprehensively. The massively parallel reporter assay STARR-seq tests genomic fragments for enhancer activity(Arnold *et al.*, 2013). STARR-seq assesses the intrinsic enhancer activity of the elements being interrogated: other approaches assess enhancer function by perturbation of the test element's endogenous locus. Large-scale CRISPR screens tile across a locus containing an essential gene, so that cell growth is compromised when a potential target is critical for the gene's expression(Fulco *et al.*, 2016; Gasperini *et al.*, 2016; Diao *et al.*, 2017). Recently, this approach has been adapted to single-cells, so that enhancer activity can be tested at non-essential genes and in a way that does not make assumptions about which gene is the target(Dixit *et al.*, 2016). Thus, between these functional tests of enhancers and efforts to map the three-dimensional organization of the genome, a major goal of the field is to develop a comprehensive map of the functional connections between distant loci in order to understand regulation at gene promoters. The nature of enhancers varies depending on the specific context of individual genes, and the assay and criteria used to define enhancers. However, some common themes emerge from all of these definitions which will be touched upon throughout this dissertation.

Steps in the Transcription Cycle Important in This Dissertation

Initiation

Pol II initiation occurs at tens of thousands of sites across the genome, only a small fraction of which is at mRNA promoters. Efforts to map initiation comprehensively have utilized the ends of mature transcripts (Carninci *et al.*, 2006), mapping of polymerase genome-wide (Kim *et al.*, 2005), and more recently mapping of nuclear RNAs' 5' ends (Nechaev *et al.*, 2010; Kruesi *et al.*, 2013) to infer initiation sites. In mammals, Pol II generally initiates bidirectionally from separate core promoters oriented outward (Core, Waterfall and Lis, 2008; Preker *et al.*, 2008; Seila *et al.*, 2008). These upstream divergent promoters were unexpected and unique to mammals (Kwak *et al.*, 2013), and present an intriguing way of regulating gene expression by influencing promoters' local environment. Upstream divergent transcription originates from separate core promoter elements, with separate PICs as evidenced by separate peaks in high resolution mapping of PIC components by ChIP-Exo (Leighton J. Core *et al.*, 2014; Pugh and Venters, 2016). Though it originates from a separate core promoter from the stable transcript, upstream divergent transcription is often not completely autonomous. In a genome-wide reporter screen of autonomous promoter activity, the van Steensel group found that upstream divergent transcription relies on TF binding sites within the stable transcript's promoter, while the primary transcript from a divergent pair can often initiate transcription without sequence from the upstream divergent promoter (Arensbergen *et al.*, 2016). Upstream divergent transcription's primary function is likely to alter the chromatin environment around promoters. It creates a large nucleosome free region that can make a wide region of transcription factor binding sites available (Scruggs *et al.*, 2015).

Another unexpected class of Pol II transcribed elements is enhancers(Kim *et al.*, 2010). Our group has found that the architecture of transcription initiation at enhancers and promoters is virtually identical(Leighton J. Core *et al.*, 2014). The major difference between promoters for stable transcripts like mRNA and lncRNA promoters and those of unstable transcripts like upstream divergent and enhancer promoters is after the entry into productive elongation. Unstable transcripts rapidly terminate, while stable transcription proceeds for up to many hundreds of kilobases. Much of this is due to the enrichment of 5' splice sites versus cleavage and polyadenylation (CPA) sites. Sites with an abundance of splice sites and a lack of CPA sites enter productive elongation, while an enrichment of CPA sites early on promotes rapid cleavage and polyadenylation and termination(Almada *et al.*, 2013). This U1/CPA axis can explain much of this stability(Leighton J. Core *et al.*, 2014). Chromatin environment may also help to specify the stability of transcripts. In flies, a first nucleosome with the histone variant H2A.Z is associated with a much lower barrier to passage of Pol II(Christopher M Weber, Ramachandran and Henikoff, 2014), illustrating the ability of chromatin environment to influence the activity of Pol II. In summary, initiation of Pol II transcriptions occurs in many contexts, which serve different purposes in carrying out gene regulation.

Capping

The addition of a 5' inverted methylguanosine cap is essential for mRNA export from the nucleus, stability, and translation(Ramanathan *et al.*, 2016). Capping occurs leading up to and during pausing(Figure 1. 2)(Rasmussen and Lis, 1993; Nilson *et al.*, 2015). The cap is added in three main steps. First the RNA 5' triphosphate is reduced to a diphosphate. Next, an inverted GTP is added in a 5' to 5' triphosphate linkage. Finally, the G cap is methylated at carbon 7. This particular cap structure is unique to Pol II transcription(Grohmann *et al.*, 1978). It is found

at virtually all Pol II transcripts, including mRNAs, lncRNA, upstream divergent transcripts, and eRNAs (Leighton J. Core *et al.*, 2014). Therefore, sequencing of the capped ends of RNAs has been used extensively for mapping sites of Pol II initiation with precision (Carninci *et al.*, 2006; Nechaev *et al.*, 2010; Kruesi *et al.*, 2013; Andersson *et al.*, 2014). Capping occurs very soon after initiation, as promoter proximally paused Pol II is associated primarily with capped transcripts (Rasmussen and Lis, 1993). After initiation, it is likely that DSIF binds polymerase almost immediately (Figure 1. 2B), and then in turn helps recruit capping enzyme (Figure 1. 2C) (Wen and Shatkin, 1999; Mandal *et al.*, 2004) and NELF (Figure 1. 2D) (Missra and Gilmour, 2010; Li *et al.*, 2013). Cdk7 mediated phosphorylation of Ser5 of the CTD could play a role in stimulating the addition of the cap (Nilson *et al.*, 2015). Therefore, although capping occurs very early, it is not immediate. In fact, capping can occur as soon as RNA emerges from polymerase in an *in vitro* transcription system subjected to a high salt wash, and is actually delayed when nuclear extract is present (Nilson *et al.*, 2015). Therefore it is possible that in the milieu of the nucleus, capping is delayed from its earliest possible occurrence by some other factor present during the transition from initiation to pausing, or simply by the normal order of assembly of the capping and pausing machinery.

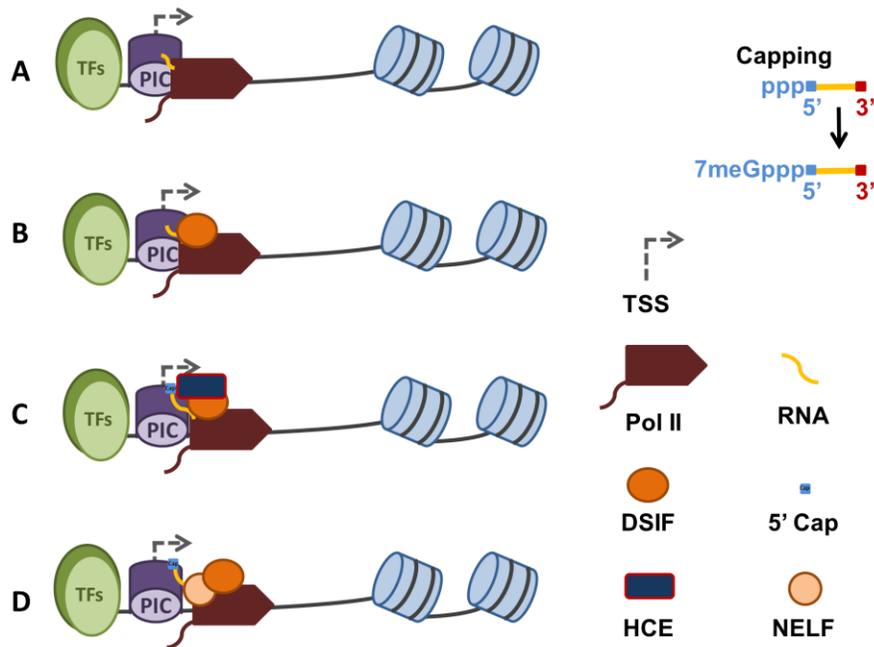


Figure 1. 2: Steps leading to promoter proximal pausing

Capping of the nascent RNA and promoter proximal pausing occur after initiation, and are intimately coupled as will be discussed in Chapter 4. **A.)** After initiation, the uncapped (triphosphate) 5' end of the nascent transcript first emerges from the nascent RNA. **B.)** DSIF is recruited to polymerase rapidly, partly through an interaction with the nascent RNA. **C.)** The capping enzyme (HCE, Human Capping Enzyme) is recruited through interactions with DSIF and Pol II. Ser5 CTD phosphorylation mediated by Cdk7 enhances its activity. **D.)** NELF is recruited through interactions with Pol II, DSIF, and nascent RNA. TFs such as GAF can aid recruitment.

Promoter Proximal Pausing

The mechanism by which the pause is induced is a major focus of this work. Paused polymerase was first characterized by UV-ChIP at heat shock genes in *Drosophila* (Gilmour and Lis, 1986). It was later found to be capable of resuming transcription in a radioactive run-on experiment, showing that it is transcriptionally engaged (Rougvie and Lis, 1988). Paused polymerase was then shown to be associated with predominantly capped transcripts from 20-40 nt in length (Rasmussen and Lis, 1993). The first genome-wide mapping of Pol II with ChIP-Chip showed a strong enrichment at the 5' end of genes, raising the intriguing possibility that

promoter proximal could be a general feature of Pol II transcription(Kim *et al.*, 2005). Mapping nuclear run-ons genome-wide in human cells later showed that pausing is indeed widespread(Core, Waterfall and Lis, 2008). Pause release is often the rate limiting step for determining level of transcription from a promoter: thus, it is one of the most critical steps in the transcription cycle for understanding gene regulation. It frequently changes in response to developmental signals(Levine, 2011), or in cellular responses such as heat shock(Dig B. Mahat *et al.*, 2016; Duarte *et al.*, 2016).

Paused polymerase is transcriptionally engaged (and therefore capable of resuming elongation), has transcribed ~20-60 nucleotides, and has had a guanosine cap added to the transcript's 5' end(Rougvie and Lis, 1988; Rasmussen and Lis, 1993; Saunders, Core and Lis, 2006; Fuda, Ardehali and Lis, 2009; Adelman and Lis, 2012). Most measurements show that the half-life of most paused Pol II is on the order of several minutes and that extreme cases remain stably paused for several hours (Henriques *et al.*, 2013; Buckley *et al.*, 2014; Chen, Gao and Shilatifard, 2015; Shao and Zeitlinger, 2017). However, some recent measurements show shorter half-lives for heat shock genes(Krebs *et al.*, 2017) and globally than expected(Nilson *et al.*, 2017). Many genes, especially those which are capable of rapid induction, are held in this paused state, poised for transcription elongation(Muse *et al.*, 2007; Zobeck *et al.*, 2010; Galbraith *et al.*, 2013). However, all genes go through some version of a pause, as evidenced by the fact that flavopiridol, an inhibitor of Positive Transcription Elongation Factor b (P-TEFb) subunit Cdk9, prevents elongation in almost all genes, even those without an observable promoter proximally paused polymerase(Chao and Price, 2001; Peterlin and Price, 2006; Jonkers, Kwak and Lis, 2014; Gressel *et al.*, 2017). P-TEFb is the major factor involved in alleviating promoter proximal pausing(Wada, Takagi, Yamaguchi, Watanabe, *et al.*, 1998;

Peterlin and Price, 2006). Therefore, the pause is an important, ubiquitous step in the transcription cycle. It serves as a major point of regulation at many genes (Core, Waterfall and Lis, 2008; Core *et al.*, 2012), and at the very least serves as a checkpoint for the assembly of a polymerase that is fully capable of productive elongation even at genes without an observable pause. This regulatory checkpoint is likely universal. Even in organisms like yeast with no observable promoter proximal pausing, perturbation of the pausing and elongation factor DSIF results in a slow moving Pol II, inefficient splicing, and increased noncoding antisense and convergent transcription (Booth *et al.*, 2016; Shetty *et al.*, 2017). A host of different factors associates with polymerase in the short space from initiation to pause release (Adelman and Lis, 2012), including PIC components (Kwak *et al.*, 2013), DSIF (Li *et al.*, 2013), capping enzyme (Mandal *et al.*, 2004; Nilson *et al.*, 2015), NELF, splicing machinery, elongation factors, nucleosome chaperones, histone modifying enzymes, topoisomerases (Baranello *et al.*, 2016; Miller *et al.*, 2017), and termination factors (Brannan *et al.*, 2012).

The establishment and release of a promoter proximal pause involve the interplay of many factors. During the pause, the net effect of these factors is to regulate the amount of time that Pol II remains in a paused state before transcribing through the gene. NELF, DSIF, and P-TEFb are three protein complexes that play pivotal roles in this process (Saunders, Core and Lis, 2006; Adelman and Lis, 2012). Negative ELongation Factor (NELF) and 5,6-Dichloro-1- β -D-ribofuranosylbenzimidazole (DRB) Sensitivity Inducing Factor (DSIF) bind to a Pol II with a hypophosphorylated CTD and cooperatively maintain the promoter proximal pause (Saunders, Core and Lis, 2006). This means that when these factors bind, Ser5 has been phosphorylated by the Cdk7 subunit of the general transcription factor TFIIF, but Ser2 has not yet been phosphorylated. These two complexes were initially identified in a biochemical fractionation

approach by their ability to render transcription elongation sensitive to DRB inhibition in *in vitro* transcription reactions(Wada, Takagi, Yamaguchi, Ferdous, *et al.*, 1998; Yamaguchi *et al.*, 1999). The nucleoside analog DRB is a cyclin-dependent kinase inhibitor, so the fact that it inhibits entry into elongation indicates that a cyclin-dependent kinase is responsible for transitioning from a paused to elongating polymerase. Therefore, DRB represses transcription by inhibiting the step that causes inhibitory factors to leave the polymerase, meaning that NELF and DSIF are unable to be rearranged to produce an elongating polymerase. P-TEFb is the DRB inhibited cyclin-dependent kinase that is most significant in facilitating pause escape(Marshall and Price, 1992). It binds the promoter proximally paused polymerase complex and facilitates pause escape by phosphorylating Ser2 of the CTD and DSIF(Peng, 1996, 1998; Lis *et al.*, 2000; Yamada *et al.*, 2006; Zobeck *et al.*, 2010). This results in a rearrangement of the complex that favors elongation. At this point, NELF leaves the polymerase, and DSIF remains, but acts as a positive elongation factor(Hartzog *et al.*, 1998; Peterlin and Price, 2006). In summary, the promoter proximal pausing stage of the transcription cycle is accentuated by the NELF and DSIF complexes, and relieved by P-TEFb.

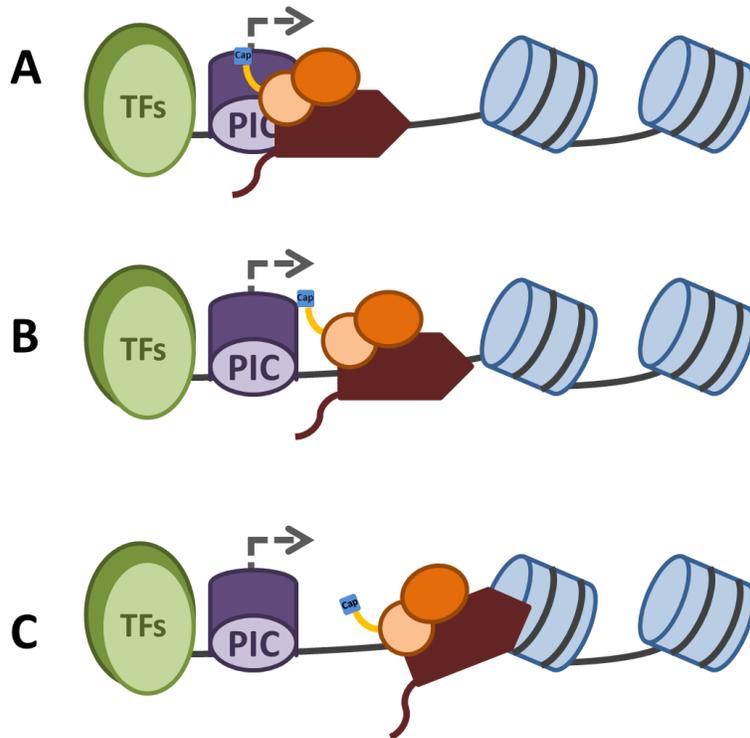


Figure 1. 3: Status of promoter proximally paused polymerase

Three possible states of promoter proximal pausing are determined by the primary barrier to transcription there. **A.)** At very proximally paused polymerase, interactions with the PIC contribute to the pause. **B.)** At intermediate distance, the position of the pause is determined by a combination of the energy landscape of DNA sequence, and the rate at which pause factors bind. **C.)** At some sites, the energy barrier of transcribing through the first nucleosome is responsible for a small fraction of pausing.

The precise location of the promoter proximal pause depends upon the sequence composition of the DNA template as well as the action pause factors, the PIC, and chromatin. In flies, Pol II that is paused very proximal to the TSS has not broken all of its interactions with the PIC components (Figure 1. 3A)(Kwak *et al.*, 2013). More distally paused polymerase is no longer associated with the PIC and is still upstream of the first nucleosome; thus it is held by interactions with DSIF and NELF (Figure 1. 3B). Interactions with the first nucleosome constitute a third potential pause-inducing boundary (Figure 1. 3C). Though this has been postulated as a major mechanism of promoter proximal pausing(Jimeno-González, Ceballos-Chávez and Reyes, 2015), only the extreme distal portion of the pause is associated with the

energetic barrier of the nucleosome(Kwak *et al.*, 2013). The barrier constituted by the first nucleosome is context specific: nucleosomes that have incorporated the histone variant H2A.Z present a significantly lower barrier *in vivo* than non-H2A.Z nucleosomes(Christopher M Weber, Ramachandran and Henikoff, 2014). The location of the DSIF and NELF dependent promoter proximal pause is likely determined by a combination of sequence and factor recruitment. The pause in flies is associated with a strong RNA-DNA hybrid within polymerase, with a weaker hybrid just downstream, consistent with a tendency to pause at a strong RNA-DNA hybrid and possibly backtracking from a weak hybrid to reach that pause(Nechaev *et al.*, 2010), a pattern which is conserved in mammals(Gressel *et al.*, 2017). The rate of pause factor recruitment also plays a role in pausing. In flies, GAF bound genes are typically very highly paused, and devoid of nucleosomes(Fuda *et al.*, 2015). GAF is able to recruit NELF to promoters independent of transcription(Li *et al.*, 2013). Therefore, GAF primes promoters for pausing partially by increasing the availability of NELF. These sequence and pause factor driven models of pause site choice are not mutually exclusive: the DNA sequence can create an energy landscape that slows polymerase, which determines when pause factors are recruited. This concept is explored in Chapter 4. In summary, promoter proximal pausing can be caused by a variety of interrelated mechanisms, which can all collaborated to provide a regulatory checkpoint before entering productive elongation.

The Role of DSIF

DSIF is the first pause-inducing protein to associate with an initiated polymerase(Missra and Gilmour, 2010). It consists of two proteins, SPT4 and SPT5, thus named because they are homologues of two transcription factors of the same name in yeast(Wada, Takagi, Yamaguchi, Ferdous, *et al.*, 1998). This complex is rapidly recruited to initiated Pol II(Missra and Gilmour,

2010). DSIF seems to track Pol II from the pause region through the gene body(Andrulis, 2000). This is because it binds paused polymerase, acting as a repressor of elongation, but remains bound after entry into productive elongation, now enhancing the activity of the polymerase(Andrulis, 2000). The elongation factor potential of DSIF was most eloquently demonstrated in *in vitro* studies when it increased the transcription rate of nucleotide deprived or mutant polymerase(Wada, Takagi, Yamaguchi, Ferdous, *et al.*, 1998). This activity is likely carried out by a direct interaction between DSIF and Pol II. The two protein complexes have been found to coimmunoprecipitate in extracts(Wada, Takagi, Yamaguchi, Ferdous, *et al.*, 1998), suggesting that they have an affinity for each other. However, Gilmour and colleagues only see an interaction between DSIF and Pol II when the polymerase is engaged with a transcript longer than 18 nt by Electrophoretic Mobility Shift Assay (EMSA)(Missra and Gilmour, 2010). Regardless of the precise timing and extent of the interaction, DSIF certainly binds transcribing polymerase, and this interaction is likely strengthened by the presence of a transcript. DSIF's RNA binding potential is conferred by the SPT5 subunit. Human SPT5 contains four tandem KOW motifs; these known RNA binding motifs have a high degree of homology with NusG, a bacterial elongation factor which also binds RNA(Wada, Takagi, Yamaguchi, Ferdous, *et al.*, 1998). Furthermore, *Drosophila* SPT5 has been cross-linked to nascent RNA. Again, this interaction was only observed if more than 18 nt had been transcribed(Missra and Gilmour, 2010). The literature lacks a characterization of this interaction, however. It appears to lack sequence specificity, as the crosslinking experiment was done using an artificial G-less cassette, rather than a transcript of biological origin.

The Role of NELF

NELF is a pausing factor that binds after DSIF to accentuate promoter proximal pausing(Missra and Gilmour, 2010). Its association with the polymerase strongly inhibits entry into productive elongation(Yamaguchi *et al.*, 1999; Wu *et al.*, 2003, 2005). NELF is recruited to the paused polymerase complex after DSIF. It is a complex composed of four proteins: NELF A, B, C/D, and E. It is generally most strongly associated with polymerase at the pause site, as it is recruited to the paused polymerase during the pause, and is expelled once Pol II enters productive elongation(Lee *et al.*, 2008). In coimmunoprecipitation and electro-mobility shift assays, it shows no appreciable binding to DSIF or Pol II alone, but does bind to DSIF-Pol II complexes(Missra and Gilmour, 2010). Thus, in contrast with DSIF, it seems that NELF's sole role is to inhibit entry into productive elongation. NELF is also strongly associated with the transcription factor GAGA Factor (GAF); at highly paused genes it seems that GAF recruits NELF before transcription initiation, so that NELF can be quickly loaded onto the polymerase(Lee *et al.*, 2008; Li *et al.*, 2013). Thus, there are multiple mechanisms by which NELF is recruited to newly transcribing polymerase, ensuring that a promoter proximal pause is induced.

Much like DSIF, NELF contains an RNA binding domain. The NELF-E subunit contains an RNA Recognition Motif (RRM). In humans, this domain has been shown to be important for proper regulation of elongation by NELF; with a series of mutations, it has been shown that the NELF-E RRM likely binds in a sequence specific way, and has a strong preference for RNA(Yamaguchi *et al.*, 2002). However, Gilmour and colleagues did not find any interaction between NELF-E and paused transcripts in an *in vitro* transcription assay, when they were able to find a direct interaction between the SPT5 subunit of DSIF and the nascent transcript(Missra

and Gilmour, 2010). This could be because of the sequence specificity of the NELF-E RRM meant that the artificial G-less cassette in their experiment was a poor substrate for NELF-E binding. Recently, the Lis lab has made a strong case for a biological role of NELF-E RRM binding the nascent transcript in promoter proximal pausing. A Systematic Evolution of Ligands by EXponential enrichment (SELEX) against the NELF-E RRM revealed a seven nucleotide binding motif. This motif, the NELF Binding Element (NBE) is strongly associated with tight binding by NELF-E. Furthermore, the motif is enriched near the pause site of highly paused genes(Pagano *et al.*, 2014). This enrichment is found both before and after the pause site, suggesting that a mechanism more complicated than just a single binding event is taking place. We hypothesize that this additional site could either serve as a redundant site to ensure pausing, or that it could be involved in removal of NELF from the polymerase after pause escape. In addition, our lab has shown that human NELF-E binds a NBE like sequence in the HIV-1 TAR RNA(Rao *et al.*, 2006; Pagano *et al.*, 2014). In addition to the RRM of NELF-E, NELF has been found to associate with RNA through its other subunits *in vivo*(Vos *et al.*, 2016). Within the nucleus, other RNAs besides the nascent transcript may regulate NELF's activity by competing for binding, as enhancer RNAs have been shown to do(Schaukowitch *et al.*, 2014). The interaction between NELF and RNA may also be modulated by post-translational modification of NELF: NELF-E has been identified as a target for poly(ADP) ribosylation by PARP-1, and this modification reduces the affinity of NELF for RNA *in vitro*(Gibson *et al.*, 2016). Taken together, these data suggest that an interaction between NELF and nascent RNA is a widespread mechanism by which promoter proximal pause is maintained, and that this phenomenon is conserved from flies to humans. Furthermore, this interaction can be regulated, both through the intrinsic specificity of NELF binding RNA, and through modulation of NELF's ability to bind

nascent RNA. This is just one of the many interactions that function cooperatively to ensure that NELF is delivered to newly transcribing polymerase so that it can contribute to the promoter proximal pause.

P-TEFb Mediated Pause Escape

P-TEFb is the protein complex most involved in releasing the promoter proximal pause (Peterlin and Price, 2006). It exists *in vivo* as a heterodimer of cyclin-dependent kinase 9 (Cdk9) and a cyclin (either cyclin T or K in *Drosophila* (Peng, 1998), and cyclin T1, T2, or K in mammals) (Peng *et al.*, 1998). Cdk9 is the catalytic subunit of this complex, while the cyclin subunit generally acts as a regulatory subunit. It functions by binding the paused polymerase complex, and phosphorylating the Pol II CTD at Ser2, DSIF at SPT5 (Kim and Sharp, 2001; Yamada *et al.*, 2006), NELF at NELF-E (Fujinaga *et al.*, 2004), and itself. After these phosphorylation events the polymerase enters productive elongation; NELF is lost from the polymerase, DSIF begins to act to enhance the rate of transcription, and P-TEFb remains bound to the transcribing polymerase. P-TEFb is regulated by its association with the 7SK snRNA and HEXIM protein in the nucleoplasm: when engaged in this complex it is inactive (Nguyen *et al.*, 2001; Yang *et al.*, 2005). Some support exists for a variety of mechanisms of activating P-TEFb and targeting it to paused polymerase. Typically, these involve another factor such as a coactivator like Mediator (Galbraith *et al.*, 2013), a transcription factor such as c-Myc (Rahl *et al.*, 2010) or HSF (Dig B. Mahat *et al.*, 2016; Duarte *et al.*, 2016; Vihervaara *et al.*, 2017). Active P-TEFb is associated with a group of other factors that can also facilitate its role in pause release and elongation. In the promoter proximal region, it has long been known to interact with the potent chromatin interacting activator bromodomain containing factor 4, Brd4 (Yang *et al.*, 2005). In addition, P-TEFb is a component of the super elongation complex (SEC), which

promotes rapid and efficient Pol II transcription through gene bodies(Luo, Lin and Shilatifard, 2012). The plurality of methods for targeting P-TEFb to polymerase demonstrates the importance of this step in transcription.

Strategies for Studying Gene Regulation

Almost any study in science adheres to the principle of observe, perturb, re-observe. The philosophy behind this is that in order to fully understand a system, one must first observe it in its normal state. But, to elevate understanding, it is critical to understand how the status of the system changes in response to particular perturbations. It is this response to perturbations that provides real insight into the ways in which the components of a biological system function. However, when a lot is already known about how a system works, a new perspective of one aspect of it is equally valuable. This is the state that we find ourselves in in the new era of genomics. In a well-used cell line such as *Drosophila* S2 or human K562, hundreds of genome-wide datasets are available, enabling a savvy researcher to gain important new insights from comparison to these reference datasets. All of the work in this dissertation is related to one of these three simple steps of observe, perturb, re-observe. The rest of this section provides background on some of the key strategies that will be drawn upon.

Observe (and Re-observe)

High-throughput Sequencing

Illumina sequencing instruments are the workhorses of the genomics revolution. These instruments borrow from the principles of Sanger sequencing(Sanger, Nicklen and Coulson, 1977). They follow the incorporation of reversible chain terminating nucleotides at DNA

clusters of around 1000 molecules of identical sequence covalently linked to a glass flowcell(Bentley *et al.*, 2008). Isothermal PCR is used to generate these clusters: after a single DNA molecule hybridizes to a lawn of DNA oligos covalently linked to a glass surface, the lawn of oligos is used as primers during subsequent annealing and extension steps to generate a cluster of copies of the original DNA molecule. This scales up the single DNA molecule to a larger cluster unit that is more easily interrogated by microscopy. The DNA is made single stranded by chemically cleaving one strand away. The sequence of the cluster is then determined by incorporating single fluorescently labeled nucleotides at a time. A primer is annealed and a mixture of all 4 reversible chain terminating dNTPs with different fluorescent labels is incorporated. The sequencer is an automated TIRF microscope that images the nucleotide added to each cluster. The reversible chain terminator and fluorescent dye are then removed chemically. This process of incorporation, imaging, and reversal is repeated to walk across the sequence of each cluster in a process known as sequencing by synthesis. The instrument is thus repeatedly imaging dye labeled nucleotide incorporation, identifying clusters, and calling basepairs added using fluorescence intensity measurements. This process can be used to determine the sequences of millions of short DNA fragments in a single run. The sophisticated microfluidics and optics of these instruments makes them an attractive platform for developing new assays the measure interactions between fluorescently labeled molecules and DNA(Perkel, 2015). Inspired by a method that used a GAIIX to image fluorescently labeled transcription factor binding to DNA to solve K_{d} s of its interaction with tens of thousands of DNA sequences at once(Nutiu *et al.*, 2011), I developed a method to quantitatively solve thousands of K_{d} s for interactions between RNA and a protein(Tome *et al.*, 2014).

The GAIIX, an older instrument that I modified to image fluorescent protein interaction with clusters in Chapter 2, could sequence up to 240 million 150 bp reads in a single run. The NextSeq 500, a newer instrument used for sequencing PRO-seq libraries in Chapters 3 and 4, can sequence up to 400 million 150 bp reads in a single run, but at around 1/10th the cost of a GAIIX run. Therefore, in addition to their usefulness in sequencing genomes *de novo* and identifying unknown DNA sequences, a litany of genome-wide assays use high-throughput sequencing instruments to count DNA fragments in a library where the reference genome is known. These counts are then mapped back to the reference, so that each position across the genome can be assigned a value for its enrichment in the library. This principle underlies all genome-wide analyses presented in this dissertation.

PRO-seq

Precision Run-On Sequencing, or PRO-seq is the latest in a series of methods used to measure the locations of actively transcribing polymerases, genome-wide. Nuclear run-ons have long been used to determine whether RNA polymerases are transcriptionally engaged (Weber, Jelinek and Darnell, 1977). This involves isolating nuclei from cells, washing away native nucleotides, and then providing labeled nucleotides under conditions that allow for polymerase to resume transcription briefly so that the nascent transcripts incorporate the label. This principle was used to determine that the Pol II promoter proximally paused at the 5' end of the *Hsp70* gene in flies is transcriptionally engaged (Rougvie and Lis, 1988). In this case, radiolabeled nucleotides were incorporated in the run-on, and the resulting run-on RNA was used as the probe in a Southern blot to show that the *Hsp70* gene contained a large amount of engaged promoter proximally paused polymerase (at a much higher density than the gene body).

Subsequent modifications have adapted these run-ons to high throughput sequencing, so that transcriptionally engaged Pol II can be mapped genome-wide. Global Run-On Sequencing (GRO-seq) was the first method to do this. It uses incorporation of bromouridine during a ~50-100 bp run-on to label nascent transcripts, which are then affinity purified with an anti-BrU antibody between consecutive enzymatic steps during the preparation of a stranded RNA-seq library. This first genome-wide look at the location of engaged Pol II showed that human promoters are divergently transcribed, and, comprehensively for the first time, that promoter proximal pausing is nearly ubiquitous (Core, Waterfall and Lis, 2008). PRO-seq is a higher resolution version of GRO-seq (Kwak *et al.*, 2013; Dig Bijay Mahat *et al.*, 2016). PRO-seq uses biotin-11-NTPs to label nascent transcripts during the run-on. However, unlike radioactive and bromouridine labels, Pol II is only able to incorporate a single biotin NTP, on average, before the bulk of biotin prevents subsequent nucleotide addition (Kwak *et al.*, 2013). Therefore, in the strand specific RNA-seq library of PRO-seq, the first nucleotide in the read corresponds to the single nucleotide incorporated during the run-on, and thus the exact location of the active site of Pol II in the nuclei before the run-on. PRO-cap and GRO-cap are modifications of GRO- and PRO-seq that map only initiation sites of transcription (Kruesi *et al.*, 2013; Kwak *et al.*, 2013). Here, the 5' end of nascent transcripts is mapped after selecting enzymatically for capped nascent RNA specifically. Critically, the RNA is not fragmented in PRO-cap, so that the biotin incorporated in the run-on remains linked to the 5' cap. Thus, only nascent RNAs short enough to be incorporated as the insert in a Illumina sequencing library will be detected. This means that PRO-cap detects initiation sites with both extreme precision and high sensitivity. It maps initiation sites with basepair resolution, and is it is extremely sensitive because it has no

background from accumulated RNA as CAGE would, and does not waste reads on the bodies of genes as PRO-seq would.

Perturb

Both Chapters 2 and 3 deal with my efforts toward the development of RNA aptamers as a new type of inhibitor for highly specific perturbations of biological systems. Aptamers are nucleic acids that have high affinity for a target molecule due to their three dimensional structure(Ozer, Pagano and Lis, 2014). They are selected from large libraries of random nucleotides by an iterative process of binding to the target, elution, and amplification know as Sequential Evolution of Ligands by EXponential enrichment, or SELEX(Ellington and Szostak, 1990; Tuerk and Gold, 1990). Aptamers act much like an antibody in their ability to recognize target molecules, but offer several advantages. The process of SELEX offers greater control than raising antibodies: the selection can be designed to target a particular domain of a factor and nothing else. Aptamers can be selected in several weeks, and do not require the sacrifice of an animal. Once identified, they can be cheaply synthesized. RNA aptamers can be expressed in cells, and bind their target *in vivo*. They often recognize larger epitopes than antibodies, and can thus recognize targets with high specificity. They are often easier to select to nucleic acid domains, thus making them an attractive inhibitor of interactions with nucleic acids.

Our group has developed a suite of improvements to the SELEX process. We have developed an efficient column based binding and elution scheme(Latulippe *et al.*, 2013), sped up enrichment by binding and eluting multiple times between amplification steps(Szeto *et al.*, 2013), multiplexed selections so that dozens can be carried out in tandem(Szeto *et al.*, 2014), characterized non-specific cooperative binding to targets(Ozer *et al.*, 2013), and integrated these

changes in a fast, streamlined, multiplex selection scheme(Reinholt *et al.*, 2016). The characterization of the resulting aptamer libraries from these selection schemes has remained a bottleneck in this process. To that end, I developed a method to quantitatively measure the binding affinity of thousands of aptamers to their target in single assay called HiTS-RAP(Tome *et al.*, 2014), which is the focus of Chapter 2. Thus, HiTS-RAP falls under the observe phase of the strategy for studying transcription, but with the purpose of developing reagents to perturb.

Chapter 3 focuses on my efforts to develop RNA aptamers as a potent and specific inhibitor of biological interactions *in vivo*. Our group has long been interested in this strategy, as it has the potential of rapidly and specifically perturbing one surface of a factor, without affecting its stability or interactions with other factors, as compared with a knockdown which takes several days and results in loss of the entire protein. We have used aptamers as an inhibitor of the splicing factor B52 in whole flies(Shi, Hoffman and Lis, 1999), of HSF in cell culture in both fly(Salamanca *et al.*, 2011) and human cells(Salamanca *et al.*, 2014), and of TBP in *in vitro* transcription reactions(Sevilimedu, Shi and Lis, 2008). However, none of these studies used the power of genomics to interrogate the effects of these targeted perturbations.

Chapter 2: Comprehensive Analysis of RNA-Protein Interactions by High Throughput Sequencing-RNA Affinity Profiling¹

RNA-protein interactions have critical roles in gene regulation. However, high-throughput methods to quantitatively analyze these interactions are lacking. We adapted an Illumina GAIIx sequencer to make several millions of these measurements with a High-Throughput Sequencing – RNA Affinity Profilng (HiTS-RAP) assay. Millions of cDNAs are sequenced, bound by the *E. coli* replication terminator protein Tus, and transcribed *in situ*, whereupon Tus halts transcription leaving RNA stably attached to its template DNA. The binding of fluorescently-labeled protein is then quantified in the sequencer. We used HiTS-RAP to measure the affinity of mutagenized libraries of GFP-binding and NELF-E binding aptamers to their respective targets and thereby identified regions in both aptamers that are critical for their RNA-protein interaction. We show that mutations additively affect the binding affinity of a NELF-E binding aptamer, whose interaction depends mainly on a single-stranded RNA motif, but not that of the GFP aptamer, whose interaction depends primarily on secondary structure.

Introduction

RNA-protein interactions are ubiquitous in biology and critical at many regulatory steps of gene expression and various stages of development(Lee, 2012). Defects are involved in a variety of disease conditions(Lukong *et al.*, 2008; Esteller, 2011). A quantitative analysis of interactions between RNA and RNA Binding Proteins (RBPs) is necessary for a more profound understanding of basic biological mechanisms(König *et al.*, 2011). Several methods exist for

¹ Most of this chapter was published in a paper was co-written with Abdullah Ozer and John Lis: Tome, J. M., Ozer, A., Pagano, J. M., Gheba, D., Schroth, G. P., & Lis, J. T. (2014). Comprehensive analysis of RNA-protein interactions by high-throughput sequencing–RNA affinity profiling. *Nature Methods*. <https://doi.org/10.1038/nmeth.2970>

determining RNA-protein affinities, though the quantitative methods are low-throughput and laborious(Wong and Lohman, 1993; Katsamba, Park and Laird-Offringa, 2002; Ryder, Recht and Williamson, 2008; Salim and Feig, 2009; Pagano, Clingman and Ryder, 2011). High-throughput techniques such as SELEX(Campbell *et al.*, 2012) or RNA-MITOMI(Martin *et al.*, 2012) can be used to determine binding motifs, but are either not quantitative or are difficult to scale genome-wide. Existing genome-wide assays, such as RIP-Seq(Tenenbaum *et al.*, 2000), CLIP(Licatalosi *et al.*, 2008), and PAR-CLIP(Hafner *et al.*, 2010) provide a great deal of information by identifying RNA molecules that interact with a protein. Nonetheless, these methods have significant limitations: they depend on antibodies, cannot quantitatively measure the affinities of the interactions, are difficult to multiplex, and they generally depend on crosslinking, which may not be uniform for every target and can be affected by the presence of other proteins. Here, we developed a high-throughput, quantitative RNA-protein interaction assay that does not require an antibody, can be scaled genome-wide and, in principle, applied to interactions of RNA with molecules other than proteins.

The Illumina platform allows the simultaneous sequencing of hundreds of millions of DNA clusters, which are each derived from the amplification of a single molecule with primers that are covalently linked to the glass flowcell(Bentley *et al.*, 2008). These sequencers can also be used as versatile and programmable instruments to automatically image and analyze the binding of fluorescently labeled proteins to these DNA clusters as was done in the creative HiTS-FLIP protocol(Nutiu *et al.*, 2011). A corresponding assay for RNA binding to protein is conceivable, but has not been realized due to the difficulty of converting the DNA of clusters into RNA that is retained at the cluster(Evanko, 2011).

Here, we devise a method called HiTS-RAP to transcribe DNA at each cluster of an Illumina flowcell and use it to measure the affinity of mutants of two well characterized RNA aptamers at an unprecedented scale. We measured the binding constants of 1,875 mutants of the GFP aptamer (GFPapt)(Shui *et al.*, 2012), and 9,832 mutants of NELFapt(Pagano *et al.*, 2014), an RNA aptamer that binds *Drosophila* NELF-E (Negative Elongation Factor E), an RNA binding subunit of a protein complex involved in maintaining promoter proximally paused RNA Polymerase II(Yamaguchi *et al.*, 1999, 2002, Wu *et al.*, 2003, 2005). We then analyzed these data to determine the features of these two aptamers that are most important for their protein interactions, and to gain insight into their complex structures.

Results

Tus Stably Halts Transcription

A critical feature of the HiTS-RAP strategy is to transcribe DNA on the Illumina instrument and have the RNA transcript stably retained at each DNA cluster. We have used the *E. coli* replication terminator protein Tus as a roadblock to transcription to achieve this. Tus prevents DNA replication through sites of replication termination in an orientation-specific manner by binding to a 32 bp sequence element (ter) with high affinity, specificity, and stability(Kamada *et al.*, 1996; Mohanty, Sahoo and Bastia, 1996; Mulugu *et al.*, 2001; Mulcair *et al.*, 2006). In addition, Tus has been shown to stop many RNA polymerases including T7 RNA polymerase (T7 RNAP) from transcribing through Tus-bound ter sites in the non-permissive orientation, resulting in either terminated or halted transcription(Mohanty, Sahoo and Bastia, 1996; Guajardo and Sousa, 1999). T7 RNAP produces only truncated transcripts on our DNA

templates used for HiTS-RAP, many of which remain anchored to the DNA, when Tus is bound to DNA in the non-permissive orientation (Figure 2. 2a). In contrast, T7 RNAP produces mostly free full-length run-off transcripts when Tus is bound in the permissive orientation.

Electrophoretic Mobility Shift Assay (EMSA) with radiolabeled DNA shows that after transcription halting, nearly every DNA is engaged in a complex of intermediate mobility containing an RNA transcript (Figure 2. 1a). RNase treatment results in a lower mobility complex, consistent with the loss of charged RNA and the added mass of T7 RNAP to the DNA-Tus complex (Figure 2. 2b). Thus, T7 RNAP transcribing into a non-permissive Tus-ter complex halts upstream of the ter site with an RNA transcript still bound to the DNA through the polymerase.

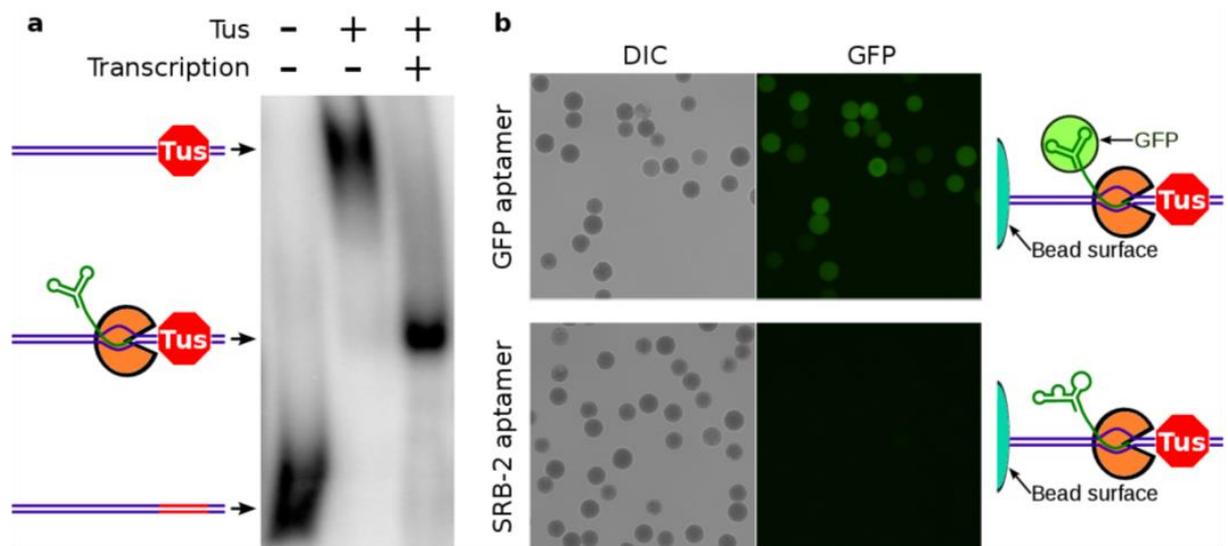


Figure 2. 1: T7 RNA polymerase halting with Tus gives stable complexes containing DNA and functional RNA

a. EMSA was carried out with radiolabeled DNA that encodes the GFP aptamer and has a binding site for Tus. Naked DNA runs with high mobility. Binding Tus to the ter DNA element (red segment) retards DNA mobility. After transcription, nearly every DNA participates in a complex of intermediate mobility, containing Tus, T7 RNA polymerase (orange), and RNA (green). This band is very sharp, indicating that each DNA-Tus-T7RNAP-RNA complex is mainly of homogeneous composition. **b.** 454 Life Sciences polystyrene beads covered in covalently linked DNA templates for transcription. After transcription halting, beads were incubated with EGFP and then washed. EGFP bound to beads presenting halted GFP aptamer RNA, but not to beads presenting SRB-2 aptamer RNA.

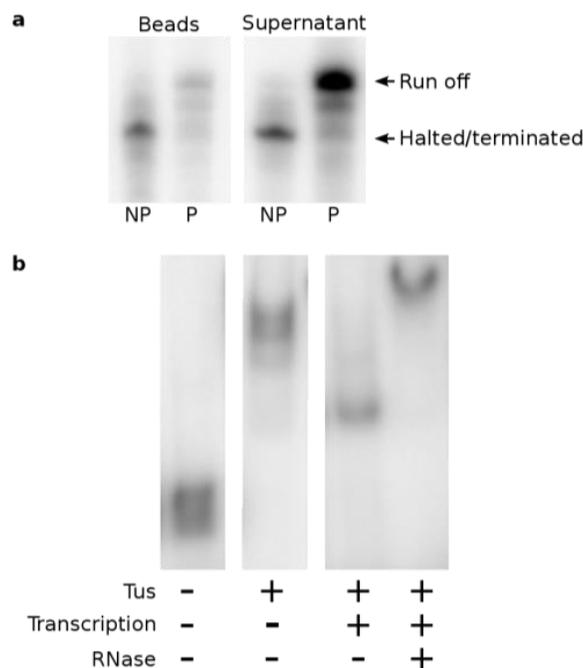


Figure 2. 2: Transcription halts or terminates at Tus-bound ter sites in the non-permissive orientation

a. Radiolabeled *in vitro* transcription of Tus-bound DNA templates. Biotinylated GFP aptamer DNA templates contained two ter sites, either in the non-permissive (NP) or permissive (P) orientation. The DNA templates were bound to streptavidin-agarose beads (Pierce) and then bound by Tus protein. After transcription for 30 minutes at 37°C in the presence of radiolabeled CTP, beads and supernatant were separated and run on an 8% denaturing PAGE (7 M urea). Most transcription does not proceed past ter site in the non-permissive orientation. 16% of the transcripts produced in the non-permissive ter reaction remained bound to beads in halted transcription complexes, while the rest terminated and released to the supernatant. RNA transcripts of a single, distinct length, consistent with halting or termination at the ter site but not before, produced in these reactions can be explained by forced termination of leading T7 RNAP(s) stopped at the ter site by the trailing ones, thus leaving a single halted T7 RNAP on the template at the ter site (Epshtein *et al.*, 2003). RNAs in the supernatant were produced by polymerases that terminated at the Tus-bound ter site or passed through and ran off the end of the template. In the permissive orientation, the vast majority of the transcripts produced are run off transcripts, and thus are found in the supernatant. Only 6% of the transcripts remained bound to the beads, and most of these are run off transcripts, indicating that they are likely due to non-specific binding to beads by full-length RNAs which are no longer engaged in halted transcription complexes. **b.** EMSA of halted transcription complexes with radiolabeled DNA. Complexes produced after transcription halting contain RNA. This experiment is identical to that of Figure 2. 1A, except that no SUPERase In (Ambion) was used and the addition of a lane that was treated with RNase cocktail (Ambion) after transcription halting. The RNase treatment caused the band of intermediate mobility created by transcription halting to slow to a mobility even less than that of Tus bound DNA, presumably a consequence of the loss of full length charged RNA and presence of bound T7 RNAP (the lanes are all from a single gel, thus mobilities can be directly compared).

In this work, we used an RNA aptamer that has high affinity to and specificity for GFP and its derivatives (i.e. EGFP)(Shui *et al.*, 2012) to develop the assay. As an initial test of our DNA anchored RNA-protein binding assay, we attached DNA to beads and generated halted transcription complexes. EGFP binds to halted transcription complexes presenting GFPapt but not to a negative control SRB-2 aptamer(Holeman *et al.*, 1998) (Figure 2. 1b). To show that the interaction with GFPapt complexes is due to the presence of the full length RNA, a similar experiment was carried out with single round transcription. EGFP bound to beads with GFPapt template after transcription had been chased to the Tus-ter complex to produce a full length aptamer, but not after initiation only (Figure 2. 3). The initiated complexes contain every component of the chased complexes with the exception of the full length GFPapt RNA. Together, this demonstrates that halted GFPapt transcription complexes bind EGFP through the specific interaction between the full-length RNA and EGFP.

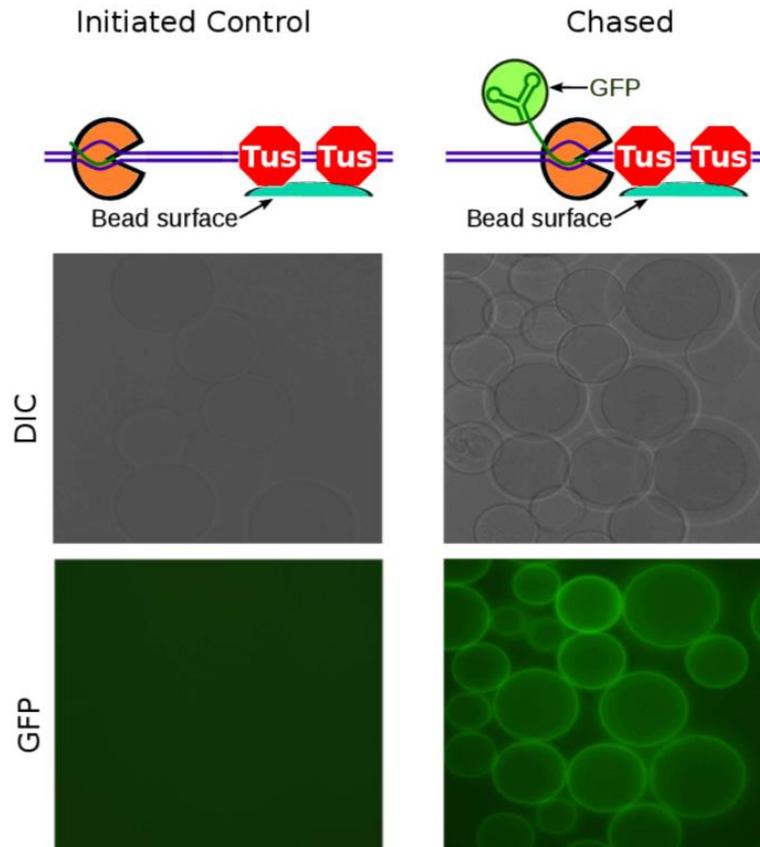


Figure 2. 3: EGFP only binds to transcription complexes with full-length GFP aptamer RNA

GFP aptamer halting templates with two ter sites were immobilized on glutathione beads through the GST tagged Tus protein. Single round transcription was initiated by incubating beads in a transcription solution lacking CTP. One sample was then chased with all four nucleotides including CTP, allowing T7 RNAP to transcribe up to the Tus bound ter site, while the other sample was not, leaving an initiated polymerase with an 11 nt transcript. The transcription complexes present in both samples contain identical components of DNA template, GST-tagged Tus protein, T7 RNAP, and the glutathione beads, but differ in that the ‘initiated control’ has only an 11 nt transcript while the ‘chased’ has the full aptamer RNA. EGFP only binds to ‘chased’ sample that presents the full-length GFP aptamer RNA. Thus, EGFP labeling of halted transcription complexes is due to an interaction with the full-length RNA transcript.

Transcription Halting with Tus is Functional on the GAIIX Sequencing Instrument

To couple Tus-dependent halting of T7 RNA polymerase transcription with sequencing on an Illumina GAIIX, we used DNA templates that contain a sequence encoding the RNA to be transcribed flanked by a T7 promoter immediately upstream and by the ter sequence downstream (Figure 2. 4). Figure 2. 5a shows a schematic of the HiTS-RAP assay. All steps of HiTS-RAP are carried out automatically. After sequencing, a new second DNA strand is generated at all clusters in the flowcell. Transcription halting is then carried out, presenting the RNA encoded by the millions of DNAs that have been sequenced. The binding properties of these RNAs of known sequence can then be probed by allowing an mOrange labeled protein to interact with the halted RNAs(Shaner *et al.*, 2004; Bentley *et al.*, 2008). Illumina’s software is used to image the flowcell at equilibrium with different concentrations of protein and measure fluorescence intensity of bound mOrange fusion protein at each cluster. The TIRF microscopy of the sequencer enables this equilibrium measurement, as excess protein in solution does not interfere with imaging of protein bound to clusters(Bentley *et al.*, 2008; Nutiu *et al.*, 2011).



Figure 2. 4: Schematic of HiTS-RAP templates

Different sequence elements needed in DNA templates for HiTS-RAP are labeled. Illumina adaptors located at the extreme 5’ and 3’ ends are used to generate clusters on the sequencer. The sequence of interest is flanked by a T7 RNA polymerase promoter at the 5’ end, and the Illumina sequencing primer and ter site at the 3’ end. Thus, during transcription, T7 RNAP initiates at its promoter transcribes through the sequence of interest (the GFP aptamer in this case), and then begins to transcribe the Illumina sequencing primer before it encounters the non-permissive Tus bound ter site and halts transcription. Thus, the entire GFP aptamer RNA should have emerged from the polymerase. The approximate positions on the template of Tus protein and T7 RNAP in a halted transcription complex are shown.

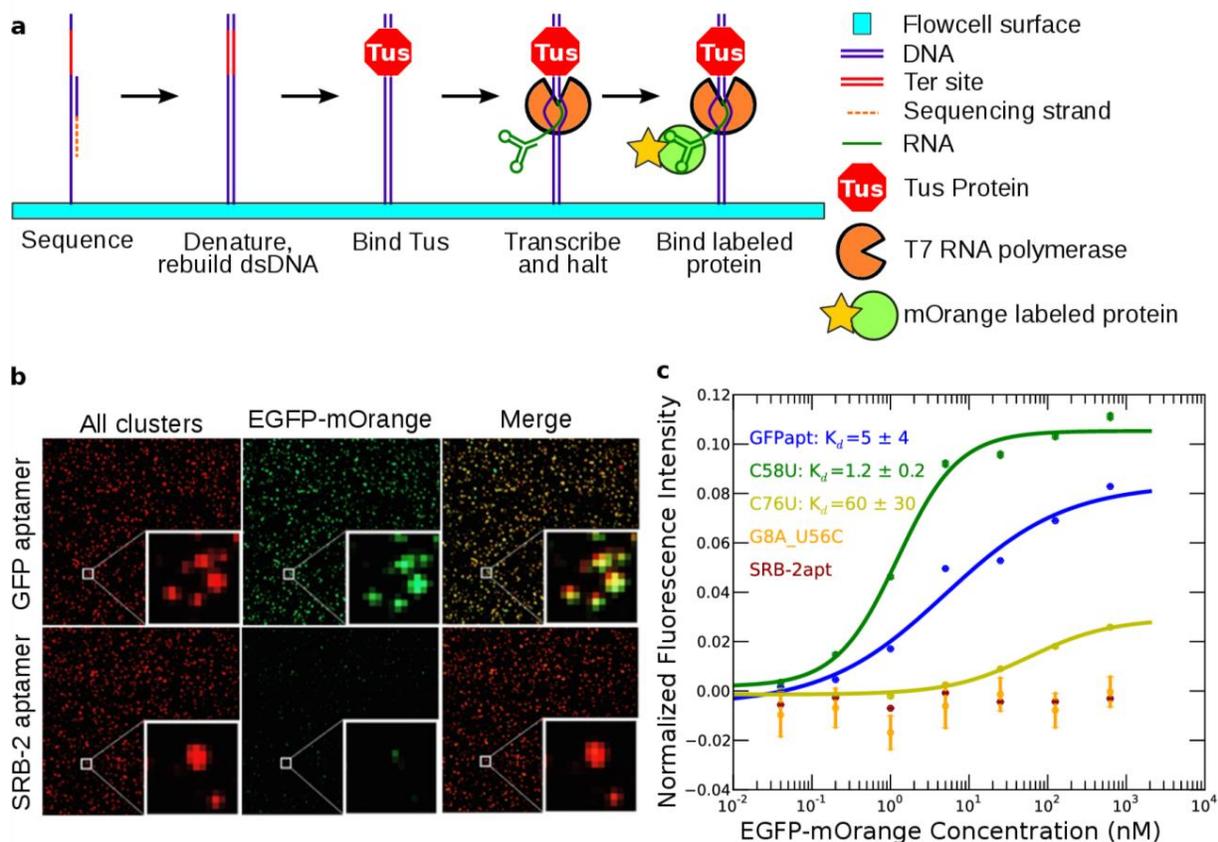


Figure 2. 5: RNA-protein interactions can be assayed by HiTS-RAP on an Illumina GAIIX instrument

a. HiTS-RAP schematic. Sequencing is done following the standard Illumina workflow. The strand synthesized during sequencing is then stripped away, a primer is annealed and the second strand is regenerated with Klenow enzyme. Tus is then bound to the ter site, and DNAs on the flowcell are transcribed. T7 RNAP initiates at its promoter, transcribes through the sequence of interest, and halts just upstream of the Tus bound ter site. The RNA transcript is stably linked to its DNA template through the polymerase. Fluorescently labeled protein is then bound to the RNA and imaged. **b.** Sample images from a HiTS-RAP run with GFP and SRB-2 aptamers. ‘All clusters’ (left panels) are labeled during sequencing and shown as a maximum intensity projection of the four channels. After transcription halting and EGFP-mOrange binding, the flowcell is imaged at equilibrium with 625 nM EGFP-mOrange. Clusters are labeled by mOrange in a lane containing all GFP aptamer halting templates, but not in a lane containing only SRB-2 aptamer templates. **c.** Binding curves for the GFP aptamer ($n = 2,665,064$), and mutants C58U ($n = 3,833$), C76U ($n = 4,758$), and G8U_U56C ($n = 29$), and the SRB-2 aptamer ($n = 1,588,404$). G8U_U56C and SRB-2 aptamer are scored as not binding. Data are representative, from one lane of the sequencer (SRB-2 aptamer is from a separate lane). Intensities are the average of all clusters of each sequence in the lane, normalized by dividing by the average sequencing intensity and subtracting the average intensity at no EGFP-mOrange. Error bars represent standard error. Error of fitted K_d s are the square root of variances returned by the fitting algorithm.

The GFP aptamer and a population of point mutants and control RNA were assayed by HiTS-RAP for their affinity to an EGFP-mOrange fusion protein. EGFP-mOrange was used because EGFP is not compatible with the optics of the sequencer (Shaner *et al.*, 2004; Bentley *et al.*, 2008). Figure 2. 5b shows that the vast majority of GFPapt DNA clusters produce halted RNA capable of binding to EGFP-mOrange, while those in a lane where all clusters encode the SRB-2 aptamer negative control RNA do not. To measure the stability of the Tus halted transcription complexes, three lanes containing only GFPapt clusters were imaged repeatedly in the presence of 625 nM EGFP-mOrange. This concentration is well above the K_d (Shui *et al.*, 2012), so all RNAs should be bound by EGFP-mOrange. Using the characteristic lifetime of the halted RNA complex measured on the sequencer we estimate that the assay is sensitive to measurements above background for the first 48 cycles, or 72 hours, given an approximate cycle time of 1.5 hours (Methods, Figure 2. 6a). Thus, the halted transcription complexes are sufficiently stable to carry out the several sequential measurements at various target protein concentrations, which are necessary to determine dissociation constants (K_d s). This decay rate was used to correct all intensities used for determining K_d s with HiTS-RAP (Online Methods, Figure 2. 6a).

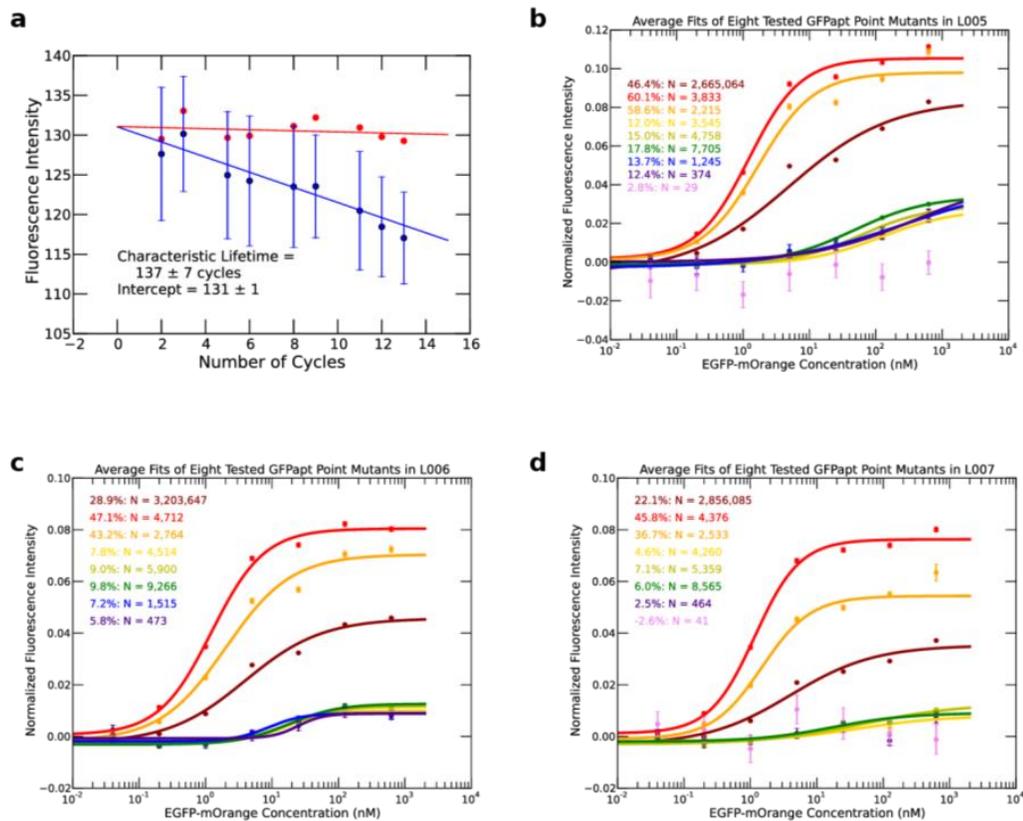


Figure 2. 6: Corrections applied to HiTS-RAP data

a. Stability of halted transcription complexes modeled by linear approximation of exponential decay. GFP aptamer RNA presenting clusters were imaged at equilibrium with 625 nM EGFP-mOrange nine successive times following a single transcription halting reaction. The average raw, non-normalized fluorescence intensity of bound EGFP-mOrange at canonical GFP aptamer clusters in three lanes (~2.5 million clusters per lane) are plotted in blue, with error bars representing the standard deviation between the three lanes. The reported decay rate is the average \pm standard deviation of the fits from the three lanes. The fitted line is plotted as a straight blue line. Decay rate corrected intensities (see Online Methods) were plotted in red together with a line fit to these intensities (red straight line). **b-d.** HiTS-RAP binding curves for GFPapt and the eight mutants verified by EMSA (see Figure 2. 7) in each of the three lanes (L005-L007). Not every mutant is plotted in every lane: it is excluded if `scipy.optimize.curve_fit` failed to converge on a fit. The number of clusters and percent increase in signal is reported for each lane. Percent increase is the percent increase in fluorescence signal between the average of the first and last two measured intensities in the average binding curve. Intensities decrease across the flowcell due to photobleaching. Variants with lower affinity tend to saturate at lower fluorescence intensities. These data were used to determine the threshold for signal increase considered as not binding, for assigning sequences in each lane affinities of >125 nM. G8A_U56C (violet), a variant which was confirmed as having an affinity of >125 nM by EMSA has a 2.8% increase in signal in lane 5 and a 2.6% decrease in lane 7 (and was not present in lane 6). G68C (purple), a mutant with a confirmed affinity of 75 nM, increased by 12.4, 5.8, and 2.5 % in lanes 5, 6, and 7. Thus, we determined that sequences with an increase of less than 3% in signal should be scored as not binding within the detection limit of our assay.

Measuring Equilibrium Binding Constants for the GFP Aptamer and Mutants with HiTS-RAP

HiTS-RAP was carried out using a flowcell with three lanes containing a GFPapt template that was subjected to many rounds of PCR and thereby accumulated a population of mutants in addition to the canonical GFP aptamer. Seven protein concentrations increasing in five-fold increments from 0.04 to 625 nM were used to measure the K_{ds} of interactions between RNAs and EGFP-mOrange. We identified 1,875 mutant GFP aptamer sequences that had at least 10 copies in at least one of the three lanes with quality scores greater than 25 at the mutated residue. We measured a K_d of $4.27 \times / 1.11$ nM (geometric mean $\times /$ (times/divide) geometric standard deviation(Limpert and Stahel, 2011)) for the EGFP-GFP aptamer interaction, which agrees well with its published affinity of between 5 and 15 nM(Shui *et al.*, 2012). The SRB-2 aptamer negative control RNA shows no appreciable binding in HiTS-RAP (Figure 2. 5c).

Most GFPapt mutants do not differ significantly in sequence from the canonical aptamer, and therefore bind EGFP with similar affinity. However, some showed altered affinity. To verify HiTS-RAP measured K_{ds} , we picked eight mutants, three with higher affinity, and five with lower affinity including some with no appreciable binding, and measured their binding affinity by EMSA(Ryder, Recht and Williamson, 2008) (Figure 2. 7a). Affinities of these eight mutants measured by HiTS-RAP and EMSA show a good correlation (Figure 2. 7b). Many of the GFPapt mutants that we identified were represented by only a few dozen clusters, a very small fraction of the total number on the flowcell, demonstrating that HiTS-RAP can reliably measure K_{ds} of even low copy number sequences in a library. These results show that such a library of mutagenized RNA can be used to determine the affinity of variations of a particular sequence to a target protein at a scale that would not be possible by conventional methods.

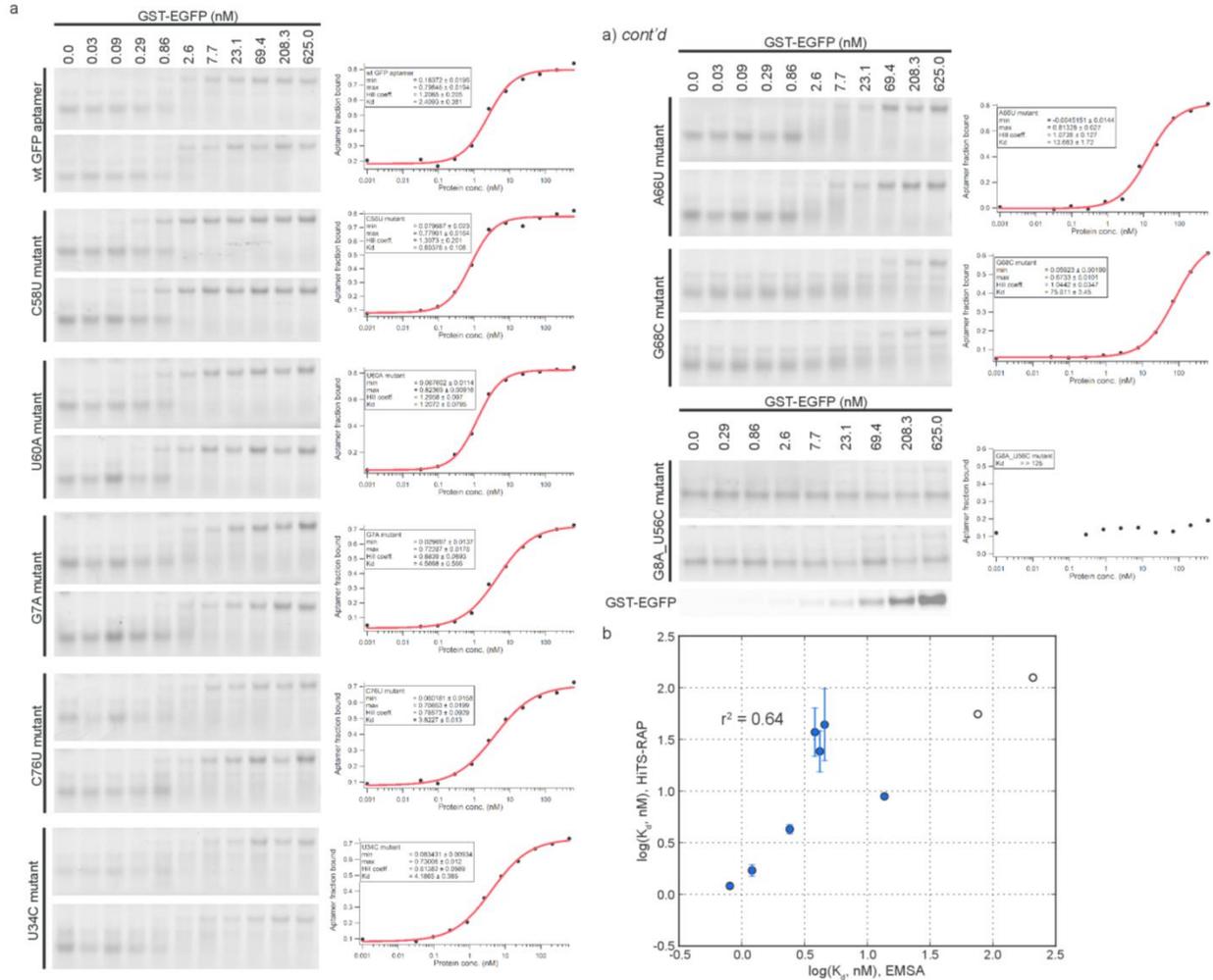


Figure 2. 7: Confirmation of HiTS-RAP measured affinities by EMSA

a. EMSA of fluorescently-labeled canonical GFP aptamer and eight mutants with varying affinities. Duplicate EMSA (left panels) were done with varying GST-EGFP concentration in three-fold increments from 0.03 to 625 nM. The bottom band corresponds to unbound RNA, and the top band to EGFP bound RNA. A representative scan of the GST-EGFP used in EMSA, visualized by EGFP fluorescence, is shown for the last EMSA gel (bottom left panel). Fraction of aptamer bound by EGFP, quantified from gels, are plotted (black dots in right panels) and fitted to the Hill equation (red solid line) to determine the K_d s. Minimum and maximum fraction of aptamer bound, Hill coefficient and the K_d of the fits are reported in the figure legend for wt (canonical), C58U, U60A, G7A, C76U, U34C, A66U, G68C, G8A_U56C mutant GFP aptamers.

b. Agreement between HiTS-RAP and EMSA. Affinities measured by EMSA in (a) are plotted against their HiTS-RAP counterpart as $\log(K_d, \text{nM})$. Mutants scored as not binding in at least one lane (or in all lanes as is the case with G8A_U56C) are plotted as open circles. The r^2 of 0.64 is for all eight mutants and the GFP aptamer.

Comprehensive mutagenesis of the GFP Aptamer

We have used our library of mutants to carry out an in-depth analysis of the 82 nt of GFPapt that we have sequenced. Of the 246 possible single base substitution mutants, 236 are present in our data set. Many of these mutations resulted in affinities that were too low to be measured by the EGFP-mOrange concentrations used in this experiment (Figure 2. 6b-d, Methods): therefore they were assigned a K_d of >125 nM (the second highest protein concentration used). If this occurred in only a subset of lanes, it was assigned a K_d of 125 nM for the lanes where it did not bind, and averaged with the rest. Figure 2. 8a shows all of the measured affinities of single point mutants. Most have a negative effect on the binding affinity of the GFP aptamer ($\log_2(K_{d_mut}/K_{d_wt}) > 0$), consistent with its highly optimized nature. However, two notable exceptions to this are C58U ($K_d = 1.21 \times / 1.03$ nM) and U60A ($K_d = 1.71 \times / 1.13$ nM). Both of these mutations do not alter the predicted secondary structure of GFPapt, but they reduce the free energy of folding (Figure 2. 9a), so they likely make the overall structure more flexible. The K_d s of both of these mutants were verified by EMSA (Figure 2. 7).

To determine the contribution of each nucleotide in GFPapt to its interaction with EGFP, we calculated the average effect of all measured mutations at each position (Figure 2. 8b). In agreement with the structural and binding affinity characterization reported in the original study (Shui *et al.*, 2012), the majority of the high impact positions (average $\log_2(K_{d_mut}/K_{d_wt}) > 3$) are located in either stem-loop #2 or #3. Secondary structure predictions of two representative single point mutants which have K_d s >125 nM, G68C and G46C are shown in Figure 2. 9b. Consistent with their effects on binding affinity, both of these mutations cause gross alterations in the predicted structure of the GFP aptamer.

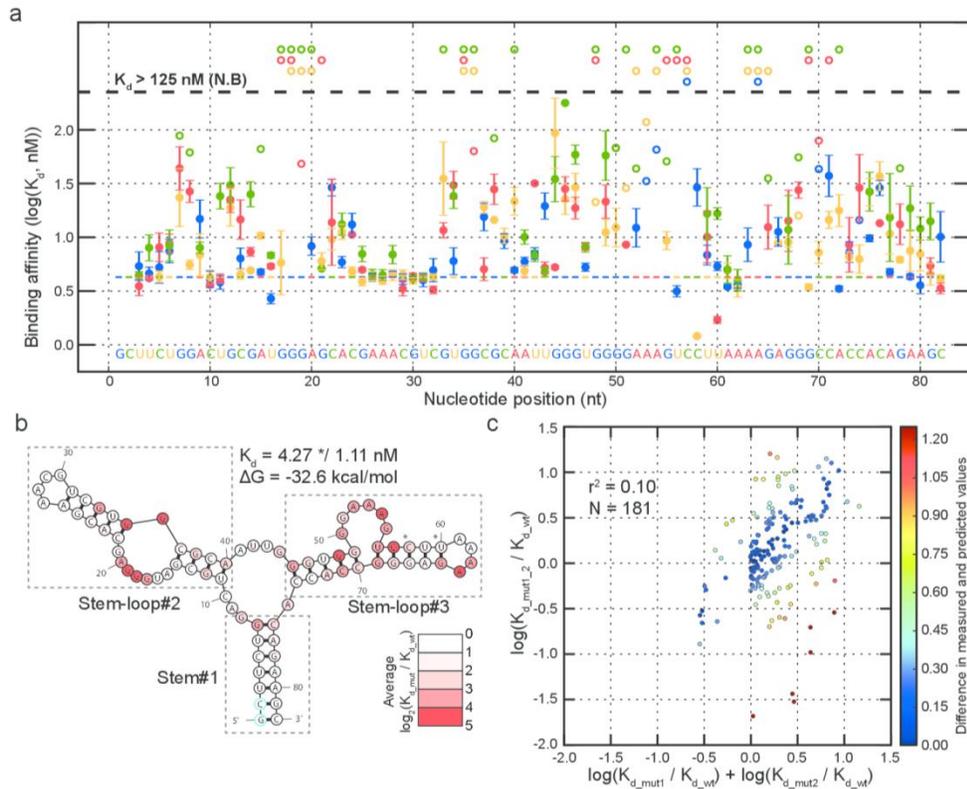


Figure 2. 8: Analysis of GFPapt by HiTS-RAP.

a) GFP binding affinities of all 236 measured single-point mutants of GFPapt. Binding affinities in nM are plotted in logarithmic scale. Mutations at each position are color-coded. Wild-type GFPapt binding affinity is indicated by the colored dashed line and the sequence is shown at the bottom of the graph. Error bars represent the standard deviations in $\log(K_d)$. 175 mutants qualify as binding in all three lanes. Single point mutants that qualify as not binding and are thus assigned an affinity of 125 nM in at least one lane are plotted with open circles, with no error bars. Those that do not bind in all three lanes are plotted at the top of the plot. **b)** Predicted secondary structure of GFPapt. GFPapt is predicted to fold into a three stem-loop structure connected by a central 3-way junction. Each position is colored by the average absolute effect $|\log_2(K_{d_mut}/K_{d_wt})|$ of all of its measured mutants. Mutations that qualify as not binding were assigned an affinity of 125 nM for this plot. Positions where the average effects are greater than 4 (>16-fold effect in affinity) are colored red. Most mutations have a negative effect or less than a 2-fold positive effect on binding affinity, except C58U and U60A (indicated by asterisk). **c)** Correlation between measured and predicted effects of GFPapt double mutants. Measured $\log_{10}(K_{d_mut1_2}/K_{d_wt})$ is plotted against the value predicted based on single-point mutants ($\log_{10}(K_{d_mut1}/K_{d_wt}) + \log_{10}(K_{d_mut2}/K_{d_wt})$). Points are colored based on the difference between the measured and the predicted effects. There is a positive but small correlation ($r^2 = 0.10$) between the measured and predicted effects, and they differ as much as 1.98, or ~100-fold.

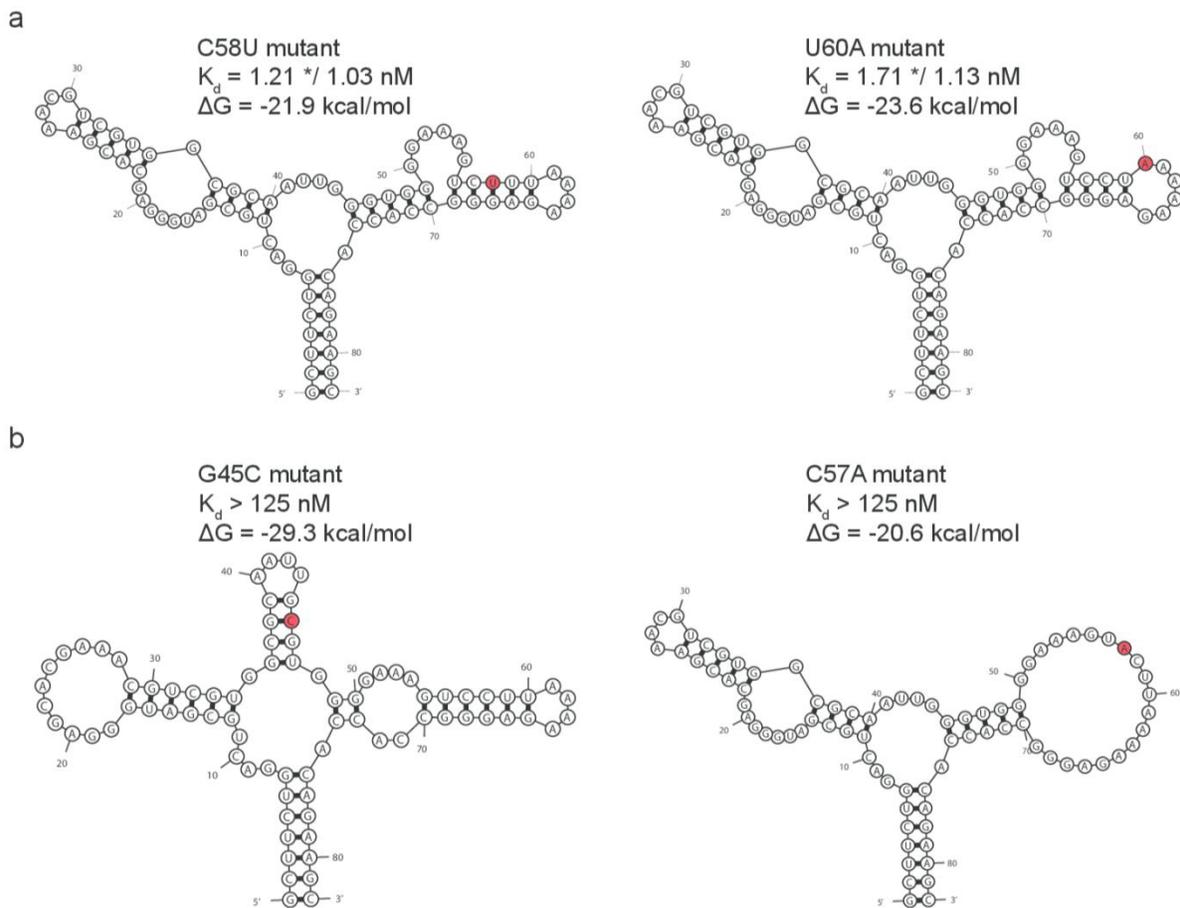


Figure 2. 9: Secondary structure predictions of GFP aptamer single point mutants with higher or lower affinities

a) Predicted secondary structures of the two highest affinity GFP aptamer single point mutants. C58U ($1.21 \times / 1.03 \text{ nM } K_d$) and U60A ($1.71 \times / 1.13 \text{ nM } K_d$) mutants are predicted to have structures that are very similar to that of wild-type GFP aptamer. Compared to the wild-type aptamer (Figure 2. 8b), U60A mutant has a slight expansion of the terminal loop (6 nt vs. 4 nt) and a concurrent shortening of the preceding stem (4 bp vs. 5 bp) in stem-loop#3. **b)** Predicted secondary structures of the two lower affinity GFP aptamer single point mutants. Both G45C ($>125 \text{ nM } K_d$) and C57A ($>125 \text{ nM } K_d$) single point mutants are predicted to have significantly altered secondary structures when compared to the wild-type aptamer structure (Figure 2. 8b). Mutated nucleotides are colored in red. Folded structures and the folding free energies are predicted by mFold.

To gain further insight into the relationship between the binding affinity and sequence of the GFP aptamer, the relationship between the effects on affinity of double mutants and their two corresponding single point mutants was analyzed. This analysis was restricted to base

substitutions, and to the 76 nt region in the center of the aptamer so that the ends could be matched to GFPapt to ensure against insertions and deletions. We identified 181 double mutants (of the 60,516 possible) where the double mutant and corresponding two single mutants bound EGFP above the threshold for signal increase and had high confidence EGFP-binding affinities. If two single point mutations affect binding independently of each other, the binding of a double mutant would be predicted by the additive effect of the two individual mutations. However, in cases where two mutations affect the interaction of the aptamer with its protein target in either a cooperative or antagonistic manner, the measured effect of the double mutant would differ from the additive effect of the two individual mutations. We used a metric analogous to $\Delta\Delta G$ (the difference in Gibbs free energy change) to make this comparison: the effect of a mutation is represented by the log ratio of the mutant and canonical aptamers' K_{dS} ($\log(K_{d_mut} / K_{d_wt})$). There is only a weak correlation ($r^2 = 0.10$) between the measured and the predicted effect of the GFPapt double mutants (Figure 2. 8c).

Double mutants with significantly higher or lower affinity than predicted by the single point mutants are common, and can provide insights into the structural components important for the interaction with GFP. For example, A23G_U34C reconstitutes an AU base pair as a GC base pair in the predicted secondary structure and binds with higher affinity than either mutant alone, while the same U34C in combination with the C58U mutation fails to rectify the structural perturbation caused by U34C and has lower affinity than either single mutant (Figure 2. 10). Overall, mutational analysis and structural predictions indicate that the interaction between the GFP aptamer and its target protein EGFP is complex in nature, depending upon an intricate structure dictated by its sequence.

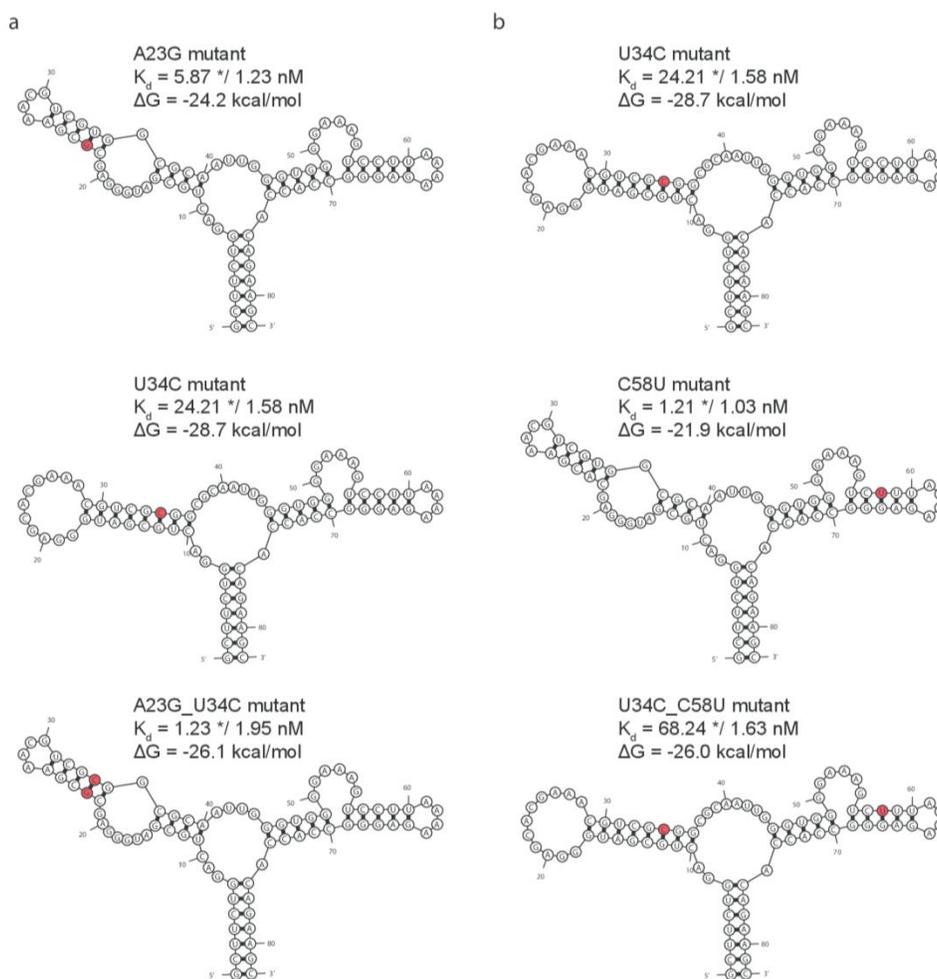


Figure 2. 10: Secondary structure predictions of GFP aptamer double point mutants with significantly higher or lower affinities than predicted by the single point mutants

a) Predicted secondary structures of the two single point mutants and the corresponding double mutant GFP aptamer with better affinity than predicted from single point mutants. A23G single point mutant ($5.87 \times / 1.23 \text{ nM Kd}$) and A23G_U34C double point mutant ($1.23 \times / 1.95 \text{ nM Kd}$) are predicted to have wild-type GFP aptamer-like structures, whereas U34C single point mutant ($24.21 \times / 1.58 \text{ nM Kd}$) is predicted to have altered structure. A23G_U34C double point mutant is expected to have $33.3 \times / 1.6 \text{ nM Kd}$ based on A23G and U34C single point mutants' effect on affinity. **b)** Predicted secondary structures of the two single point mutants and the corresponding double mutant GFP aptamer with worse affinity than predicted from single point mutants. U34C single point mutant ($24.21 \times / 1.58 \text{ nM Kd}$) and U34C_C58U double point mutant ($68.24 \times / 1.63 \text{ nM Kd}$) are predicted to have altered secondary structures, whereas C58U single point mutant ($1.21 \times / 1.03 \text{ nM Kd}$) has a wild-type GFP aptamer-like structure. U34C_C58U double mutant is expected to have $6.85 \times / 1.28 \text{ nM Kd}$ based on U34C and C58U single point mutants' effect on affinity. Mutated nucleotides are colored in red. Folded structures and the folding free energies are predicted by mFold.

High-Throughput Affinity Profiling of *Drosophila* NELF-E

As a second example, we examined an interaction between NELF-E, an RNA binding protein with a highly conserved RNA recognition motif (RRM)(Pagano *et al.*, 2014), and a target RNA using HiTS-RAP. Our group has recently selected an RNA aptamer, NELFapt, which binds *Drosophila* NELF-E with high specificity and affinity. A truncated version, NELFapt min, of this aptamer binds NELF-E with ~20-50 nM affinity measured by EMSA or Fluorescence Polarization assay(Pagano *et al.*, 2014). Within NELFapt, NELF-E recognizes an 7-8 nucleotide motif (CUGAGGA(U)), called the NELF-E Binding Element (NBE), which is located within a putative k-turn motif that forms a sharp bend between two stems to present the NBE on a single stranded loop region(Pagano *et al.*, 2014). To further characterize this interaction, we mutated NELFapt through error prone PCR (epPCR) (Methods), and performed HiTS-RAP to probe NELF-E binding using a single lane of an Illumina GAIIx flowcell. We measured high confidence binding affinities for 9,832 mutants. The binding constant for the full length NELFapt was measured to be 5.2 nM, which is in good agreement with the 8.5 nM affinity measured by EMSA (Figure 2. 11).

The effects of single point mutants were analyzed to identify regions of the aptamer that are critical for its interaction with NELF-E. All of the 210 possible single base substitution mutants within the 70 nt long NELFapt were identified, 206 with high confidence fits for K_d (Figure 2. 12a). The majority of single point mutations did not have a significant effect on binding affinity, and none had a significantly higher affinity than the canonical aptamer. Based on the average effect of all three possible mutations at each position, mutations within the NBE were among the most disruptive to the interaction between the aptamer and NELF-E (Figure 2. 12b), with affinities as low as 76 nM for A43C (~15-fold weaker). Interestingly, the bases in the

loop region opposite of the NBE in the predicted secondary structure (GAUU, nucleotides 58-61) were also found to be important for binding (Figure 2. 12b). The most critical residues outside of the NBE were located at positions G58 and A59 (with G58C and A59C mutations having 65 nM and 110 nM K_d s, respectively). The deleterious effects of these two mutations were confirmed by EMSA (Figure 2. 11). These residues likely interact with G45 and A46 through non-Watson-Crick base pairing (Leontis, Stombaugh and Westhof, 2002) characteristic in k-turn structures. The observed effects of the single point mutations strongly support the putative k-turn structure of this aptamer and the presentation of the NBE as a single stranded loop.

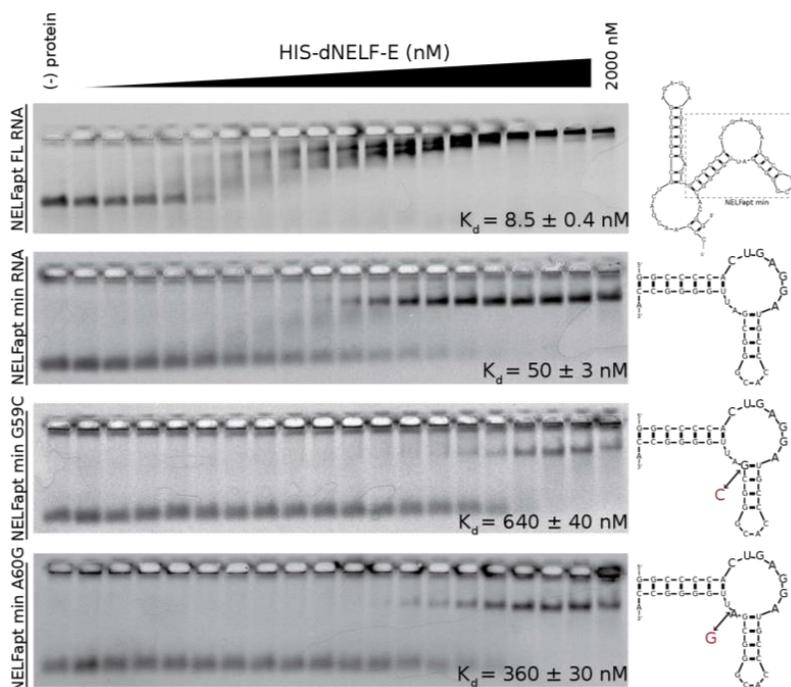


Figure 2. 11: EMSA confirmation of HiTS-RAP affinities for NELFapt

Affinities of two NELFapt mutations within the k-turn were verified by EMSA. These experiments were carried out using the minimal (35 nt) aptamer with one GC base pair added due to the need for a G rich region at the beginning of a T7 transcription template, as was done in the original publication (Pagano *et al.*, 2014). The minimal aptamer binds with lower affinity than observed by HiTS-RAP when measured by EMSA (50 nM for the minimal aptamer vs 8.5 nM for the full length by EMSA). Consistent with HiTS-RAP, A59G binds NELF with lower affinity (48 nM for the full length by HiTS-RAP, 360 nM for the minimal version by EMSA), and G58C binds with even lower affinity (65 nM for the full length by HiTS-RAP, 640 nM for the minimal version by EMSA). Thus, these two bases are as important as those within the NBE.

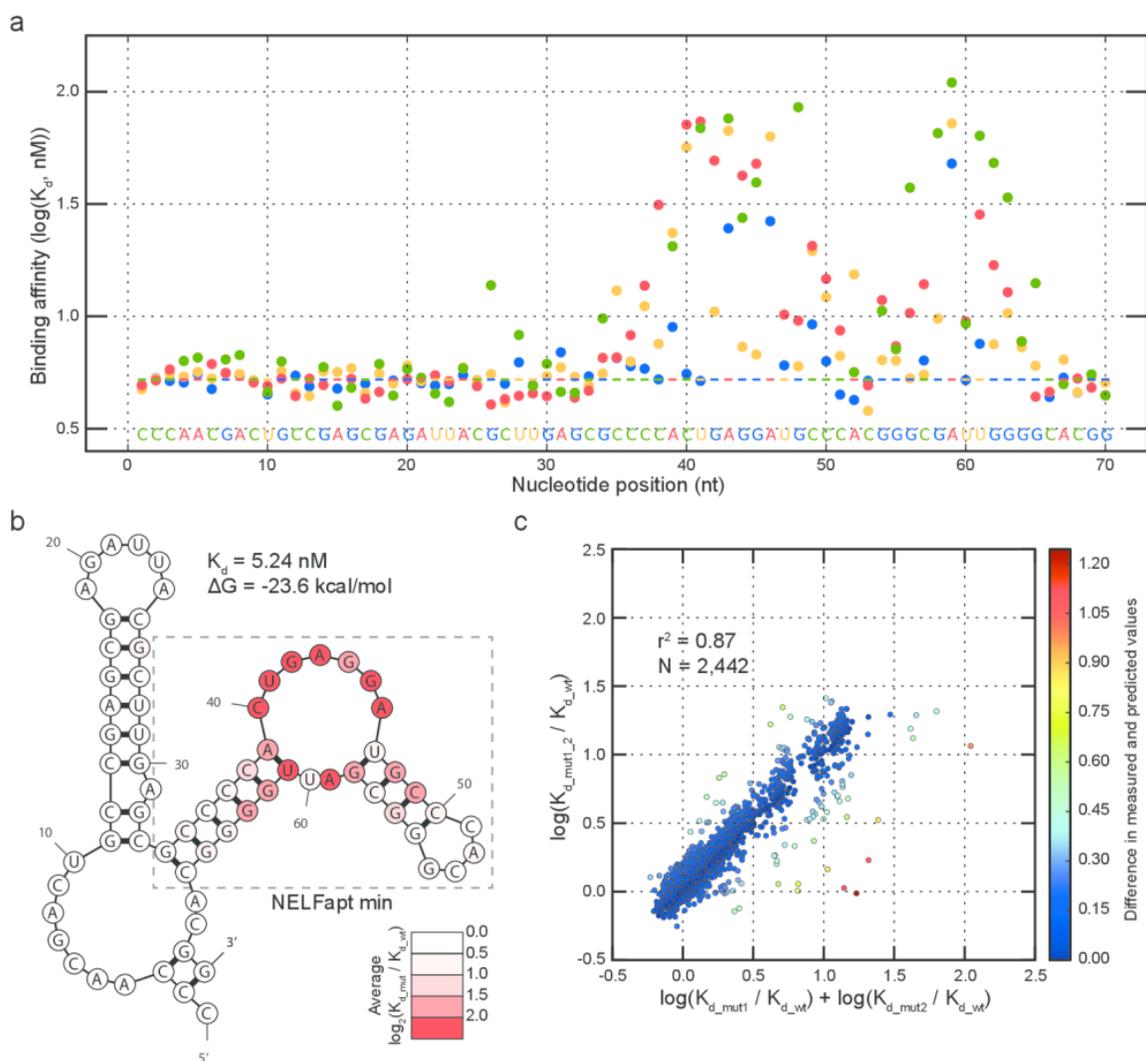


Figure 2. 12: Analysis of NELFapt by HiTS-RAP

a) NELF-E binding affinities of 206 single-point mutants of NELFapt. Binding affinities in nM are plotted in logarithmic scale. Mutations and the canonical aptamer sequence shown at the bottom of the graph are colored as in Figure 2. 8a. **b)** Predicted secondary structure of NELFapt. NELFapt is predicted to have 2 stem-loops connected via a loop-stem-loop structure in between. Predicted secondary structure is drawn and colored as in Figure 2. 8b. Most mutations show a less than 2-fold effect ($\log_2(K_{d,\text{mut}}/K_{d,\text{wt}}) \leq 1$) on binding affinity. **c)** Correlation between measured and predicted effects of NELFapt double mutants ($n = 2,442$). Measured fold effect of double mutants is plotted against the fold effect predicted by single mutations as in Figure 2. 8c. In this case, there is a strong positive correlation ($r^2 = 0.87$) between measured and predicted fold effect.

To assess the interplay of different nucleotides, we compared affinities of mutant NELF-E aptamers with single and double point mutations. We identified 2,442 double mutants with high confidence K_d s. In contrast with the GFP aptamer, the affinities of NELFapt double

mutants were predicted well by the affinities of their individual single mutants ($r^2 = 0.87$) (Figure 2. 12c). The majority of single mutations had less than a two-fold effect on affinity, and their effects were additive. Thus, most double mutants have small effects, and are well predicted by an additive model. Many of the double mutants that do not follow this trend are notable. Some are compensatory: for example, both A39G_U61C and A39U_U61A bind NELF-E far better than predicted by their corresponding single mutants' effect on affinity, presumably because the double mutants reconstitute a predicted AU base-pair in NELFapt as GC and AU base-pairs, respectively, at the end of a critical stem immediately adjacent to the NBE (Figure 2. 13a). Others, such as A39G_G63A bind with lower affinity than predicted by individual mutations because they may result in occlusion of the NBE within a double stranded region (Figure 2. 13b), which was shown to be deleterious for NELF-E interaction (Pagano *et al.*, 2014). Other double mutants that deviate from additive behavior for less obvious reasons have both mutations within the NBE, such as C40G_U41G. Altogether, this analysis shows that most mutations within NELFapt affect its interaction with NELF-E in an additive way with only a few notable and informative exceptions.

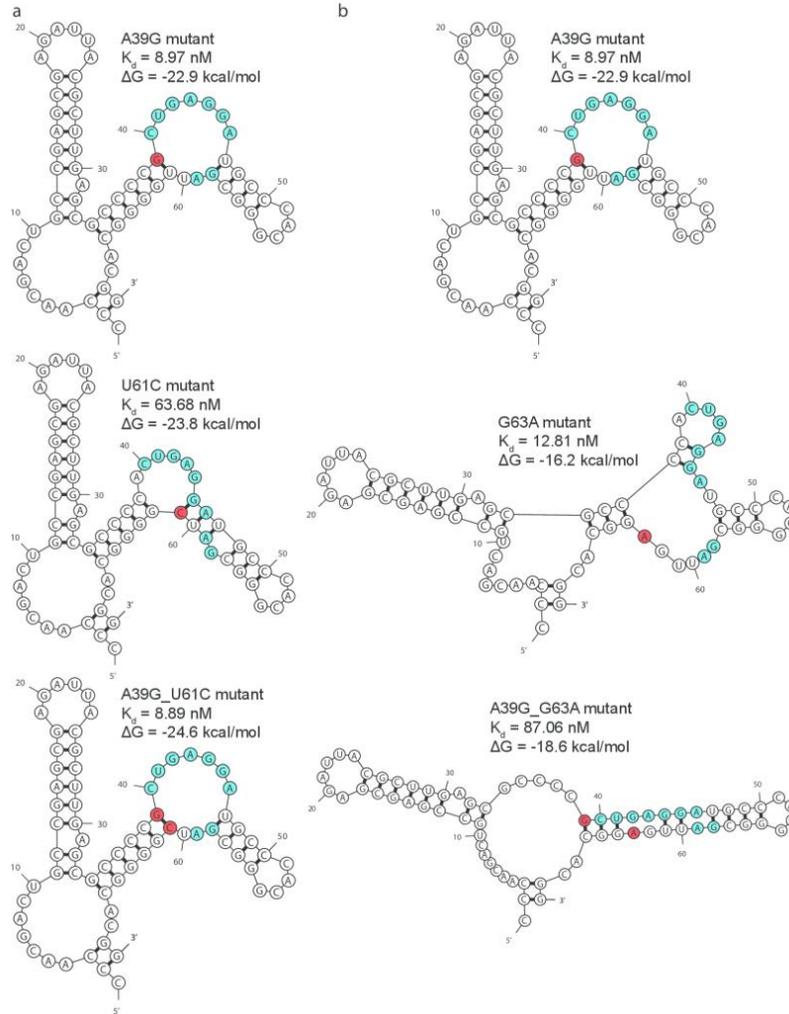


Figure 2. 13: Secondary structure predictions of NELF-E aptamer double point mutants with

a) Predicted secondary structures of the two single point mutants and the corresponding double mutant NELF-E aptamer with better affinity than predicted from single point mutants. A39G single point mutant (8.97 nM K_d) and A39G_U61C double point mutant (8.89 nM K_d) are predicted to have wild-type NELF-E aptamer-like structures, whereas U61C single point mutant (63.68 nM K_d) is predicted to have altered structure where 3'-end of the NBE and the k-turn inducing GA nucleotides, colored in blue, are in a stem. A39G_U61C double mutant is expected to have 109.04 nM K_d based on A39G and U61C single point mutants' effect on affinity. **b)** Predicted secondary structures of the two single point mutants and the corresponding double mutant NELFapt with worse affinity than predicted from single point mutants. A39G single point mutant (8.97 nM K_d) is predicted to have wild-type NELF-E aptamer-like structure. Both G63A single point mutant (12.81 nM K_d) and A39G_G63A double point mutant (87.06 nM K_d) are predicted to have altered secondary structures where the NBE and the k-turn GA are partially or completely in a stem. A39G_G63A double mutant is expected to have 21.93 nM K_d based on A39G and G63A single point mutants' effect on affinity. Mutated nucleotides are colored in red, and NBE and k-turn inducing nucleotides are colored in cyan. Folded structures and the folding free energies are predicted by mFold.

Discussion

Here, we developed a new method that combines high-throughput binding affinity measurements for RNA and sequencing in one assay, which we named HiTS-RAP. After a normal sequencing run, all additional manipulations for HiTS-RAP are carried out automatically by incorporating these steps into the .xml recipe used for the run, so that it adds very little to cost and hands-on time, but generates a large and useful data set. As presented here, this technique is streamlined enough that it could be carried out by anyone familiar with the GAIIX instrument. The fact that the sequencing chemistry used is identical across all of Illumina's instruments means that this would be possible in the future with HiSeq and MiSeq instruments as well, for increased data output or faster small-scale readout, respectively. Using this assay, we have accurately measured the affinities of two previously selected aptamers that bind GFP and NELF-E. In addition to the canonical aptamers, affinities of thousands of mutants of each aptamer were accurately measured. The effects of these mutations on affinity provided insight into the structure of the aptamers and identified regions that are critical for their RNA-protein interactions.

The interaction between GFPapt and EGFP is based upon the complex three dimensional structure of the RNA aptamer. Analysis of all possible single base substitutions showed that most positions critical for EGFP binding reside within stem-loops #2 and #3, while the rest of the aptamer is still required for proper folding (Figure 2. 8b). Of particular interest, two GFPapt mutants with higher affinities than the canonical aptamer were identified, highlighting the potential impact of HiTS-RAP for in-depth analysis and optimization of a pre-selected aptamer in a cost-effective high throughput manner.

In contrast, the interaction between NELFapt and NELF-E is primarily due to a short consensus motif presented on a single-stranded region. This analysis can also be used to identify

a minimal aptamer: in fact, HiTS-RAP shows that the sequence outside of the 35 nt minimal aptamer defined in the original publication (Pagano *et al.*, 2014) does not influence binding appreciably (Figure 2. 12b). The strong effect of mutations in bases opposite the NBE sequence motif in the predicted secondary structure shows that the presentation of the NBE within NELFapt, in the single stranded region of a k-turn, is just as important as its presence for recognition by this RNA binding protein. Thus, the sequence specified recognition of RNA by the RRM of NELF-E has a strong structural component as well.

The behavior of double mutants shows that the interaction between EGFP and GFPapt is fundamentally different from that of NELF-E with NELFapt. Multiple mutations within the intricate structure of GFPapt confound each other, so that the affinity of a double mutant is poorly predicted by the effects of the two individual mutations. In contrast, most double mutants of NELFapt are well predicted as the additive effect of their individual mutations, as we would expect for an interaction where only a small fraction of the total RNA is indispensable for binding. Moreover, many of the double mutants that deviate from this trend occur within the NBE, suggesting that mutations within this short consensus motif are less likely to affect the interaction additively. Sequences with two mutations within the NBE that have higher affinity than predicted by their single mutants could be used to identify an alternative NBE that deviates from the canonical motif, but could not be identified by single mutations alone. Such insights are enabled only by the large number of quantitative binding constants determined by HiTS-RAP. With a larger library of mutants, it may be possible to derive a comprehensive three dimensional model of an RNA-protein interaction.

HiTS-RAP has as its foundation a technique of halting transcription by sequence-specific binding of Tus to *ter* sites so that the RNA transcript emanating from the polymerase (i.e. T7

RNA polymerase) remains stably halted on DNA template and competent for interaction with fluorescently-labeled molecules for many hours. This could effectively be used to convert assays deemed to be restricted to DNA, such as microarrays, to measure properties of RNA as well. Additionally, although we have used HiTS-RAP to measure affinities of RNAs to a protein, its utility extends to any entity that can be fluorescently labeled (e.g., small molecules and peptides). Lastly, DNA libraries tailored for many applications, including random sequence libraries, aptamer libraries generated by SELEX, random genomic fragments, or targeted genomic libraries (such as nascent RNA, mRNA, ncRNA, enhancer regions, or pre-mRNA) can be easily adapted or constructed *de novo* for HiTS-RAP. Thus, HiTS-RAP could facilitate genome-wide, direct, quantitative measurement of RNA affinity for regulatory proteins with known or yet to be discovered RNA interactions. Overall, we have shown here that HiTS-RAP can determine quantitatively the affinity of specific RNA-protein interactions at an unprecedented scale and resolution.

Methods

Purification of proteins

The gene encoding Tus protein was PCR amplified from a vector provided by the D. Bastia lab, and then inserted between BamHI and XhoI restriction sites in the expression vector pGST-|| (Sheffield, Garrard and Derewenda, 1999). GST-Tus was then overexpressed in BL21 (DE3) RIPL bacteria for 4 hours after induction with 1mM IPTG at 37°C. The protein was purified using glutathione coupled agarose resin (Pierce), dialyzed into 40 mM Tris-HCl pH 7.5, 40 mM NaCl, 2 mM EDTA, 10 mM β -mercaptoethanol (Mohanty, Sahoo and Bastia, 1996), mixed with an equal volume of 80% glycerol, and stored at -80°C.

An EGFP-mOrange construct with a 4x GGS flexible linker was produced by overlap extension PCR, and inserted between the BamHI and XhoI restriction sites of the expression vector pHis-|| (Sheffield, Garrard and Derewenda, 1999). 6xHis-EGFP-mOrange was overexpressed in BL21 (DE3) RIPL bacteria for 4 hours after induction with 1mM IPTG at 37°C. The protein was purified using Ni-NTA coupled agarose resin (Pierce), dialyzed into 10 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 5 mM β -mercaptoethanol, allowed to mature at 4°C for one month, mixed with an equal volume of 80% glycerol, and stored at -80°C.

NELF-E mOrange was made following a protocol similar to EGFP-mOrange. mOrange was first PCR amplified with BglII and BamHI sites at either end, digested with those enzymes, and inserted into the BamHI site of pHis-||, conserving its entire multiple cloning site. NELF-E was then subcloned from pHis-|| into the mOrange containing pHis-|| using EcoRI and SpeI restriction sites. Expression and purification were carried out exactly as for EGFP-mOrange.

Library preparation

Figure 2. 4 shows a schematic of a HiTS-RAP template. In the case of NELFapt, error-prone PCR (epPCR) was first carried out as described elsewhere (Mccullum *et al.*, 2010). DNA templates were prepared for sequencing and transcription halting by sequential addition of primers by PCR. First, a T7 promoter was added to the 5' ends of the GFP, Sulforhodamine B (SRB), and NELF-E binding aptamers, and the Illumina sequencing primer site to the 3' ends. Then, adaptors complementary to oligos on the Illumina flowcell were added to either end along with a ter site immediately 3' of the Illumina sequencing primer site. A third PCR step was used to add adaptors for the 454 Life Sciences sequencing platform for templates used to make polystyrene beads with covalently linked halting templates. Sequences of primers used in this work are listed in Table 3.1.

Final halting templates for HiTS-RAP (T7 RNAP promoter in *italics*, aptamer sequence underlined, ter sequence in **bold**):

GFPapt halting template used for HiTS-RAP:

5' -

CAAGCAGAAGACGGCATAACGAGATCGGT *GATAATACGACTCACTATAGGGAATGGATCCACATC*
TACGAATTCAGCTTCTGGACTGCGATGGGAGCACGAAACGTCGTGGCGCAATTGGGTGGGGAAA
GTCCTTAAAAGAGGGCCACCACAGAAGCTAGATCGGAAGAGCGTCGTGTAGGGAAGAGTGTAA
ATTAGTATGTTGTAAGTAAAGTCACGTCATGAGATCTCGGTGGTCGCCGTATCATT-3'

SRB-2 aptamer halting template used for HiTS-RAP:

5' -

CAAGCAGAAGACGGCATAACGAGATCGGT *GATAATACGACTCACTATAGGGAATGGATCCACATC*
TACGAATTCGGAACCTCGCTTCGGCGATGATGGAGAGGCGCAAGGTTAACCGCCTCAGGTTCCA
GATCGGAAGAGCGTCGTGTAGGGAAGAGTGTAA**ATTAGTATGTTGTAAGTAAAGTCACGTCAT**
GAGATCTCGGTGGTCGCCGTATCATT-3'

NELFapt template used for epPCR:

5' -

*GATAATACGACTCACTATAGGGAATGGATCCACATCTACGAAT*CCCAACGACTGCCGAGCGAG

ATTACGCTTGAGCGCCCCACTGAGGATGCCACGGGCGATTGGGGCACGGCTTCACTGCAGACT
TGACGAAGCTT-3'

NELFapt halting template used for HiTS-RAP:

5' -

CAAGCAGAAGACGGCATAACGAGATCGGTGATAATACGACTCACTATAGGGAATGGATCCACATC
TACGAATTCCCAACGACTGCCGAGCGAGATTACGCTTGAGCGCCCCACTGAGGATGCCACGGG
CGATTGGGGCACGGCTTCACTGCAGACTTGACGAAGCTTATGGCTAGATCGGAAGAGCGTCGTG
TAGGGAAAGAGTGTA**AATTAGTATGTTGTA**ACTAAAGTC**CAGTCATG**AGATCTCGGTGGTCGCC
GTATCATT -3'

EMSA of halted transcription complex

An electrophoretic mobility shift assay (EMSA) was performed to resolve DNAs engaged in different complexes with Tus, T7 RNA polymerase, and RNA. The DNA template contains the GFP aptamer template flanked by a T7 RNA polymerase promoter at the 5' end and a *ter* element at the 3' end. This was then end labeled with ³²P using T4 Polynucleotide Kinase and γ ³²P- ATP as its substrate. The resulting radiolabeled-DNA was purified using a P30 size exclusion column (BioRad), mixed with either 1x T7 transcription buffer or GST-Tus at ~100 times the concentration of DNA in 1x T7 transcription buffer (30 mM HEPES pH 7.8, 80mM Potassium Glutamate, 15mM MgAc, 0.25 mM EDTA, 5 mM DTT, 0.05% Tween-20, 2 mM Spermidine), and incubated at 37°C for 30 minutes. Then, an equal volume of 1x T7 transcription buffer was added to DNA alone and DNA + Tus protein samples, and an equal volume of 2x transcription reaction was added to another DNA + Tus sample to make the Tus + DNA + transcription sample, and incubated at 37°C for another 30 minutes. The final transcription reaction consists of 1x T7 transcription buffer, 0.5 mM NTPs, 3 ng/ μ L T7 RNAP, 0.001 unit/ μ L YIPP (New England Biolabs), 0.2 units/ μ L SUPERase In (Ambion), 0.5 μ M GST-Tus). Glycerol was then added to a final concentration of 10%, and equal volumes of the

resulting reactions were run on a 4% native polyacrylamide gel. The gel was then dried, exposed to a PhosphorImager screen, scanned by a Typhoon 9400 Imager, and analyzed by ImageQuant software.

Transcription halting on 454 beads

Halting DNA templates compatible with the 454 sequencing platform were made by adding 454 adaptors via PCR to the halting templates used for the HiTS-RAP assay. These were used to coat 454 polystyrene beads with either SRB-2 aptamer or GFP aptamer templates by PCR (without an emulsion) using the bead-bound 454 Primer A and in solution 454 Primer B. An aliquot of these beads was washed with 1x T7 transcription buffer, incubated with 1 μ M GST-Tus in 1x T7 transcription buffer for 30 minutes at room temperature, washed in transcription buffer, and then resuspended in a transcription reaction. After transcription (30 minutes at 37°C), beads were washed with GFP aptamer binding buffer (1x PBS, 5 mM MgCl₂, 0.01% Tween-20), and incubated with 1 μ M 6xHis-EGFP. After 20 minutes of binding at room temperature, beads were washed with GFP aptamer binding buffer and imaged with a Zeiss Axioplan II epifluorescence microscope using a FITC/Fluo filter set (Chroma Technology Corp. Cat # 41001). Both DIC and fluorescence images are taken.

Transcription halting on glutathione beads

DNA templates were prepared with the GFP aptamer template flanked on the 5' end by a T7 promoter and an 11 nt C-less cassette, and two ter sites on its 3' end. First, these templates were

bound to GST-Tus in 10x molar excess in 1x T7 transcription buffer for 30 minutes at 37°C. The resulting Tus-DNA complexes were then incubated with glutathione coupled agarose beads (Pierce) in 1x T7 transcription buffer for 30 minutes at 37°C. An equal volume of 2x transcription reaction lacking CTP was then added to the resulting bead slurry, and transcription was allowed to proceed for 30 minutes at 37°C. At this time, the one aliquot of beads was washed three times in 1x T7 transcription buffer and then mixed with an equal volume of 2x transcription reaction mix lacking polymerase but containing all four NTPs. Transcription was allowed to proceed for another 30 minutes at 37°C. Both bead treatments were then washed with GFP aptamer binding buffer, incubated with 1 μM 6xHis-EGFP for 30 minutes at RT, washed again, and imaged as described before for 454 beads.

HiTS-RAP

A protocol paper for HiTS-RAP has been published: (Ozer *et al.*, 2015)

Libraries for sequencing were prepared to contain an RNA template insert flanked by a 5' T7 promoter and 3' ter binding site with the appropriate adaptors complimentary to the lawn of oligos on the Illumina flowcell at the extreme 5' and 3' ends. The ter site was positioned 3' of the Illumina sequencing primer site to ensure that the target RNA is fully emerged from the polymerase upon halting and so that sequencing begins with the target RNA. A standard sequencing run was performed with these libraries on an Illumina GAIIx using an 82 cycle read length on a paired end flowcell. The same .xml recipe used for the sequencing run included all subsequent steps to effect transcription halting and binding of mOrange labeled protein to the sequenced DNA clusters, so that fresh solutions are added at once for all steps through

transcription halting, and then once for the binding curve. Thus, reagents for HiTS-RAP are loaded onto the Paired End Module (PEM) twice. The recipe program is details the exact reagent delivery used for HiTS-RAP.

Table 3. 1: Oligonucleotides Used in HiTS-RAP

PCR Step	Library	Name	Sequence
epPCR	NELFapt	Temp DNA N70 Library FOR	GATAATACGACTCACTATAGGGAATGGATCCACATCTACGAATTC
epPCR	NELFapt	Temp DNA N70 Library REV	TACACTCTTTCCCTACACGACGCTCTTCCGATCTAAGCTTCGTCAAGTCTGCAGTGAA
1	NELFapt	Temp DNA N70 Library FOR	GATAATACGACTCACTATAGGGAATGGATCCACATCTACGAATTC
1	NELFapt	IllumFORSeq-BC2-AptLibConsREV	TACACTCTTTCCCTACACGACGCTCTTCCGATCTATGGCTAAGCTTCGTCAAGTCTGCAGTGAA
1	GFPapt	GFP Temp DNA FOR	GATAATACGACTCACTATAGGGAATGGATCCACATCTACGAATTCAGCTTCTGGACTGCGATGGGAG
1	GFPapt	Temp DNA REV	TACACTCTTTCCCTACACGACGCTCTTCCGATCTGTCTTCTGTGGTGGCCCTCTTTAAGG
1	SRB2apt	Temp DNA SRB-2 FOR	GATAATACGACTCACTATAGGGAATGGATCCACATCTACGAATTCGGAACCTCGCTTCGGCGATGA
1	SRB2apt	Temp DNA SRB-2 REV	TACACTCTTTCCCTACACGACGCTCTTCCGATCTGGAACCTGAGGCGGTAAACCTTGC
2	All	Illum IllumAdaptor T7	CAAGCAGAAGACGGCATACGAGATCGGTGATAATACGACTCACTATAGGGAATGGATCC
2	All	IllumFORAdapt T1 IllumFORSeq	AATGATACGGGACACCGAGATCTCATGACGTGACTTTAGTTACAACATACTAATTTACTACTCTTTCCCTACACGACGCTCTTCCGAT
3	All	454RevPrimer IllumFORAdapt	CTATGCGCCTTGCCAGCCCGCTCAGAATGATACGGCGACCCCGAATCT
3	All	454PrimerA IllumREVLawn	CGTATCGCCTCCCTCGGCCATCAGCAAGCAGAAGACGGCATACGAGATCGGT

2.25 mL of each solution was loaded onto its own position on the PEM of the GAIIX. The .xml recipe delivered reagents in the proper sequence, and set the flowcell temperature using the Peltier heater in the instrument. During each reaction or binding step, 75 µL of solution is flowed through each lane, and then the flowcell is incubated for 30 minutes to equilibrate. During the incubation, 15 µL of fresh solution is delivered to each lane every 5 minutes. The second strand generated during sequencing was stripped away and 1 µM primer (IllumFORAdapt_T1_IllumFORSeq, Table 3. 1) for double stranded DNA regeneration annealed as per the standard Illumina protocol. Excess primer was then washed away with 1x NEB buffer 4 with 0.01 % Tween-20 (New England Biolabs). DNA was then made completely double stranded by flowing in a Klenow exo- enzyme reaction mix (1x NEB Buffer 4, 0.01% Tween-20,

0.2 mM dNTPs, Klenow exo-(New England Biolabs)) and incubating for 30 minutes at 37°C. The flowcell containing double stranded DNA clusters was then equilibrated with 1x T7 transcription buffer. Tus was allowed to bind the DNA templates' ter elements by equilibrating with 1 μ M GST-Tus in 1x T7 transcription buffer for 30 minutes at 37°C. The flowcell was then equilibrated with a transcription reaction (1x T7 transcription buffer, 0.5 mM NTPs, T7 RNAP, YIPP, Superase In, ~0.5 μ M GST-Tus). Transcription and halting was allowed to proceed for 30 minutes at 37°C.

After transcription halting, the flowcell was equilibrated with GFP or NELF (10 mM HEPES, pH 7.5, 100 mM NaCl, 25 mM KCl, 5 mM MgCl₂, 0.02% Tween-20) aptamer binding buffer at room temperature. It was imaged immediately, just as during sequencing, to measure the background intensity at every cluster. It was then equilibrated successively with increasing concentrations of mOrange labeled protein in binding buffer and imaged. Imaging at each concentration was carried out in equilibrium binding after 30 minutes equilibration at room temperature. Concentrations of EGFP-mOrange varied from 0.04 nM to 625 nM, increasing in fivefold increments. The images in Figure 2. 5b are from .tif files collected by the Illumina SCS during this binding. HiTS-RAP for NELF-E mOrange was identical to EGFP-mOrange, except that concentrations of protein varied from 0.064 to 1000 nM.

Sequencing and data extraction

Sequencing, transcription, and protein binding are executed as a single run on a GAIIx. In the case of NELF-E, transcription and protein binding were carried out twice. The first binding curve was used for all analyses. Intensity data were collected and basecalling done using the Illumina SCS version 2.9. During the run, .cif and .bcl files were saved, along with 5% of the

raw .tif images taken. For labeled protein binding steps, intensities are measured for all four channels for every cycle at every cluster, and saved in binary .cif files, just as during the sequencing portion of the run. We used scripts generously provided by Robin Friedman to extract intensities from these files corresponding to mOrange fluorescence for every cluster. This gives the coordinates of each cluster, together with its protein binding intensities (in the T channel) and average intensity during sequencing. These data were then matched by cluster coordinates to their sequence from .qseq.txt files generated by the Illumina OLB version 1.9.4. Only clusters that passed filter were included. If a cluster contained a mutation (deviating from NELFapt or GFPapt), it was only included if the mutated base had a quality score greater than 25 (<0.003 probability of an incorrect base call). A threshold of quality score >20 is common practice for SNP calling algorithms (Nielsen et.al Nature Review Genetics 2011 Vol 12, page 443).

Loss of signal correction

All development of HiTS-RAP was done using the interaction between the GFP aptamer and EGFP-mOrange as a model RNA-protein interaction. After the binding curve, the flowcell was imaged 9 successive times after reequilibrating with 625 nM EGFP mOrange. After three of these cycles, imaging was not carried out, but rather the flowcell was allowed to sit in 625 nM EGFP-mOrange for the time that it would take for one equilibration and binding cycle. Thus, the 9 imaging steps span the time that it would take for 12 equilibration and imaging steps. A similar rate of decay was observed between successive and staggered imaging cycles, indicating that most loss of signal is due to time rather than photobleaching caused by the number of times that the flowcell was imaged (Figure 2. 6). We find that in the regime where our imaging takes place (i.e. where time is much less than the characteristic lifetime), the observed decay is

described well by a linear approximation to exponential decay. Thus, we took the average intensities of clusters with canonical GFP aptamer sequence through the nine cycles imaged and used `scipy.optimize.curve_fit` to perform a weighted linear least squares regression, using weights determined from the standard deviation of the intensities at each time point. The equation used was:

$$I_{observed} = I_0 \left(1 - \frac{t}{\tau}\right)$$

where $I_{observed}$ is the measured mOrange intensity, I_0 is the initial intensity, t is the time in units of cycles, and τ is the characteristic lifetime in cycles. I_0 and τ were solved for the GFP aptamer clusters in each of the three lanes (each containing ~ 2.7 million clusters with canonical GFP aptamer sequence); the final characteristic lifetime is the average of these three fitted lifetimes. In this regime, halted complexes are decaying at a rate corresponding to a characteristic lifetime of 137 ± 7 cycles, or 206 ± 11 hours (at an average cycle time of ~ 1.5 hours). This lifetime simply describes the rate at which the halted transcription complexes are decaying, independent of the actual intensity. Given that the background intensity in these lanes is approximately 85, this means that fluorescence signal will be above background for 48 cycles, or 72 hours (at the decay rate we have measured, it takes 48 cycles to decay to the background intensity of 85 from the initial intensity of 131).

The characteristic lifetime was used to apply a correction factor to each fluorescence intensity measured in binding curves, using the equation:

$$I_{actual} = \frac{I_{observed}}{\left(1 - \frac{t}{\tau}\right)} = \frac{I_{observed}}{1 - 0.00730 \times t}$$

where I_{actual} is the fluorescence intensity used in the fit for K_d , $I_{observed}$ is the measured intensity, 0.00730 is the decay rate in cycle⁻¹ (the inverse of τ), and t is the time since transcription, in cycles. Figure 2. 6 shows intensity data from three lanes, together with the same intensities after applying the correction factor.

K_d calculation

Sequential measurements of the corrected mOrange intensities at increasing concentrations give a binding curve for each cluster. Intensities were normalized for cluster size and position in the tile by dividing each intensity by the average sequencing intensity (Nutiu *et al.*, 2011). This correction is applied so that all intensities from a lane can be averaged, to be representative of each sequence. Clusters with the same sequence in each lane were matched, and average binding curves were generated by taking the mean of their normalized intensities for each protein concentration. This gives an average binding curve representative of each sequence. If more than 10 clusters had a given sequence in a lane, this binding curve was fit to a Hill equation, solving the equilibrium dissociation constant (K_d) of the interaction (Barlow and Blake, 1989). We have used the following Hill equation:

$$I = b + \frac{m - b}{1 + \left(\frac{K_d}{C}\right)^n}$$

Where b is the background fluorescence intensity at the cluster, m is the maximum fluorescence intensity, K_d is the dissociation constant, n is the Hill coefficient, and C is the concentration of target protein. The intensities measured by imaging at several different concentrations are then used to solve m , b , K_d , and n in a weighted non-linear least squares regression. We have found

that letting these four parameters vary gives reliable fits; this also means that the K_d that we solve is independent of the intensity values, so that it shows the inflection point in the binding curve. Only fits for which the `scipy.optimize.curve_fit` algorithm of the NumPy Python package returned a variance of less than 1,000,000 were considered to be high confidence and used in these analyses.

The GFPapt run contained three lanes of GFPapt. To generate a single data set from all three lanes, K_d s for every unique sequence in each lane were determined by fitting to average binding curves. The K_d s for all unique sequences in the flowcell were then determined by geometrically averaging the fitted K_d s across the three lanes: therefore the K_d values are reported as the average $K_d \times/$ (multiply or divide by) standard deviation.

The GFPapt data have the added complication of very low background binding. Thus, there are sequences in this data set which do not bind GFP measurably, while for NELF-E, background binding for this RNA binding protein is high enough that every sequence is expected to bind to some extent. To mitigate this problem, only sequences which show a 3% increase (determined in Figure 2. 6b-d) between the first and last two measured intensities were considered as binding: all others were called as not binding, meaning that they effectively have a K_d greater than 125 nM. We set the limit at 125 nM because this was the second highest concentration probed, so we do not expect to be able to measure affinities greater than this. For analysis of single mutants, any mutant scored as not binding based on its intensity increase was assigned an affinity of 125 nM; this value was averaged with other real measurements from other lanes if it was called as binding there. This results in a K_d measurement of greater than that average. Such mutants were excluded from analysis of double mutants.

EMSA of GFPapt and NELFapt and mutants

EMSA was used to verify the HiTS-RAP measured binding affinities of wild-type and mutant GFP binding aptamers to EGFP. To this end, in vitro transcribed aptamers were 3' end-labeled with AlexaFluor647 Hydrazide (Invitrogen) as described elsewhere (Pagano, Clingman and Ryder, 2011) and quantified by Qubit Fluorometer (Invitrogen). Fluorescently labeled aptamers were mixed with recombinant GST-EGFP protein at 25°C for 45 min. The GST-EGFP concentration in the binding reaction was varied as a 2/3 dilution series starting from 500 nM, and a no protein control. The final 20 µl binding reaction was composed of 1X PBS, 5 mM MgCl₂, 0.4 U of Suprase In, 1 µg of yeast tRNA, 0.005% NP-40, and 5 nM fluorescently-labeled aptamer. After addition of Bromocresol Green containing 30% glycerol to 6% final glycerol concentration, binding reactions were loaded on a 6% polyacrylamide slab gel (0.5X TBE, 5 mM MgCl₂) that was pre-equilibrated to 4°C and pre-run at 120V for 10 min at 4°C. Loaded gels were run at 120V for 90 min at 4°C, and then imaged with a Typhoon 9400 scanner using Cy5 settings. Images were quantified by ImageQuant5.2 software, and data were fitted to Hill Equation to determine the K_d values using Igor software.

EMSA was carried out with fluorescein labeled minimal NELFapt as described elsewhere (Pagano *et al.*, 2014).

RNA secondary structure predictions

The average absolute effect of all three mutations at each position was calculated using the following equation: $[|\log_2(K_{d_mut1}/K_{d_wt})| + |\log_2(K_{d_mut2}/K_{d_wt})| + |\log_2(K_{d_mut3}/K_{d_wt})|]/(\text{Number of point mutants observed at that position})$.

We used Kinefold(Xayaphoummine, Bucher and Isambert, 2005) to predict the folded structures and the associated folding free energies. Kinefold, unlike other programs which predict secondary structures based on minimal free energy, predicts the structure of RNA as it is being synthesized, and thus recapitulates what is happening in HiTS-RAP, where the RNA folds co-transcriptionally. This predicted structure is consistent with the published secondary structure of the GFP aptamer which was supported by significant mutational analysis(Shui *et al.*, 2012). Perhaps minimal free energy based programs would be more suited if the aptamer had undergone a heat denaturation and renaturation cycle following synthesis in its entirety.

Predicted secondary structures and the Gibbs Free Energy (ΔG) of the folded structures are obtained from Kinefold, unless indicated otherwise, however the actual drawings are obtained from mFold(Zuker, 2003), because they are easier to manipulate.

Notes on HiTS-RAP²

Efficiency of halting

While halted transcription complexes contain a single RNA per DNA after washing, we see that during the transcription reaction, T7 RNAP is likely transcribing until it reaches Tus, whereupon it either halts or terminates. In the situations presented here, we have separated the transcription complexes from the other components of the transcription reaction that allow the cycle of halting and termination to continue; in the EMSA the unincorporated NTPs are removed from the complexes quickly due to their high mobility through the gel, in the cases of halting on beads, the beads were washed, and on the GAIIX, the flowcell was washed. In these cases, we have shown that halted complexes with functional RNAs are very stable. However, we see that during transcription, RNA produced by polymerase that terminates at the *ter* site accumulates. When we transcribe DNAs bound to beads, a large amount of transcript remains anchored to beads, but an even larger fraction is released into the supernatant (Figure 2. 2a). Bacterial RNA polymerases have been shown to terminate through a cooperative forward translocation mechanism when presented by roadblocks (Epshtein *et al.*, 2003). Thus, removal of the components of the transcription reaction from the halted transcription complexes is likely to be at least partially responsible for the stability that we observe.

The experiment in Figure 2. 2a also demonstrated the polarity of Tus. In a parallel reaction, the templates used were identical to those in the first except that their *ter* sites were in the opposite orientation. In this case, T7 RNAP is able to break through the barrier of Tus and run off the end of the template in most cases; this template had two *ter* sites in the permissive

² Supplementary text from (Tome *et al.*, 2014)

orientation and only minor bands corresponding with each ter site are seen (Figure 2. 2a).

Therefore, when Tus is bound to DNA in the non-permissive orientation, nearly every T7 RNAP either halts or terminates, while in the opposite orientation, most polymerases proceed past the Tus/ter complex.

We believe that the technique for transcription halting presented in this work will be suited to a wide range of applications. It is the first technique, to our knowledge, capable of producing stable complexes containing both a DNA molecule and its RNA transcript. We have presented one case where it was coupled with high-throughput DNA sequencing technology to analyze RNA properties at the time of sequencing.

Strategies for solving Kds

We have determined that a single weighted nonlinear regression to an average binding curve representative of a unique sequence is the best way to solve dissociation constants. In this way of fitting, the intensities of each individual clusters' binding curve are normalized for cluster size and position (Nutiu *et al.*, 2011) by dividing by the average intensity of that cluster during sequencing. After this normalization, each average intensity is representative of each sequence's behavior at a given concentration. The standard deviation of the clusters' intensity gives a good measure of confidence in how that average intensity is truly representative of the RNA's behavior; thus, it is used as the weight in the fit. We find that median and average binding curves give similar dissociation constants, especially considering that this is a case where we are measuring point mutants, so we expect many of the measured affinities to be near that of the canonical aptamer.

We have also used a method of fitting where we solve a dissociation constant for each cluster individually in a nonlinear regression. This method truly exploits the data to its full potential, as each cluster is itself an independent binding event. Rather than giving each unique sequence only one opportunity to converge on a good fit to the Hill equation, this method gives these sequences as many fits as there are clusters. The dissociation constant for a unique sequence is then the geometric mean of these fitted K_{dS} . While we find that the geometric mean of all fitted K_{dS} is usually accurate, individual clusters tend to contain a large amount of noise and therefore give poor fits. Because of this, we use only those whose fits return a low covariance when taking the mean for a unique sequence. A major disadvantage of this method, however, is that it is computationally expensive. When doing fits to average binding curves, we take only unique sequences with a certain number of clusters in the flowcell. This amounts to only several thousand fits per lane of the flowcell. Individual fits, however, require several million fits per lane. The fact that these two methods give similar results means the extra time and complication of individual fits is unnecessary.

HiTS-RAP as a Tool for Identification, Characterization, and Optimization of Aptamers

HiTS-RAP can be used to optimize aptamers that are selected by SELEX. Although they have enormous diversity (up to 10^{16} unique sequences), libraries used for SELEX are far from having complete coverage of the potential sequence space (70 nt random library = 1.4×10^{42} unique sequences). Therefore, SELEX identifies the highest affinity aptamers present within the starting library, but the true highest affinity aptamer could be a related sequence that was not present. In addition, even the most enriched aptamer pools at the end of a SELEX process with 10 rounds of selection often have a large complexity with thousands to millions of different

aptamers(Schütze *et al.*, 2011). The highest affinity aptamers in such pools are not necessarily the most abundant ones either, therefore, a high-throughput analysis method, like HiTS-RAP, would be useful in identifying the highest affinity aptamers for downstream applications. While large-scale, quantitative characterization of DNA aptamer libraries has been carried out(Cho *et al.*, 2013), no such method has been developed for RNA libraries. HiTS-RAP could be used to rapidly identify the true highest affinity RNA aptamers. This point is well demonstrated by identification of GFP aptamer mutants that have higher affinity than the canonical aptamer (i.e. C58U and U60A). The original selection of the GFPapt was exceptionally difficult, requiring a laborious two stage selection process with 27 total rounds of selection, followed by mutational analysis, and structure minimization. Therefore, the starting, unmutated GFP aptamer that was used in this study had already undergone a substantial optimization. Also, HiTS-RAP could be modified in ways to select aptamers for additional functions such as disruption of a specific interaction of the target molecule. Such applications would be valuable given the fact that many aptamers which disrupt their target proteins' function or interaction with other molecules are in clinical trials(Ni *et al.*, 2011; Sundaram *et al.*, 2013).

Overview of HiTS-RAP³

RNA encoding DNA Library.

Any library suitable for Illumina sequencing could be adapted for HiTS-RAP. Thus, biological assays such as PAR-CLIP or ChIP-seq could be used to enrich for specific candidate regulatory regions of the genome, which could then be interrogated for their RNA transcript's affinity for a protein. The quality of the DNA templates used for HiTS-RAP is very important. Since multiple PCR steps are used for creation of the DNA template, we optimized the number of PCR cycles for each step and monitored the quality of the PCR product by polyacrylamide gel electrophoresis at each step. PCR products are purified by PCR purification kit or gel extraction (especially for the final template) to ensure the best quality and sufficient removal of primer dimers and other PCR byproducts. Shorter fragments can easily dominate the flowcell due to better amplification efficiency during cluster generation, and as a result diminish the number of full-length templates that can be analyzed. We use home-made Taq or Phusion polymerases for PCR amplification. Depending on the purpose (i.e. mutational analysis of a single RNA-single protein or a library of RNAs with a single protein) the choice of DNA polymerase and PCR conditions (e.g. buffer composition, number of PCR cycles) in each PCR step can be modified to increase or to minimize the rate of PCR introduced mutations using Taq polymerase under error-prone PCR conditions (McCullum *et al.*, 2010) or high fidelity Phusion polymerase, respectively. Quantification of DNA template concentration, for which we use Qubit dsDNA HS Assay, is also critical to get the optimum number of cluster on the flowcell. Each type of library produces

³ This section is an excerpt from a paper co-written with Abdullah Ozer and John Lis: Ozer, A., Tome, J. M., Friedman, R. C., Gheba, D., Schroth, G. P., & Lis, J. T. (2015). Quantitative assessment of RNA-protein interactions with high-throughput sequencing-RNA affinity profiling. *Nature Protocols*, 10(8), 1212–33. <https://doi.org/10.1038/nprot.2015.074>

clusters with different efficiencies, so we have found that it takes some experience to reliably reach an optimal cluster density on the flowcell.

T7 RNA polymerase is used for transcription because it has a short, specific promoter sequence, and is widely available commercially or easily prepared in-house, and very active.

The DNA templates that we have used in HiTS-RAP so far are designed for single-read sequencing of a single template and mutants derived from it; however, conceptually it is possible to do paired-end sequencing by incorporating TruSeq or Paired-end sequencing oligos into the template design. This would allow analysis of protein interactions of longer RNAs that are derived from the genome or transcriptome. In such a template the RNA encoding DNA would be flanked by the two Illumina paired-end sequencing oligos on either end, then by a T7 RNA polymerase promoter and Tus binding Ter site, and finally the Illumina flowcell adaptors. The portion of the Illumina sequencing oligo that becomes part of the transcript can be used to quantitate the amount of RNA transcript in each cluster using a fluorescently labeled complementary oligo as has been done in RNA-MaP (Buenrostro *et al.*, 2014). A barcode sequence can be introduced between the T7 RNA polymerase promoter and Illumina sequencing oligo and can be read by an index read as part of multiplex sequencing protocol, and allow simultaneous processing and analysis of multiple RNA libraries in a single-lane of a GAIIX flowcell, thus minimizing experimental differences between the two libraries. This might be particularly useful for analyzing interactions of a protein with RNA libraries prepared from different sources or under different conditions, and SELEX enriched aptamer libraries from various rounds or perform under different conditions.

Critical considerations in designing such a template: We always design libraries so that sequencing is done toward the flowcell, and transcription proceeds away from the flowcell surface. We reason that this increased accessibility of RNA transcript for the fluorescently labeled protein of interest. Sequencing toward the flowcell allows us to place the Illumina sequencing primer just upstream of the Tus binding site, so that it serves as a stalling site during transcription. Limitations on the length of DNA template (~500 bp max total) restricts the length of the RNA encoding DNA region to ~400 bp. Longer DNA are ideal templates for neither sequencing nor bridge-amplification during cluster generation. Transcription halting by Tus-bound Ter site is orientation specific; that is, the Tus-bound Ter site is permissive to read-through transcription in one orientation but not in the opposite direction (similar to DNA polymerases).

We strongly recommend a testing the materials used for a HiTS-RAP run in vitro before applying to the sequencer. This includes testing transcription, with and without halting, of the template DNA by simply transcribing overnight and running RNA on a denaturing gel, and EMSA, respectively. We also like to test a known interaction of the labeled RBP by EMSA. The cost and labor associated with the HiTS-RAP assay is immense for a failure, which could have been avoided by these tests. Assessing a rough estimate of K_d is also important for deciding what protein concentrations to use for HiTS-RAP protein binding steps.

Writing an .xml recipe for a GAIx run

1. Recipes begin by setting the tile selection, in the TileSelection definitions section. This defines which tiles are imaged in every step of a recipe. The same tile selection must be used in each step (i.e. it is not possible to sequence a whole flowcell and then image protein binding to a

subsection of it within the same recipe). There are two types of tile selection. Incorporation defines the tiles used during sequencing and protein binding, and ReadPrep defines the tiles to be used in the first cycle of sequencing for finding focus and defining the edges of the flowcell.

If one were only using a subsection of the flowcell, as in (Gravina, Lin and Levine, 2013), the Incorporation definition would be changed to indicate this.

2. Chemistry definitions. Chemistry attributes are defined at the beginning of a recipe.

a. Each attribute definition begins by defining the name, which can be used to call the chemistry during the protocol.

b. Next, set the temperature to be used for the attribute with the Temp command. This uses the Peltier heater in the instrument to warm or cool the flowcell. If no temperature is defined, the flowcell will be at ambient temperature. A Duration element can be used here to set the length of time that the flowcell will be held at the specified temperature.

c. After that, solutions can be pumped through the flowcell with the PumpToFlowcell attribute. The Solution element for this attribute indicates which port on the GAIIX or PEM the instrument will pull solution from. AspirationRate sets the rate of flow, in $\mu\text{l}/\text{min}$. We leave the DispenseRate at 2500 $\mu\text{l}/\text{min}$. Finally, the Volume element indicates how much of the solution, in μl , will be pumped through each of the eight lanes of the flowcell (thus, it is one eighth of the total solution used).

d. We typically use a Wait command between PumpToFlowcell commands to set incubation times. The Duration element of this command sets the time of the incubation, in milliseconds.

e. The chemistry definition that we use for transcription is included below as an example. First, we set the temperature of the Peltier heater to 37 C. Next, we pump 75 μ l of transcription reaction through the flowcell (the port at position 11 on the PEM was loaded with 2.25 ml of transcription reaction). The flowcell is then incubated for 5 minutes. Then, 15 more μ l of solution is flowed through each lane, followed by a five minute incubation, several times so that 150 μ l of solution is flowed through each lane of the flowcell and it is incubated for 30 minutes total for transcription. We carry out every step with periodic addition of more solution in this way. Finally, 75 μ l of the next solution (transcription buffer only) is flowed through each lane, through position 14 of the PEM. The TempOff element removes the Peltier heater from the flowcell. Finally, we close the attribute definition with </Chemistry>.

```
<Chemistry Name="Transcription">
    <Temp Temperature="37" Duration="120000"/>
    <PumpToFlowcell Solution="11" AspirationRate="50" DispenseRate="2500"
Volume="75" />
    <Wait Duration="300000"/>
    <PumpToFlowcell Solution="11" AspirationRate="50" DispenseRate="2500"
Volume="15" />
    <Wait Duration="300000"/>
    <PumpToFlowcell Solution="14" AspirationRate="60" DispenseRate="2500"
Volume="75" />
    <TempOff />
</Chemistry>
```

3. Protocol section. This section contains instructions for carrying out the run. It utilizes the attributes defined in the TileSelection and ChemistryDefinition sections to carry out the run. Instructions are carried out in the order that they are listed in the protocol. Our HiTS-RAP recipes have all started with a complete sequencing run, which we leave unchanged from the standard Illumina recipes. We then perform the protein binding (i.e. RNA Affinity Profiling (RAP)) steps at the end. We have utilized three types of commands: UserWait, ChemistryRef, and Incorporation.

a. UserWait commands are used to pause the run until the user loads solutions required for the next step. Text can be entered in the Message element, which then directs the user to click OK to continue with the next step in the protocol.

b. ChemistryRef commands are used to carry out a chemistry step, as defined in ChemistryDefinitions, without imaging. The Name element is used to indicate the chemistry step used.

c. Incorporation commands are used for steps where the flowcell is to be imaged. ‘Incorporation’ indicates the tile selection that was defined in the TileSelection section. The ChemistryName element indicates which chemistry step to carry out before imaging. The four Exposure elements (ExposureA, G, T, C) indicate exposure times to be used during imaging, in milliseconds, for each channel.

We typically wait to prime a refrigerated (enzyme or protein containing) solution until just before it is used. Priming fills the fluidic lines with the solution; thus, solution that is primed early will sit in the line, at room temperature, before being pumped into the flowcell. The dead

volume within the lines from the PEM to the flowcell is ~550 μL [1], which is a large fraction of the total solution used for each step.

The following is an example from a protocol section of one of our recipes. It starts with a chemistry reference that primes reagent port 11 on the PEM (which contains transcription reaction). Another chemistry reference then carries out transcription, without imaging. A comment, contained within `<!-- "comment" -->`, then tells readers of the recipe what should have happened at this point in the recipe. A `UserWait` attribute then pauses the run so that the protein solutions can be loaded on to the instrument. It includes a `Message` element which tells the operator of the instrument what should have happened up to this point and how to proceed. Next, there is another comment, indicating the imaging cycle number (this is helpful when carrying out analysis). Last is the incorporation step carried out immediately after the `UserWait`. Incorporation indicates that the tile selection defined as `Incorporation` in the `TileSelection` section will be used for imaging. The `ChemistryName` element dictates that the `TargetWash` chemistry definition will be used before imaging (this washes the flowcell with the protein binding buffer). Exposure elements are then used to define what exposure times to use for each channel in the imaging.

```
<ChemistryRef Name="Prime11x3" />
<ChemistryRef Name="Transcription" /> <!-- At this point, all clusters are transcribed and
      halted. No target has been added yet-->
<UserWait Message="Transcription complete. Replace chilled reagents on PEM with protein
      target dilutions,
with concentration increasing from positions 16 to 20 and then 9 to 13. Click OK to Proceed, or
      CANCEL to Stop." />
<!-- Cycle 83 -->
<Incorporation ChemistryName="TargetWash" ExposureA="500" ExposureC="350"
      ExposureG="200" ExposureT="175" />
```

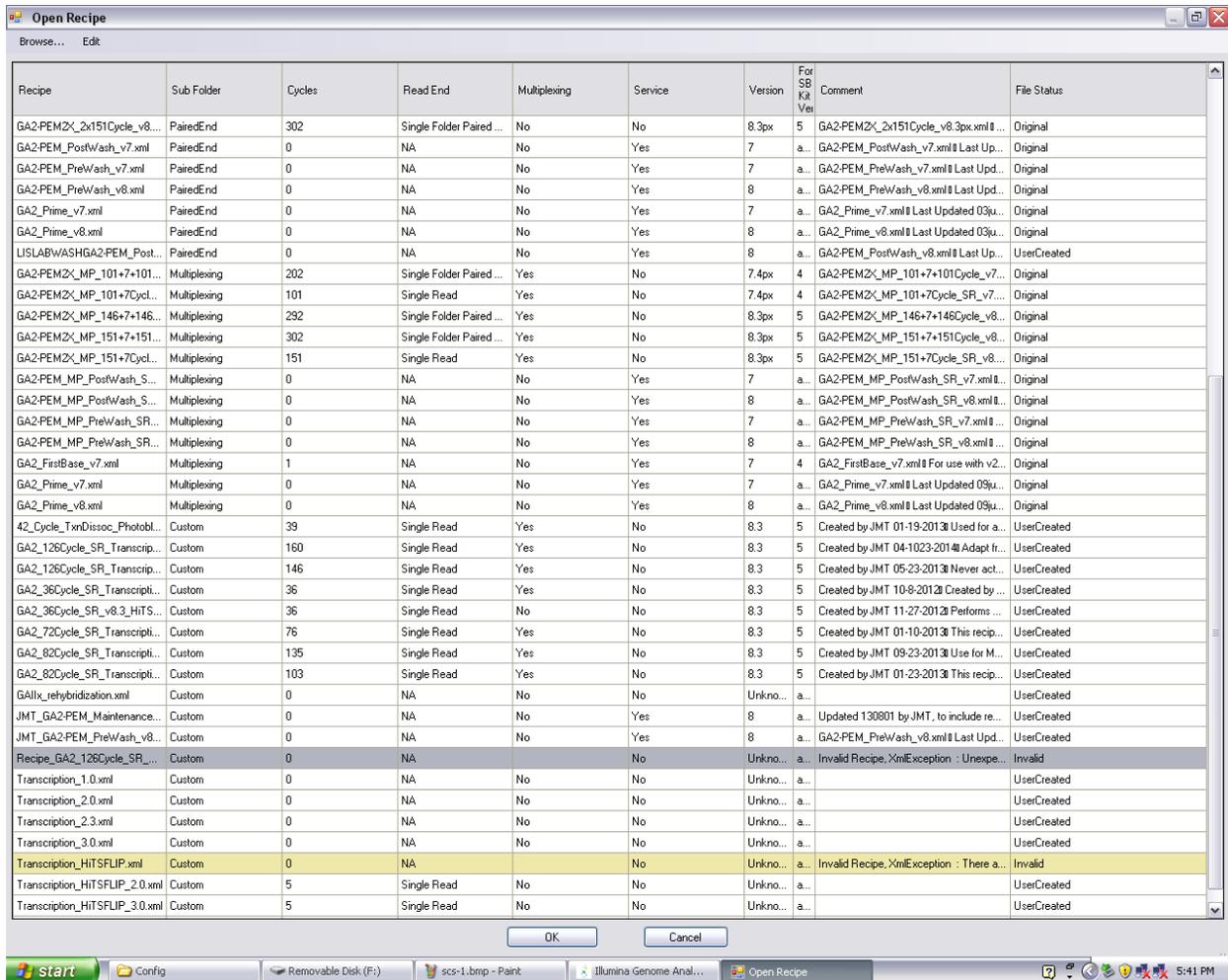
4. Check that the recipe contains no errors and can be read by SCS 2.9 software.

a. To do this, first place the modified recipe into the custom recipes folder on the computer controlling the GAIIX instrument

(C:\Illumina\SCS2.9\DataCollection\bin\Recipes\Custom).

b. Open the recipe from within the SCS2.9. Select the Open Recipe command under File tab. Select the recipe to open.

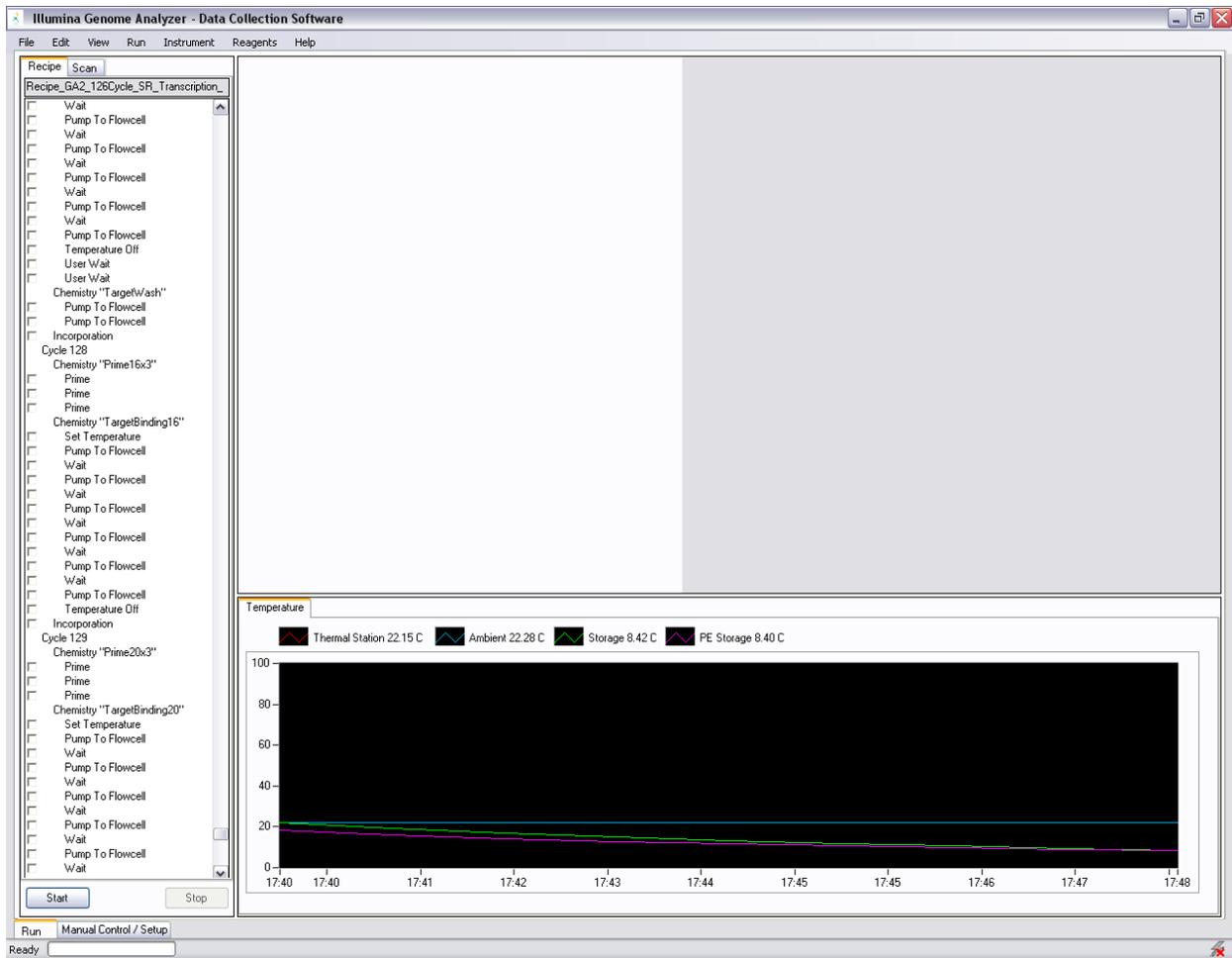
c. If the recipe contains an error, it will not open. Type of the error and the line containing the error will be indicated under the Comment column of the open recipe dialogue box, and the recipe will be marked as “Invalid” under File Status column, as shown below.



d. Correct the error in the recipe file using NotePad++. Special characters such as #, &, ! have specific uses, therefore their use in comments and messages should be avoided. Once corrected, the first comment in the recipe will be displayed under the Comment column, and the file status will be displayed as “UserCreated”, as shown below.

Recipe	Sub Folder	Cycles	Read End	Multiplexing	Service	Version	For SBS Kit Version	Comment	File Status
GA2_126Cycle_SR_Transcrip...	Custom	160	Single Read	Yes	No	8.3	5	Created by JM...	UserCreated
GA2_126Cycle_SR_Transcrip...	Custom	146	Single Read	Yes	No	8.3	5	Created by JM...	UserCreated
GA2_36Cycle_SR_Transcripti...	Custom	36	Single Read	Yes	No	8.3	5	Created by JM...	UserCreated
GA2_36Cycle_SR_v8.3_HITS...	Custom	36	Single Read	No	No	8.3	5	Created by JM...	UserCreated
GA2_72Cycle_SR_Transcripti...	Custom	76	Single Read	Yes	No	8.3	5	Created by JM...	UserCreated
GA2_82Cycle_SR_Transcripti...	Custom	135	Single Read	Yes	No	8.3	5	Created by JM...	UserCreated
GA2_82Cycle_SR_Transcripti...	Custom	103	Single Read	Yes	No	8.3	5	Created by JM...	UserCreated
GAllx_rehybridization.xml	Custom	0	NA	No	No	Unknown	any		UserCreated
JMT_GA2-PEM_Maintenance...	Custom	0	NA	No	Yes	8	any	Updated 1308...	UserCreated
JMT_GA2-PEM_PreWash_v8...	Custom	0	NA	No	Yes	8	any	GA2-PEM_Pre...	UserCreated
Recipe_GA2_126Cycle_SR_...	Custom	133	Single Read	Yes	No	8.3	5	Created by JM...	UserCreated
Transcription_1.0.xml	Custom	0	NA	No	No	Unknown	any		UserCreated
Transcription_2.0.xml	Custom	0	NA	No	No	Unknown	any		UserCreated
Transcription_2.3.xml	Custom	0	NA	No	No	Unknown	any		UserCreated
Transcription_3.0.xml	Custom	0	NA	No	No	Unknown	any		UserCreated
Transcription_HITSFLIP.xml	Custom	0	NA		No	Unknown	any	Invalid Recipe...	Invalid

e. If desired the steps of the recipe can be visualized and inspected in the SCS2.9 program after opening it, displayed on the left panel as shown below. Hovering the mouse above a step will display all of its elements (volume, flowrate, duration, etc.).



Cluster generation and sequencing.

Upwards of 30 million clusters can be sequenced per lane of Illumina GAIIX flowcell; however, in HiTS-RAP assays we performed, we aimed at ~20 million clusters per lane to get good separation between neighboring clusters. This generally gives better quality sequencing and protein binding data. Proper cluster generation is extremely important for extracting the maximum possible amount of data from a HiTS-RAP run. Both under-clustering and over-clustering will result in reduced yields, and can ruin runs almost entirely. Unfortunately, this takes experience. Thus, we recommend erring on the side of under clustering the first time a run is carried out: we started by loading 10 pM library in our earliest runs. Every library is different. We have found that our libraries actually produce the optimal cluster density (~20 million per lane), when we load 200 pM denatured library, or about ten times the maximum concentration recommended by Illumina.

Through our experience, we have learned to accurately and reproducibly cluster flowcells using Qubit measurements. As each run is a fairly large financial investment, we recommend that new users carefully consider loading the optimal library concentration during clustering. Quantitative PCR is the most precise way to measure library concentration, as it measures not just DNA concentration, but cluster forming units, when primers identical to those on the flowcell are used. Therefore, it likely provides the most reliable measurement of library concentration for novice users.

An Illumina GAIIX flowcell has 8 lanes, each of which can be used for a different DNA template. We recommend that one of these lanes to be used for PhiX control DNA to ensure high quality sequencing in all lanes. This should be specified as the control lane when starting the Illumina SCS before a run. Inclusion of a control lane is especially important if the first few

bases read in a library are not very diverse, as this will not allow the analysis software to generate a suitable basecalling matrix. If possible and available, one of the lanes should be used for a template that has no affinity for the protein of interest (negative control), and one lane for a well-established target protein binding RNA (positive control). These can be mixed in a single lane as well and distinguished based on the sequence of the template or the barcode. Although all 8 lanes of the flowcell can be addressed individually on the cBot instrument during cluster generation, Illumina GAIIX instrument processes all of them simultaneously with the same solutions at any given time. Recently, modifications to the GAIIX instrument that enable individual processing of 8 lanes has been published(Gravina, Lin and Levine, 2013). This could be implemented to perform up to 8 HiTS-RAP assays with a single flowcell.

We perform sequencing and protein binding as a part of the same run. This means that we write a single .xml recipe program to operate the sequencer through the entire protocol. Illumina's standard software can then be use to image during protein binding, and analyze these images to get protein binding intensities. We feel that this is the simplest way of carrying out HiTS-RAP. It allows even an inexperienced user to operate the instrument, with no modifications. More complicated protocols, with more elegant ways of extracting data can be used, as in RNA-MaP, but would require much more sophisticated software for analyzing data, a much more intimate knowledge of the instrument, and engineering expertise.

dsDNA regeneration and transcription halting.

Regeneration of a clean full-length dsDNA template following sequencing is necessary for three main reasons; (i) to generate a double stranded Ter site for Tus binding, (ii) to generate a double stranded T7 promoter for transcription by T7 RNA polymerase, and (iii) to generate a

template strand for transcription that is full-length and free of any artifacts introduced during sequencing.

A Tus-bound Ter site has been shown to halt progression of many DNA and RNA polymerases in one orientation but not in the other (Mohanty, Sahoo and Bastia, 1996; Mohanty, 1998). Also, it is well-documented in the literature that an RNA polymerase that has synthesized at least a 9 nt long transcript is extremely stable when stalled because of a physical barrier or absence/limiting levels of ribonucleotides (Core *et al.*, 2012). HiTS-RAP combines these two phenomenons to tether RNA transcripts to their DNA templates in each cluster, enabling observation of interactions with the target protein. We initially tested halting of T7 RNA polymerase using a streptavidin-biotin interaction on the template strand of DNA without success. We believe this is because Tus has a specific contra-helicase activity that impedes T7 RNAP's progress, rather than acting as a simple roadblock. However, *E.coli* RNA polymerase can be halted by streptavidin bound to a biotin moiety on template strand of DNA, under single-round transcription conditions, as demonstrated with RNA-MaP (Buenrostro *et al.*, 2014).

Protein Binding.

To get an accurate measure of binding affinity (K_d), concentrations of the fluorescently-labeled target protein (in protein binding steps of HiTS-RAP) need to be centered around the expected K_d . Ideally, 5 or more concentrations below and 5 or more concentrations above the K_d (total 11 points), each separated in 5-fold increments, should be used. For proteins that are known or suspected to bind DNA, contribution of protein binding at each DNA cluster needs to be measured by performing protein binding at the same concentrations without transcription halting.

Our estimates of the halted transcription complex lifetime is ~ 72 hours, which is sufficient to perform 48 protein binding steps (Tome *et al.*, 2014). Based on this slow dissociation of the complex, a correction is applied to account for the loss in protein binding signal, and the corrected values are used for calculation of the binding affinities.

Data Analysis.

Analysis Workflow

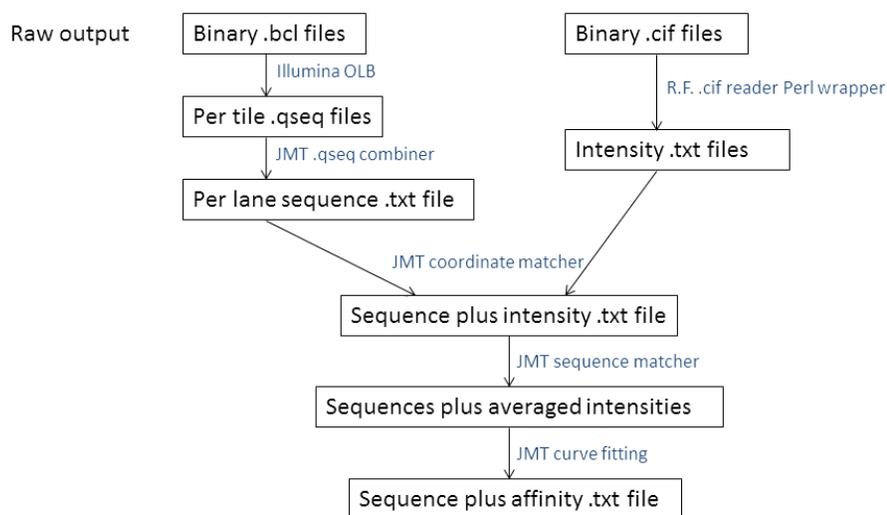


Figure 2. 14: HiTS-RAP data analysis workflow

Raw output from a sequencing and binding run must be processed through several steps to extract sequence and binding information (separately), match them, and then fit K_d s.

After a HiTS-RAP run, the run folder contains a number of files created by Illumina's Real-Time Analysis (RTA) software (Figure 2. 14). Most notably for the purpose of HiTS-RAP, the RTA analyzes the raw .tif images collected during the run to measure the fluorescence intensity of each cluster. These measurements are stored in binary .cif files. RTA uses this intensity information from all four channels to determine the base added at each cluster, at each cycle, by comparing to a basecalling matrix generated in the first cycles. This dictates which

intensity measurements across the four channels correspond to which base. While the intensity measurements recorded within .cif files are generated for the purpose of comparing relative signals to determine the most suitable basecall, they are also sufficient for comparing intensities within a single channel across cycles to make affinity measurements for HiTS-RAP. In this way, the RTA is carrying out the laborious task of identifying clusters and measuring fluorescence intensities from raw images that is necessary for making binding measurements. We use a set of scripts written by Robin Friedman to extract intensity data from .cif files into more accessible text files. These text files contain the intensity measurements for every cluster in a lane, for the specified channel and cycles of the run. Each cluster is identified by its tile number, and x and y coordinates within that tile. We use the channel that is used for T during sequencing, as this most closely fits mOrange fluorescence with the least amount of noise.

Basecalling is carried out by the Illumina RTA during the run. Thus, after a run is finished, the run folder contains this information stored in binary .bcl files. Each individual .bcl file contains basecalls for clusters within a single tile, at a single cycle. To get the full sequence of each read, the information in these files must be assembled through all cycles. This is done by the .bcl to .qseq function of Illumina's OffLine Basecaller (OLB) software, which uses .bcl, .stats, .filter, .control, .pos.txt, and the config.xml file created during the run to make .qseq.txt files. We carried this out using the setupBclToQseq.py script on a machine running CentOS Linux, with OLB version 1.9.4 installed. Each .qseq.txt file contains all of the sequence information, divided by tile. As in the intensity file created by reading .cif files, clusters are identified by lane, tile number, and x and y coordinates within the tile. In addition to sequence, the entry for each cluster contains a quality string, and a binary pass filter metric (last entry,

index 10), which are important for determining which clusters to include in downstream analyses.

The first step in analyzing the raw data is to match clusters sequences to their corresponding intensities. We do this using the coordinates of clusters contained in both files. Because this information is recorded in .bcl and .cif files differently, the x and y coordinate numbers in intensity files must be converted to match the sequence file coordinates by multiplying by 10, adding 1000, and rounding to an integer. During this matching stage, we only include clusters for which we have both sequence and intensities, and which pass Illumina's quality filter as recorded in the last field of every entry in .qseq.txt files. At this stage, the quality string can also be used to implement a more stringent quality filter. For example, when we were performing HiTS-RAP against a library of mutagenized versions of a single sequence, we controlled for sequencing errors resulting in unmutated clusters being called as mutants by requiring that any bases different than the original sequence have a Phred quality score of at least 25. Next, we prepare for determining K_d s by averaging all clusters corresponding to each sequence to generate a single binding curve that is representative of that sequence. We then do a nonlinear, least squares fit to the Hill equation to determine the K_d . We have found that this works best when the four parameters for the K_d , hill coefficient, base intensity, and maximum intensity are found to vary. We use the covariance matrix returned by SciPy's fitting algorithm to identify good fits; a covariance below ~ 1000000 for K_d generally indicates a good measurement. It is also helpful to look at the other parameters as well when assessing an individual K_d measurement. Our Hill coefficients are generally less than one, indicating negative cooperativity in binding, which we interpret as steric hindrance as saturation of the cluster is reached. In addition, when the fitted base intensity is close to the first few measured intensities,

and the fitted maximum is close to the measured values of the last intensities, this is indicative of a good fit. Fitted K_d s that are either above or below the range of concentrations assayed generally fail the covariance filter. However, if they do not, they are not reliable. We expect the dynamic range of the assay to be only within the concentrations of target protein imaged to generate binding curves.

Alternatively, we have considered fitting binding curves from individual clusters and averaging the results to give a representative K_d , rather than averaging intensities and doing a single fit per sequence. These two approaches give similar results, so we recommend the simpler procedure of averaging intensities.

Anticipated Results

The number of K_d measurements that can be made in a single HiTS-RAP experiment depends entirely on the library and the application. We have solved as many as 10,000 dissociation constants from a single lane in experiments with SELEX libraries and mutagenized aptamers. In our experience, HiTS-RAP gives reliable results for sequences represented by as few as 5 clusters in a lane, so with an ideal library (20 million reads per lane) up to ~4 million measurements per lane should be possible. Analyzing targeted libraries with long sequences approaching this maximum might be difficult. However, for other applications, such as k-mer analysis with a longer random library, the number of motifs that can be assessed with a single lane could exceed the number of clusters.

The first step in assessing the success of a HiTS-RAP run is to make sure that the sequencing portion of the run was successful. This can be determined from the run metrics in the Sequencing Analysis Viewer (SAV) software as the run progresses (though some metrics are not

available until cycle 25). Favorable run metrics indicate that the instrument is functioning properly. We generally try to reach a cluster density of 200,000-400,000 clusters/mm². Underclustering means that the potential of the assay is underutilized, and overclustering results in lower quality data, and lower pass filter rates. The percent Q30 rate serves as a measure of overall confidence in basecalling in a lane. In successful runs, at least 85 to 90% of reads are above Q30. A lower Q30 rate indicates that the instrument is not functioning properly: we have observed lower rates when lanes are overclustered and when the optics of the instrument were not functioning properly. The SAV will also report error rates for lanes containing the PhiX genome control. Error rates are typically low (below 1%). It is important to only check the Q30 and error rate values for sequencing cycles, and to keep in mind that they begin to drop precipitously with longer reads (past ~75 cycles). Once the protein binding portion of the HiTS-RAP starts, the aggregate values for these metrics will decrease drastically, as the SCS includes the measurements from these cycles in calculation of the run metrics.

Assessing HiTS-RAP results will depend on the application. When we have applied it to aptamer libraries, we expect most sequences assessed to bind the target protein, while in other applications binding could be a rare event. We first look at the intensity measurements reported in the Illumina SAV during the protein binding to assess whether protein binding is taking place. With libraries in which the majority of sequences bind the protein, the binding curve can be seen in the T and G channels. If a negative control (or PhiX) lane was included, lack of signal in that lane indicates low background and low nonspecific binding of the labeled protein to DNA and other components of a halted transcription complex (i.e. T7 RNA polymerase and Tus protein).

The results of K_d curve fitting, as described in Data Retrieval and Analysis section, will ultimately determine whether the run was successful or not. Low covariances of fitted parameters, Hill coefficients that are close to one, base and maximum values close to the first and last protein binding intensity measurements, and K_d s within the range of protein concentrations assayed are all indicators of good fits and thus a successful HiTS-RAP run.

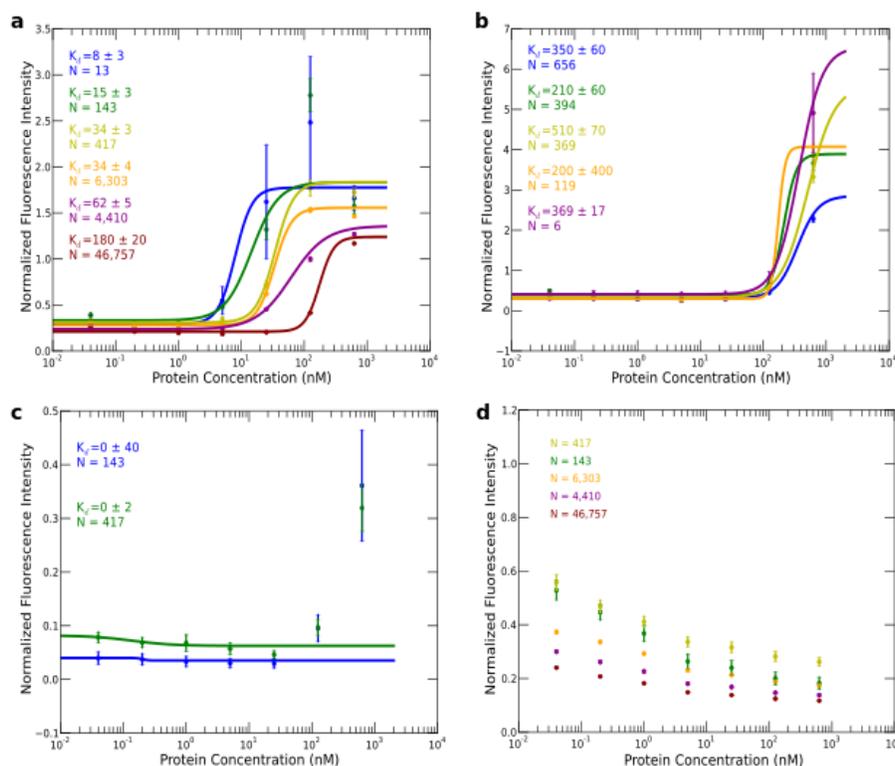


Figure 2. 15: Example binding curves.

(a) A set of binding curves from a successful experiment. These RNAs represent a range of K_d s and copy numbers. In all panels, intensities are the average of all clusters with a given sequence in one lane, normalized by dividing by average sequencing intensity. Error bars, s.e.m. Error of the K_d is the square root of the variance returned by the analysis pipeline. (b) Binding curves which only have one point above background. K_d s are able to be fitted, but are likely unreliable. (c) Binding curves with bad fits. In both of these cases, the fitted base is above the maximum, so these measurements should be disregarded. (d) Binding curves from an experiment where no binding is observed. Here, intensities decrease throughout the run. These data are from an experiment where several biochemical steps leading up to the binding measurements did not work properly.

a, c, and d are from a HiTS-RAP done to characterize HSF1 SELEX libraries. The same flowcell was reused for several binding measurements after a single run. b is from a separate run to characterize BEAF-32 SELEX libraries.

The protein binding curves used to determine K_{ds} should also be inspected to determine whether a run has worked well and the results obtained from data analysis are consistent. In the lowest concentrations of protein, fluorescence intensities of protein binding should be near background, similar to that of no protein control. As the protein concentration increases, intensities should also increase. We have found that the highest intensity measurements are generally noisier than lower intensity measurements. Transition to protein binding is often easily identifiable in raw HiTS-RAP data, but the measurements after that point tend to be noisy. In successful runs, we are able to solve reliable K_{ds} through a range of affinities. Figure 2. 15a shows a series from one such run, in which different sequences, all from the same lane, show low fluorescence in the first few protein concentrations, and then saturate binding at different points. Seeing a range of affinities in one lane is the ultimate indicator of a successful HiTS-RAP run, and the binding measurements observed are due to real binding, rather than a systematic bias.

Binding curves that do not match the calculated affinities are indicators of a failed run. There could be many reasons for a measurement to be unreliable. Sometimes, the fitting algorithm simply fails to find parameters that are a good match for the data (Figure 2. 15b): such a fit will usually have a high covariance value. Other times, only one binding measurement rises above background (Figure 2. 15c). This occurs in cases where the range of protein concentrations assayed were not within the range of the affinities of the RNAs probed. We have also observed instances where the intensity data show no binding at all (Figure 2. 15d). These data are from a run where several different protein binding series were carried out on the same flowcell (sequenced once) after denaturing and remaking double stranded DNA, transcribing and halting, and introducing each labeled protein successively. In that run, we suspect that at least one of the

biochemical steps leading up to the halted RNA complex was unsuccessful in the later protein binding series.

Chapter 3: In Vivo Disruption of NELF-E's Interaction with Nascent RNA by NELFapt Expression

Introduction

The NELF complex is critical for establishing a promoter proximally paused polymerase in *Drosophila* (Wu *et al.*, 2005; Fuda, Ardehali and Lis, 2009). This four subunit complex is recruited to polymerase by various contacts, including DSIF, Pol II, and the nascent RNA (Missra and Gilmour, 2010). NELF binds RNA with a higher level of specificity than other pausing factors (Missra and Gilmour, 2010), primarily through the RRM of NELF-E (Rao *et al.*, 2006, 2008), which has a high potential for specific interactions. Our group has previously characterized this specificity of RNA binding by NELF-E by identifying a preferred sequence motif by SELEX (Pagano *et al.*, 2014), and the importance of 3D RNA structure for proper recognition of that motif (Pagano *et al.*, 2014; Tome *et al.*, 2014). NELF may have many contacts with the nascent RNA, as some support exists for each subunit to have some capacity for RNA binding. In a thorough characterization of the human NELF complex's architecture (Vos *et al.*, 2016), the Cramer lab found that the NELF A/C subunits form the core of the NELF complex, and together form a positively charged pocket that binds RNA primarily through NELF-C. The NELF-B subunit also binds RNA, and serves to tether NELF-E to the A/C core. Therefore, interactions with the nascent RNA likely constitute an important facet of NELF's association with and influence upon paused Pol II.

Previous efforts to characterize the role of NELF in pausing genome-wide have utilized RNAi to deplete NELF (Gilchrist *et al.*, 2008, 2010; Core *et al.*, 2012). These perturbations result in the loss of multiple subunits of the NELF complex, and are typically carried out over the

several days(Zhou *et al.*, 2013), allowing secondary effects to accumulate. These and other studies(Missra and Gilmour, 2010; Li *et al.*, 2013) implicate NELF in the fine-tuning of pausing, as some pausing is still observed when NELF is perturbed. RNA aptamers are an attractive alternative to degradation by RNAi, as they can be expressed inducibly(Shi, Hoffman and Lis, 1999; Salamanca *et al.*, 2011), and selected to specifically target single domains of individual factors(Ozer, Pagano and Lis, 2014), rather than depleting the entire protein. Our group has previously used RNA aptamers as inhibitors both *in vitro*(Shi, Hoffman and Lis, 1999; Zhao *et al.*, 2006; Sevilimedu, Shi and Lis, 2008) and *in vivo*(Shi, Hoffman and Lis, 1999; Salamanca *et al.*, 2011, 2014). Therefore, expressing the NELF aptamer (NELFapt), an RNA which binds the RRM of NELF-E with high affinity and specificity(Pagano *et al.*, 2014; Tome *et al.*, 2014), is an attractive alternative to RNAi for understanding the role that NELF plays in pausing. Because NELFapt binds only the RRM of NELF-E, the aptamer should be able to specifically prevent NELF from interacting with nascent RNA with NELF-E's RRM. Thus, it allows us to ask what the role of this interaction is in establishing the level of paused polymerase near promoters: NELFapt inhibition should not affect the stability of the entire NELF complex, or any other of the many interactions by which NELF is recruited to polymerase. PRO-seq(Kwak *et al.*, 2013) was used as a highly specific readout for the highly specific perturbation of NELF-E by NELFapt. PRO-seq maps transcriptionally engaged RNA polymerase, genome-wide, by sequencing the RNA after a single nucleotide biotin run-on. Therefore, it allows assessment of both the precise position and level of promoter proximally paused polymerase.

In this chapter, I describe the strategy that we used to express NELFapt as an inhibitor of NELF's ability to use the interaction between NELF-E and nascent RNA as it plays its role in promoter proximal pausing. NELFapt binds NELF-E well in the context of the transgene, and

was expressed to a very high level in the nucleus. I carried out two sets of PRO-seq experiments to assess the effect of this inhibition. Altogether, these experiments show that the aptamer did affect promoter proximal pausing, and that this effect is mitigated at promoters with GAF. Though some common biology seems to be revealed in my two PRO-seq experiments, a few issues limit my ability to draw many concrete conclusions: some of the effects of the aptamer were not reproduced in the two experiments, and a control scrambled RNA showed the same effect as NELFapt.

Results

Overview of NELFapt Inhibition Strategy

To understand the role of interactions between NELF and nascent RNA, I expressed the NELF-E aptamer in *Drosophila* S2 cells to specifically occlude its RRM while leaving the rest of the four subunit complex intact. I used PRO-seq (Kwak *et al.*, 2013) as the readout, to measure the active site of engaged polymerase genome-wide. This enables interrogation of the exact site used for promoter proximal pausing, and precise measurement of the level of pausing, genome-wide. In total, I carried out two rounds of PRO-seq experiments (Figure 3. 1); the first with just stably selected aptamer transfected cells and untransfected cells as a control (before and after aptamer induction), and the second with the aptamer, a scrambled aptamer control, and an empty vector control (all before and after induction).

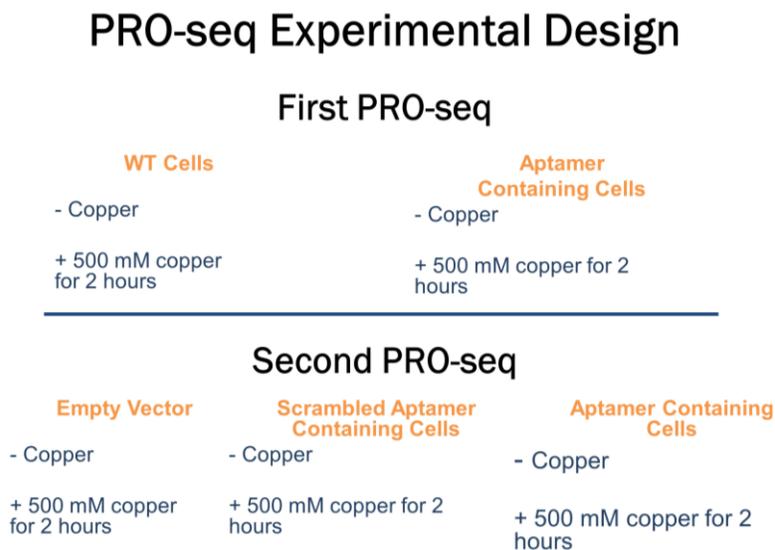


Figure 3. 1: NELFapt PRO-seq Experimental Design

Two rounds of PRO-seq were carried out. All cell lines in the second PRO-seq were new stable transfected lines (different NELFapt lines were used). PRO-seq was carried out for each cell line (orange) with both copper treated and no copper treatments (blue), with two replicates each.

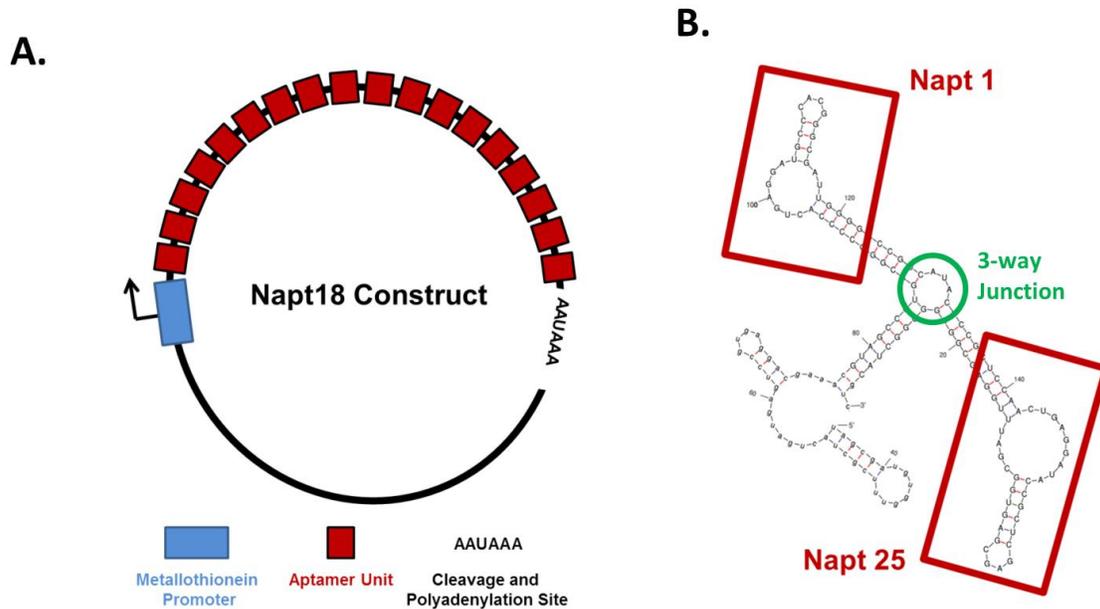


Figure 3. 2: NELFapt in vivo expression strategy

A.) pRMHa3 expression vector. 18 NELFapt units are arranged in tandem in the expression vector. This places them under the control of the copper inducible metallothionein promoter, and with the 3' UTR from the *Adhr* gene to facilitate 3' end formation (cleavage and polyadenylation, termination). **B.)** Each aptamer unit is composed of two different NELF aptamers(Pagano *et al.*, 2014) with similar features; both have an NBE within a putative k-turn. The hammerhead ribozyme forms the third arm: it self-cleaves so that multiple units can be arranged in tandem to amplify expression and ensure that most units are neither capped nor polyadenylated. Cleaved units have some tendency to circularize(Canny, Jucker and Pardi, 2007), which could further enhance stability. These three arms are separated by a separate 3-way junction.

The strategy used for expressing the aptamer borrows from previous *in vivo* aptamer inhibitions and inducible transgene expressions done in the lab(Shi, Hoffman and Lis, 1999; Lis *et al.*, 2000; Salamanca *et al.*, 2011, 2014). The NELFapt vector, originally made by John Pagano in the lab, contains 18 total aptamer units under the control of the *Drosophila* metallothionein promoter. The expression vector used, pRMHA3(Bunch, Grinblat and Goldstein, 1988), contains the metallothionein promoter, a multiple cloning site, and 3' UTR from the *Adhr* gene (Figure 3. 2A). The metallothionein promoter is induced up to 100 fold upon addition of CuSO₄ to the media, and cleavage and polyadenylation of the transgene is carried out at the end of the *Adhr* 3' UTR. Each unit of the 18-mer (Figure 3. 2B) contains two

NELF aptamer arms(Pagano *et al.*, 2014), and a hammerhead ribozyme arm(McCall, Hendry and Jennings, 1992; Scott, Finch and Klug, 1995), separated by a central three-way junction(Germer, Leonard and Zhang, 2013). Two aptamer arms were included to ensure that the resulting unit has extremely high affinity for NELF-E. Each NELFapt arm had its GC rich stem extended slightly to ensure that the units are well separated after folding. This, coupled with the highly stable 3-way junction, means that the aptamer RNA folds into an exceptionally tight structure. The hammerhead ribozyme efficiently self-cleaves as the 18-mer is transcribed, so that transcription once through the transgene produces 18 separate NELFapt units. Such cleaved units are retained in the nucleus and resistant to degradation, as they have neither a 5' cap nor a poly(A) tail, as shown in our lab previously(Shi, Hoffman and Lis, 1999). I chemically transfected *Drosophila* S2 cells with this transgene, with a blasticidin resistance marker cotransfected to enable selection of cells that took up the vector. After selecting stably transfected lines, this should result in cells with the transgene stably inserted in random locations in the cells' genome.

For my second round of PRO-seq I made a scramble aptamer control and empty vector transfection control (so that the cells underwent the same selection and were all grown with blasticidin in the media). The scrambled aptamer control was similar to the NELFapt construct, but with the aptamer arms of each unit scrambled (with the same base composition). To facilitate gene synthesis, I used 9 different scrambled versions, and assembled two scrambled 9-mers to make an 18-mer. Because each arm of each unit is different, this means that the control contains 18 different scrambled aptamers. I checked that none of them had matches to the NBE by FIMO(Grant, Bailey and Noble, 2011), and manually inspected secondary structure

predictions for the presence of NELFapt like stem-loops(Zuker, 2003). Each scrambled unit leaves the hammerhead ribozyme and 3-way junction intact.

Characterization of the NELFapt Construct

To assess the NELFapt and Scramble's ability to bind NELF and act as an inhibitor, I carried out EMSA with NELF-E using RNA transcribed from a linearized expression vector. In vitro, cleavage by the ribozyme is highly efficient, as the T7 transcription reaction from the 3000 bp vector produced mostly 160 nt monomer RNA (Figure 3. 3A). NELF-E has high background affinity for RNA as a nucleic acid interacting factor: some binding is observed no matter what the sequence or structure of the RNA. The N70 library of random 70-mers served as a control for non-specific binding. The three RNAs tested all bind at some level to NELF-E. The RNAs' mobility changes in steps as it binds RNA (Figure 3. 3B, top, right of gel). First, free, unbound RNA moves through the gel with high mobility. Next, RNA bound by a single NELF-E molecule moves with intermediate mobility. Two intermediate mobility bands are seen in the NELFapt EMSA (Figure 3. 3B, top), likely representing RNA monomers bound by one and two NELF-E molecules (as the RNA contains two aptamer units). This behavior is unique to NELFapt, as NELF-E binds the aptamer even at very low concentrations, before the formation of aggregates. Finally, aggregates fail to enter the gel and are seen in the well. Aggregation is more of an issue in these vertical gels than in slab gels in (Figure 2. 11): it has often been my experience that EMSA is cleaner in a slab gel, though slab gels are more cumbersome to dry when using radioactivity as readout. Thus, there is a trade-off between the sensitivity and wide dynamic range of radioactivity, and the ease of handling and cleaner behavior of fluorescent labeled EMSA in slab gels.

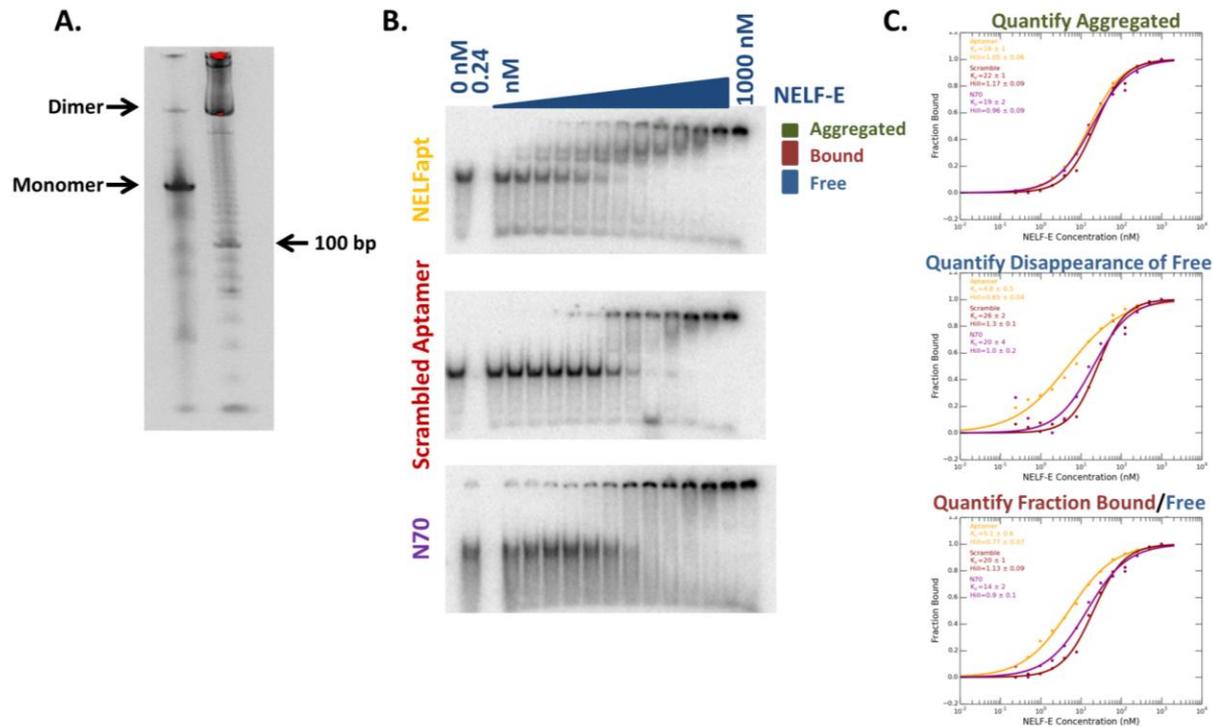


Figure 3.3: In vitro characterization of the NELFapt transgene

A.) *In vitro* transcription of the transgene. A version of the construct with a T7 promoter between the *MtnA* promoter and first NELFapt repeat was linearized by cutting with a restriction enzyme after the last NELFapt repeat, and then used as the template for *in vitro* transcription. The transcription reaction was treated with DNase, run on 7 M Urea 8% PAGE with a 10 bp DNA ladder as a reference, and stained with ethidium bromide. Most of the RNA has been self-cleaved to a 159 nt monomer. **B.)** EMSA for the NELFapt, and Scrambled transgene RNAs, with N70 random 70-mer as a non-specific control. All are P³² labeled RNA transcribed from linearized T7 promoter containing vector, in 4% vertical PAGE. No protein control is the furthest left lane, then NELF-E increases in 2-fold increments from 0.24 to 1000 nM. Binding modes are indicated to the left of the NELFapt (Aptamer) gel. Aggregates remain stuck in the well. Bound complexes (NELFapt + NELF-E) enter the gel, and runs with low mobility. Two binding steps are seen here, corresponding to the two NELFapt arms of the construct. Unbound RNA runs with high mobility. **C.)** Quantifications of binding from B, with the three RNAs color coded. K_d of the three RNAs is indicated on the plot. At top, fraction bound is calculated as a fraction of the maximum amount of signal seen in the well. At middle, fraction bound is calculated by setting 0 nM NELF-E as 0, and the minimum signal in the unbound region as 1. At bottom, fraction bound is calculated by setting 0 nM NELF-E as zero, and 1 as maximum signal in the bound+aggregated regions.

To quantify binding, signal in three different regions of the gel was used. First, binding is assessed by just appearance of aggregates of RNA and NELF-E. All three RNAs aggregated at similar NELF-E concentrations (Figure 3.3C, bottom), indicating that this behavior is not

specific in this assay. Next, disappearance of free RNA alone was used (Figure 3. 3C, middle). As the amount of RNA used is constant and protein amount varies across the EMSA, the total amount of signal should remain the same. So, to use disappearance of free, I simply use the amount of signal in the unbound region in the zero protein well as 100%, and quantify the amount in the free region of the other wells as a fraction of that. Using this method, a K_d of 5 nM is measured for NELFapt, in good agreement with HiTS-RAP and previous EMSA experiments, while the scrambled aptamer and N70 library do not bind with affinity higher than the nonspecific aggregation measured with the first quantification scheme. Finally, binding is assessed as fraction of RNA bound (intermediate mobility) vs fraction free (Figure 3. 3C, bottom). The bound fraction is specific to the NELFapt construct, as binding is observed almost from the beginning of the dilution curve. A very small amount of bound RNA is observed for the scrambled RNA, but not the N70 random RNA control (Figure 3. 3B, middle and bottom, respectively), indicating that one of the scrambled arms may exhibit some specific binding with NELF-E (this may explain similar effects seen after expressing the scrambled and aptamer, as discussed later in this chapter). This small amount of bound RNA was not enough to significantly affect K_d measurements for the scrambled aptamer mixture, so again this method gave a K_d of 5 nM for NELFapt but no measurable affinity above nonspecific aggregation for the others. Altogether, these results show that incorporation of NELFapt into the construct, with the 3-way junction and hammerhead ribozyme did not adversely affect its affinity for NELF-E. The high affinity of the NELF aptamer construct indicates that it could function as a promising inhibitor of NELF-E's interactions with endogenous RNA, as those are unlikely to contain the distinct tertiary structure and RNA motif present in the aptamer.

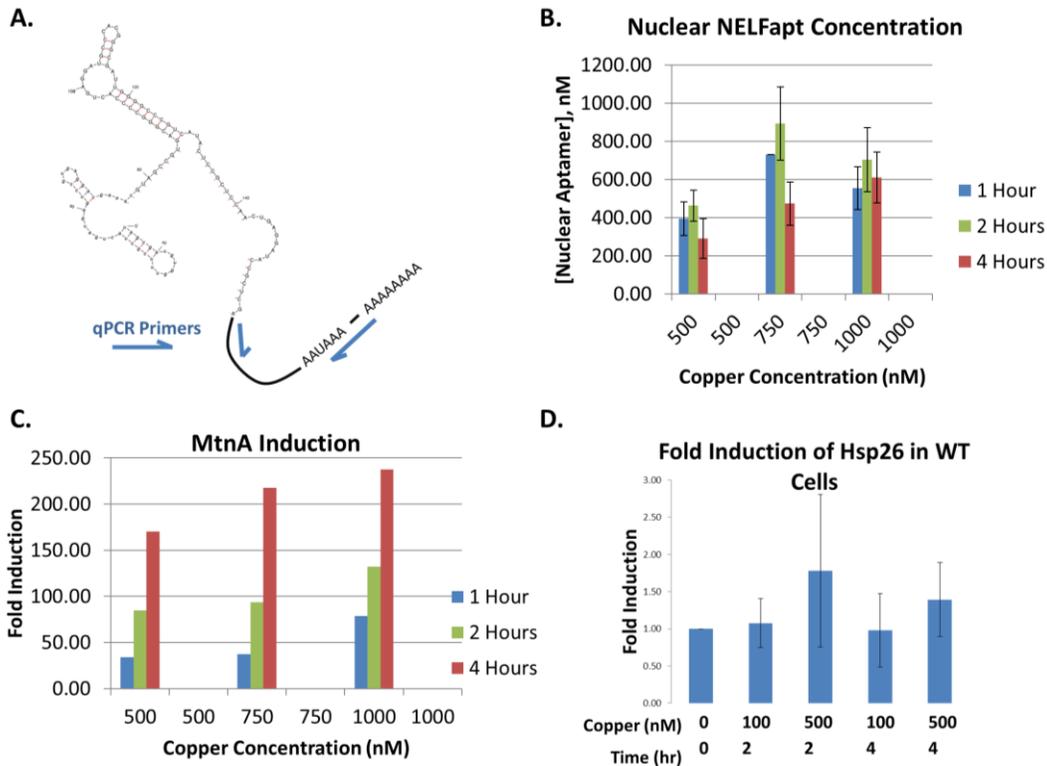


Figure 3. 4: Quantification of *in vivo* NELFapt expression by qPCR

A.) Strategy used for qPCR. NELFapt was quantified using RT and qPCR primers at the 3' end of the last unit (*Adhr* 3' UTR). Because the endogenous *Adhr* gene is not induced by copper, new expression beyond the 20 nM basal level is likely all due to aptamer induction. **B.)** Nuclear NELFapt concentration, nM, after induction with varying amount of CuSO₄ for 1, 2, and 4 hours. **C.)** Induction of the endogenous *MtnA* gene in the same time course as B. **D.)** *Hsp26* induction in a separate time course, with varying copper concentration, measured with RT-qPCR. Minimal stress response is observed up to 500 nM CuSO₄.

Quantifying aptamer expression after transfecting into *Drosophila* S2 cells proved to be difficult. Because of the extremely stable secondary structure of the 3-way junction and GC rich stems of NELFapt, reverse transcription of the cleaved units was extremely inefficient. To get accurate quantifications, I used two alternative strategies; Northern blot and using the 3' UTR of the transgene for qPCR. The last unit of the 18-mer transgene includes a cleaved hammerhead ribozyme, two aptamer units, the 3' UTR of *Adhr* from the pRMHa3 vector, and a poly(A) tail. Therefore, I could simply do reverse transcription and qPCR from the *Adhr* 3' UTR (Figure 3. 4A) to measure the last unit of the transgene, which constitutes 1/18th of the total aptamer within

cells. Even though this only detects one of the 18 units of the transgene, it proved to be much more sensitive than RT-qPCR for the core aptamer portion of the repeats. Through careful tracking of number of cells used per assay and calibration to an *in vitro* transcribed aptamer standard, I was able to approximate a nuclear aptamer concentration of 500 nM after 2 hours of induction with CuSO₄, and 20 nM before (Figure 3. 4B). This copper concentration and time point has been used with this vector previously (Shi, Hoffman and Lis, 1999; Salamanca *et al.*, 2011), and was chosen as it gives strong induction of both the transgene and the endogenous *MtnA* gene (Figure 3. 4C), with no detectible stress response (Figure 3. 4D), and in a brief enough induction that secondary effects should have minimal time to accumulate.

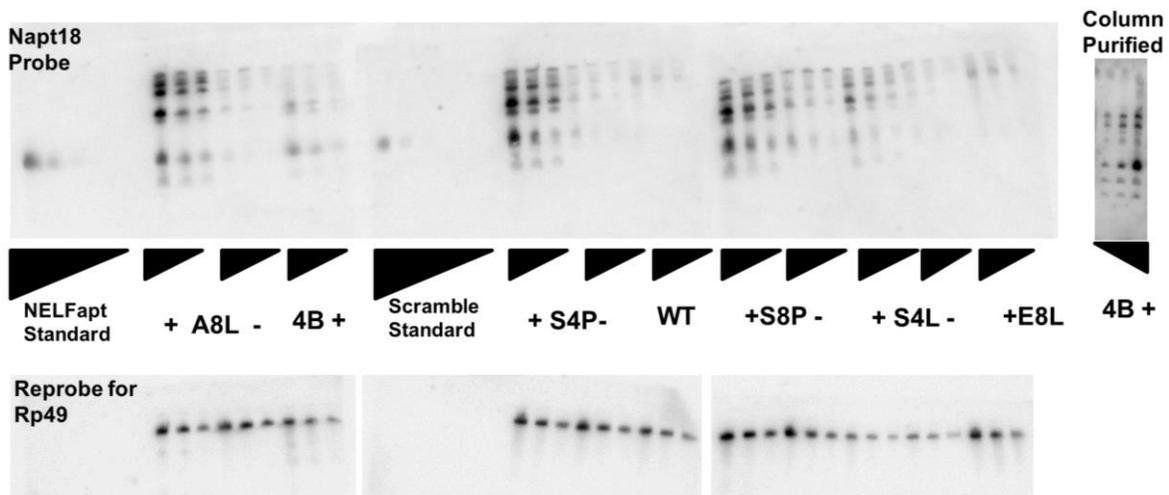


Figure 3. 5: Quantification of *in vivo* NELFapt expression by Northern Blot

Various aptamer containing cell lines are assayed for expression before and after. Cell lines are named as follows: First letter, vector transfected (A for NELFapt, S for scramble, E for empty vector). Number in middle, amount of vector transfected (100 ng) per 1 million cells. Last letter, P for circular plasmid transfected, L for linearized plasmid. + indicates copper induction at 500 nM for 2 hours, - is no induction. The probe used is a PCR amplified single aptamer monomer, end labeled with P³². Total cellular RNA was isolated with Trizol, run as a 2-fold dilution series on 7M Urea 8% PAGE, and transferred to Biodyne B nylon membrane. NELFapt (Napt18) and Scramble (Napt18 Scr) *in vitro* transcribed standards produced as in Figure 3. 3A. 4B is John Pagano's NELFapt line, used in the first PRO-seq. A8L is my NELFapt line used in the second. S8P is the scrambled RNA line used in the second PROseq. E8L, empty vector line used in second PRO-seq. Aptamer sized RNAs (aligned to standard) are unique to NELFapt and Scramble containing cell lines. The same amount of RNA (ng) was used for the top of a 2-fold dilution curve for each cell line, except S4L.

The membrane was stripped and reprobbed for Rp49 as a control, scan at bottom.

Northern blot also proved to be an efficient and sensitive way of assaying NELFapt expression *in vivo* despite the RNA's difficult secondary structure. I used radiolabeled, *in vitro* transcribed NELFapt construct as a probe. Therefore, this one probe worked for NELFapt and Scramble detection through common sequences like the hammerhead ribozyme, junction, and portion of methallothionein promoter in the first aptamer unit and the 3' UTR in the last. In addition to confirming the level of expression measured by RT-qPCR, Northern blots provide information about the size of the RNAs being detected. Thus, I see that in contrast with the *in vitro* transcribed multimer where most units were monomers, only ~60% of units have self-cleaved to monomers with the hammerhead ribozyme (Figure 3. 5), with a large fraction existing as larger units, presumably dimers and trimers (PAGE purified standard is monomer size). The fraction cleaved to monomer was much higher in an earlier experiment where I used a column purification kit with DNase digestion to isolate RNA rather than Trizol (Figure 3. 5, far right), likely because the DNase step afforded the ribozyme more time to cleave *in vitro*. Thus, the lower efficiency compared to the *in vitro* transcribed multimer is likely due to more cleavage during handling. Northern blots were also useful because, in my hands, their sensitivity was higher than qPCR. With Northern, I robustly detect basal aptamer expression in uninduced cells (Figure 3. 5, series labeled -). Because Northern also shows that the size of these RNAs is consistent with the cleaved aptamer, I can rule out the possibility that the basal expression measured from qPCR is from the endogenous *Adhr* 3' UTR. From these results, I picked cell lines for the second round of PRO-seq. Rather than using John Pagano's cell line from the first PRO-seq, I picked a new aptamer transfected line (Figure 3. 5 A8L line) that had been selected alongside the Scramble (Figure 3. 5 S8P line) and Empty vector (Figure 3. 5 E8L line) lines (and had thus been treated equally both in number of passages and growth conditions). In Figure 3. 5,

the same amount of total cellular RNA (in ng) was loaded in each well, except for the S4L scramble line which grew more slowly. Thus, aptamer expression in the 4B line and my A8L line can be compared: the A8L line expresses much more aptamer. The S8P and E8L lines were chosen as controls as they had been transfected with the same amount of vector per cell as the A8L line, and S8P showed similar levels of transgene expression as A8L.

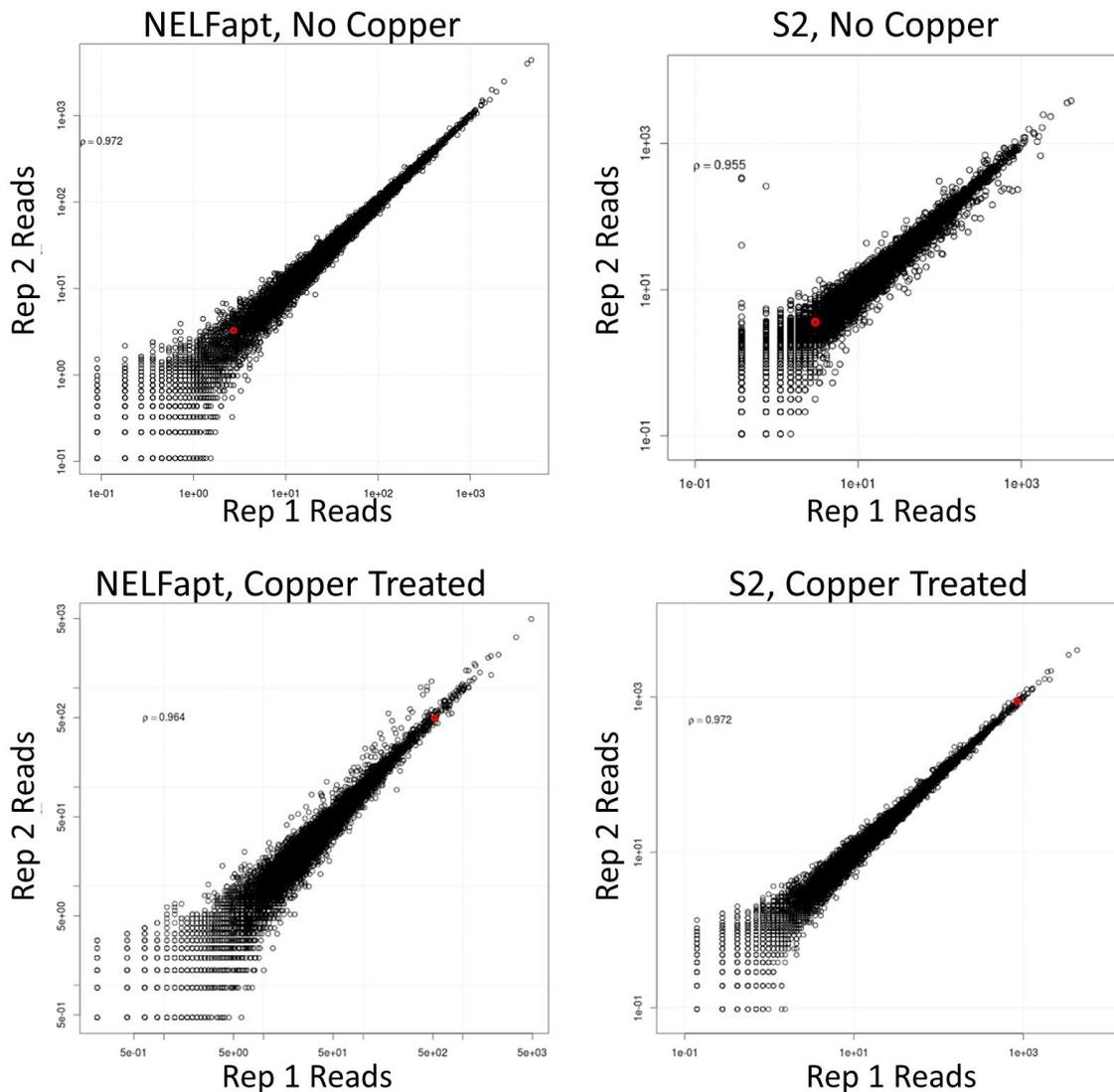


Figure 3. 6: Correlation between replicates in the first PRO-seq experiment

Gene body read number is highly correlated between replicates of the first PRO-seq experiment, in the four treatments used (Aptamer and S2 WT, before and after copper). Reads collected from +300 of the annotated start site to -300 of the annotated CPA site. Induction of the *MtnA* gene can be seen by highlighting it in red in all four plots: its expression is much higher after adding copper. Spearman rank correlation indicated on each plot. N = 9,452

NELFapt Inhibition in the First Experiment Caused a Downstream Shift in Pausing

I carried out two rounds of PRO-seq to measure the effect of occluding NELF-E's RRM with NELFapt on promoter proximal pausing. In both rounds, each cell line was assayed before and after induction of with CuSO₄, with replicates for each treatment. In both sets of experiments, replicates were highly correlated, both when considering signal across the entire gene body and when considering only signal in the pause region. The high level of induction from copper treatment can be observed at the endogenous *MtnA* gene, which is highly induced in all copper treated samples. This is seen both by highlighting this gene in the plot assessing correlation of replicates (Figure 3. 6) and in the genome browser (Figure 3. 7A). Furthermore, induction of the transgene can be seen by examining reads that map to the *Adhr* gene (Figure 3. 7B). There, signal from the 3' UTR of the transgene aligns to the endogenous gene, and is specific to the copper induced, transfected cells, as expected.

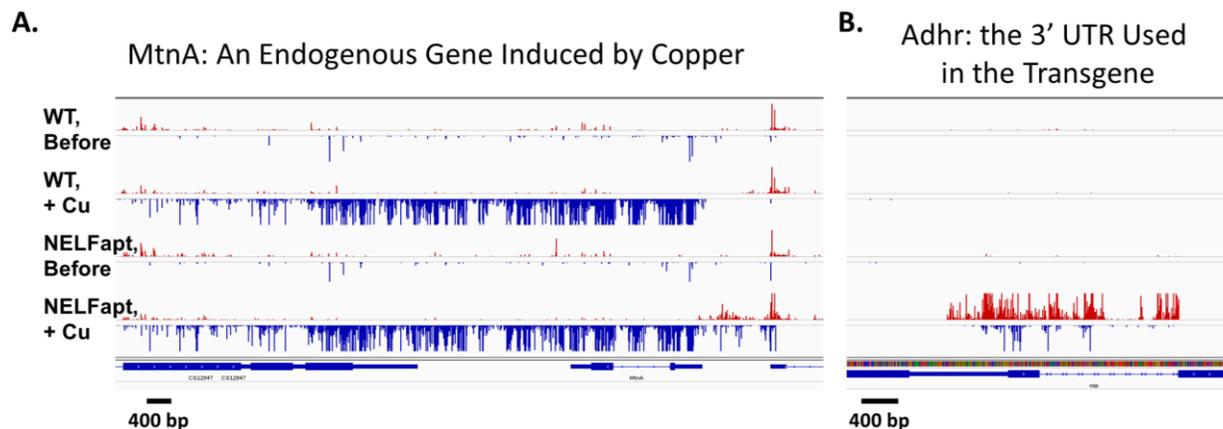


Figure 3. 7: Effects of copper treatment in S2 cells

A.) IGV browser shot of the metallothionein (*MtnA*) gene, whose promoter was used in the NELFapt transgene. In these PRO-seq data from the first set of experiments, strong induction of this gene is seen, in both WT and aptamer containing cells treated with copper. Nascent transcription beyond the CPA is apparent. **B.)** PRO-seq at the 3' UTR of the *Adhr* gene, which was used for its poly(A) site in the expression vector. Here, nascent transcripts from the NELFapt transgene map to the endogenous copy of *Adhr*. Thus, the strong induction at the NELFapt site can be seen here only in the NELFapt cells after copper treatment. All tracks are RPM normalized, and displayed at a fixed scale from -2.5 to 2.5.

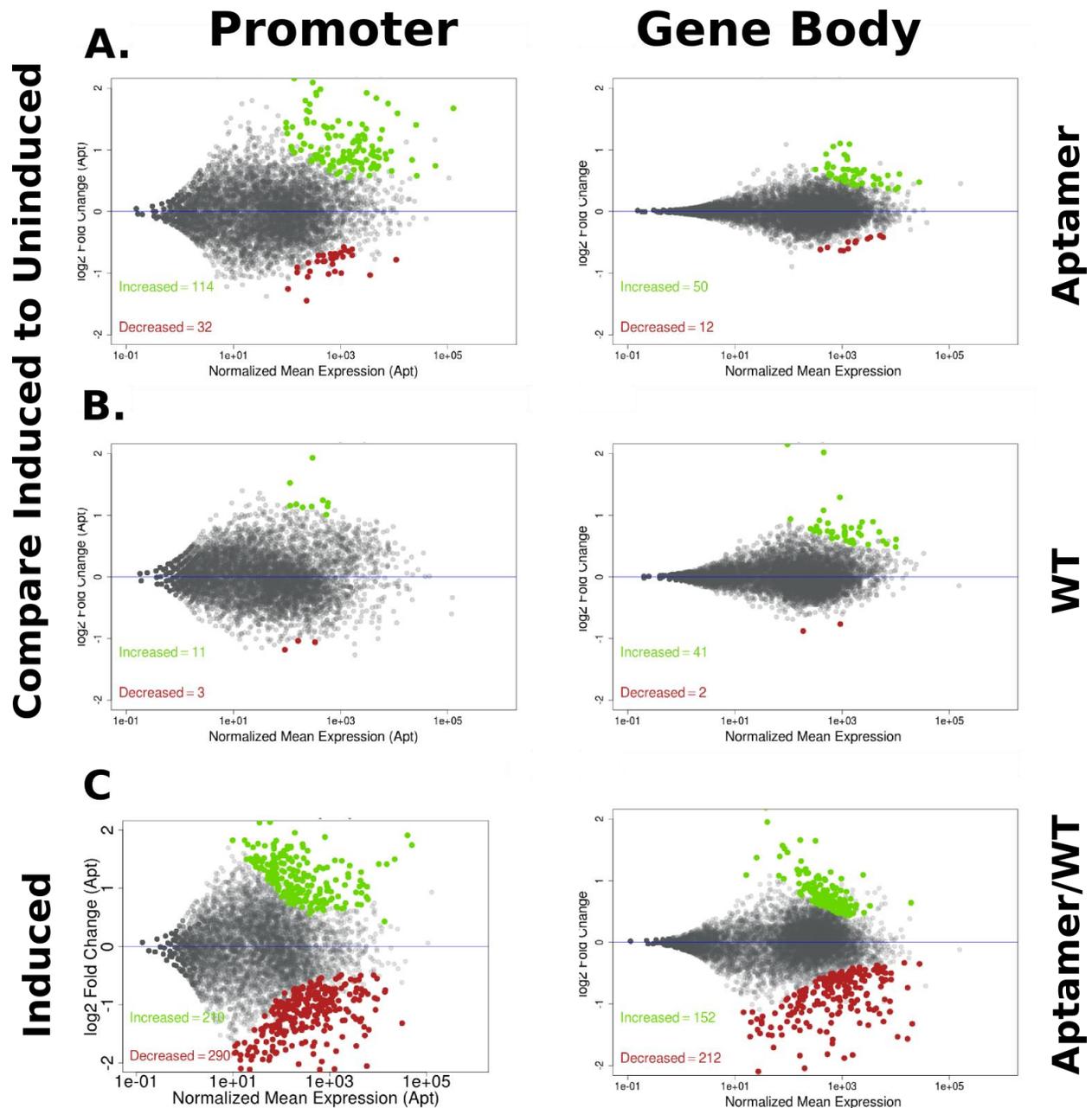


Figure 3. 8: Differences between WT and NELFapt are greater than either upon induction DESeq2 normalized MA plots comparing PRO-seq signal from the first round of experiments in various libraries (rows), in promoter proximal pause regions (TSS to +300, left column) and gene bodies (+200 from TSS to -200 from CPA, right column). Significantly changed genes ($p_{adj} < 0.01$) indicated in green (increased) and red (decreased). $N = 9,452$ **A.**) Comparison of NELFapt uninduced to copper treated cells. Promoter proximal pause regions are much more changed than gene bodies **B.**) Comparison of WT uninduced to copper treated cells. **C.**) Comparison of copper treated empty vector to NELFapt. More changes are seen here than in copper induction of NELFapt.

I first used DESeq2 to assess changes in transcription at both the promoter proximal pause region (Promoter, Figure 3. 8, first column) and the bodies of genes (Gene Body, Figure 3. 8, second column). DESeq2(Love, Anders and Huber, 2014) assesses significance of changes in sequencing density at genes by asking whether changes observed are greater than the variation between replicates. Critically, when no normalization is provided, it assumes that the bulk of expression is unchanged (thus it will miss shifts in the distribution of the data). In the first PRO-seq experiment, overall PRO-seq levels in the pause region (defined as a fixed window from annotated TSS to +300) become noisier in the aptamer expressing cells upon copper induction (Figure 3. 8A, left), with 146 promoters called as significantly changed. However, these changes are not much greater than the changes seen upon induction in the WT cells (Figure 3. 8B). The greatest change in pausing is seen when comparing aptamer transfected cells to WT cells, with 500 promoters called as significantly changed (Figure 3. 8C). Overall, most changes are in the pause region, with much less change in the aptamer expressing cells' gene bodies (Figure 3. 8, compare first and second column).

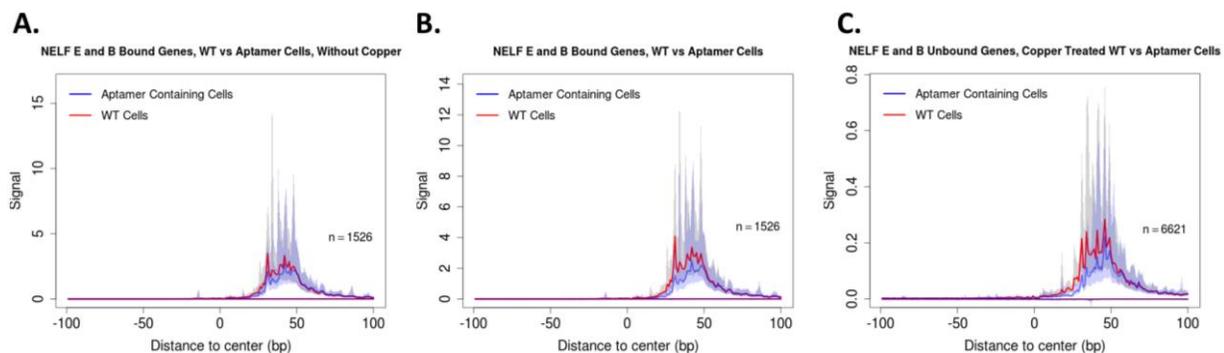


Figure 3. 9: Pausing at NELF Bound genes

A.) Metaplot of PRO-seq in the pause region of genes called as bound by both NELF-E and NELF-B by ChIP-chip(Lee *et al.*, 2008). Centered on annotated TSS. NELFapt cells in blue, WT in red. Both are for cells without copper induction. Bootstrapped estimates of median (solid lines) with 12.5 to 87.5% confidence intervals are shown for all metaplots in this chapter **B.)** Same promoter regions as A., after copper induction. **C.)** Metaplot of pause region of genes not called as binding NELF-B or NELF-E, in copper induced NELFapt and WT cells.

Analysis of the most changed promoters in aptamer expressing cells yielded few results: genes with increased pausing were enriched for GAF, but little else could be gleaned from the 500 changed genes. I also examined genes with features that could make them more dependent upon the interaction between NELF and nascent RNA for proper regulation of pausing. When I examined genes that were called as bound by NELF-B by ChIP-chip(Lee *et al.*, 2008), I see that pausing is slightly reduced there in aptamer containing cells when compared to WT, both before (Figure 3. 9A) and after (Figure 3. 9B) aptamer induction, however, this change is not greater than the global reduction in pausing as a similar change is seen at genes where NELF was not called as binding at the promoter (Figure 3. 9C). Therefore, this change is not specific for a greater reliance on NELF. Genes not called as having NELF binding have much less PRO-seq signal in the pause region (Figure 3. 9C), though they are still paused overall as the signal in the promoter proximal region is much higher than the gene body. NELF likely plays a role in an obligatory checkpoint before polymerase enters elongation, as stated in the publication whose NELF ChIP calls were used(Lee *et al.*, 2008). Therefore, the genes not called as binding are simply not expressed enough that NELF was detectable above background with ChIP-Chip, even if it associates with polymerase at a frequency similar to more highly expressed genes that were called as NELF bound.

Previously, our lab showed that an NBE-like element is enriched downstream of gene promoters with higher levels of pausing(Pagano *et al.*, 2014). To assess whether such genes are more dependent upon NELF-E interacting with the nascent RNA for establishing the level of pausing, I compared genes with a good match to the NBE (Figure 3. 10A) downstream of the TSS (where it is transcribed and the RNA could help recruit NELF) to a separate control set with a match to the NBE 25 bp or greater upstream of the TSS by FIMO (where the NBE could not

predispose the nascent RNA to associate with NELF). Both sets of promoters show a slight reduction in pausing in induced NELFapt expressing cells when compared to CuSO₄ treated control cells (Figure 3. 10B and C). Therefore, genes with strong matches to the NBE do not appear to be any more dependent upon a specific interaction between NELF-E and nascent RNA for proper regulation of the pause.

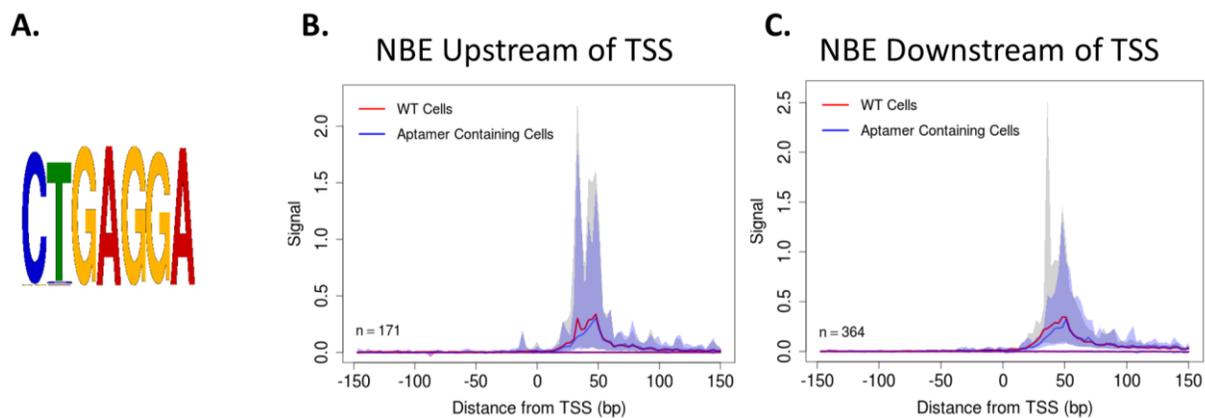


Figure 3. 10: Effect of NELFapt inhibition in pause regions with NBEs

A.) Position Weight Matrix (PWM) of the NELF Binding Element (NBE) determined by SELEX(Pagano *et al.*, 2014) with Meme(Bailey *et al.*, 2009), used to identify NBE containing genes. **B.)** Metaplot of PRO-seq signal in the pause region of genes with a match to the NBE by FIMO from -25 to -175 upstream of the TSS. Here, the NBE is not transcribed, so it could not play a role in recruiting NELF through the nascent RNA. Compare pause levels in WT (red) and NELFapt containing cells (blue). **C.)** PRO-seq signal in the pause region of genes with an NBE match between +25 to +175 downstream of the TSS. Here, the NBE could help recruit NELF through the nascent RNA.

While the level of pausing did not drastically change with NELFapt expression, the first round of PRO-seq experiments showed a dramatic shift of the pause further downstream at a subset of genes that was not reproduced in the second round of experiments. *CG5854* exemplifies this shift (Figure 3. 11). In WT cells, there are two peaks of pausing separated about 50 bp, with more signal at the first peak. In aptamer expressing cells, the first peak is reduced while the second peak remains the same.

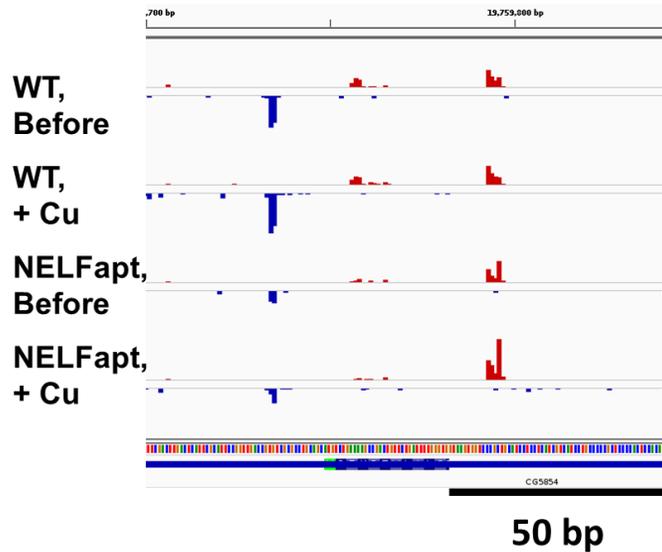


Figure 3. 11: *CG5854* shows a downstream shift in pausing in the first PRO-seq IGV browser shot of the *CG5854* pause region in Aptamer expressing and WT cells, before and after induction. RPM normalized PRO-seq, scale set from -1.3 to +46 (group autoscale).

To quantify these shifts genome wide, I used a cumulative distribution function (CDF) of to describe the distribution of PRO-seq signal within the pause region (Figure 3. 12A for PRO-seq Signal at *CG5854*, Figure 3. 12B for the corresponding CDF). The CDF tracks the cumulative fraction of total signal within the window seen up to each position. Pause regions where most signal is seen very close to the TSS will saturate early, and the CDF value will be close to 1 beyond the pause peak. Therefore, the area under the CDF is much higher at promoters where all signal occurs early in the pause region than it is at those where it is later. A downstream shift in pausing thus results in a reduction of area under the CDF. I did not use area under the CDF alone to identify shifts, as some small changes in dispersed pause peaks manifest as a large change in CDF area even though rearrangements are minor, as exemplified by *CG7324* (Figure 3. 13A). Therefore, in addition to requiring that both the copper treated and untreated aptamer containing cells show a reduction of CDF area of at least 10 (as the shift is observed both before and after induction, I used the untreated and treated as replicates), I also require that a reduction of CDF area of at least 4.5 occur in at least one single 20 bp window (Figure 3. 13A,

bottom, and Figure 3. 13B). 503 genes met these criteria for being called as significantly shifted out of 2,913 genes filtered for expression above background and lack of run-through transcription from upstream genes (Figure 3. 13C). 391 genes exhibiting a minimal redistribution of signal within their promoter proximal region were identified as an unshifted control.

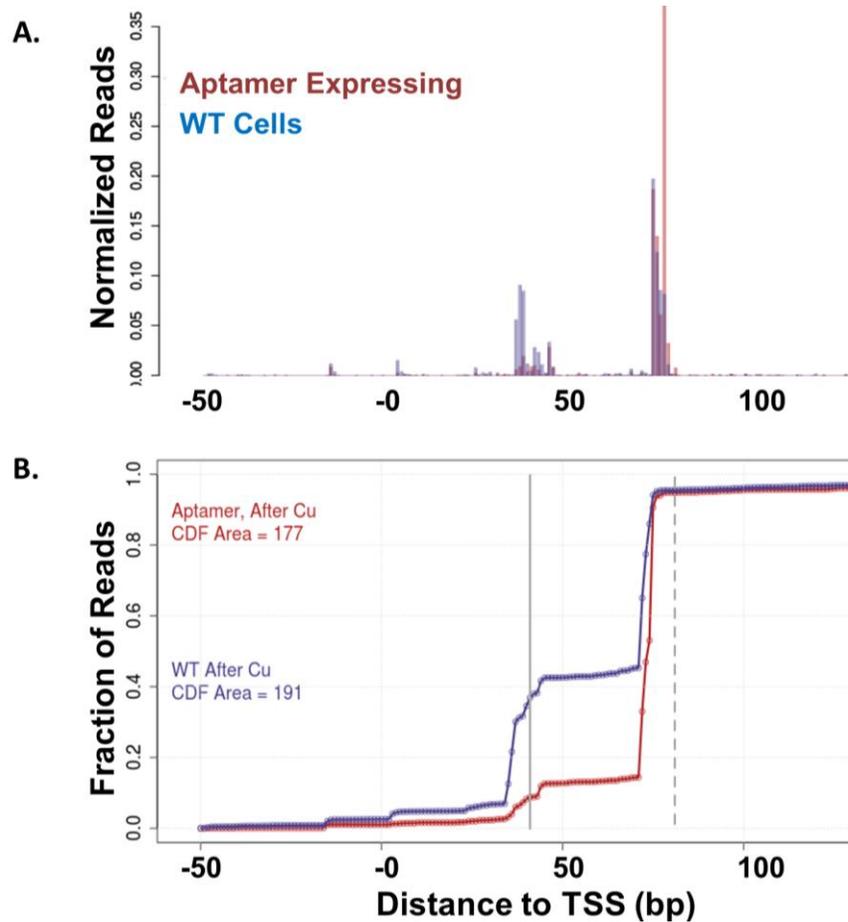


Figure 3. 12: Quantifying shifts in pause distributions with CDF area.

A.) *CG5854* PRO-seq, in CuSO_4 induced NELFapt and WT cells. Normalized as fraction of total from -50 to +250 of TSS. **B.)** CDF of the *CG5854* pause region. The region from -50 to +250 of the start site is used. At each bp, the cumulative fraction of total reads within the window seen up to that point is calculated. The total area under the CDF curve is a measure of the center of the pause distribution. At *CG5854*, ~35% of the total pause is observed from 40 nt to 45 nt downstream of initiation in WT cells, but in NELFapt cells, only ~5% of the pause is located there. Because the pause distribution shifts later in the NELFapt cells, the total area under the CDF curve is reduced from 191 in WT cells to 177 in NELFapt cells. Approximate positions of shifts (gray vertical lines) were called as the center of the 20bp window with the biggest decrease (solid) or increase (dashed) in cumulative fraction.

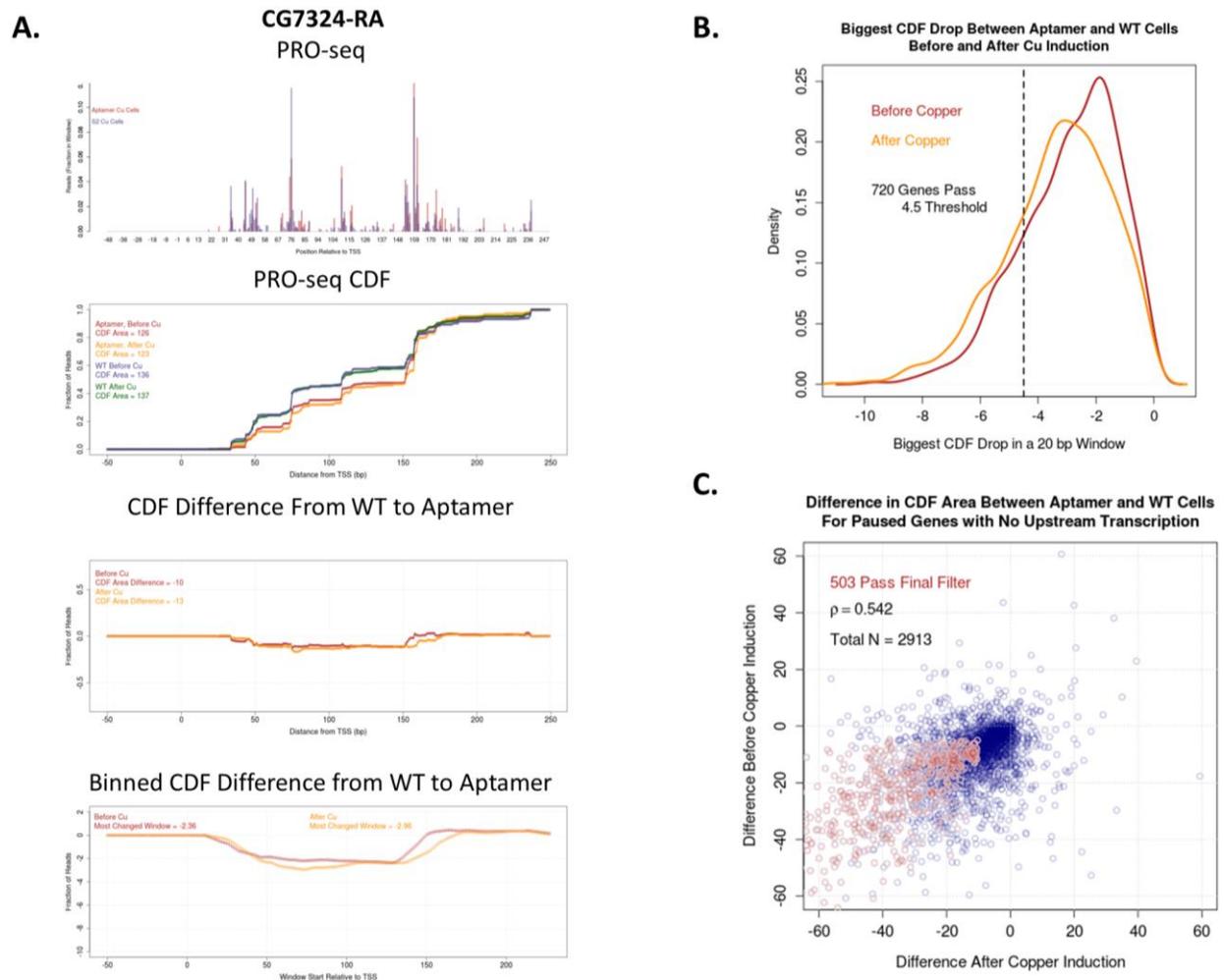


Figure 3. 13: Calling genes with shifts in pause distribution in the first PRO-seq

A.) *CG7324-RA* pause region PRO-seq, with different features plotted. This is a problematic pause region, with many subtle changes over a long distance that could potentially be called as a downstream shift. Top, PRO-seq in the pause region normalized as in Figure 3. 12A (-50 to 250 to TSS). 2nd, PRO-seq CDF, made as in Figure 3. 12B. 3rd, difference between NELFapt and WT CDF value at each position in the pause region, before (red) and after (orange) copper induction. Total difference indicate at top. Bottom, binned difference in CDF area. At each position, this is the difference in the CDF area (as in 3rd) in the 20 bp bin centered on that position. **B.)** Density plot of the biggest binned CDF drop between aptamer and WT cells (i.e. the minimum value of A, bottom for *CG7324-RA*). This is a measure of how abrupt the drop in CDF area is. A cutoff of 4.5 was chosen for calling a promoter as shifted. This corresponds to a 22.5% drop in cumulative sum sustained for 20 bp. **C.)** Correlation in CDF area change (between WT and NELFapt cells) before and after copper induction. 2,913 genes filtered for expression and lack of readthrough from upstream transcription are shown. 503 were called as shifted (red) by having a decrease in total CDF area both before and after induction (as in A, 2nd from top), and by having a minimum abruptness of change (as in B, and A bottom).

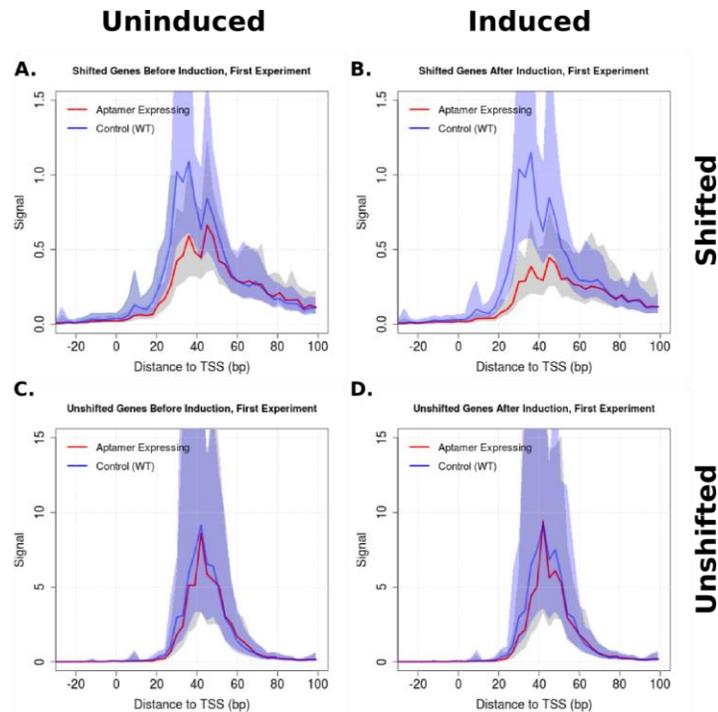


Figure 3. 14: A shift in pause distributions with NELFapt the first PRO-seq experiment Metaplots of RPM normalized PRO-seq in the pause region of genes called as shifted (N = 503) or unshifted (N = 391). Comparison of NELFapt (red) to WT (blue) before (left column) or after (right column) copper treatment. **A.)** Shifted genes before induction. **B.)** Shifted genes after induction. The early part the promoter proximal pause peak is strongly reduced. This effect is present before induction, but amplified with induction. **C.)** Unshifted genes before induction. **D.)** Unshifted genes after induction. Levels of promoter proximal pausing are much higher at unshifted genes, and less affected by the aptamer.

Shifted genes show a reduction in the more proximal part of the pause with little change in the downstream portion, overall (Figure 3. 14), and have less total paused polymerase on average when compared with unshifted genes (Figure 3. 14A and B compared to C and D). Transcription factors known to play a role in pausing like GAF, M1BP, and BEAF-70 were differentially enriched at shifted genes. Unshifted genes are strongly enriched for GAF, while genes with the shift are more likely to be bound by M1BP or BEAF-70 (Figure 3. 15). Consistent with their enrichment for GAF, which is known to recruit nucleosome remodelers to keep GAF bound genes relatively nucleosome free (Fuda *et al.*, 2015), genes with no shift have reduced nucleosome positioning and occupancy while shifted genes have a highly bound first

nucleosome, with strong downstream phasing(Gilchrist *et al.*, 2010). Figure 3. 14 looks at PRO-seq in the pause region aligned to the annotated TSS in order to represent the distance that polymerase has likely transcribed after initiation before pausing. Thus, the information surmised from such a metaplot depends upon the perspective from which the loci being summarized are aligned. In Figure 3. 16, I plot the same PRO-seq data as Figure 3. 14, but aligned to the dyad of the first nucleosome (from the subset of genes that had a nucleosome call(Kwak *et al.*, 2013)). From this perspective, it becomes apparent that shifted genes (Figure 3. 16B) have a reduction in pausing far upstream of the first nucleosome, but not near it, while unshifted genes again do not show this reduction (Figure 3. 16A). Thus, the shift is in reality a proximal decrease, without an accompanying distal decrease that shifts the balance of the pause distribution.

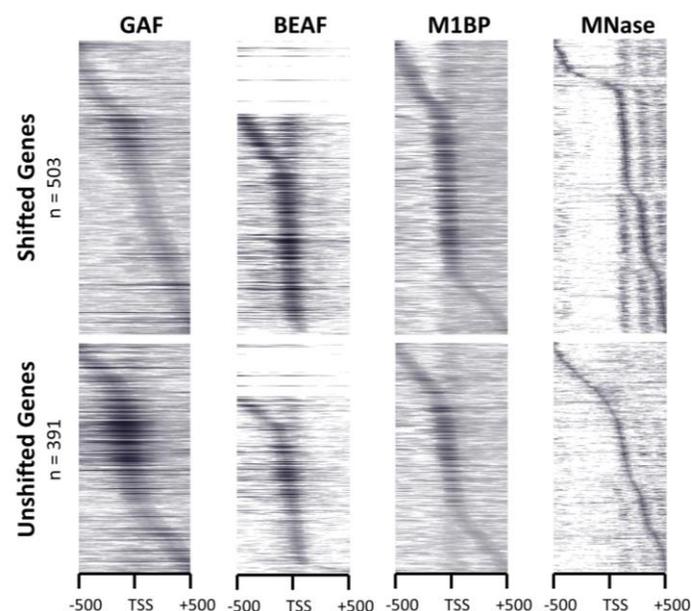


Figure 3. 15: Genes with shifts in pause distribution are depleted for GAF, but have high nucleosome occupancy and positioning.

ChIP-seq for GAF(Fuda *et al.*, 2015), BEAF(Yang, Ramos and Corces, 2012), M1BP(Li and Gilmour, 2013), and MNase-seq(Gilchrist *et al.*, 2010) around the TSSes of genes whose pause distribution was called as shifted (top) or not shifted (bottom). 1000 bp window centered on the TSS. All data are sorted by the position of the bin with the greatest signal. Thus, intensity shows levels of occupancy, and the shape of the curve through the plot shows preference for certain locations relative to the TSS.

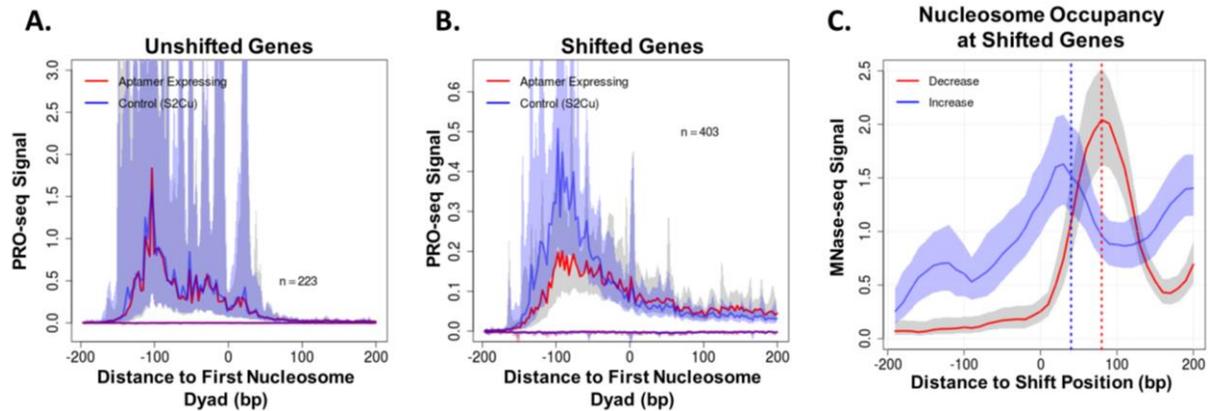


Figure 3. 16: Shifted pause distributions are a result of reduction of proximal pausing with no change in pausing at the first nucleosome.

A.) Metaplot of PRO-seq centered on the dyad of the first nucleosome at unshifted genes. All data are after copper induction, NELFapt (red) and WT (blue). N = 223 shifted genes with calls for first nucleosome dyad(Kwak *et al.*, 2013). B.) Same as A, but for shifted genes. N=403 C.) Metaplot of MNase at shifted genes, centered on the position of the reduction (red) or the increase in fraction signal (blue). Thus, the same data are plotted from two different perspectives. Vertical lines at 40 bp from the increase (blue) and 80 bp from the decrease (red).

A similar trick of switching perspective in making a metaplot from the same sites reveals the relationship of the nucleosome to the shift. I called the location of the shifts as the center of the 20 bp bin with the largest change in CDF area (both the proximal decrease and distal increase in CDF, Figure 3. 12 for an example). When MNase is plotted around the decrease, I see that the position of the decrease is near 80 bp from the nucleosome dyad (Figure 3. 16C, red). When the same MNase data from the same genes is plotted from the perspective of the gain in fraction signal portion of the shift, I see that it is positioned closer to -40 bp from the nucleosome dyad, where the greatest energetic barrier to unwinding nucleosomes occurs(Hall *et al.*, 2009). The position of the decrease is consistent with the previously reported average distance of the pause peak from the nucleosome dyad, while the position of the gain in fraction signal is consistent with a small fraction of the distal portion of the pause occurring at the barrier of the first nucleosome(Kwak *et al.*, 2013). Thus, the shift that I have observed is the result of a decrease in pausing at the normal location, upstream of the energetic barrier of the nucleosome, while the

amount of polymerase slowing through the energetic barrier of the first nucleosome is unchanged.

NELF Aptamer Inhibition in a Second Experiment Confirmed GAF's Role in Recruiting NELF

The second round of NELFapt PRO-seq was designed to have more robust controls. In the first, the control was normal *Drosophila* S2 cells, with the same copper treatment as the aptamer containing cells. This means that the aptamer was under selection for blasticidin resistance, while the control was not, and the two cell lines were on different passages. The main goal for this second round of experiments was to include a scrambled aptamer control to ensure that effects seen are not due to a response to expression of high levels of a short RNA, but also to use a new no aptamer control that was transfected with the empty pRMHa3 expression vector to ensure that all cells used were on the same passage number, and cultured under the same conditions (with blasticidin selection). Replicates were very highly correlated in this second experiment, whether considering only pause regions (Figure 3. 17A) or gene bodies (Figure 3. 17B). In this second round of experiments, I did not observe the shift in pause distributions seen in the first, though I did see a global reduction in pausing that was more pronounced at genes with features similar to those that showed the shift in the first experiment. Furthermore, the effects seen were the same in the scrambled aptamer control and NELFapt cells: this is either due to one or several of the scrambled units still having a specific interaction with NELF-E or to a lack of specificity of the response seen (a change in pausing is part of a cellular response to RNA accumulation in the nucleus). However, the results of this inhibition recapitulate known aspects of NELF's function, and its mechanism of recruitment.

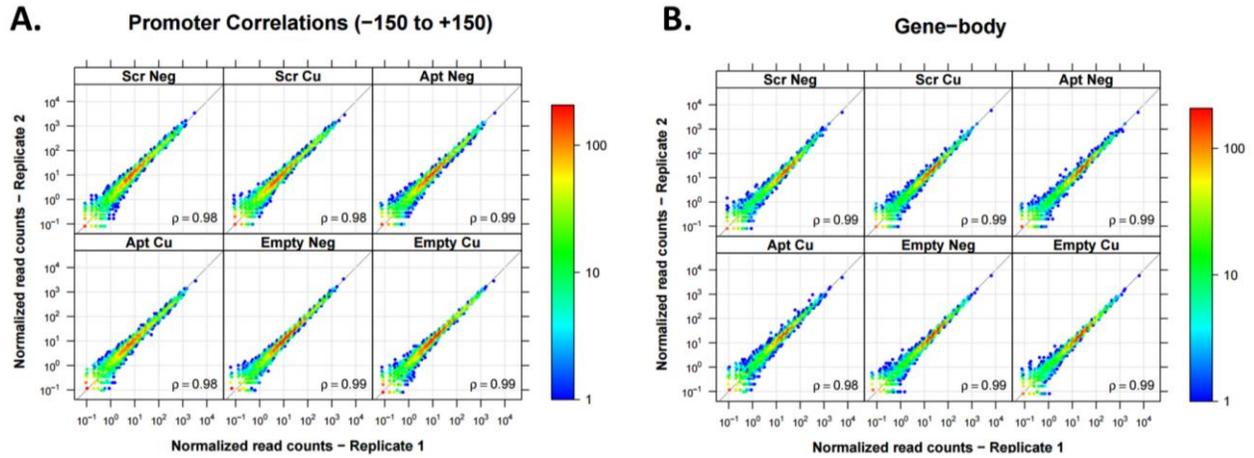


Figure 3.17: Correlation between replicates in the second PRO-seq experiment
A.) Promoter correlations, from -150 to +150 of the annotated start site. This is designed to capture pausing, to assess its reproducibility. N = 9,452, Spearman rank order correlation reported on each plot **B.) Gene body correlation, as in Figure 3.6.**

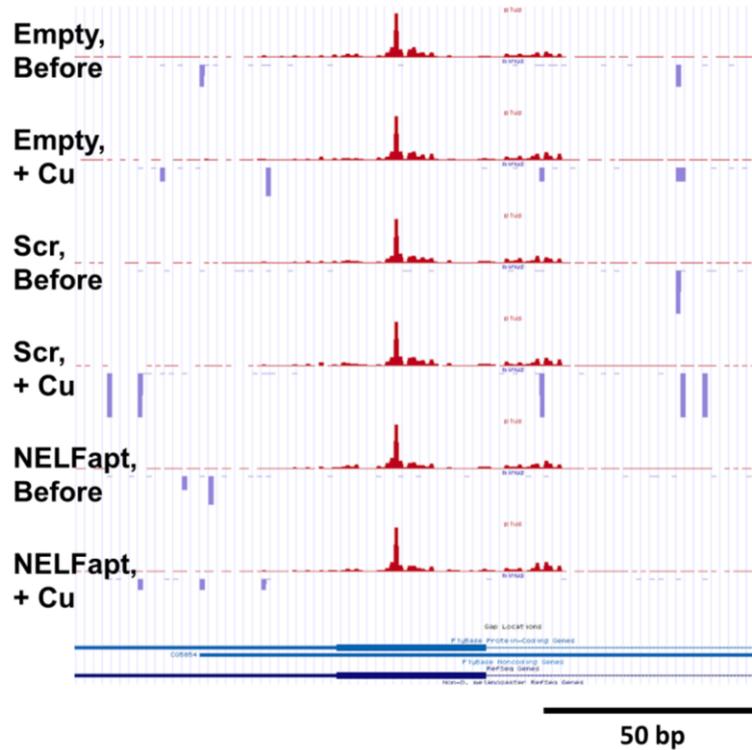


Figure 3.18: *CG5854*'s downstream shift is not reproduced in the second PRO-seq
 UCSC genome browser shot of the *CG5854* pause region in NELFapt, Scramble, and Empty vector cells, before and after induction. RPM normalized PRO-seq, scale set to autoscale.

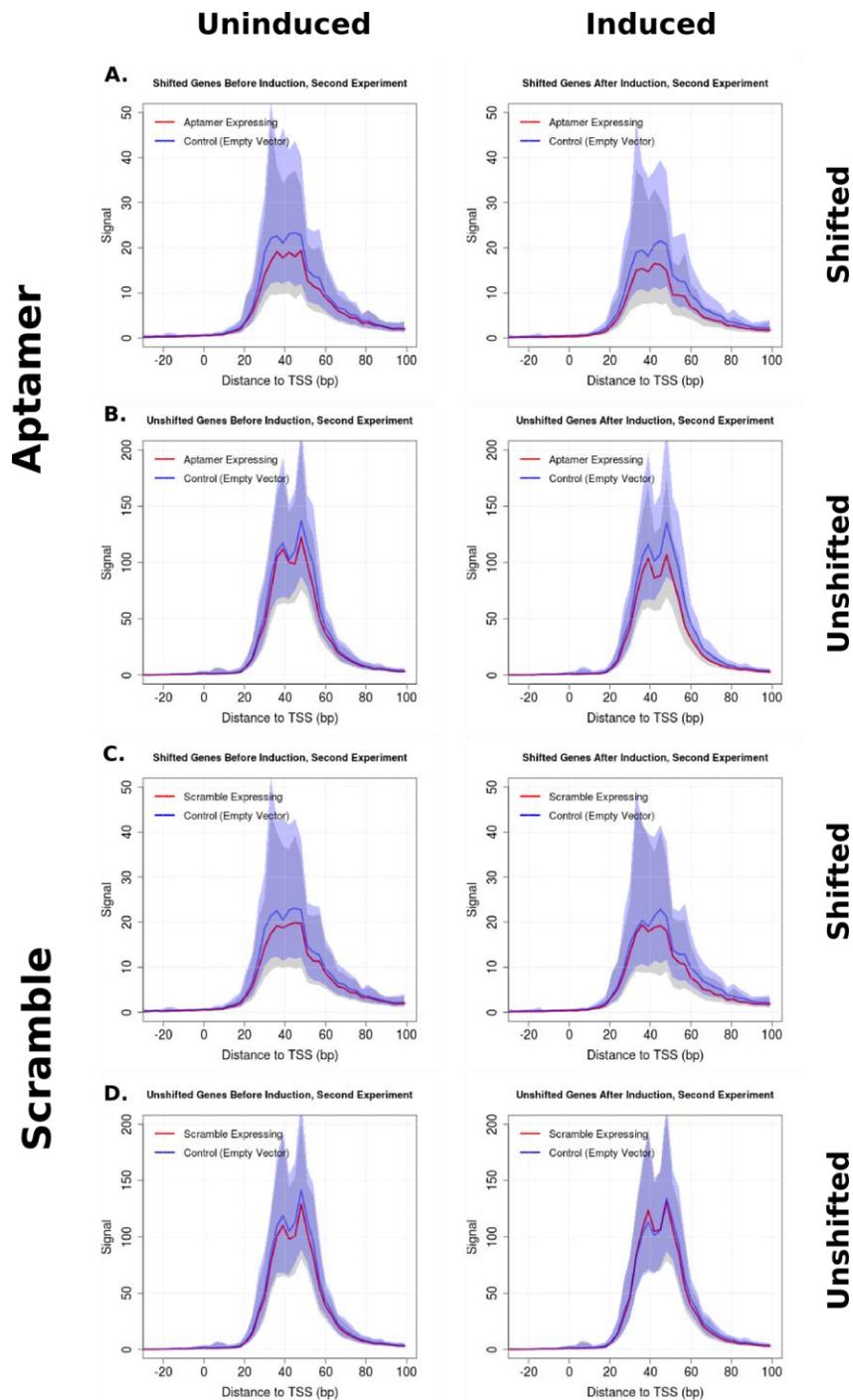


Figure 3. 19: The shift in pause distributions with NELFapt was not reproduced in the second PRO-seq experiment

Metaplots of RPM normalized PRO-seq in the pause region of genes called as shifted (N = 503) or unshifted (N = 391) in the first experiment. Comparison of NELFapt or Scramble to Empty vector before (left column) or after (right column) copper treatment. **A.)** Shifted genes, comparing NELFapt (red) to empty vector (blue) **B.)** Genes with no shift, comparing NELFapt (red) to empty vector (blue). **C,D.)** Same as A and B, but for Scrambled aptamer.

The shift in pausing observed in the first NELF inhibition was not reproduced in the second. This is seen at both individual genes such as *CG5854* where a strong shift was seen in the first experiment (Figure 3. 11 for first PRO-seq, Figure 3. 18 for the second), and in aggregate with metaplots of the genes with shifts (Figure 3. 19). As before, there is a global reduction in pausing that is more pronounced in genes called as shifted in the first experiment (Figure 3. 19A, C) than genes called as unshifted (Figure 3. 19B, D). This reduction is similar when comparing both NELFapt (Figure 3. 19A, B) and Scramble (Figure 3. 19C, D) to the empty vector cells, both before (Figure 3. 19, first column) and after (Figure 3. 19, second column) copper induction. In the first experiment, proximal pausing was strongly reduced at this subset of genes, while this reduction was not seen in the same region of the pause at a set called as not having a shift (Figure 3. 14A, B compared to Figure 3. 14C, D). That change in the shape of the pause distribution from control to aptamer expressing cells is not seen here.

Though the distal shift in pause distributions was not seen in the second round of experiments, the aptamer did have a profound effect on pausing. There is a nearly global decrease in the amount of paused polymerase in the promoter proximal region of genes. Figure 3. 20 shows MA plots normalized by DESeq2 for both promoter regions (left column) and gene bodies (right column). Though very few genes are called as significantly changed in the promoter proximal region, the distribution has an unusual shape where the most paused genes are either slightly up or unchanged, and the rest of the genes show a slight decrease in pausing. Despite these changes in the promoter proximal region, gene bodies remain largely unchanged. This change amplified upon copper induction in aptamer containing cells (Figure 3. 20 A), and is apparent when comparing both aptamer expressing (Figure 3. 20A, B) and scramble cells (Figure 3. 20C) to empty vector cells. NELFapt and scrambled aptamer cells have very similar levels of

paused polymerase across the distribution (Figure 3. 20D). Therefore, this change is nearly the same in both aptamer and scrambled aptamer expressing cells.

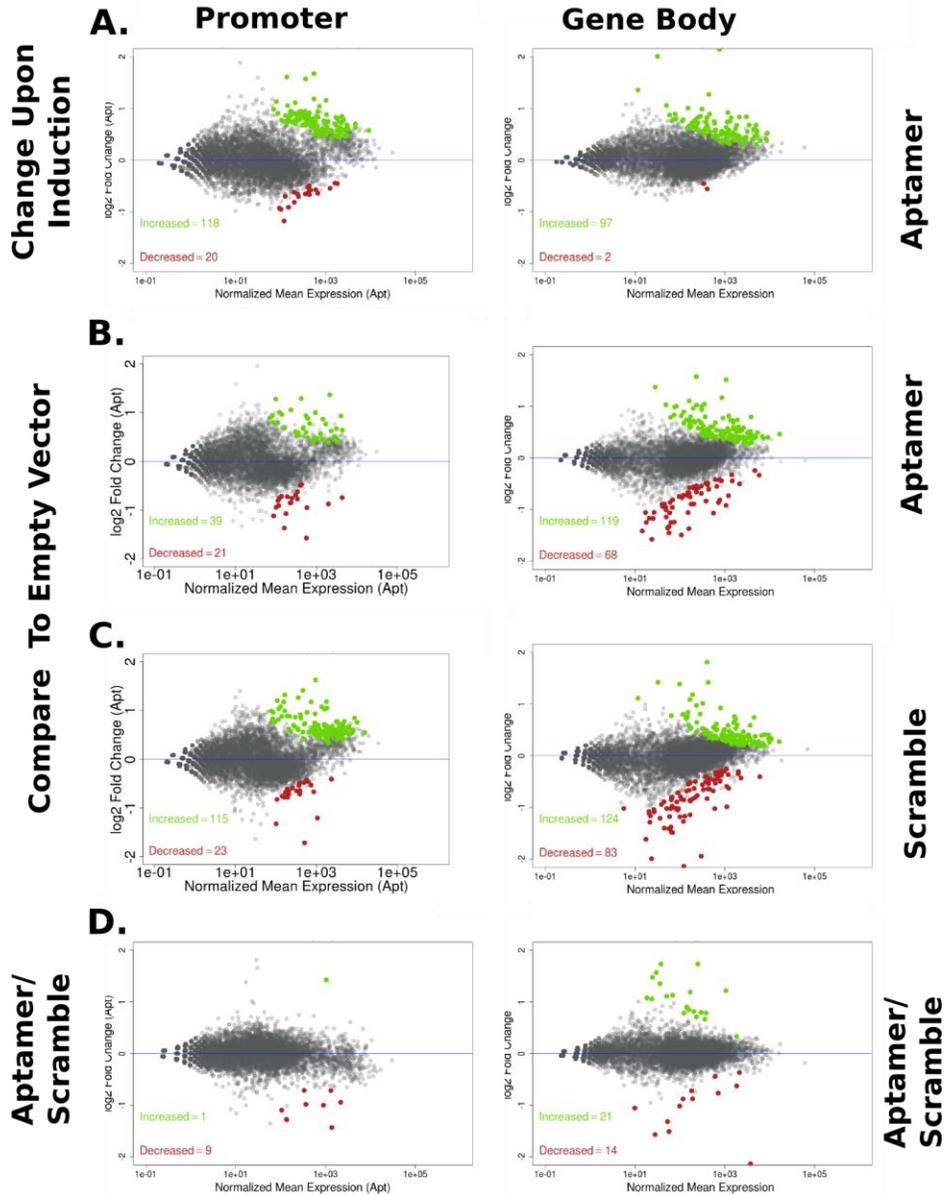


Figure 3. 20: NELFapt and scrambled control have very similar effects

DESeq2 normalized MA plots comparing PRO-seq signal from the second round of experiments in various libraries (rows), in promoter proximal pause regions (TSS to +300, left column) and gene bodies (+200 from TSS to -200 from CPA). Significantly changed genes ($p_{adj} < 0.01$) indicated in green (increased) and red (decreased). $N = 9,452$ **A.)** Changes in NELFapt cells upon induction. **B.)** Difference between induced NELFapt and empty vector cells **C.)** Difference between induced Scramble and empty vector cells **D.)** Differences between induced NELFapt cells and Scramble cells are minimal.

DESeq2 normalizes genes to account for variation between replicates, assuming that the whole distribution has not changed (Love, Anders and Huber, 2014). Thus, the importance of slight variations in lowly expressed sites is appropriately minimized (Figure 3. 20, low normalized mean expression). Though useful for interpreting changes where low signal means that observed variation is not reliable, this feature could mask systematic shifts. Therefore, to characterize the full magnitude of the aptamer's perturbation of pausing, I made MA plots using just the RPM normalized PRO-seq data (Figure 3. 21). Plotted this way the systematic reduction in levels of paused polymerase at most genes is more apparent. Now, the bulk of promoters show reduced pausing (negative \log_2 fold change), while the most highly paused promoters are unchanged (high expression promoters are near 0 \log_2 fold change). DESeq2's normalization averaged out this shift, so that the most highly paused promoters appeared to increase. The reduction in pausing occurs in both NELFapt containing (Figure 3. 21A, B) and scramble containing (Figure 3. 21C, D) cells when compared to empty vector, and is present both before (Figure 3. 21A, C) and after (Figure 3. 21B, D) copper induction, though it is amplified after induction in both cell lines (Figure 3. 21B, D). In these plots where variation between replicates and expression level are not considered in normalization, wide variation in promoters with low counts is seen as wide scattering at the low expression end of the plot. There is little global change in the amount of polymerase transcribing through gene bodies (Figure 3. 21, right column), confirming that the effect seen here is primarily on pausing. RPM normalization takes into account all transcribing polymerase seen with PRO-seq. This is not significantly affected by the reduction in pausing as paused polymerase is still a minority of total engaged Pol II. Thus, RPM normalization uses the lack of change in gene bodies to reveal the change in the levels of paused polymerase.

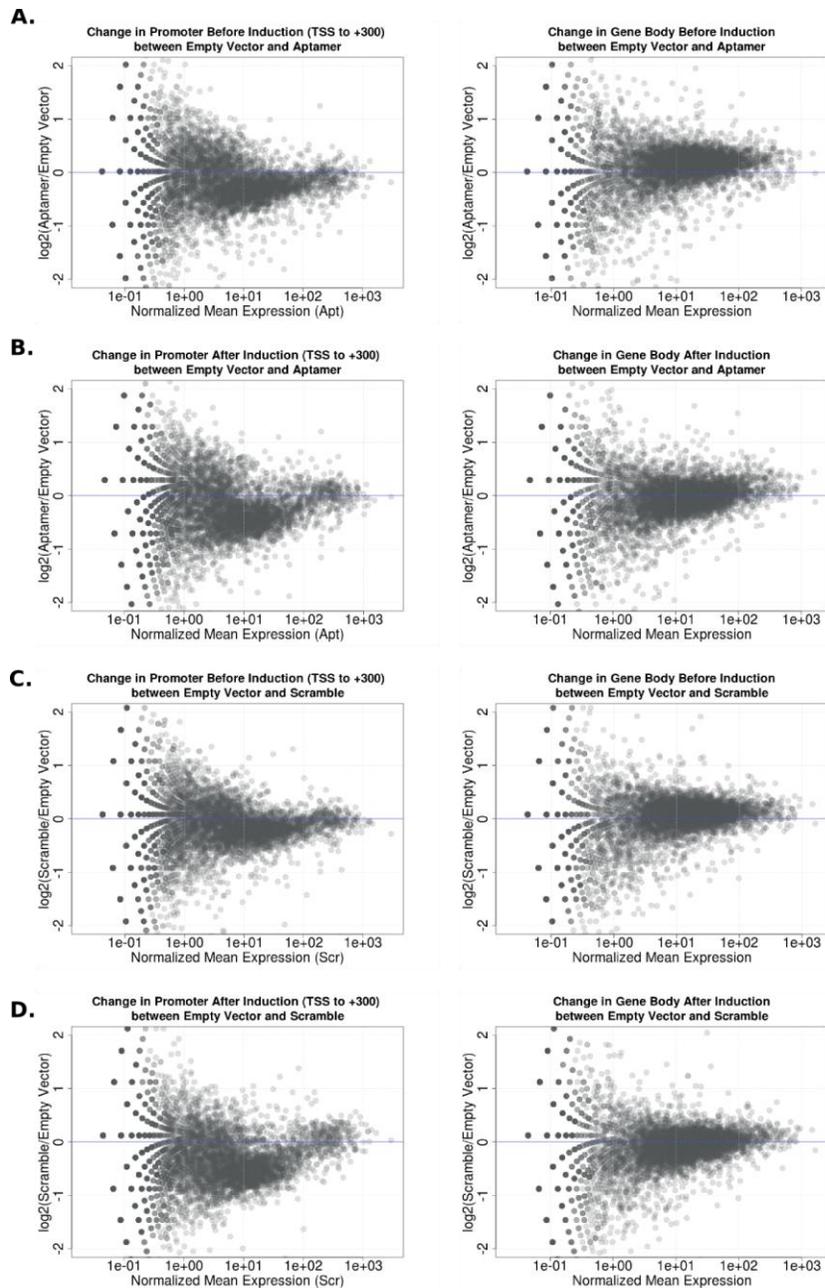


Figure 3. 21: NELFapt and scrambled NELFapt cause a global reduction in pausing at most promoters that is amplified upon induction.

RPM normalized MA comparing PRO-seq signal from the second round of experiments in various libraries (rows), in promoter proximal pause regions (TSS to +300, left column) and gene bodies (+200 from TSS to -200 from CPA). Global changes are more apparent here than with DESeq2 normalization. N = 9,452 **A.)** Comparison of uninduced NELFapt to empty vector **B.)** Comparison of copper treated NELFapt to empty vector. The gene body is largely unchanged. The pause at moderately paused genes is reduced, while highly paused genes show no reduction. **C.)** Comparison of uninduced Scramble to empty vector **D.)** Comparison of copper treated Scramble to empty vector. Changes in the pause region are almost identical to NELFapt.

To understand features of promoters that made them susceptible to the reduction in pausing after NELFapt expression, I selected genes whose total expression and \log_2 fold change fell in certain regions of the DESeq2 normalized distribution. Reduced moderate pause genes (Figure 3. 22A, red) show reduced pausing in the NELFapt and scramble expressing cells, and have moderate levels of paused polymerase. Moderate Unchanged genes are taken from the same range of pause levels as the reduced set, but do not have a reduction in pausing. Finally, highly paused unchanged genes fall at the high end of pause levels, but show no reduction upon aptamer expression. A metaplot confirms that the moderate unchanged genes do not show a reduction in pausing between aptamer expressing and empty vector cells in aggregate, while the changed genes do have a marked reduction in pausing (Figure 3. 22B). Furthermore, even though they were selected from the same range of pause levels, the unchanged genes have higher pausing. High unchanged genes are strongly enriched for GAF (Figure 3. 22C), moderate unchanged genes have some GAF binding, while the moderate reduced genes show almost no GAF binding. M1BP(Li and Gilmour, 2013), another transcription factor associated genes with high promoter proximal pausing is enriched in all three classes. BEAF-32(Yang, Ramos and Corces, 2012; Satya Prakash Avva and Hart, 2016), an insulator factor also implicated in regulation of transcription, is enriched at moderate reduced promoters (Figure 3. 22C). Finally, nucleosome positioning and occupancy is highest at moderate reduced promoters and lowest at high unchanged promoters (Figure 3. 22C, MNase), consistent with GAF's association with low nucleosome occupancy.

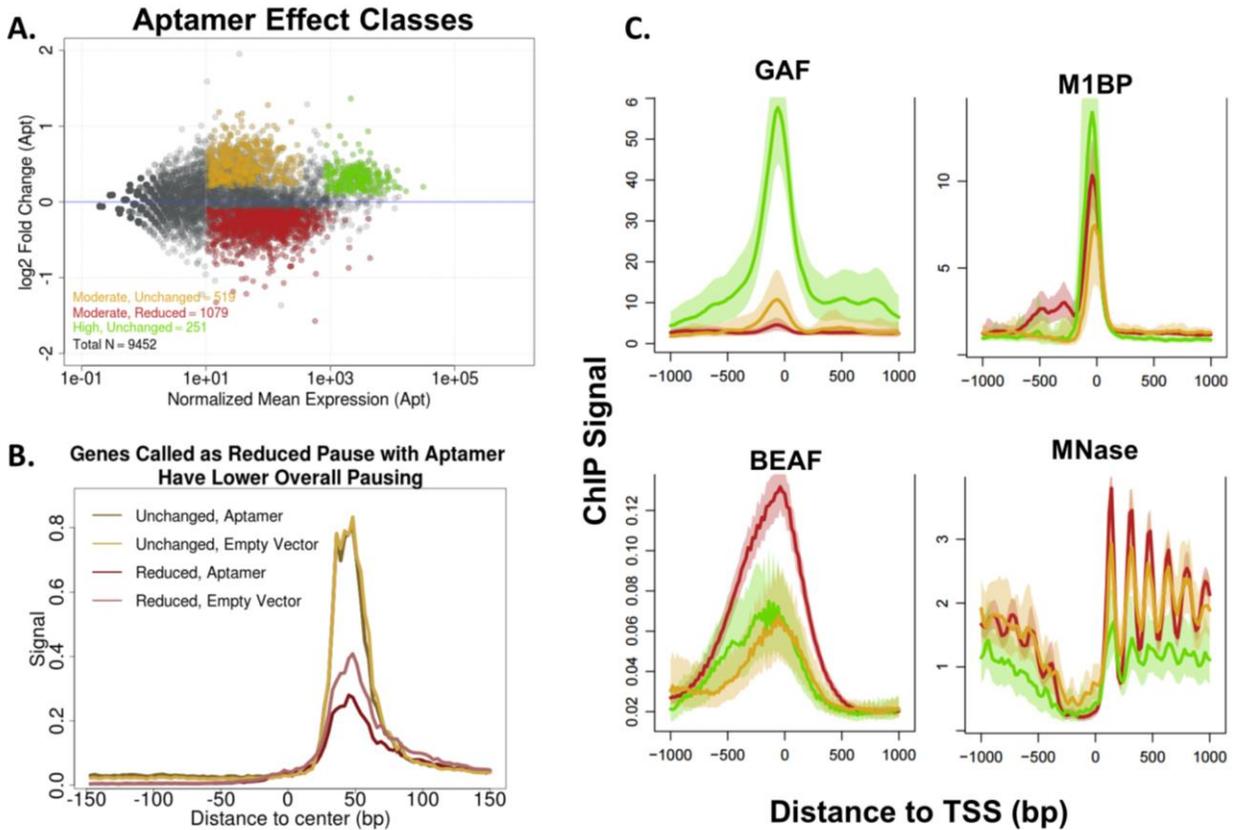


Figure 3. 22: Pause regions affected in the second PRO-seq are depleted for GAF, and have high nucleosome positioning and occupancy.

A.) DESeq2 normalized MA plot of PRO-seq from 0 to 300 bp of the TSS in copper treated NELFapt cells compared to empty vector cells. Genes were divided into classes based on DESeq2 normalized mean expression (for NELFapt or Scramble cells) and log₂ fold change. Moderate unchanged genes (yellow) have moderate levels of pausing and do not have a decrease in pausing between empty vector and NELFapt or Scramble containing cells. Highly paused unchanged genes (green) are highly paused and do not show a decrease in pausing. Moderate, reduced pause genes (red) are moderately paused, and show a reduction in pausing between empty vector and NELFapt or Scramble. **B.)** Metaplot of PRO-seq signal in NELFapt (dark) and empty vector (lighter) cells, treated with copper. Moderate reduced pause genes (red, N = 1,079) and moderate unchanged pause genes (yellow, N = 519). **C.)** Metaplot of ChIP-seq for GAF, M1BP, BEAF, and MNase-seq around the sets of genes indicated in A. Color scheme identical to A.

Conclusions and Discussion

NELF-E's Interaction with Nascent RNA Is More Important at Moderately Paused Genes

Though the two PRO-seq experiments using NELFapt as a specific inhibitor have issues with reproducibility of some of the results, and with critical controls, the effects that I see fit well with what is known about NELF's mechanisms of recruitment, and refine our understanding of formation and maintenance of a promoter proximally paused RNA polymerase. In both sets of experiments, GAF bound genes were largely unaffected by NELFapt, indicating that the presence of GAF means that NELF recruitment is less dependent upon an interaction between the nascent RNA and NELF-E. GAF is able to recruit NELF independent of many of NELF's other interactions (such as DSIF and Pol II)(Li *et al.*, 2013); thus, it makes sense that GAF bound genes are less dependent upon NELF's other interactions for its role in orchestrating pausing. In this way, I have refined our understanding of NELF's mode of recruitment by inhibiting just one of the many interactions that can bring it to paused polymerase.

In the first experiment, the most prominent effect of NELFapt was a shift in the distribution of the pause at many genes (Figure 3. 23 for a summary). At genes with more moderate levels of pausing the more proximal pause is reduced, while the distal portion is largely unaffected. This distal portion is associated with the energetic barrier to transcribing through the first nucleosome. Promoter proximal pausing can be composed of heterogeneous set of complexes at some genes, especially when pausing is dispersed (in different cells, as these polymerases would be too close to exist on the same strand of DNA). These complexes can vary in their composition and their mechanism. As I will show in chapter 4, some are more likely to be associated with a capped

nascent transcript than others. As seen by others (Kwak *et al.*, 2013; Christopher M. Weber, Ramachandran and Henikoff, 2014), I found that some of those polymerases are in a canonical paused state, likely through the action of DSIF and NELF, while others are paused at the energetic barrier to transcribing through the first nucleosome. When the paused complex is reduced, the total fraction of promoter proximally paused polymerase that is contacting the first nucleosome increases.

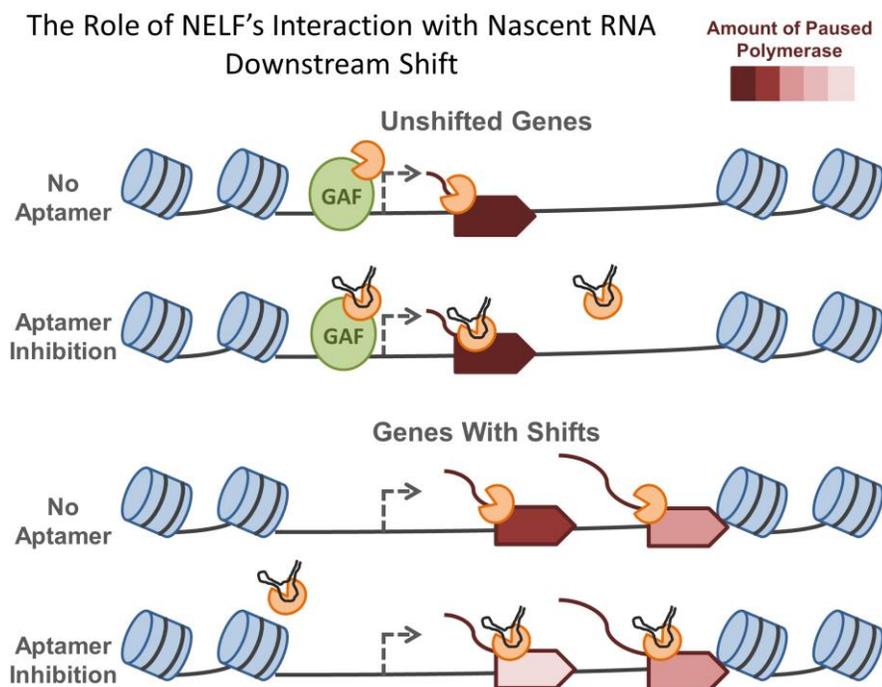


Figure 3. 23: Aptamer inhibition of NELF-E results in a reduction of pausing at genes where NELF-E to nascent RNA interactions are critical.

Schematic of NELF's role in pausing, and NELFapt's effect in the first PRO-seq experiment. Level of paused polymerase is indicated by the shading off the polymerase shape. GAF helps to recruit NELF independent of some of its other interactions. Thus, promoters with GAF are less dependent upon a free NELF-E RRM to establish the normal level of promoter proximal pausing. At other promoters, NELF's recruitment and the proper level of pausing are more dependent upon the interaction between NELF-E and nascent RNA. However, types of pausing that are not dependent upon NELF, like the pause seen at the energetic barrier to transcription through the first nucleosome, are not affected.

In my second set of experiments, I did not reproduce this shift. However, I saw a reduction in pausing at moderately paused promoters where GAF is not present (Figure 3. 24 for a

summary). Thus, both of these observations are at similar types of genes, where GAF is not present and NELF is more dependent upon interactions with nascent RNA for recruitment. Because of its role in recruiting NELF and modulating chromatin environment, GAF is, in many ways, a potent pausing factor in flies (Duarte *et al.*, 2016). GAF bound genes have much lower nucleosome occupancy (Fuda *et al.*, 2015), therefore, in both of these experiments, the genes most affected have higher nucleosome occupancy by virtue of a lack of GAF and lower transcriptional activity. Thus, the genes most affected by NELFapt, in both experiments, are those that are most predisposed to nucleosomes presenting a barrier to elongation: this reconciles the observations made in the two experiments, to some extent.

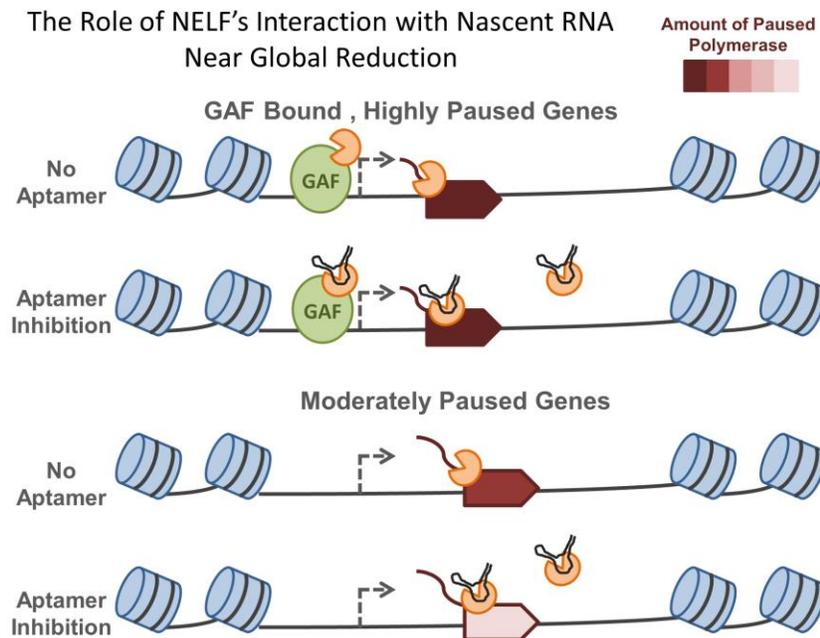


Figure 3. 24: NELF-E's interaction with nascent RNA has a greater contribution to NELF recruitment in the absence of GAF.

Similar to Figure 3. 23. Schematic of NELF's role in pausing, and NELFapt's effect in the second PRO-seq experiment. GAF helps to recruit NELF independent of some of its other interactions. Thus, promoters with GAF are less dependent upon a free NELF-E RRM to establish the normal level of promoter proximal pausing. At other promoters, NELF's recruitment and the proper level of pausing are more dependent upon the interaction between NELF-E and nascent RNA.

Difficulties in Determining the Effect of NELFapt from These Data

Though some interesting biology seems to shine through in both sets of PRO-seq data, some major issues prevent us from publishing these data as is. The effects that I see are modest, inconsistent between the two PRO-seq experiments, not fully inducible despite placing the transgene under a copper inducible promoter, and the same in my scrambled RNA control. The first major issue is that the ‘shift’, a reduction in proximal pausing with no change in polymerase at the barrier of the nucleosome, was not reproduced. I have not been able to find an experimental reason for this. Initially, I worried that a band cutting artifact from the PAGE purification step was responsible: the most proximally paused polymerase amplicons are only ~20 bp longer than the adapter dimer which we are trying to eliminate, so cutting too late could skew the pause distribution. However, I’ve found that the pause lengths, in aggregate, are not different between libraries, and the unshifted genes have proximally paused polymerase where I see no reduction in pausing, but are of a similar length to those that were reduced with the aptamer. The fact that short nascent RNAs are not globally affected alleviates concerns about a systematic problem with these libraries. The biggest difference in the two sets of experiments is the stable transfect S2 cell line used. The first used an aptamer expressing line made by John Pagano, and the second was made by me. My line expresses much more aptamer per cell, as seen by both Northern (Figure 3. 5), and in the fraction of uniquely mappable reads that align to the aptamer in the two (0.4% of total in the first experiment, 3% of total in the second).

The second major issue is that the scrambled control has the same effect as the aptamer. The aptamer tips only constitute around 40% of the total sequence of each unit. So the scrambled control leaves most of the structure intact. Thus, if the GC-richness predisposes an RNA to interact with NELF-E(Li *et al.*, 2013), it could be interacting with the core structure at some

level above background. I included 9 different scrambled sequences, as this made gene synthesis possible. This may increase the likelihood of a lucky scrambled sequence that binds NELF-E with some specificity. I may have one such lucky scrambled sequence: though it was difficult to quantify, the scrambled RNA does have a faint shifted band by EMSA (Figure 3. 3) similar to the aptamer that is not present in the N70 random 70-mer control. Even NELFapt binding would likely not be readily apparent if it were only 1/9th of a mixture. This is especially worrisome considering that aptamer expression is so high. Before induction, the aptamer is already around 20 nM in the nucleus (above the K_d of the aptamer-NELF interaction) in the cell line from the first experiment, and likely closer to 40 nM in the second. Copper induces expression, so that the aptamer reaches 500 nM in the cell line from the first experiment, and around 2 μM in the second, where it accounts for 3% of Pol II transcription by PRO-seq. Thus, even a weak interaction with NELF-E could result in full occupancy. Many RNA binding proteins' background affinity to RNA is below 500 nM; NELF-E binds to the N70 library with 120 nM affinity, so at 500 nM and higher, non-specific effects have to be a concern. Thus, more aptamer is not necessarily better, as non-specific effects likely begin to accumulate at a certain aptamer concentration. Therefore, without a new scrambled RNA control, or similar non-specific RNA control, I cannot rule out the possibility that the effect that I have seen here is not specific to occluding the RRM of NELF-E, but rather would be observed any time a small RNA is expressed at a sufficiently high level.

The effects of the aptamer were not fully inducible. Both the shift in pausing seen in the first experiment and the reduction in pausing at moderately paused genes seen in the second were observed by comparing aptamer expressing cells to control (either WT or empty vector containing), and were seen both before and after aptamer expression was induced by addition of

copper. Both effects were amplified upon induction, however. This makes sense given the amount of leaky expression seen before induction: even in the first cell line, with lower expression, the measured nuclear aptamer concentration is above NELFapt's K_d . If the problem with the scrambled control showing the same effect as NELFapt is due to expression being too high, then the lack of inducibility does not bode well for attempting to titrate aptamer expression for specific effects. One of the major goals of inhibition with aptamers is to observe effects shortly after expression of the aptamer, before secondary effects accumulate. Knockdowns in *Drosophila* S2 cells take 3-6 days to take effect, while the goal here was to look two hours after the start of inhibition. However, it takes several weeks to select a stable transfected cell line, so if basal expression has almost the full effect, then secondary effects have had much longer to accumulate and the cells have reached a new equilibrium.

The cell lines that I used to express NELFapt were populations of cells selected for integration of the transgene by selection for blasticidin resistance. The resistance marker is on a separate vector, cotransfected with the aptamer construct at a 1:20 ratio. This means that this is a heterogeneous mixture. Each cell likely has the transgene integrated at several locations, though it is possible that some cells have blastidin resistance but no aptamer. Therefore, it is difficult to interpret differences in aptamer expression. While it is likely that the population of cells in the second PRO-seq express more aptamer per cell than the first, it is possible that expression is similar, but a higher fraction of cells are transfected.

Finally, the effects of NELFapt are modest. In the first experiment, about 20% of my filtered set of genes show a downstream shift. The criteria used to identify shifted genes correspond to a drop in 25% of the cumulative fraction seen sustained for at least 20 bp when comparing NELFapt to WT cells. In the second experiment, the reduction in pausing at

modestly paused genes is, in aggregate, around a 30% reduction in the total amount of pausing, but only at genes that already had modest levels of pausing. These are subtle changes, observed broadly. This makes them difficult to quantify, as they are not readily apparent in bulk summarizations of the data, but rather require a nuanced interpretation of a targeted comparison (i.e., I have to compare RPM normalized pause levels between empty vector and NELFapt, but the effect is the same in scramble). NELF likely serves to fine-tune the control of promoter proximal pausing. Thus, these effects are very different than what is seen when a critical factor's perturbation affects viability, as is the case for M1BP knockdown(Li and Gilmour, 2013), or where one factor is critical in the regulation of a small fraction of genes, where a strong effect is seen upon its perturbation, such as HSF(Duarte *et al.*, 2016). PRO-seq is an extremely sensitive assay, with basepair resolution. Thus, there is always some worry of over-interpreting subtle differences.

Future Directions

In retrospect, the experimental design used here was not ideal: under the control of a copper inducible metallothionein promoter, we use a bivalent aptamer, expressed as an 18-mer whose self-cleaving hammerhead ribozyme arms should result in the release of 18 bivalent aptamer molecules for each time the construct is transcribed. While the strategy of bivalent aptamers to increase the apparent affinity of the resulting aptamer for its target has some merit, there is some worry that one aptamer unit could recruit additional NELF once its first aptamer arm binds. Furthermore, self-cleavage of units within the 18-mer is not complete: *in vitro*, I see ~90% cleavage, but *in vivo*, I see ~60% single units, with the rest primarily 2 or 3 unit multimers by Northern. Thus, there is some possibility that NELFapt could cause additional recruitment and

aggregation of NELF-E at promoters. In addition, as addressed in the previous section, leaky expression was enough to give a partial inhibition, and the scrambled RNA control showed the same effect as NELFapt either because one of the scrambled monomers binds NELF, something in the core 3-way junction and ribozyme binds NELF, or abundant small RNAs in the nucleus have a non-specific effect.

Widespread use of RNA aptamers will require extensive investment in development of a suitable system for carrying out this specific inhibition. I would rather see the aptamer truly expressed inducibly, as a monomer with minimal extra sequence, with an ironclad control RNA, at a level that is specifically titrated to bind the target while minimizing off-target effects, targeting an interaction whose disruption will give a strong and easily discernable phenotype, and in a stable clonal cell line where we also have a wealth of genome-wide data to aid in our interpretation of results. This effort is only worthwhile when a lot stands to be learned when inhibiting with the aptamer being used. NELFapt seemed like a perfect case: it binds NELF-E's RRM with high affinity, and would allow us to ask what the role of this single domain is in pausing, when the complex being studied has many interactions whose various contributions are hard to quantify. The aptamer was extremely well characterized, as we know what surface of NELF it binds, and in what way (Pagano *et al.*, 2014; Tome *et al.*, 2014), thus facilitating interpretation of results. NELF is instrumental in promoter proximal pausing, and we have PRO-seq as a highly sensitive readout of pausing.

Many of the issues outlined in the previous section are potentially addressable. However, this would require considerable investment. An extremely optimistic timeline would be for design, synthesis, cloning, and characterization of a new scrambled control to take around two months. Selecting stable cell lines takes around six to eight weeks to get lines that grow near the

normal rate under selection. Northern blots for picking clones take around a week. Scaling up cultures, preparing cells for PRO-seq, and carrying out PRO-seq takes two to four weeks. Sequencing and analysis would take one month, if the sequencing queue is short, the exact question being addressed is simple and well defined, and pipelines are in place. Thus, in total, each new iteration of aptamer inhibition experiments in S2 cells with PRO-seq as the readout takes at a minimum six months and costs around \$3,000 in reagents and sequencing, if you know what you're doing. If a well-functioning expression system were in place, cell line selection would still be a bottleneck. Selection of clonal cell lines in S2 is difficult, and would add considerable time and effort (Cherbas and Cherbas, 2007). Thus, for an experiment that is close to working, but not quite there, the time and effort taken for each iteration was simply too great for this to be carried to fruition in the time that I had available. Future students should be wary of this, and start with an experiment with no known design flaws and only after a thorough cost-benefit analysis. Sometimes, it makes sense to proceed with an imperfect experiment in the hope of learning something and then perfecting the approach. This strategy works well with basic biochemistry such as protein purification or EMSA, where each iteration takes around a week. However, a new mentality is needed for experiments involving the generation of a stably transfected cell line and genomics.

The idea of blocking a single domain of a protein with a highly specific and well characterized aptamer is still extremely attractive. To regulate transcription with exquisite precision, regulatory factors are often involved in many complex macromolecular interactions. Knockdowns are a blunt instrument for depleting the entire factor, but more specific inhibitors are needed to understand the role of individual interactions. Thus, aptamer inhibitors could provide important new insights. But, methods for using these inhibitors are not at a point where

they can routinely be used to study biology. Considerable investment in this strategy as a technology development oriented project is required before aptamers' widespread adoption as inhibitors.

Materials and Methods

Transfection and Stable Cell Line Generation

Drosophila S2 cells were cultured in M3 + BPYE medium with 10% FBS at room temperature, as described (Roberts, 1998). Cells were transfected with pRMHa3 vectors containing either no insert, an 18-mer NELFapt insert, or an 18-mer scrambled aptamer insert using Effectene (Qiagen, cat. 301425), following the manufacturer's protocol. 800 ng vector was transfected per 1 million cells, along with 40 ng of the pCO-Blast to enable antibiotic selection. After 72 hours, the cells were spun down and resuspended in new media with 6 µg/mL blasticidin. Cells were grown under blasticidin selection until the transformed cultures grew under selection at a similar rate to WT cells in normal media (about 6 weeks). As a control, no growth occurred for WT cells under selection in the same span of time.

Northern Blot

Selected S2 cell lines or WT control were grown to 3 to 5 million cells/mL, and treated with CuSO₄ for varying amounts of time. Total RNA was isolated as indicated (either Qiagen RNeasy, cat. 74104 or Trizol, ThermoFisher, cat. 15596026). RNA was quantified with nanodrop, and then 300 ng was run on 7M urea, 8% PAGE in 0.5x TBE. RNA was transferred to Biodyne B membrane with a semi-dry transfer apparatus, and UV crosslinked with energy set to 1200. The rest of the protocol followed the instructions of the ThermoFisher NorthernMax kit

(cat. AM1940). The probe used was a PCR amplified aptamer unit from the NELFapt vector, end labeled with P³² using PNK, following the manufacturer's instructions (NEB cat. M0201S). To reprobe for Rp49, the membrane was stripped by adding 100 mL boiling 0.1% SDS and allowing the container to cool to room temperature, and reprobbed with an end labeled Rp49 probe made by amplifying genomic DNA with the lab's stock qPCR primer.

Blots were kept damp while being exposed to a PhosphorImager screen, scanned by a Typhoon 9400 Imager, and analyzed by ImageQuant software.

EMSA

P³² body labeled aptamer was made by linearizing the appropriate pRMHa3 vector with *KpnI*, and in vitro transcription with T7 RNA polymerase, as described in Chapter 3. Single unit sized products were purified from a PAGE band as described in Chapter 3.

EMSA was carried out essentially as described in Chapter 3, with 0.5x TBE vertical 4% PAGE. Dried gels were exposed to a PhosphorImager screen, scanned by a Typhoon 9400 Imager, and analyzed by ImageQuant software.

PRO-seq

Cultures grown to 3 to 5 million cells/mL. For copper treated cells, CuSO₄ was added to the media to 0.5 mM two hours before harvesting. In the first set of experiments, nuclei were isolated, and in the second, permeabilized cells were used. PRO-seq was carried out as described(Dig Bijay Mahat *et al.*, 2016).

Libraries were sequenced on an Illumina NextSeq 500, and processed and aligned to the dm3 build of the *Drosophila* genome as described(Duarte *et al.*, 2016).

Chapter 4: Human transcription characterized by single molecule coupling of Pol II initiation and active site⁴

In higher eukaryotes, the timing and level of transcription at gene promoters by RNA Polymerase II (Pol II) is specified largely by the sum of information from the promoter itself and from distal enhancers. Transcription and Pol II pausing has been observed at enhancers, suggesting Pol II may be a ubiquitous nexus of regulatory signaling. To explore this idea, we sequenced nascent RNAs at single-molecule resolution to identify Pol II initiation, capping, and regulated pause sites. Our analyses reveal distinct sequence-specified pause classes associated with differences in RNA capping dynamics. Integrated analysis of nearby chromatin and transcription factors suggests a model of gene regulation in which Pol II initiation provides a biophysical scaffold to create and maintain regulatory domains, with pause checkpoints to limit transcription of non-coding RNAs.

Introduction: Nascent RNA Sequencing for Dissecting Gene Regulatory

Mechanisms

Recent studies have revealed widespread transcription at active regulatory sites in mammalian genomes, raising numerous questions about the role of transcription in gene regulation and chromatin architecture (Kim *et al.*, 2010; Leighton J. Core *et al.*, 2014). Existing methods track the active site of RNA Polymerase II (Pol II) (Core, Waterfall and Lis, 2008; Nechaev *et al.*, 2010; Kwak *et al.*, 2013; Mayer *et al.*, 2015a; Nojima *et al.*, 2015) and have identified promoter-proximal pausing as a key regulatory step for Pol II gene expression. Similarly, variants of these

⁴ This chapter is a draft of a manuscript co-written with Nate Tippens and John Lis

methods(Andersson *et al.*, 2014; Leighton J. Core *et al.*, 2014; Scruggs *et al.*, 2015) map sites of transcription initiation with high sensitivity, revealing an unexpected abundance of initiation from enhancer loci. Previously, most studies have relied on efficient cap selection and aggregate molecular profiles to identify Pol II initiation and pause sites. Here, we developed paired-end PRO-seq with selection for distinct 5' RNA capping states (Uncapped, Capped, or Capped+Uncapped; see Methods), with the goal of providing coordinated information about 5' and 3' events during early Pol II transcription genome-wide (Coordinated Precision Run-On and sequencing, or CoPRO).

Results

CoPRO Facilitates Identification of Transcription Initiation Sites

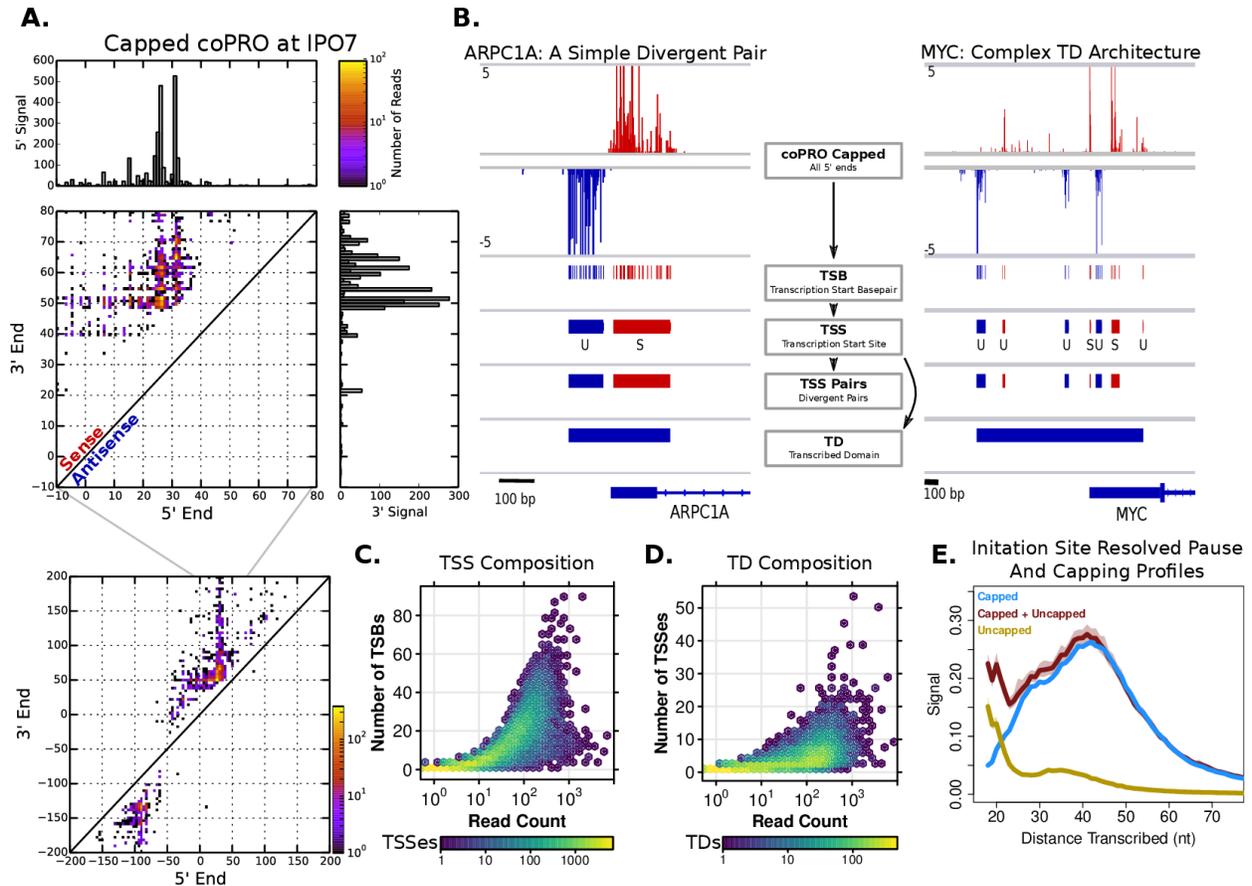


Figure 4. 1: coPRO simultaneously measures initiation site and active site of engaged RNA polymerase II genome-wide.

A.) coPRO capped RNA paired plot for the *IPO7* gene. 0 is the RefSeq annotated start site. Each bin represents a unique pairing of 5' end of RNA (initiation site, x axis), and 3' end (polymerase active, y axis), and is colored according to the number of reads observed with that pairing. Expanded view of the same site below. B.) Hierarchy of transcription initiation. coPRO capped (top) is used to call individual nucleotides where initiation occurs (TSN), TSNs within 60 nt are grouped into TSSes, divergently oriented TSSes within 300 bp are grouped into TSS pairs, and TSSes within 750 bp are grouped into TIDs. C.) Hexbin of TSN usage vs total coPRO capped expression at TSSes, N = 75,057. D.) Hexbin of TSS usage vs total coPRO capped expression at TIDs, N = 24,171 E.) True initiation site resolved pause profile of capped, uncapped, and capped + uncapped coPRO at TSNs that are uniquely mappable at 18 nt, N = 133,622. Bootstrapped median of signal, solid line, 12.5% and 87.5% confidence intervals, shaded.

CoPRO provides a distinct elongation profile for each base in the genome (Figure 4. 1A). Polymerases elongating from the same transcription start nucleotide (TSN) make a vertical line in a heatmap of coPRO for a single locus, while Pol II initiating from different TSNs but pausing at the same nucleotide create a horizontal line. This characteristic “initiation profile” – a common RNA 5’ end with numerous 3’ ends – provides a sensitive method to discriminate real initiation sites from background, independent of capping efficiency and enzymatic biases. As an example, the *IPO7* gene promoter initiates at multiple positions within ~50 nt, with strong preference for the most used two (Figure 4. 1A). We often detect polymerase at multiple positions (vertical lines) both within and beyond the pause. Notably, transcription from nearby positions tends to pause at the same nucleotide. A wider view (Figure 4. 1A, inset) reveals elongation beyond the pause (within the insert size limit of sequencing), as well as antisense transcription. Thus, CoPRO can deconvolve the interplay between initiation, capping, and pausing to provide a single-molecule view of transcription initiation genome-wide.

We organized transcription initiation into a functional hierarchy to facilitate subsequent analyses (Figure 4. 1B, Methods). First, individual nucleotides where initiation occurs (Transcription Start Nucleotides, or TSNs) are identified by having at least 5 distinct 3’ ends. We cluster nearby TSNs into Transcription Start Sites (TSSes), as they are likely driven by the same transcription factors (TFs) and pre-initiation complex (PIC). Nearby TSSes are also clustered into higher-order structures: Divergent Pairs are pairs of TSS transcribing in opposite directions <300 bp apart (Leighton J. Core *et al.*, 2014; Scruggs *et al.*, 2015; Chen *et al.*, 2016). Finally, Transcription Initiation Domains (TIDs) are larger TSS clusters (<750 bp apart) on either strand, that are thus likely to cooperatively position or modify nearby nucleosomes. We use CAGE data to determine which TSNs and TSSes give rise to stable RNAs such as mRNAs

and lncRNAs. However, most TSSes produce unstable RNAs (Leighton J. Core *et al.*, 2014) of unknown function (e.g., eRNA, ncRNA, upstream antisense RNA). In total, CoPRO identified 503,631 TSNs, 75,057 TSSes, 17,925 divergent pairs, and 24,171 TIDs.

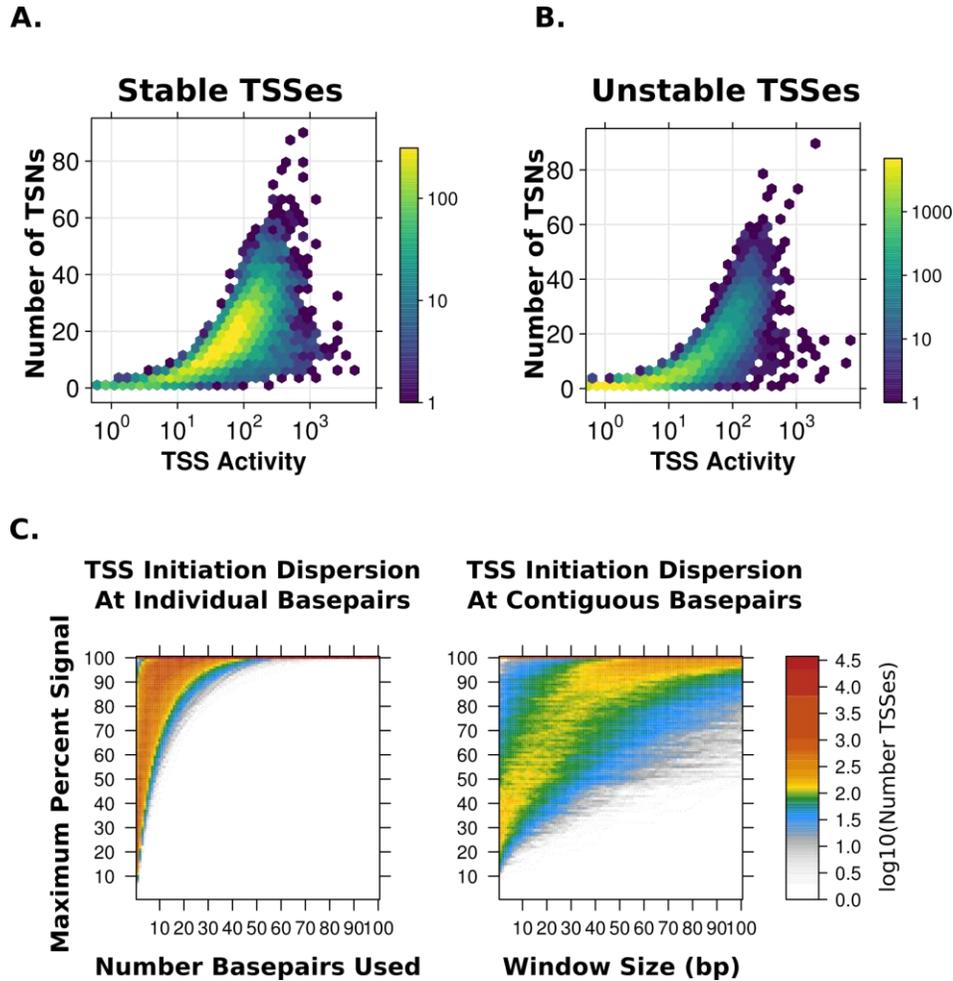


Figure 4. 2: Dispersion of initiation within TSSes

A.) TSNs used plotted against TSS activity at TSSes that give rise to stable transcripts by CAGE. $N = 11,768$ TSSes **B.)** Same as a.), but for TSSes that do not give rise to stable transcripts. $N = 63,289$ TSSes **C.)** Dispersion of initiation within TSSes is best described as individual basepair spikes, not a smooth Gaussian. Both plots use the top 50% most expressed TSSes, $N = 37,522$. This analysis tests the maximum amount of signal at each TSS that can be seen either at different numbers of basepairs (left) or windows of different size (right), which is the x axis for both. Y axis is maximum percent of signal seen with a window of size x. Color is number of TSSes described by that x,y combination. Each column thus sums to total number of TSSes ($N = 37,522$). **D.)** The left plot (single basepairs) saturates much faster than the right (windows). Thus, initiation is primarily situated at a small number of non-contiguous basepairs at TSSes.

As previously described (Carninci *et al.*, 2006; Vo Ngoc *et al.*, 2017), most TSSes initiate at multiple TSNs and TSN number increases with TSS activity (Figure 4. 1C), though the degree of dispersion varies widely (Figure 4. 2C). Extremely focused TSSes are primarily housekeeping genes, such as *ACTB* (Figure 4. 3) and *HIST1H1*, where initiation remains focused at a small number of TSNs despite very high activity (Figure 4. 2A), while initiation at unstable promoters (i.e. upstream divergent promoters, putative enhancers) tends to be more dispersed (Figure 4. 2B). Consistent with previous work, the basic unit of initiation is a divergent pair, each with its own PIC (Leighton J. Core *et al.*, 2014; Duttke *et al.*, 2015; Scruggs *et al.*, 2015; Arensbergen *et al.*, 2016; Chen *et al.*, 2016). Interestingly, most TIDs, especially those producing stable transcripts, are composed of a constellation larger than a divergent pair of TSSes (Figure 4. 1D). Previous work was limited to Stable TSS constellations such as nearby alternative TSSes and head to head genes (Chen *et al.*, 2016), but CoPRO reveals that these constellations are much more prominent when all unstable TSSes are considered: stable TIDs (i.e. promoters), on average, contain 1.3 stable TSSes paired with 4.1 unstable TSSes.

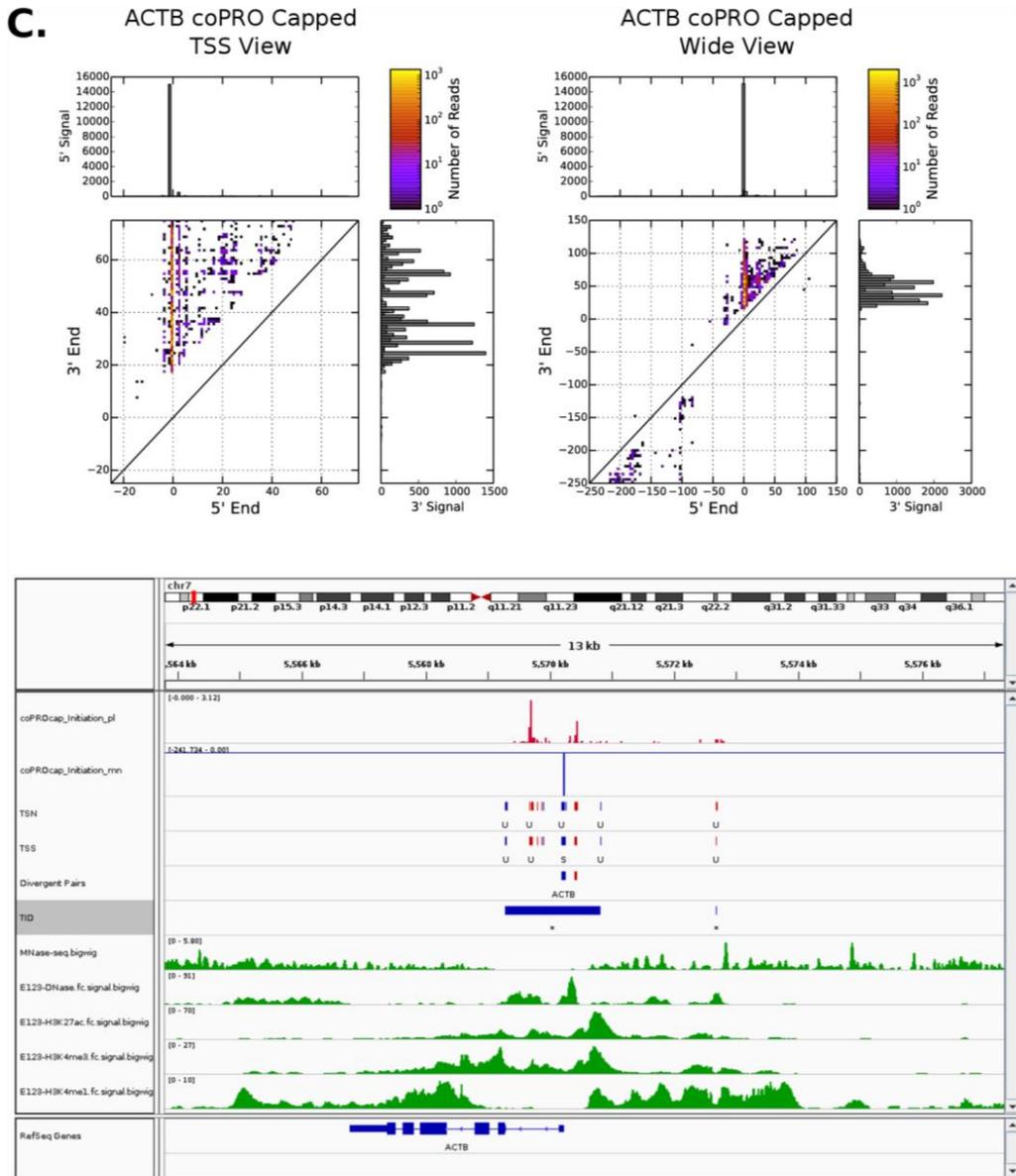


Figure 4. 3: *ACTB*, an extremely focused promoter.

The *ACTB* promoter is seen from three viewpoints:

Top left, 2D coPRO capped plot around the promoter. 0 is the annotated start site for all 2D plots. **Top, right**, 2D coPRO showing divergent transcription and higher order organization in TIDs. **Bottom**, IGV browser shot showing coPRO capped (set to autoscale independently for each strand, scale indicated at left), transcription initiation calls (TSN, TSS, Divergent pair, TID) colored by strand (red for plus, blue for minus) and with transcript stability calls from CAGE for TSNs and TSSes, MNase from ENCODE, DNase-seq, H3K27ac ChIP-seq, H3K4me3 ChIP-seq, and H3K4me1 ChIP-seq from Epigenomics Roadmap. RefSeq gene annotation at the bottom. *ACTB* is an extremely highly expressed, focused TSS. The most used TSN is used 29-fold more than the next most used. The TID is highly directional as well; the sense TSS is expressed 75-fold more than its upstream divergent. The RefSeq annotation is perfect for *ACTB* in K562 cells.

The majority of Capped RNAs at TSNs are within the pause region, reported to occur between 20-120 bp (Kwak *et al.*, 2013; Leighton J. Core *et al.*, 2014; Gressel *et al.*, 2017). At single-molecule resolution, pausing is restricted to 20-60 nt (Figure 4. 1E, blue). In contrast, Uncapped 5' ends from TSNs are found associated with very short RNAs (Figure 4. 1E, gold), confirming that capping begins as the 5' triphosphate first emerges from polymerase (~18 nt), and expanding gene-specific observations (Rasmussen and Lis, 1993; Nilson *et al.*, 2015).

Two Types of Promoter Proximal Pausing

Sequence Largely Determines Pause Position

A histogram of the most abundant pause position per TSN (maxPause) reveals a bimodal distribution: one population from 20-32 nt (Early), and another from 32-60 nt (Late; Figure 4. 4A), though most initiation sites have some pausing in both windows (Fig 2B). After grouping TSNs by their maxPause, the sequence determinants of pausing become clear (Fig 2C). Early pause TSNs are more AT-rich overall, whereas Late pause are GC-rich +5 from the initiation site to 8 nt before the pause site. AT-richness is seen -30 to the initiation site, particularly in the Early class, reflecting a stronger enrichment for the TATAWR motif of early pausing TSNs (Figure 4. 4, A/T richness at -30). All sites show a strong preference for initiation on CA dinucleotides (the Inr motif (Vo Ngoc *et al.*, 2017)), and pausing on cytosine with some enrichment for guanidine just before (Gressel *et al.*, 2017) (Figure 4. 4, top). This strong preference at the exact pause base is likely due to the fact that CTP is the least abundant nucleotide (Traut, 1994) and to the high energetic cost of stacking interactions between GC dinucleotides. For both Early and Late pause TSNs, the sequence at the exact pause position is very similar (Figure 4. 5A and B), indicating that a similar mechanism determines the precise position of the pause. On a larger scale, the sequence determinants of these two pause types are very different (Figure 4. 4C). From this, we

conclude that the lower availability of CTP strongly influences the selection of pause sites on a basepair scale. The pause catches polymerase during incorporation of a CTP as this is a slow step, so CTP is the last nucleotide incorporated at the pause. Elsewhere, when polymerase is transcribing, diffusion of CTP into the active site is slower due to its lower abundance, so polymerase is observed just before incorporating CTP more frequently than other nucleotides (Figure 4. 5C). Overall, these results refine previous kinetic models for pausing: the position of the pause is dictated by the energy landscape of nucleotide incorporation, the transcriptional bubble, and the strength of the RNA-DNA hybrid(Nechaev *et al.*, 2010; Gressel *et al.*, 2017), and the rate of pause factor recruitment(Missra and Gilmour, 2010; Li *et al.*, 2013).

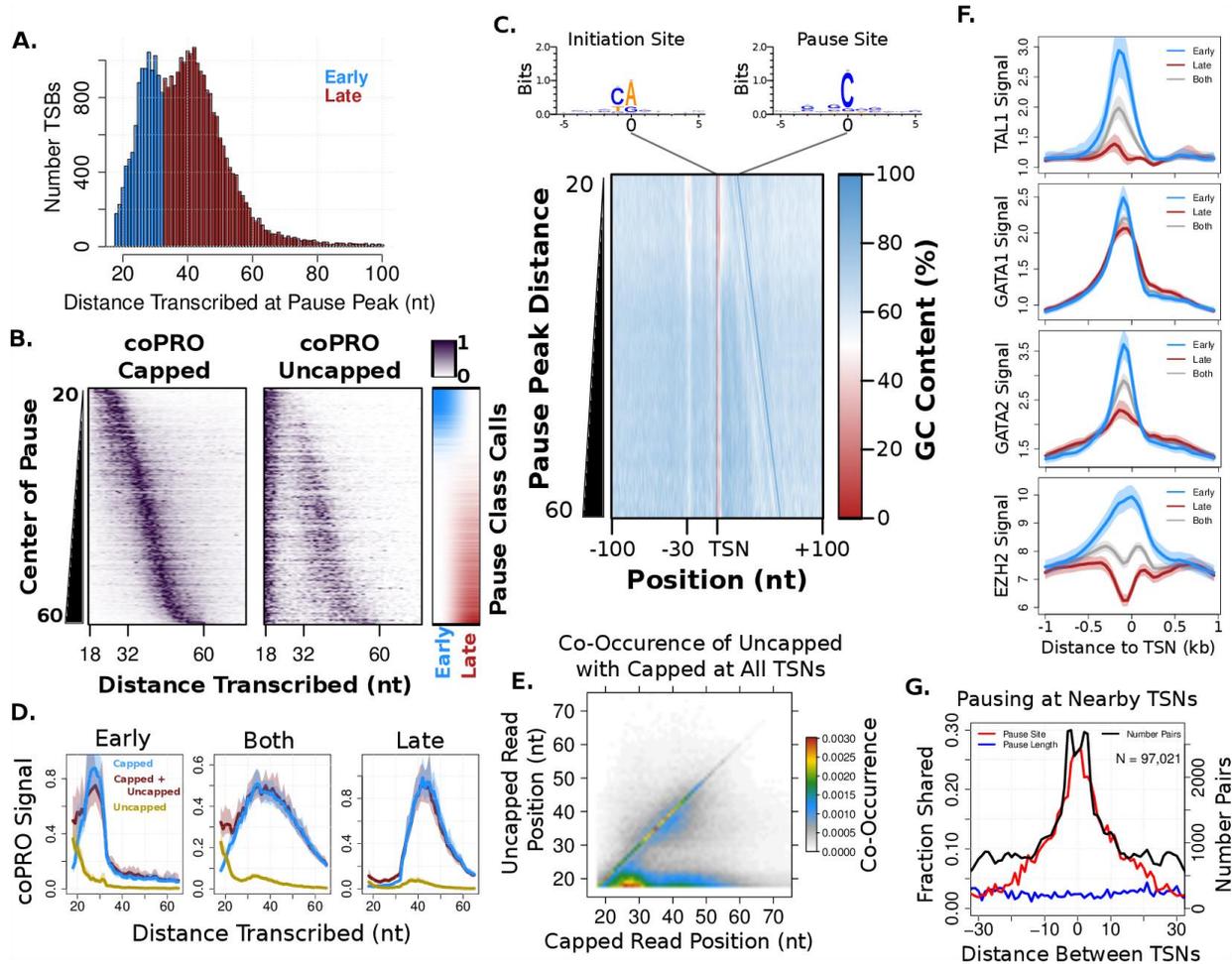


Figure 4. 4: Two distinct modes of promoter proximal pausing exhibit different capping dynamics.

A.) Most used pause site at the most used TSN in each TSS (maxTSN), $N = 31,464$ **b.)** coPRO capped and uncapped at maxTSNs, sorted by the center of their pause. TNSs are identical for capped and uncapped. Row normalized so that the highest value for each row is 1. Early/late called TSNs are indicated at the far right. $N = 31,464$ **C.)** Average GC content around maxTSNs, sorted by distance transcribed at max pause. Each row is the average of all maxTSNs with each possible maxPause in the pause region defined in a.), from 20 to 60 nt. Sequence logo for initiation site and pause site at top. $N = 31,464$ **D.)** Three cap state selected libraries at maxTSNs mappable at 18 nt in pause classes, same maxTSNs as c). Bootstrapped median, 12.5% and 87.5% confidence intervals. Early $N = 5,907$, Both $N = 17,047$, Late $N = 8,510$ **E.)** Joint probability distribution of uncapped coPRO with capped coPRO signal at the same TSN, from 18 to 100 nt. All maxTSNs mappable at 18 nt were used ($N = 31,464$). Each TSN is normalized to sum to 1 in both Capped and Uncapped before calculating joint probability, thus, the entire distribution sums to 1. **F.)** ChIP-seq for factors enriched at pause classes. All data from ENCODE. Same TSNs and method as d. **G.)** Fraction of auxiliary TSNs that pause at the same site (red) and same length (blue) as their maxTSN. Number of auxiliary TSNs observed at each distance (black). $N = 97,021$ auxiliary TSNs

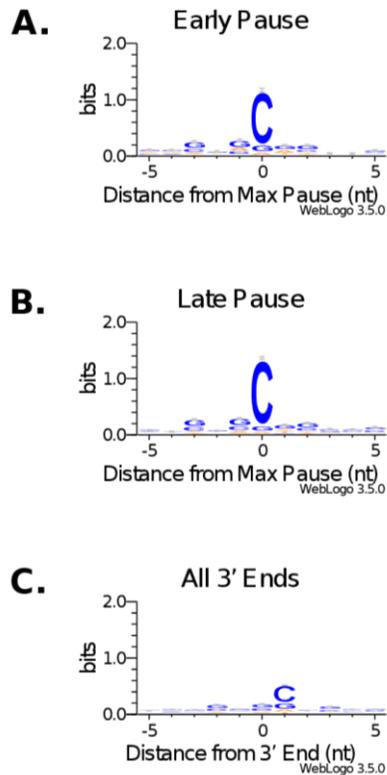


Figure 4. 5: DNA sequence around and pause sites

A.) Sequence at the pause position of early pause TSNs. 0 indicates the last nucleotide incorporated in the reads used to generate the position weight matrix. **B.)** Sequence at the pause position of late pause TSNs. **C.)** Sequence around the 3' ends of coPRO reads outside the pause region. Here, the C is just upstream of the 3' end.

Different Capping Dynamics in Early and Late Pause Classes

The Early and Late pause classes have different uncapped RNA distributions, indicating an interplay between the processes of capping and pausing. TSNs of both classes have very short uncapped RNA (less than 22 nt), but Late also have uncapped RNAs at the pause site (Figure 4. 4C, right, and Figure 4. 4D). Capped + Uncapped coPRO shows that the overall distribution of RNAs at Early TSNs is continuous (Figure 4. 4D, red), transitioning from mostly uncapped at the beginning to a mostly capped RNAs at the end (Figure 4. 4C and D). These different uncapped distributions are best understood with joint probabilities (Figure 4. 4E), which quantify co-occurrence of uncapped and capped RNAs at all positions for individual TSNs. After averaging

joint probabilities across all TSNs, it is clear that capped RNAs in the late pause (33-60 nt) mostly co-occur with uncapped RNAs of the same length (diagonal in Figure 4. 4E), while capped RNAs in the early pause (20-32 nt) are more likely to co-occur with very short uncapped RNAs (below the diagonal). Early pause TSNs are capped at shorter RNA lengths than late pause TSNs, consistent with longer residence time between 20-32 nt. By contrast, late-pausing Pol II proceeds efficiently through the first 32 nt, resulting in less time for capping and thus more uncapped RNAs at late pause sites. This is consistent with an order of assembly where DSIF is rapidly recruited to Pol II partly through an interaction with the nascent RNA (Missra and Gilmour, 2010), and then helps to recruit capping enzyme (Mandal *et al.*, 2004). These results extend previous capping analyses (Rasmussen and Lis, 1993; Nilson *et al.*, 2015) genome-wide, and reveal a new pause-dependent variability in the location of capping.

Early Pausing Is Linked to a Poised State of Chromatin

Overall, sequence specifies pause distributions from TSNs that exhibit a spectrum of pause fraction within Early and Late windows. Pausing in both windows is detectable from most TSNs (Figure 4. 4C). To understand the functional consequences of pause distance, we examined TSNs at the top and bottom quartiles of this spectrum (Methods). These pause classifications are reproducible between replicates. The enrichment for TATAWR at early pause TSNs supports a model where this type of pausing remains coupled to the PIC, similar to *Drosophila* (Kwak *et al.*, 2013), and is supported by CHIP-seq assays showing greater levels of TBP and a higher ratio of Ser5 to Ser2 CTD phosphorylation in the Early class (Figure 4. 6).

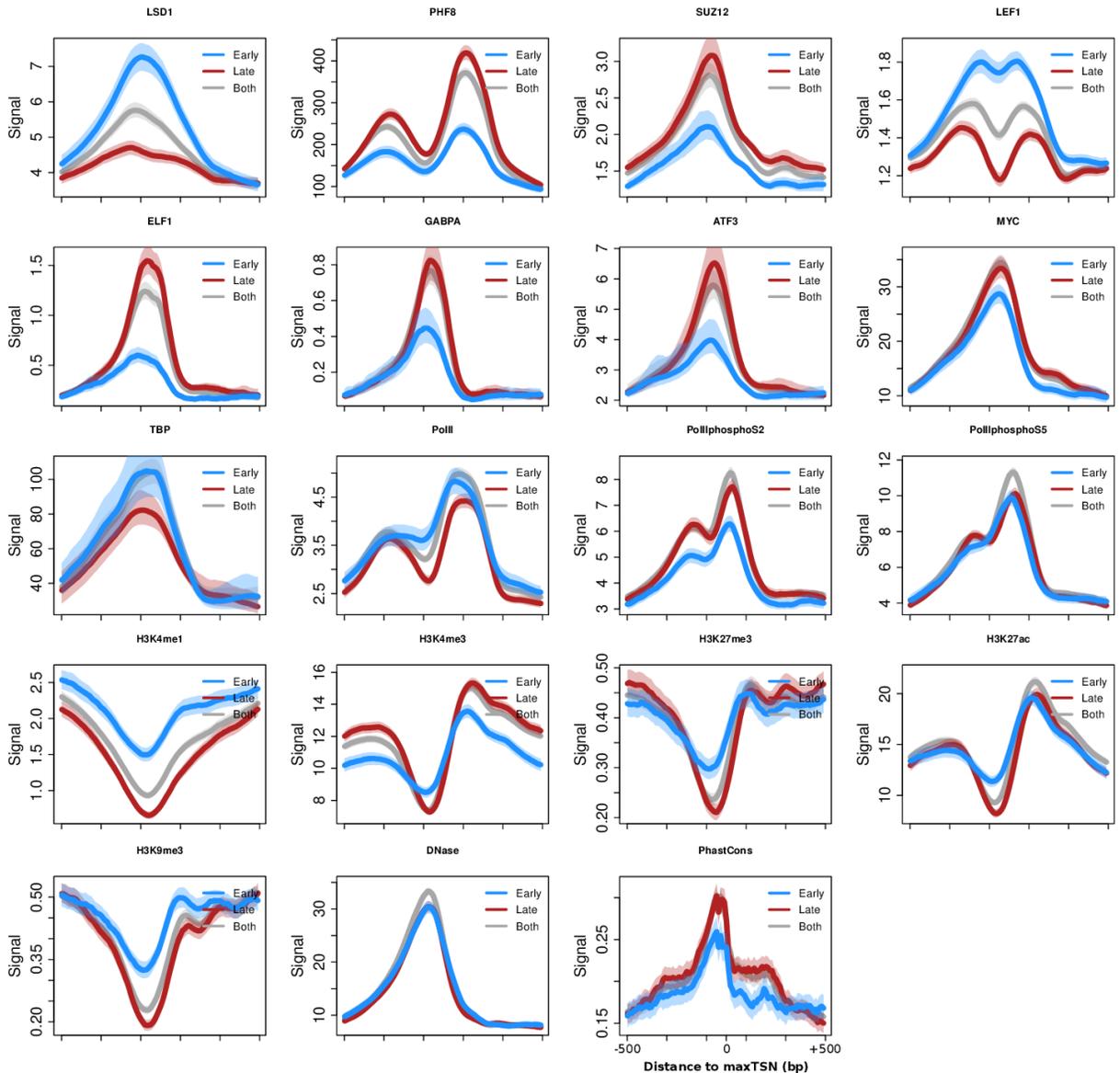


Figure 4. 6: Transcription factors, chromatin environment, and Pol II phosphorylation at Early and Late pause TSNs.

Extension of Figure 2F. ChIP-seq for factors enriched at pause classes. All data from ENCODE, except histone modification and DNase data from Roadmap Epigenomics. Early N = 5,907, Both N = 17,047, Late N = 8,510

Early pause sites are also enriched for binding of lineage-specific TFs such as GATA1, GATA2, and TAL1, the polycomb subunit EZH2 (Figure 4. 4F), and the H3K4me3 demethylase LSD1(Whyte *et al.*, 2012) (Figure 4. 6). Enhancer-specific TF binding, the CTD phosphorylation status, lower GC content, and overall less active chromatin state (Figure 4. 6),

implicates Early pausing as a novel feature of the “poised” state of RNA polymerase (Ferrai *et al.*, 2017). Early pause TSNs are less expressed than Late, and are much less likely to undergo pause release (Figure 4. 7). From this, we conclude that the Early pause counteracts productive elongation. In contrast, Late pause TSNs are enriched for binding of activating TFs such as GABPA, ATF3, ELF1, the H3K9 demethylase PHF8, and are enriched for H3K4me3 and H3K27ac (Figure 4. 6). Therefore, TSNs with late pausing are endowed with a pattern of GC content, transcription factor binding, and chromatin environment that facilitate productive elongation.



Figure 4. 7: Expression Level and Pausing at Early and Late TSNs

When compared to Late pause TSNs, Early pause TSNs are significantly more paused (ratio of reads beyond 80 nt to total), less expressed (total capped reads), and less likely to undergo pause release (total capped reads beyond 80 nt).

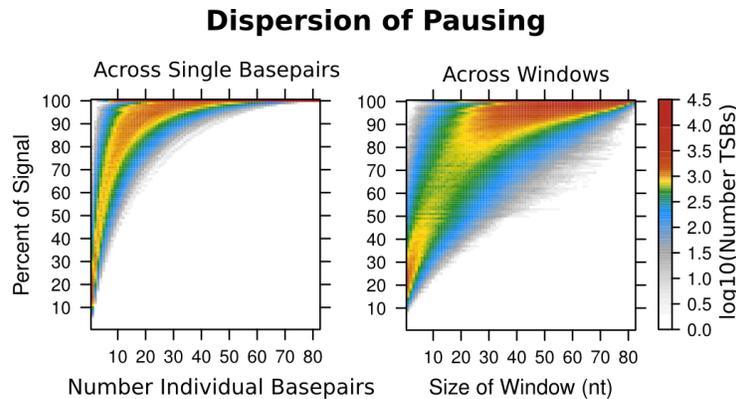


Figure 4. 8: Dispersion of Pausing

Analysis of dispersion of pause sites similar to analysis of dispersion of initiation sites within TSSes in **Figure 4. 2**. Signal saturates much faster by gathering from individual non-contiguous basepairs (left) than by gathering from windows of contiguous positions (right).

Comparison of Pause Sites Used by Nearby TSNs

Given that most TSSes initiate on multiple TSNs using the same PIC, structural insights may be gained from comparing pause and initiation distributions at TSNs within the same TSS. That is, by comparing pausing of the maximum TSN to other TSNs in the same TSS (auxiliary TSNs), we can determine the frequency of pausing on the same base and how it varies with the distance between TSNs (Figure 4. 4G). As expected from a sequence-specified pausing mechanism, nearby TSNs frequently share exact pause sites (red line), whereas a model with a fixed initiation to pause distance for each TSN is not supported (blue line). Pausing at individual TSNs is often distributed across several noncontiguous basepairs (Figure 4. 8): examination of pause profiles at individual TSNs shows that even minor pause sites are generally shared by nearby TSNs. *MAPK1* is a good example of such pause sharing (Figure 4. 9): it has an exceptionally dispersed TSS, enabling comparison of the pause profiles of multiple nearby TSNs. Nearby TSNs often pause at the same position. Even where pause sites are minor, they are shared by nearby TSNs, indicating that the mechanisms determining major and minor pause site choice are similar. Further supporting a model of multiple pause sites per TSNs, pausing is best described as occurring on several non-contiguous basepairs, rather than a smooth Gaussian of contiguous site (Figure 4. 8). Furthermore, auxiliary TSNs are clustered within 10 bp of the maximum TSN before reaching background (Figure 4. 4G, black line), likely reflecting the maximum scanning distance for the PIC (Fishburn, Galburt and Hahn, 2016; Vo Ngoc *et al.*, 2017).

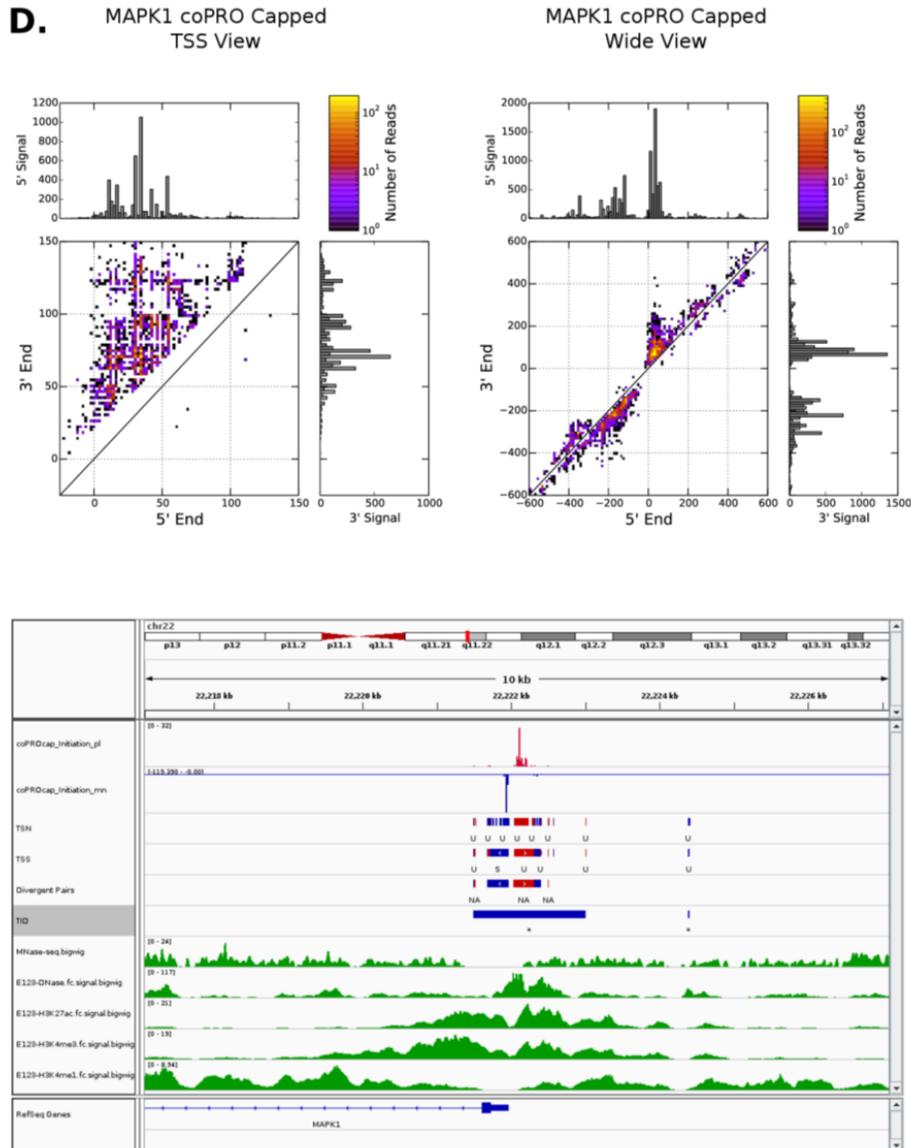


Figure 4. 9: Many TSNs Pause at Shared Sites in the *MAPK1* Promoter

The *MAPK1* promoter is seen from three viewpoints:

Top left, 2D coPRO capped plot around the promoter. 0 is the annotated start site for all 2D plots. **Top, right**, 2D coPRO showing divergent transcription and higher order organization in TIDs. **Bottom**, IGV browser shot showing coPRO capped (set to autoscale independently for each strand, scale indicated at left), transcription initiation calls (TSN, TSS, Divergent pair, TID) colored by strand (red for plus, blue for minus) and with transcript stability calls from CAGE for TSNs and TSSes, MNase from ENCODE, DNase-seq, H3K27ac ChIP-seq, H3K4me3 ChIP-seq, and H3K4me1 ChIP-seq from Epigenomics Roadmap. RefSeq gene annotation at the bottom. *MAPK1* is a highly expressed, dispersed TSS. This makes it good for visualizing sharing of pause sites by nearby TSNs (top left). The major pause site is on the same nucleotide for nearby TSNs. Even minor pauses are often shared. Pause sites are indicated by the 3' end (y axis), so a pause site shared by different TSNs makes a horizontal line on this plot; even minor pause sites have increased signal at that site for most TSNs with signal there.

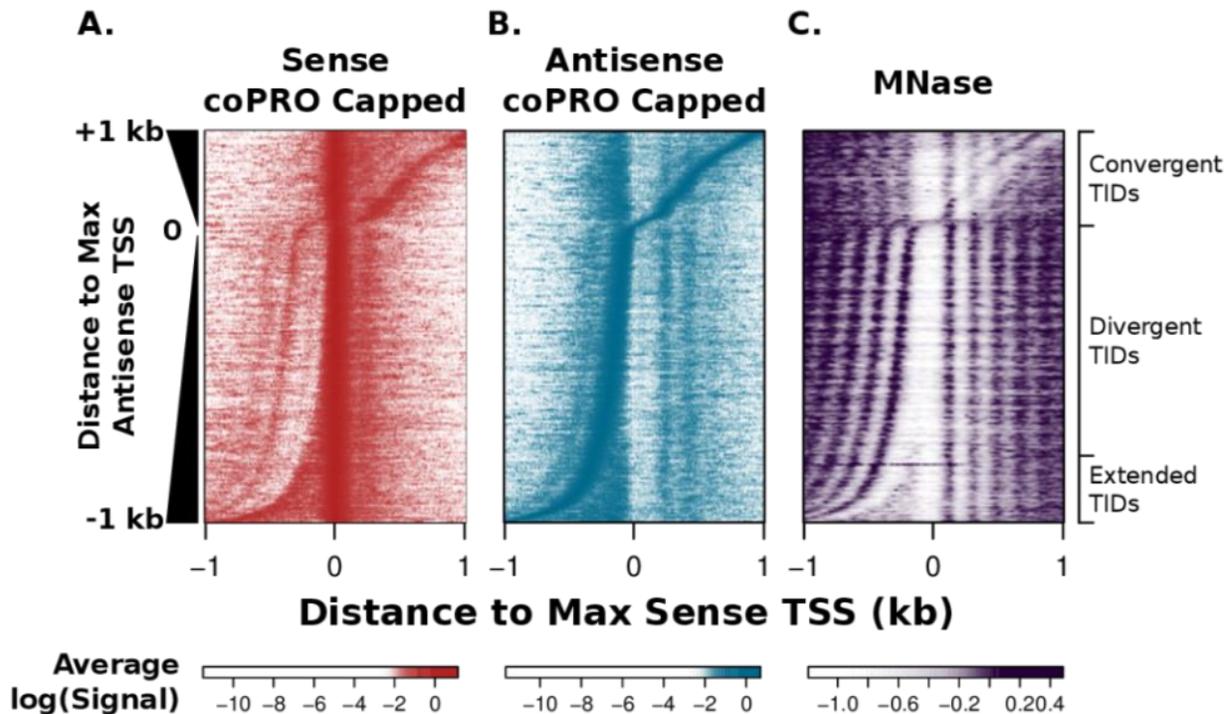


Figure 4. 10: A global view of transcription initiation shows rules for divergent pairing and reveals widespread complex organization.

a.) CoPROcap initiation centered on the strongest sense TSS in per TID. Sites are sorted by the distance to the strongest peak on the opposite strand, from +1 kb (convergent) to -1 kb (upstream divergent). Sites filtered for expression, at least 7 reads per strand. $N = 13,731$ (3,444 Convergent TIDs, 7,988 Divergent TIDs, and 2,299 Extended TIDs, labeled to right of C). Similar sites averaged to generate 200 rows in final heatmap. **b.)** coPRO capped antisense, same sites as a. **c.)** MNase (ENCODE) at same sites as a. and b.

Transcription Initiation Occurs in Clusters

Arrangement of TSSes within TIDs

While widespread bidirectional pairing of mammalian TSSes has received considerable attention (Leighton J Core *et al.*, 2014; Duttke *et al.*, 2015; Scruggs *et al.*, 2015; Chen *et al.*, 2016), larger TSS clusters, or transcription initiation domains (TIDs), remain uncharacterized. To compare and refine these two conceptual frameworks, we identified the most Pol II occupied TSS (maxTSS) on each strand of TIDs and sorted them by distance (Figure 4. 10A, B). At 25% of TIDs, the antisense maxTSS is downstream (Convergent), and at 17% the antisense maxTSS is >300 bp upstream (Extended) (Figure 4. 10B). In both cases, there is still a clear upstream

divergent TSS for both maxTSSes (sense and antisense; Figure 4. 10A, B). Thus, as recently reported for stable TSSes (Chen *et al.*, 2016), divergent pairs are separated by at most 300 bp, as TSSes separated by >300 bp are best described as two divergent pairs. Convergent transcription from intronic enhancers through promoters has been reported to attenuate gene expression (Mayer *et al.*, 2015b; Cinghu *et al.*, 2017), but the dense clustering of divergent TSS pairs within most TIDs necessarily creates convergent relationships between some TSSes within the TID. These studies are referring to convergent transcription with a more downstream origin. Elongation through TIDS by such convergent transcription with a distal origin likely results in a more repressive chromatin environment at the TID and thus lower expression (Cinghu *et al.*, 2017). A study more similar to ours found that nearby convergent TSSes can extend the active chromatin environment further downstream, and can even increase expression (Lavender *et al.*, 2016). Therefore, it is unlikely that all convergent transcription between divergent pairs is repressive. In fact, in this section we show that the many TSSes within TIDs collaborate to create an active chromatin compartment around gene promoters. Convergent relationships inevitably arise from such clusters thus may not be unique in their relationship to the promoter.

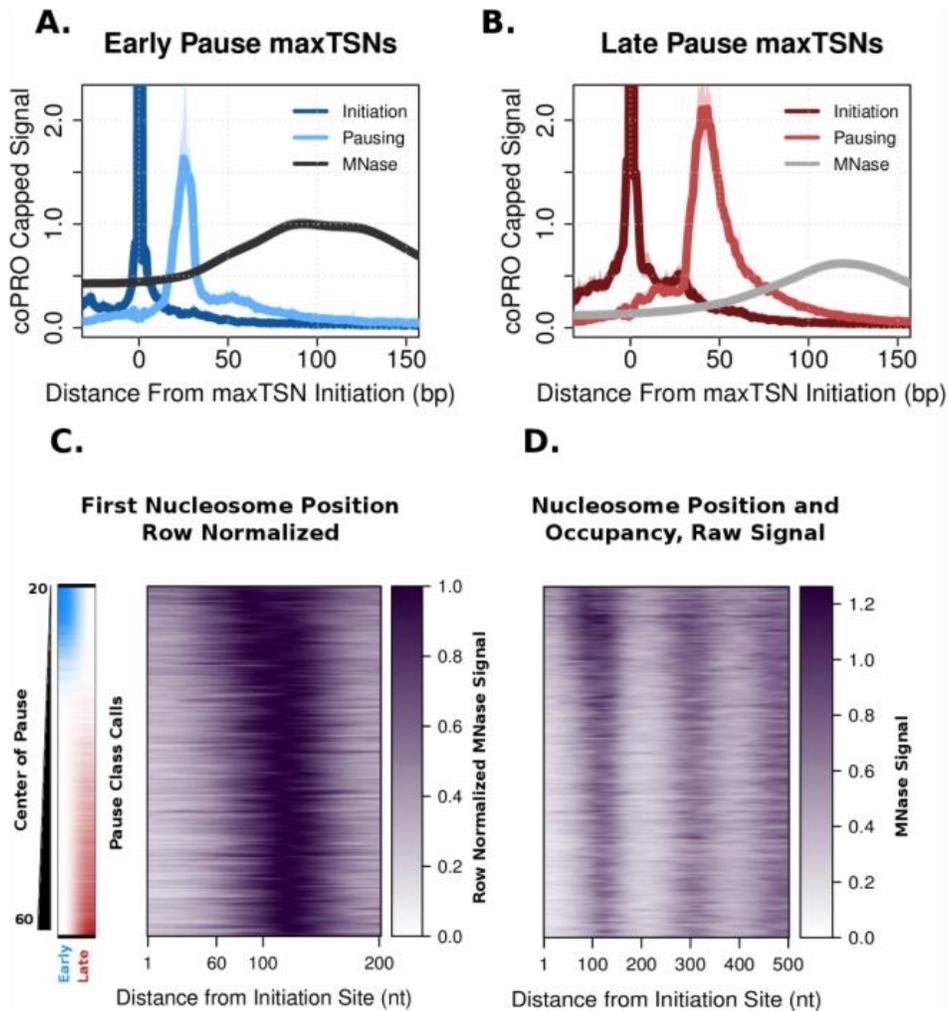


Figure 4. 11: The first nucleosome is positioned by paused polymerase

A.) Meta profile of coPRO capped initiation (dark blue), pausing (light blue) and MNase (dark gray) at TSNs called as early paused and mappable at 20 nt ($N = 6,176$) **B.)** Same as a.), for late pause TSNs ($N = 8,838$) **C.)** Row normalized MNase downstream of maxTSNs. Signal normalized to highest bin. Same elements and order as Fig 2c, $N = 31,464$ TSNs binned into 200 rows. Sorted by pause center. Early/Late called TSNs shown to right. **D.)** Same TSNs and order as c.), from initiation site to +500 bp. No row normalization to show higher nucleosome occupancy at early pause TSNs.

The Strongest TSSes Determine Nucleosome Phasing within TIDs

As reported previously (Gilchrist *et al.*, 2010; Scruggs *et al.*, 2015), we find that the maxTSS on each strand sets nucleosome phasing from Simple Divergent and Extended TIDs (Figure 4. 10A-C, 0 to 1 kb). Here, we expand this model by asking how nucleosomes are positioned within larger clusters of TSSes at the scale of TIDs. At a finer scale, we find that

nucleosomes are more closely aligned to pausing than initiation by comparing Early and Late TSNs (Figure 4. 11), consistent with paused Pol II specifying precise nucleosome positioning and occupancy. At Convergent TIDs, nucleosomes are still phased to the maxTSS between the convergent TSSes, but maxTSS phasing is disrupted beyond the convergent TSS (Figure 4. 10C, Convergent TIDs, 0 to 1 kb, Figure 4. 13). This indicates that maxTSS phasing largely supersedes phasing from weaker TSSes, but weaker TSSes can dominate phasing downstream (model in Figure 4. 13A). Nucleosome phasing from the max TSS determines where TSSes are likely to fall downstream: the distribution of convergent TSSes is serpentine, falling preferentially between phased nucleosomes while the distribution of upstream TSSes is smooth (Figure 4. 13A, sense and antisense maxTSS lines, and Figure 4. 10B). Upstream divergent TSSes do not interfere with sense maxTSS phasing and thus operate under less constraint than convergent TSSes. In Convergent and Extended TIDs, nucleosome occupancy is reduced between the multiple divergent TSS pairs (Figure 4. 10C, top and bottom portions), meaning that TSSes within TIDs collaborate to reduce nucleosome occupancy and establish phasing. Finally, weak TSSes are enriched in the gaps between nucleosomes (Figure 4. 10A, upstream and Figure 4. 10B, downstream of maxTSS, and Figure 4. 14): these TSSes are phased within the constraints of nucleosome positioning established by the stronger TSSes. In summary, nucleosome phasing appears to be preferentially specified by the strongest TSSes and constrains weaker TSSes within the TID (Figure 4. 13A); the weaker TSSes fine-tune phasing and occupancy of nucleosomes within the TID and beyond.

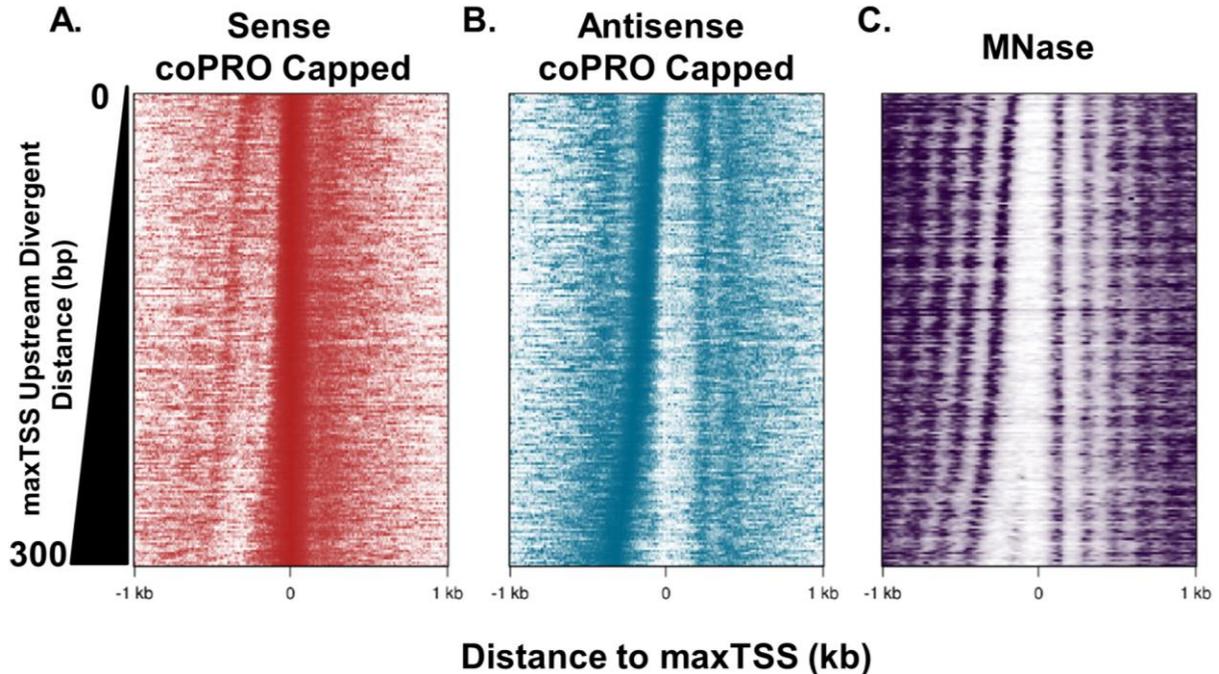


Figure 4. 12: The maxTSS is usually part of a divergent pair
 All stable TSS containing TIDs centered on the bin with the most signal, and sorted by the distance to the antisense coPRO capped bin with the most signal from -300 bp to 0 bp to the maxTSS. A.) Sense coPRO capped. Sense means the strand with the highest activity TSS. B.) Antisense coPRO capped. C.) MNase. Here, more complex organizations, such as convergent and extended TIDs in Figure 4. 10, are randomly distributed along the heatmap.

Spacing between adjacent phased nucleosomes appears to be very consistent: the variability of phasing patterns comes from this ability of new sense TSSes to reset phasing, and the fact that such TSSes arise from divergent pairs where the upstream divergent TSS (the convergent) is constrained by nucleosome phasing from the maxTSS. Thus, the variability in the distance between the divergent pairs within TIDs is strongly associated with the intricate phasing patterns within TIDs. These intricate patterns of nucleosomes around TIDs are only apparent when MNase data are sorted appropriately. For example, the strongest divergent pair in the TID explains much of the nucleosome phasing there. TIDs where other divergent pairs reset phasing are randomly distributed along the continuum of divergent pair distances, so the distance between the main divergent pair appears to explain all phasing when TIDs are sorted this way (Figure 4. 12). However, when MNase data are sorted by the location of other TSSes relative to

the max TSS, as in Figure 4. 13B, the way in which these minor TSSes refine the pattern of nucleosomes around TIDs within the constraints of phasing from major TSSes becomes apparent. In summary, the precise nucleosome organization within TIDs is set by two simple rules. First, TSSes only phase nucleosomes in the downstream direction, so upstream divergent TSSes phase independently. Therefore, the distance between divergent pairs is what determines how phasing emanates from the pair. Second, the strongest TSSes dominate in phasing. This means that downstream convergent TSSes are constrained by phasing from the strongest TSS. The sense TSS paired with the convergent can reset phasing in the downstream direction. Thus, from these two simple rules about nucleosome phasing within TIDs, we can explain the association of complex TID organization with nucleosome organization. Because nucleosomes tend to phase uniformly downstream of individual TSSes, the distance between divergent pairs is the source of complex patterns of phasing. Future work could dissect this relationship: here, we cannot definitively say whether TSSes are the primary determinant of phasing, or to what extent nucleosome phasing is established by DNA sequence composition and the action of a few transcription factors which in turn constrains TSS location. However, the constraint of convergent TSSes, and the ability of downstream sense TSSes to reset phasing implies that it is TSSes which determine the precise arrangement of nucleosomes.

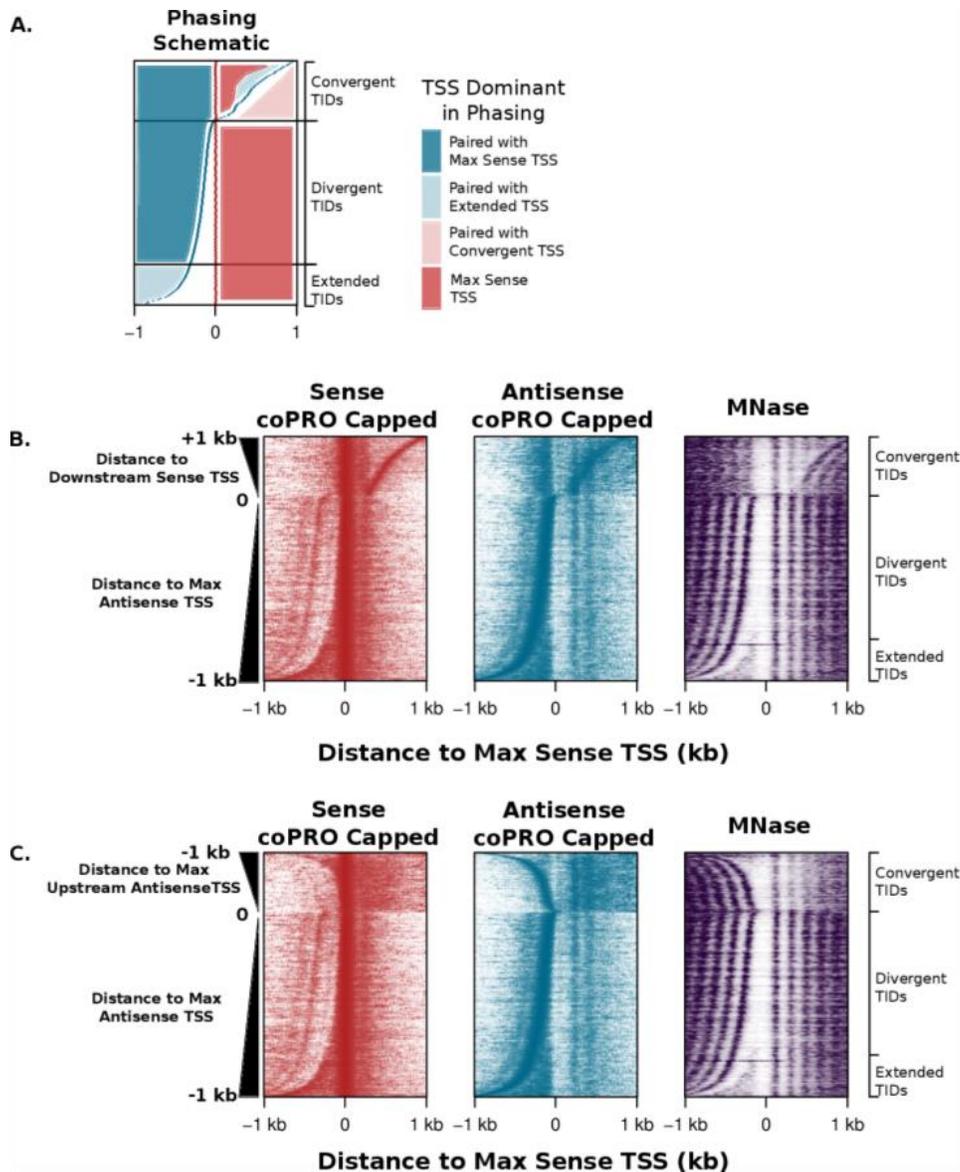


Figure 4. 13: Nucleosome phasing in complex TIDs

Same sites as Figure 4. 10, sorted in different ways to show nucleosome phasing. Proper sorting is critical for visualizing patterns in heatmaps; thus, nucleosome phasing cannot be seen in various regions at the same time, as it is oriented to different TSSes in different regions. $N = 13,731$ (3,444 Convergent TIDs, 7,988 Divergent TIDs, and 2,299 Extended TIDs, labeled to right of C) **A.**) Schematic of phasing. The sense and antisense max TSS position are indicated in red and blue lines. Nucleosome phasing relative to TSSes on the sense and antisense strands is indicated by red or blue shading respectively. Shading is dark when phasing is to the max TSS on that strand, lighter when it is a weaker TSS. **B.**) Sort Convergent TIDs by the most used Sense TSS from 300 to 1000 bp of the max Sense TSS (thus the TSS on the sense strand paired with the strong convergent TSS on the antisense strand). Now, nucleosome phasing beyond the convergent TSS can be seen; these nucleosomes are phased to the sense TSS paired with the strong convergent TSS. **C.**) Sort Convergent TIDs by the upstream divergent TSS to the max sense TSS (max antisense from -1000 to 0 to the max sense TSS). Now, upstream phasing to the upstream divergent is seen.

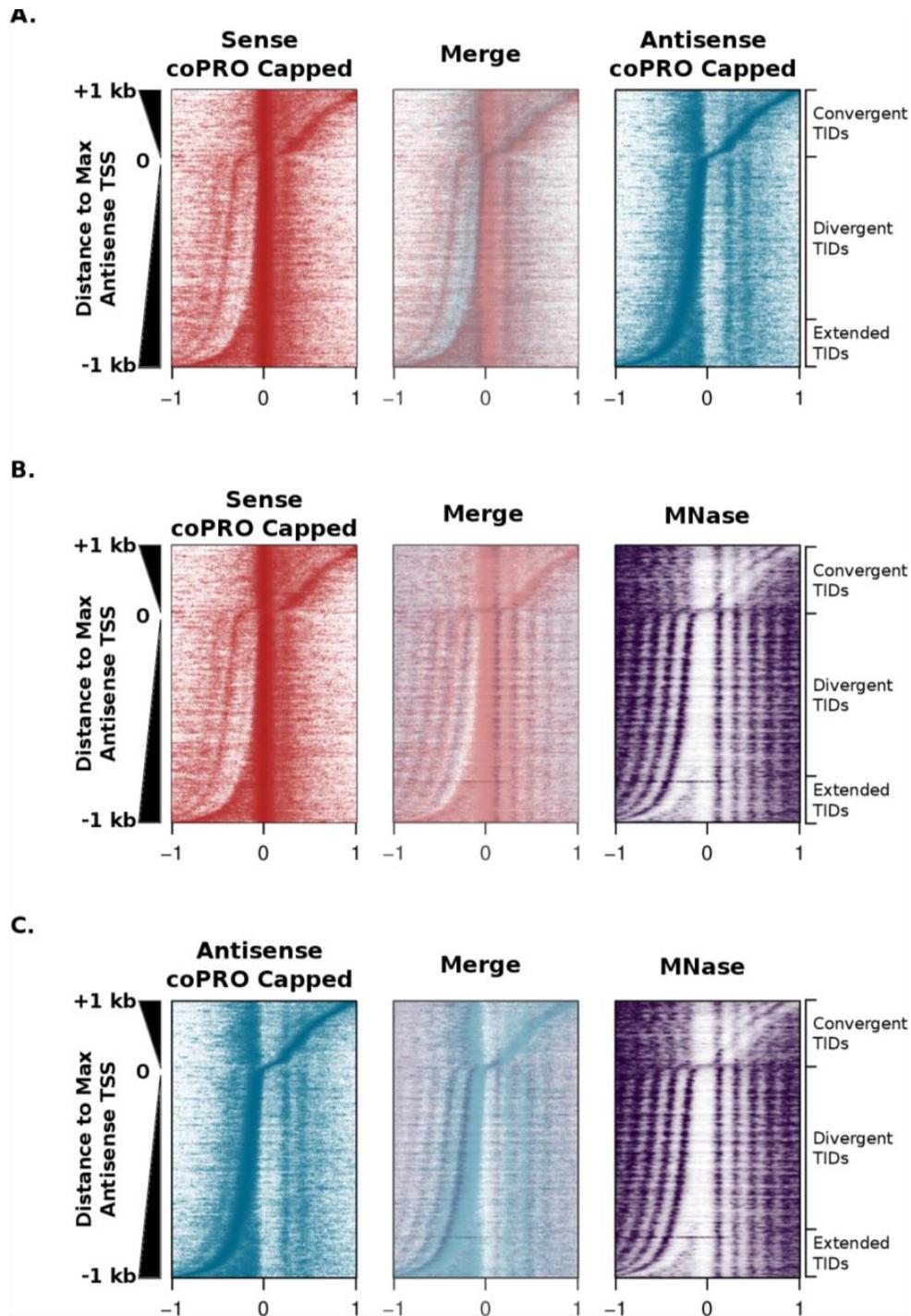


Figure 4. 14: Minor TSSes are restricted to the gaps between nucleosomes phased by stronger TSSes.

A.) Overlay of Figs. 3A and 3B. Divergent pairing at Convergent and Extended TIDs becomes apparent **B.)** Overlay of Figs. 3A and 3C. Phasing of the first downstream nucleosome to the sense Max TSS is apparent. Weak sense TSSes are phased between upstream nucleosomes **C.)** Overlay of Figs. 3B and 3C. Phasing of upstream nucleosomes to antisense transcription in Divergent TIDs is apparent. Weak antisense TSSes are phased between downstream nucleosomes

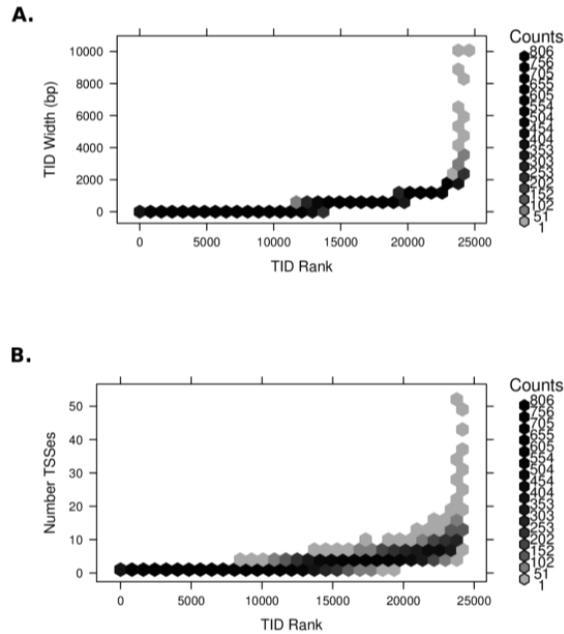


Figure 4. 15: Large TIDs Are Strong Outliers

A.) TID rank plotted against TID size in bp. N = 24,171 B.) TID rank plotted against number of TSSes in the TID. N = 24,171

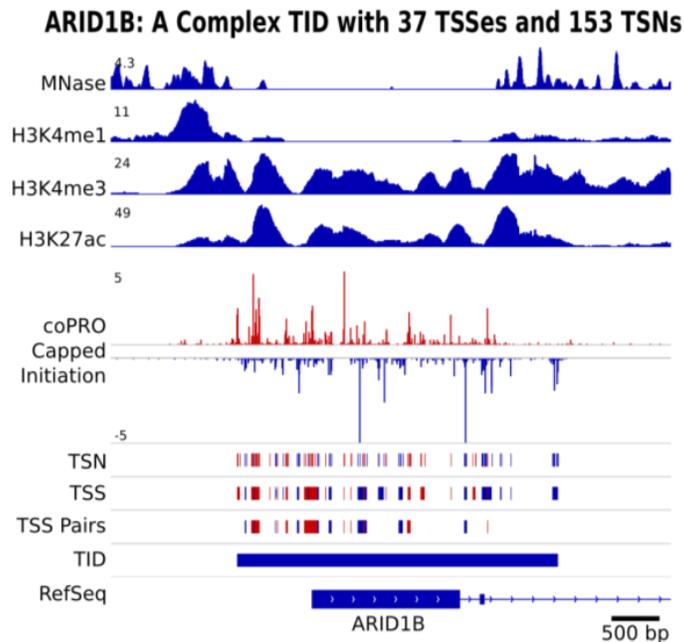


Figure 4. 16: ARID1B, an Exceptionally Large TID

The ARID1B promoter falls within one of the largest TIDs. None of its 37 TSSes have enough CAGE signal to be called as stable. Most of the TID is devoid of MNase (ENCODE), yet is enriched for ChIP signal for active histone marks H3K4me3 and H3K27ac. H3K4me3 extends past H3K27ac, especially into the gene body. H3K4me1 is depleted within the TID, similar to MNase, and enriched at its edges

Gene Promoters Are Found in Large TIDs

Simple divergent pairs are not sufficient to characterize TID architecture at the majority of sites: 42% of TIDs contain >2 TSSes, and 75% of TSSes are within a TID containing >2 TSSes. Furthermore, 77% of TIDs containing at least one stable TSS (Stable TIDs, i.e. promoters) are complex. The largest TIDs are strong outliers, both by length and the number of TSSes that they contain (Figure 4. 15). The ARID1b promoter is among the largest TIDs, spanning 3.4 kbp and containing 37 TSSes (Figure 4. 16), all of which are called as unstable due to a lack of CAGE signal. This extreme case illustrates many features of the chromatin environment observed around TIDs. TSS number is correlated with TID activity (Figure 4. 1D), indicating that, to some extent, more transcriptionally active loci utilize or generate more TSSes. The size distribution of stable TIDs, with a median width of 729 bp (and 4 TSSes), is very different from unstable TIDs (i.e. putative enhancers), with a median width of 39 bp (1 TSS), reflecting their lower overall transcriptional activity (Figure 4. 17 coPRO).

TID Structure and Chromatin Environment Are Tightly Linked

Chromatin environment and transcription are inexorably linked (Karlic *et al.*, 2010), and the comprehensive map of initiation afforded by coPRO provides a framework with which to better understand their interplay. CoPRO is enriched at TID boundaries (oriented outward) and scattered throughout their center, regardless of TID size (Figure 4. 17), a pattern which is corroborated by Pol II ChIP-seq (Figure 4. 19). TID boundaries are strongly enriched for outward transcription, consistent with TSSes generally existing as divergent pairs (sup fig). TIDs are DNase hypersensitive and conserved (PhastCons), with good agreement between boundaries (Figure 4. 18). MNase mapping of nucleosomes shows marked positioning just beyond TID boundaries, with broad depletion and intricate patterning inside (Figure 4. 17).

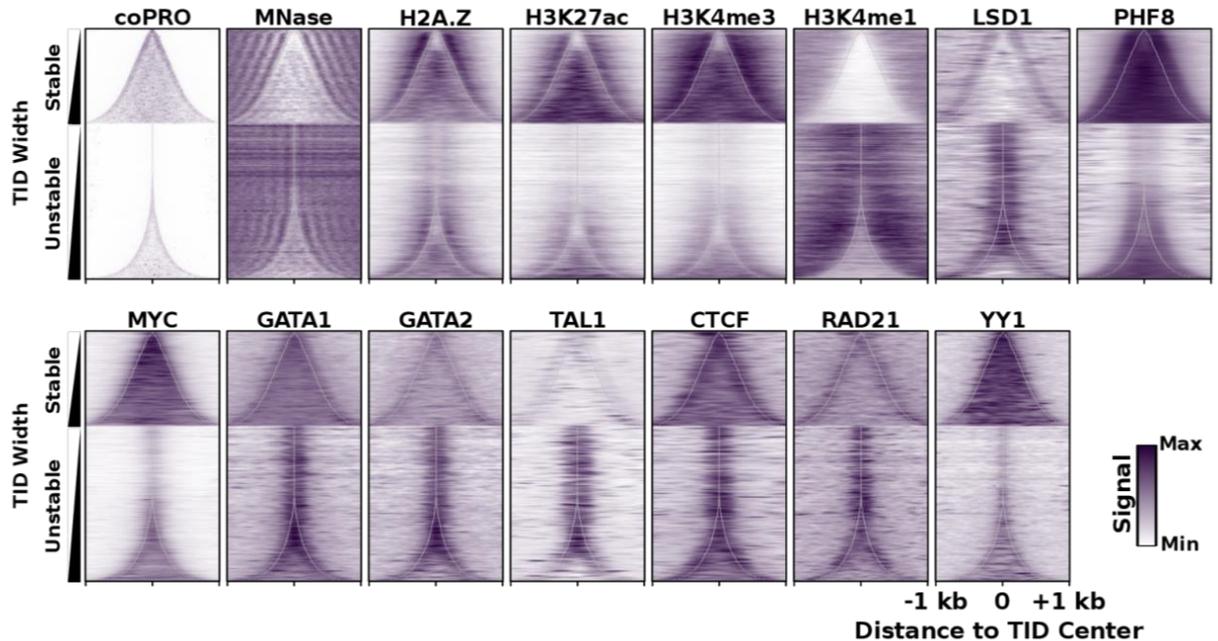


Figure 4. 17: TID organization is linked to chromatin environment

All heatmaps made as described in Methods. TIDs are sorted by width (in bp), and aligned by their center. N = 24,171 TIDs. coPRO Capped initiation is the sum of signal on both strands. Other than coPRO and histone modification ChIP datasets (Roadmap Epigenomics), all others are from ENCODE. Color scale is from the minimum to the maximum bin's signal. Histone modifications are on a linear color scale of processed fold-change values from Roadmap Epigenomics, all others are on a log color scale.

Histone modifications and variants also show diverse patterning relative to TIDs. The active histone marks H3K27ac and H3K4me3 are strongly enriched inside TIDs and just outside (Figure 4. 17), with depletion at the boundary corresponding to the nucleosome free region. Thus, these marks decorate histones that are found within TIDs, and one or two nucleosomes flanking TIDs. To better understand H3K27ac spreading, we use published Native MNase ChIP (Pradeepa *et al.*, 2016). This nucleosome-resolution data strongly suggests that H3K27ac is primarily confined to nucleosomes immediately adjacent to TSSes within TIDs (Figure 4. 20). While nucleosome-resolution data are not available for H3K4me3 in K562, two ChIP datasets show this mark extending past H3K27ac (Figure 4. 17, Figure 4. 20), and data in other cell types show that H3K4me3 peaks tend to be wider than H3K27ac peaks(Heintzman *et al.*, 2007). Thus, it is tempting to speculate that both marks are usually deposited at the first nucleosome outside

TIDs, while the second nucleosome and beyond are more likely to have H3K4me3 and not H3K27ac. Other active marks have distributions within TIDs that are very similar to H3K4me3 and H3K27ac, with variable spreading beyond the TID boundaries. H3K27ac and H3K9ac (Figure 4. 21) are strongly enriched within large TIDs, while H2A.Z (Figure 4. 17) shows stronger enrichment at nucleosomes flanking small, stable TIDs.

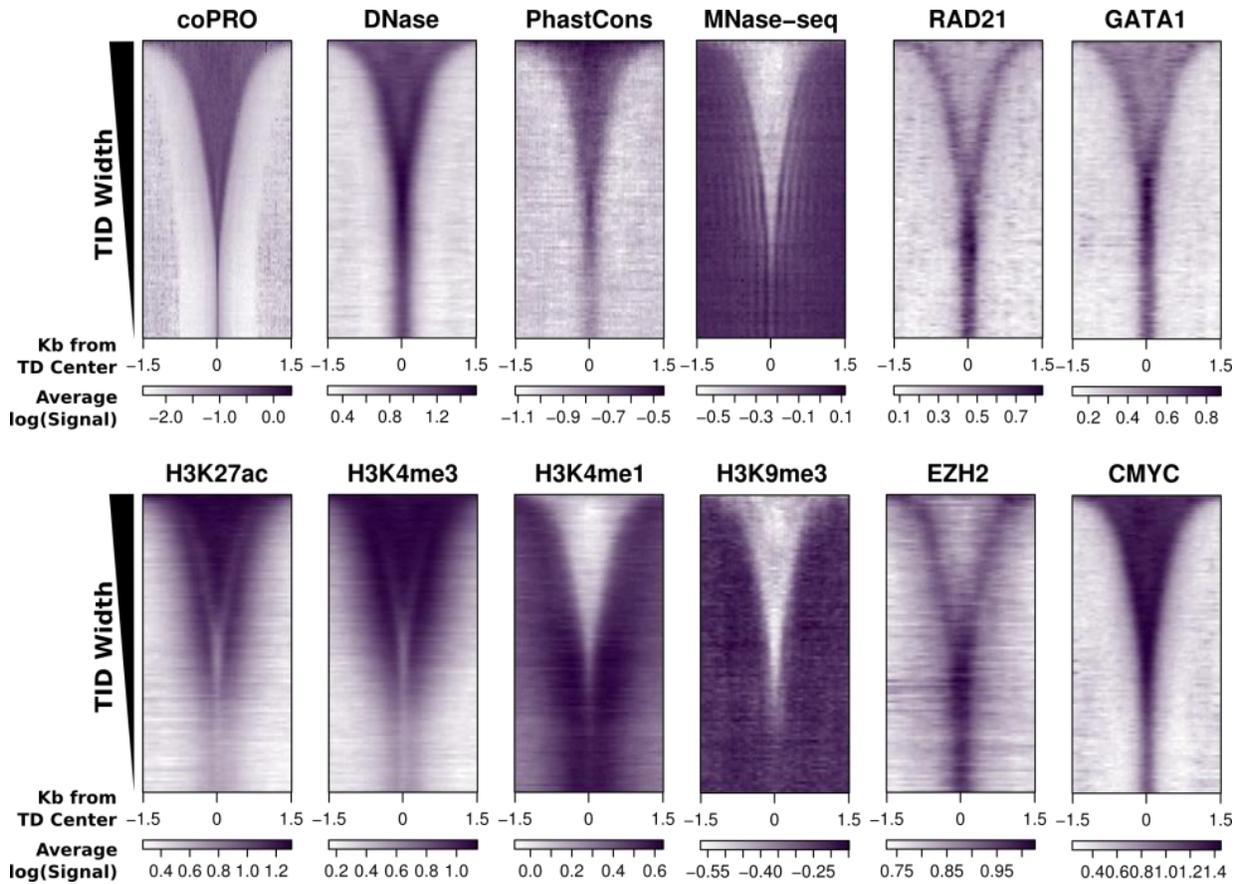


Figure 4. 18: Features of TIDs Irrespective of Transcript Stability

All 24,171 TIDs in the preceding figures, without separating by stability. Here, enrichment at small or large TIDs is more apparent.

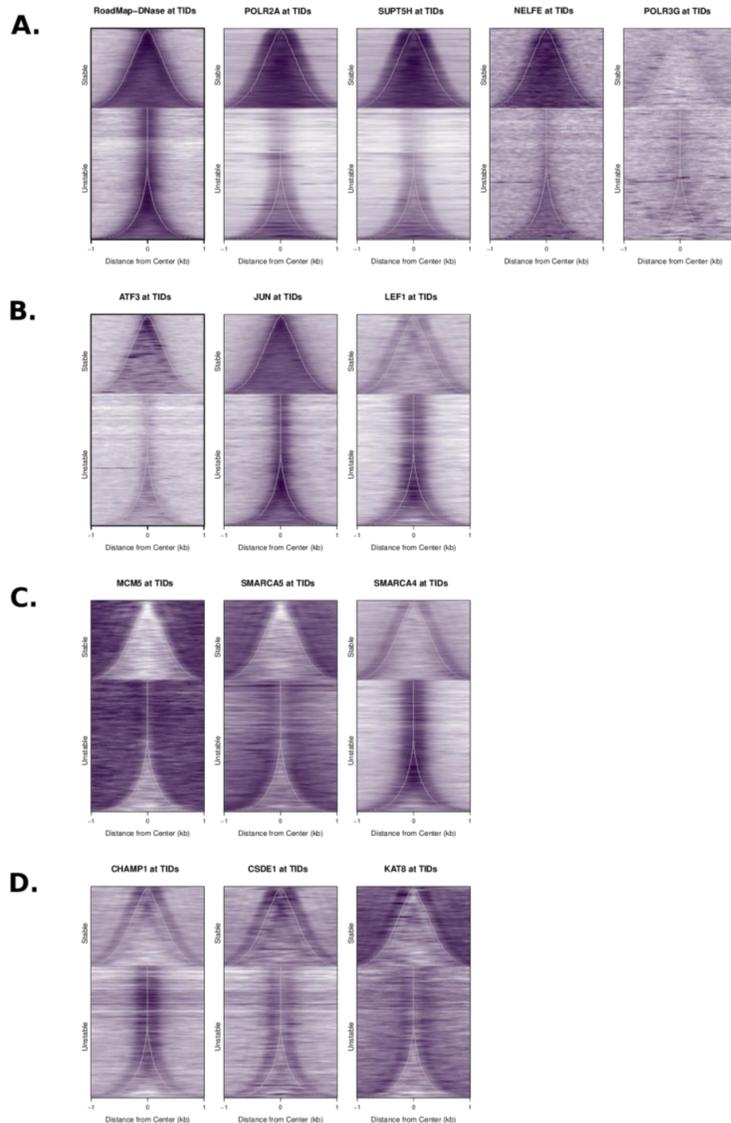


Figure 4. 19: Distinct binding modes within TIDs

A.) Polymerase-like. RNA Pol II is provided as a reference. Strongly enriched in stable TIDs with signal extending past the boundary. Pausing factors DSIF (SUPT5H, Spt5 subunit) and NELF-E show similar distributions. Pol III shows very little enrichment with respect to TIDs called from coPRO. TIDs and sorting identical to Figure 4. 17. (N = 8,568 stable TIDs, N = 13,829 unstable TIDs) **B.)** Transcription factors. ATF3 is enriched preferentially in smaller, stable TSS containing TIDs. JUN is enriched in both stable and unstable TIDs. LEF1 is bound infrequently at stable TID boundaries and frequently within narrow, unstable RNA producing TIDs. **C.)** Stable TID exclusion. DNA replication initiation factor MCM5 is excluded from TIDs, and bound outside. SWI/SNF subunit SMARCA5 is excluded from TIDs, and exhibits some nucleosome-like phasing outside. SWI/SNF subunit SMARCA4 is excluded from the interior of stable TIDs, binds at their boundaries, and is found within short unstable RNA producing TIDs. **D.)** Phased. Kinetochores associating factor CHAMP1 binds at either side of stable TID boundaries, and within unstable TIDs. Cold Shock protein CSDE1 binds at boundaries of both stable and unstable TIDs. H4K16 acetyltransferase KAT8 binds outside of TIDs, with stronger enrichment outside of stable TIDs.

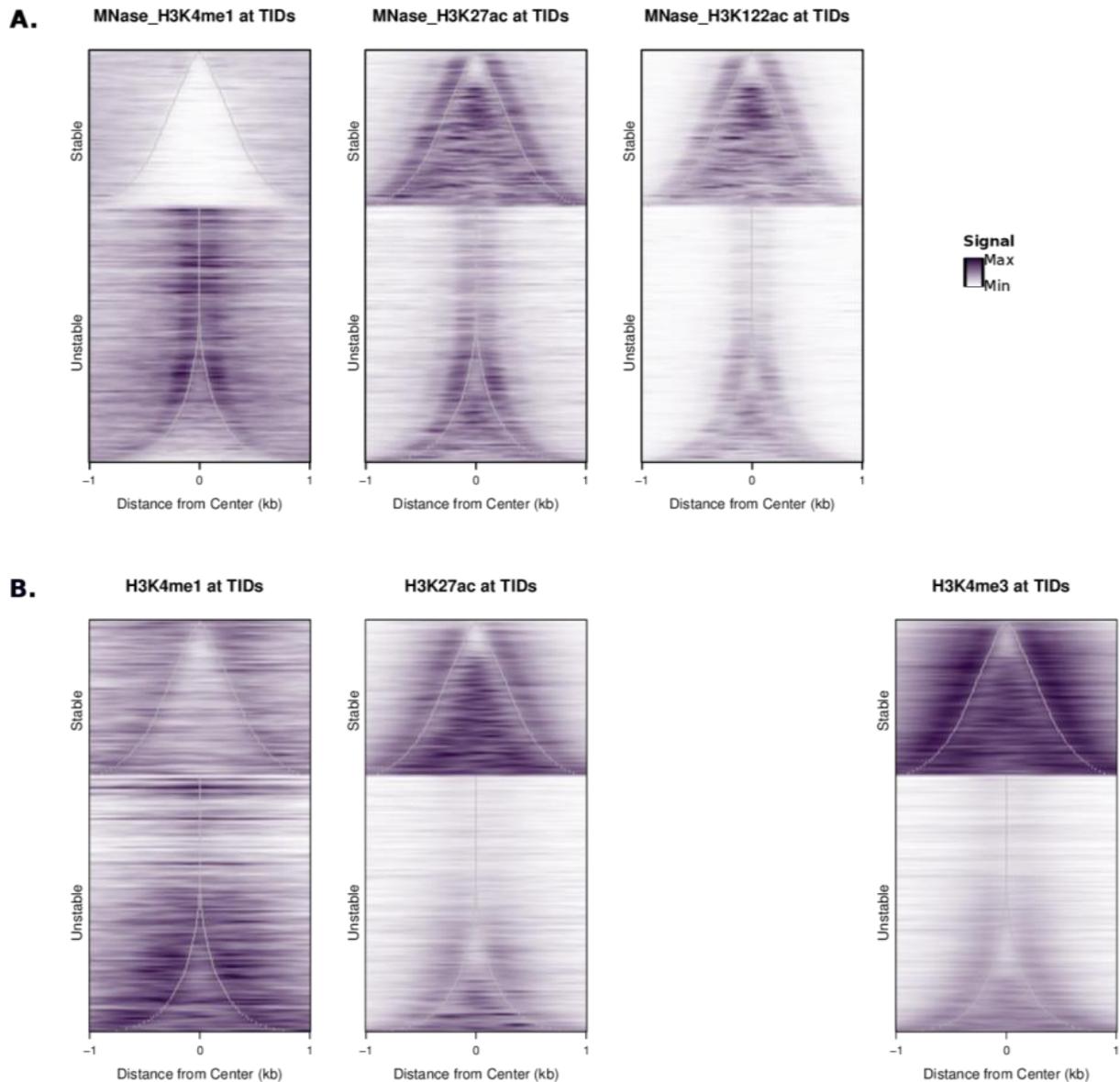


Figure 4. 20: Native MNase ChIP for histone modifications at TIDs reveals tight covariation with transcription initiation that is supported by multiple datasets
A.) Native MNase ChIP for H3K4me1, H3K27ac, and H3K122ac at TIDs in K562 cells (Pradeepa *et al.*, 2016). Order and elements are identical to figure 4 (N = 8,568 stable TIDs, N = 13,829 unstable TIDs). These data were produced without crosslinking, and thus show where the globular domains of nucleosomes with the given mark on their tails are localized.
B.) ENCODE ChIP-seq for H3K4me1, H3K27ac and H3K4me3. Validation of patterns in Figure 4. 17.

Surprisingly, H3K4me1 is almost entirely excluded from TIDs (Figure 4. 17), and instead accumulates outside their borders, with a strong preference for narrow TIDs. Narrow TIDs are more likely to be unstable (Figure 4. 17, coPRO), lowly expressed (Figure 4. 1D), and have a single nucleosome free region (Figure 4. 17, MNase), and thus lack any internal nucleosomes. This strong exclusion from TIDs is more similar to the pattern of repressive marks like H3K9me3 and H3K27me3 (Figure 4. 21) than it is to other active marks. The ratio of H3K4me3 to H3K4me1 has been proposed to distinguish promoters from enhancers(Heintzman *et al.*, 2007); we previously suggested this ratio is correlated with transcription and not exclusively promoter activity(Leighton J. Core *et al.*, 2014)(Pundhir *et al.*, 2016). Here, we reveal the patterns of these histone modifications, and many others, are intimately coupled to TID architecture and provide a novel framework for understanding genome architecture and function.

Enrichment within specific sizes of TIDs is more apparent when data unstable and stable TIDs are combined within a single plot (Figure 4. 18). All TIDs are enriched for coPRO, DNase, cMYC binding, and conservation (PhastCons 100 way). H3K4me1 is enriched outside of small TIDs while H3K4me3 and H3K27ac are enriched inside large TIDs. H3K9me3 is excluded from TIDs. GATA1 is enriched in narrow TIDs (which are more likely to be putative enhancers). Rad21 and EZH2 are enriched in narrow TIDs and at the boundaries of large TIDs. This enrichment at the boundaries of large TIDs is likely due to the fact that they are boundaries; CTCF is the archetypical boundary factor, and EZH2 limits the spread of active chromatin.

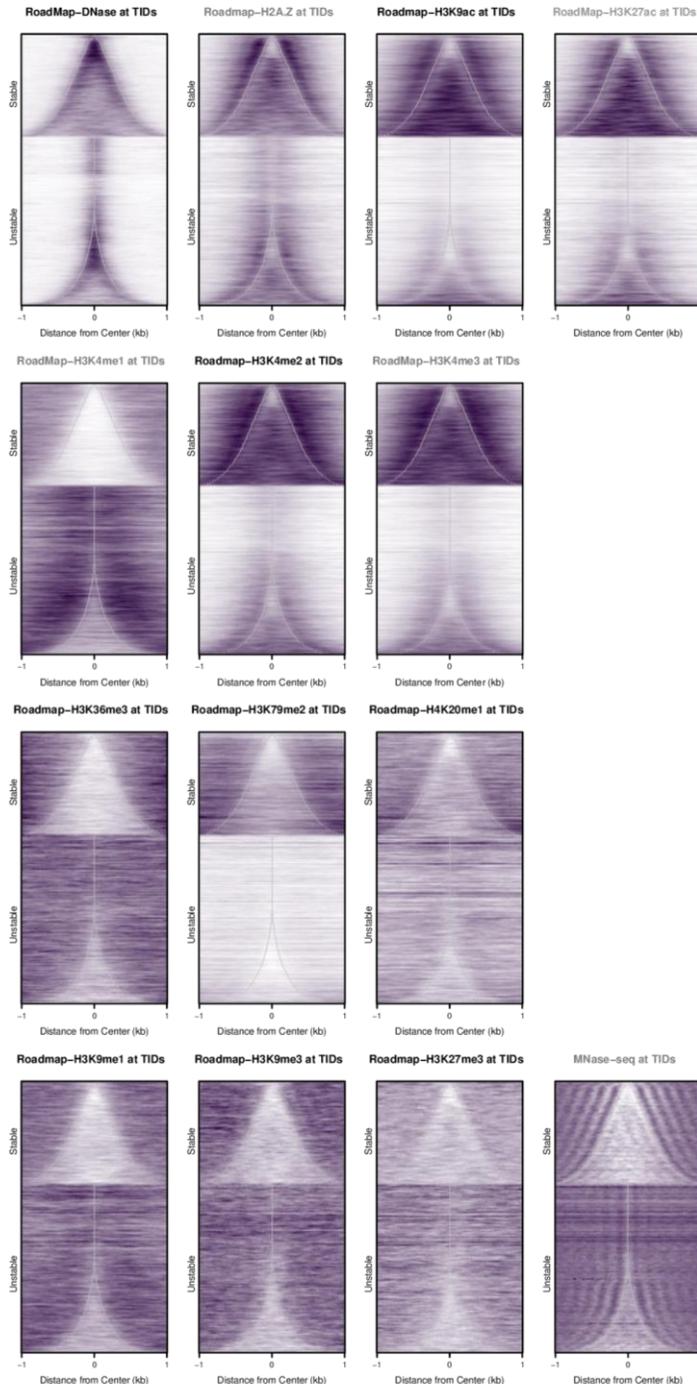


Figure 4. 21: Histone modifications show strong patterns around TIDs.

Heatmaps for all histone marks and DNase from the Roadmap Epigenomics project, with MNase from ENCODE at bottom as a visual aid. MNase is on a log scale (minimum of the matrix to maximum), the rest are linear from minimum to maximum. TIDs and sorting are identical to Figure 4. 17. Plots repeated from Figure 4. 17 indicated by gray title ($N = 8,568$ stable TIDs, $N = 13,829$ unstable TIDs). Active marks are bound within and just outside of TIDs, while repressive marks are depleted within and found outside. Marks of elongation such as H3K36me3 are also depleted within TIDs and found outside.

Transcription Factors' Distribution within TIDs Reflects Their Function

TFs are positioned within TIDs in patterns that belie their function (Figure 4. 19). Histone modifiers such as the H3K4me3 demethylase LSD1, and the H3K9me3 demethylase PHF8 are positioned commensurate with their targeting and activity (Figure 4. 17). Activating TFs such as MYC and ATF3 are enriched within stable TIDs (Figure 4. 17, Figure 4. 19). TFs associated with the transcription machinery like TBP are distributed within TIDs and accumulate moderately at their edges, similar to Pol II (Figure 4. 19). The lineage-specific TFs GATA1, GATA2, and TAL1 (Figure 4. 17) are enriched at the boundaries of small, unstable TIDs (putative enhancers). The chromatin looping factors CTCF and cohesin (RAD21) are markedly enriched at TID boundaries (Figure 4. 17), consistent with their known role in establishing chromatin boundaries. In contrast, a TF recently implicated in mediating activating distal interactions, YY1 (Beagan *et al.*, 2017; Weintraub *et al.*, 2018), is strongly enriched within TIDs (Figure 4. 17).

Possible Functional Roles of TIDs in Gene Regulation

TIDs have the seemingly paradoxical properties of low nucleosome occupancy while containing abundant active histone modifications. Internal TID nucleosomes may be too labile for efficient detection by MNase-seq, or highly modified despite low occupancy. H3K4me3 and Pol II are known to engage in a feed-forward loop that may help to explain TID creation and maintenance. The propensity of TSSes to cluster into TIDs likely serves to mark or sustain chromatin modification and TF binding, or may even drive phase-separation. The presence of disordered domains (histone tails, Pol II CTD, and Cyclin T1) with a high potential for trans-activating interactions may create an active compartment around the stable TSSes within large TIDs. This phase separation could also facilitate their co-localization with enhancers, as active

compartments are postulated to form phase separated liquid droplets within nuclei(Hnisz *et al.*, 2017). Phase separation occurs when many active enhancers and promoters are in close spatial proximity, and each constituent element is associated with many macromolecules with several multivalent interaction domains that form a network of macromolecular interactions. Similarly, TIDs form a chromatin compartment around TSSes, so that promoters are best described as the clusters of TSSes included within TIDs on a systems level. The special relationship between TID boundaries and chromatin compartments is seen across a wide range of TID widths, indicating that this phenomenon occurs across most of the spectrum of transcribed domains. Thus, the array of minor TSSes that we find near stable promoters, including the upstream divergent TSS, could serve to mediate this phase separated behavior and to precisely tune the availability of factors to stable RNA producing TSSes within large TIDs. Together, these many TSSes likely exert a level of transcriptional control that would be impossible with a single divergent pair, without necessarily relying on distal enhancers for fine tuning of expression. There must be an upper limit to the information content that a single divergent pair of TSSes can contain, so larger TIDs could serve to increase the sequence space that is available for regulation of promoters, akin to building a giant satellite dish to receive critical but subtle signals rather than relying on a standardized small dish. This is very similar to the way in which enhancers expand the regulatory potential at promoters by increasing the information content there. Consistent with these ideas, broad H3K4me3 domains have been found to be located at promoters of genes critical for cell identity, and to be associated with less cell to cell variability in transcript level(Benayoun *et al.*, 2014; Chen *et al.*, 2015). Our highly sensitive map of initiation allows us to link these domains to TID structure, as the many unstable TSSes present within these domains evaded detection in earlier work.

Methods

coPRO Experiments

K562 cells were obtained from ATCC and cultured antibiotic-free in accordance with their standards in DMEM, high glucose + HEPES (ThermoFisher cat. 12430054). Cultures were verified to be mycoplasma-free(Young *et al.*, 2010) prior to library preparation and sequencing. Two biological replicates were cultured independently, separated by two passages, with library preparations done separately for technical and biological replicates. See Figure 4. 22 for an overview of coPRO. Cells were permeabilized, and run-on reactions with all four biotin NTPs were carried out with 20 million cells per reaction as described previously(Dig Bijay Mahat *et al.*, 2016). After isolating RNA from the run-on with Trizol (ThermoFisher, cat. 10296028), three run-ons per biological replicate were pooled, and cap state specific RNA spike-ins were added to the pool (see below). Two adapter ligations and reverse transcription were performed as described(Dig Bijay Mahat *et al.*, 2016), with custom adapters detailed in Table 4. 1. Critically, no RNA fragmentation was done so that the pairing of RNA 5' and 3' ends remains biologically meaningful. The first ligation adds a sample barcode to the library; TruSeq barcodes were chosen to minimize predicted secondary structure of the adapter(Zuker, 2003). Between adapter ligations, cap state selection reactions were performed. The three cap state selections use a series of enzymatic treatments to reduce specific populations of RNAs to 5' monophosphate, making them capable of ligation to an RNA adapter by T4 RNA ligase (NEB, cat. M0204S). All steps were carried out following the manufacturer's protocol, with phenol:chloroform extraction and ethanol precipitation between steps.

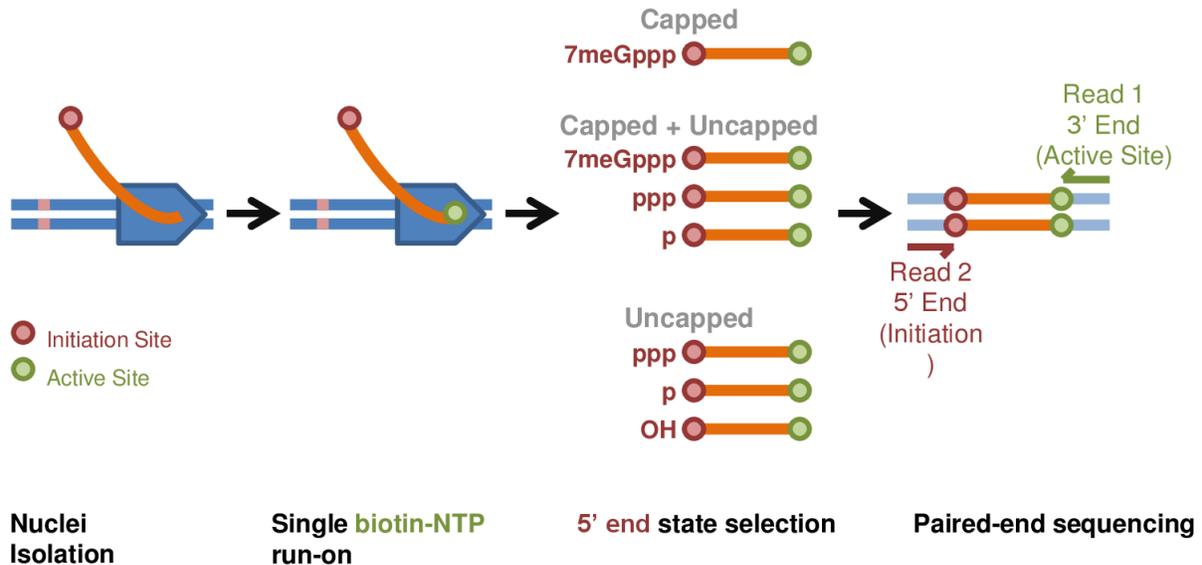


Figure 4. 22: Schematic of coPRO

coPRO is similar to a PRO-seq experiment (Kwak *et al.*, 2013), with three major differences. First, there is no RNA fragmentation step, so initiation site remains coupled to active site (detectable insert size is restricted by the DNA length limit of the Illumina platform (Fan *et al.*, 2010)). Second, we select for specific 5' end capping states. Third, coPRO uses paired end sequencing to map both initiation site and polymerase active site (pause) with basepair resolution.

Overall workflow is as follows: **i.)** K562 cells are rapidly cooled and permeabilized, washing away native nucleotides and preserving RNA polymerase in place **ii.)** Run-on with only biotin-11-NTPs in the presence of sarkosyl. Polymerases that were engaged when cells were isolated will incorporate a single biotin nucleotide, labeling the location of the active site of polymerase. The biotin is used for affinity purification three times between subsequent enzymatic steps. **iii.)** Enzymatic selections for different capping states (Figure 4. 23). Ligation of the 5' adapter for library construction requires a 5' monophosphate RNA, so all treatments are designed to specifically reduce a targeted population of RNAs to 5' monophosphate. Capped selection yields pure capped RNA (regardless of cap methylation state). Capped + Uncapped can incorporate any capped RNA, 5' triphosphate (pppRNA) or 5' monophosphate (pRNA). Uncapped captures pppRNA, pRNA, and 5' hydroxyl RNA (OHRNA). pRNA is a product of nucleolytic cleavage during termination and is thus not observed with the same 5' end created when initiation occurred. Therefore, throughout this work, we specifically observe pppRNA by calling sites of initiation from the capped library and then looking for RNAs from the uncapped library that share 5' ends with these capped RNAs. **iv.)** Strand specific RNA-seq library construction and paired end sequencing.

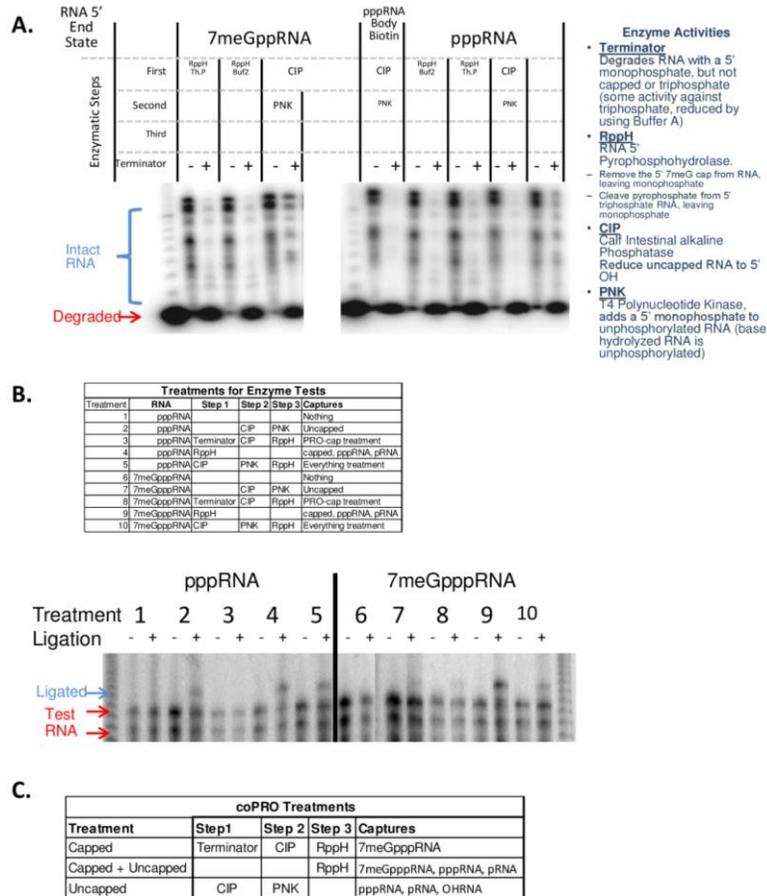


Figure 4. 23: Enzymatic steps for cap state selection in coPRO

A.) Test of enzymatic steps using degradation by terminator as a readout. All experiments use P³² body labeled RNA. In each pair of lanes, the RNA is subjected to the indicated series of enzymatic reactions, and then, the RNA is split and subjected to a terminator exonuclease treatment or mock treatment (without enzyme), and equal amounts are loaded onto a 7M urea 8% polyacrylamide gel. Enzymes' activity is listed to the right. 7meGpppRNA was capped using the vaccinia capping system (NEB cat. M2080S); separate tests showed that this is ~80% efficient. pppRNA is simply T7 transcribed RNA. pppRNA body biotin was labeled with biotin by substituting 1/2 of the CTP in the T7 transcription reaction with 14-Biotin CTP. Intact RNA indicated in brackets, degradation products (free NTPs) labeled at the bottom of the gel. Capped RNA is rendered sensitive to terminator after decapping with RppH, regardless of buffer used (NEB buffer 2 is supplied with the enzyme, NEB ThermoPol buffer is recommended for decapping). Capped RNA is much less sensitive to degradation than decapped. Most of the degradation here is due to incomplete capping by vaccinia enzyme. RNA body labeled with biotin is still sensitive to terminator; thus, this approach was used to label spike-ins during coPRO library preparation. Uncapped RNA is much more sensitive to terminator degradation after CIP and PNK treatment than capped. This treatment is designed to specifically reduce uncapped RNA to monophosphate. Labeled 10 bp ladder is in the first lane of each gel.

B.) Similar to a.), using ligation by T4 RNA ligase as the readout. If the series of enzymatic steps reduced the RNA to monophosphate, a 20 nt RNA adapter was able to be ligated (indicated to left of gel). 12% acrylamide sequencing gel. Reactions are similar to final steps used for coPRO cap selections.

C.) Final enzymatic steps used for coPRO

We designed three separate 5' state selections (Figure 4. 23):

- 1.) Uncapped RNAs were selected by treating with CIP, Calf Intestinal alkaline Phosphatase (NEB, cat. M0290S) and PNK, T4 Polynucleotide Kinase 3' phosphatase minus (NEB, cat. M0236S), in order to reduce all uncapped RNAs to 5' hydroxyl and then add 5' monophosphate, but without removing the cap from capped RNA.
- 2.) Capped RNAs were selected by treating with Terminator 5' Phosphate dependent exonuclease (Epicentre, cat. TER51020) to degrade 5' monophosphate RNA (from terminating polymerase) and CIP to reduce other uncapped RNAs to 5' hydroxyl, making them incapable of ligating to 5' adapter. 5' cap was removed with RNA 5' pyrophosphohydrolase, RppH (NEB, cat. M0356S), using ThermoPol buffer (NEB, cat. B9004S).
- 3.) Capped + Uncapped RNAs were selected by treating with RppH, reducing both capped RNA and pre-capped RNA (5' triphosphate) to 5' monophosphate.

Each library was sequenced on a NextSeq 500 as both cDNA (no PCR), and after PCR amplification and PAGE purification. See Figure 4. 24 for a schematic of library design.

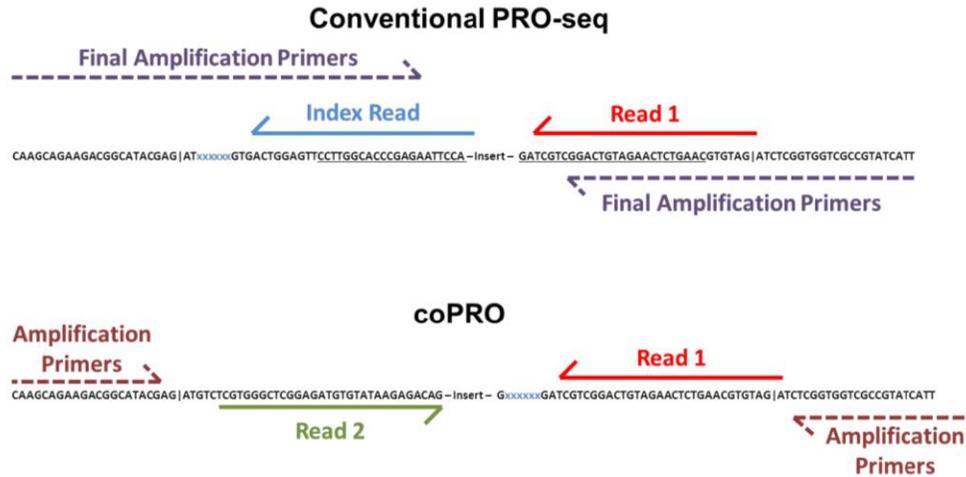


Figure 4. 24: coPRO library design schematic

A.) Conventional PRO-seq library design. TruSeq small RNA primer is used for sequencing. A separate index read is carried out on the same strand of DNA. Primers used in the final PCR amplification are indicated with a dashed line. **B.)** coPRO library design. Here, most of the adapter is added with the RNA ligation, rather than PCR as with conventional coPRO. This library design is compatible with PCR-free sequencing (for this work, half of the sequencing was done PCR-free, with just cDNA, and half after PCR amplification). The oligo used for final amplification is indicated with a dashed line. Read 1 uses the TruSeq small RNA sequencing primer, with the index sequenced in-line with read 1. Read 2 uses the Nextera read 2 sequencing primer.

Table 4. 1: Oligonucleotides used for coPRO

Purpose	Name	Sequence	Library
3' Adapter	PCRfreeRC3_BR2	/5Phos/rGrCrGrArUrGrUrGrArUrCrGrUrCrGrGrArCrUrGrUrArGrArArCrUrCrUrGrArArC/3InvdT/	Uncapped, Rep 2
	PCRfreeRC3_BR3	/5Phos/rGrUrUrArGrGrCrGrArUrCrGrUrCrGrGrArCrUrGrUrArGrArArCrUrCrUrGrArArC/3InvdT/	Capped, Rep 2
	PCRfreeRC3_BR4	/5Phos/rGrUrGrArCrCrArGrArUrCrGrUrCrGrGrArCrUrGrUrArGrArArCrUrCrUrGrArArC/3InvdT/	Capped + Uncapped, Rep 2
	PCRfreeRC3_BR6	/5Phos/rGrGrCrCrArArUrGrArUrCrGrUrCrGrGrArCrUrGrUrArGrArArCrUrCrUrGrArArC/3InvdT/	Uncapped, Rep 1
	PCRfreeRC3_BR8	/5Phos/rGrArCrUrUrGrArGrArUrCrGrUrCrGrGrArCrUrGrUrArGrArArCrUrCrUrGrArArC/3InvdT/	Capped, Rep 1
	PCRfreeRC3_BR9	/5Phos/rGrGrArUrCrArGrGrArUrCrGrUrCrGrGrArCrUrGrUrArGrArArCrUrCrUrGrArArC/3InvdT/	Capped + Uncapped, Rep 1
5' Adapter	5Adapt_Paired_ePRO	rCrArArGrCrArGrArArGrArArCrGrGrCrArUrArCrGrArGrArUrGrUrCr - UrCrGrUrGrGrGrCrUrCrGrGrArGrArUrGrUrGrUrArUrArArGrArGrArCrArG	All
Reverse Transcription	RPI	AATGATACGGCGACCACCGAGATCTACACGTTTCAGAGTTCTACAGTCCGA	All
Final Library PCR	PairEPRO_For_FinalAmp	CAAGCAGAAGACGGCATAACGAGATGTC	All
	PairEPRO_Rev_FinalAmp	AATGATACGGCGACCACCGAGATCTAC	All

Cap State Spike-ins

Spike-in RNA for different cap states were added to calculate enrichment between the different coPRO cap state treatments. RNA was made using home-made T7 RNA polymerase and buffer (30 mM HEPES pH 7.8, 80mM Potassium Glutamate, 15mM MgAc, 0.25 mM EDTA, 5 mM DTT, 0.05% Tween-20, 2 mM Spermidine), with YIPP (NEB cat. M2403S) and Suprase-In (ThermoFisher cat. AM2696) added as per the manufacturer's protocol. 2.5 mM ATP, GTP, and UTP, 1.25 mM CTP, and 1.25 mM Biotin-14-CTP (ThermoFisher cat 19519016)

were used in the transcription, so that the spike-in could be captured with the streptavidin pull-downs used in coPRO library preparation. Full-size RNA products were purified from a 7M urea 8% polyacrylamide gel. Capped spike-in was then capped using the vaccinia capping system (NEB cat. M2080S), and remaining uncapped RNA was removed by treating with CIP, PNK and then Terminator. 5' triphosphate spike-in did not require any treatment after gel purification as the product of T7 transcription has a 5' triphosphate.

For tests of coPRO treatments, radiolabeled RNA was made similar to the spike-in by incorporating P³² into the body of the RNA during transcription. Capped RNA was made by vaccinia capping, but without subsequent uncapped clean-up (so some uncapped background remains). Triphosphate requires no additional steps. A radiolabeled, biotinylated RNA was made to test terminator degradation of biotin labeled RNA. The ability of different series of treatments to selectively reduce capped and uncapped RNA to monophosphate was assessed in two ways: first by using Terminator degradation as readout as it requires a 5' monophosphate for its exonuclease activity just as adapter ligation requires a 5' monophosphate in library preparation (Figure 4. 23A), and second by using ligation of the 5' adapter from a standard PRO-seq (Dig Bijay Mahat *et al.*, 2016) as the readout (Figure 4. 23B).

Sequence alignment

Adaptor sequences were trimmed from paired-end reads with the 'cutadapt' toolkit. Internal barcodes were used to de-multiplex pooled libraries with a custom Python script. The remaining sequences were aligned with the bowtie2 --very-sensitive option, which greatly improved alignment of very short RNAs (<25 nt). We specific -X 1000 (maximum insert size), -

-no-mixed (discard unpaired reads), --no-discordant (no alignments > 1 kbp apart), and --no-unal (discard unaligned pairs). Alignments were performed against a pooled genome index that contained dm6, hg19, and spike-in RNA sequences. For most analyses, we only use reads shorter than 400 bp to avoid complications from co-transcriptional splicing.

Read summarization and normalization

Alignments (e.g. BAM files) were processed with a custom R script to summarize alignments sharing identical genomic coordinates (start and end) with an individual ‘count’ score. Normalization factors were computed as follows:

$$W_{C \cdot R} = \frac{R_{capped}}{C_{capped}} \quad \text{and} \quad W_{U \cdot R} = \frac{R_{uncapped}}{U_{uncapped}}$$

where WCR is the weight applied to Capped reads, and WUR is the weight applied to Uncapped reads. We validated this normalization scheme by quantifying the agreement between Capped, Uncapped, and RppH reads genome-wide.

Reads were additionally weighted to correct for the known length bias of Illumina sequencers using previously reported weights (Fan *et al.*, 2010).

Defining transcription start nucleotides (TSNs), start sites (TSSs), and initiation domains (TIDs)

Start bases were identified as any 5' end associated with at least 5 distinct 3' ends shorter than 60 nt, thus enriching for RNA polymerase pausing independently of total TSS activity and capping efficiency. Nevertheless, we removed TSNs with less than 33% capped reads longer than 100 nt, as these were predominantly Pol II termination or Pol I and Pol III transcription products. Sites with more than a 5-fold difference between the RppH and Capped treatments were removed (likely enzymatic bias). Chromosome M was removed from all analyses.

Using these candidate TSNs, TSSs were defined as TSN clusters on the same strand with no gaps larger than 60 nt. Similarly, TIDs were defined as TSS clusters on either strand with no gaps larger than 750 nt. Single-stranded TIDs with >90% total uncapped reads were removed as they were predominantly found in regions with high levels of terminating polymerase. Only TSS and TSNs associated with a valid TID were retained for further analysis. Oppositely oriented TSSes < 300 bp apart (oriented outward) were paired as TSS pairs.

TSNs were called as stable if at least eight CAGE reads overlapped them, as done previously (Leighton J Core *et al.*, 2014). TSSes were called as stable if they contained at least one stable TSN.

Some analyses are restricted to “maxTSN” or “maxTSS” to minimize effects from other nearby elements. We defined maxTSN as the TSN with the highest number of Capped reads within a TSS. Similarly, maxTSS are defined as TSS with the highest number of Capped TSN reads on each strand of a TID.

Pause classification

First, to determine whether the 60 nt window used to define TSNs affected pause profiles, we extended the window to 120 nt for TSN identification. The pause distributions from this set of TSNs are still heavily biased for reads <60 nt.

To better understand pause behaviors, the pause probability within the first 100 nt was computed for each maxTSN. Analysis of the individual TSNs' pause distributions revealed that pausing is stochastic and mostly occurs on 1-10 major positions at mTSNs (Figure 4. 8). Further analysis of the max pause positions revealed a bimodal distribution (Figure 4. 4A). To classify individual distributions as Early or Late pause, the probability of pausing anywhere between 20-32 nt was computed and divided into quartiles. Early is the first quartile (highest probability of pausing between 20-32 nt) and late is the fourth quartile (highest probability of pausing between 33-60 nt). This classification is robust to experimental noise, as classes were consistent across our experimental replicates.

In several plots, we sort by the center of the pause (Figure 4. 4B, Figure 4. 11C, D). The strategy here is the same as CDF area for quantifying shifts in pausing, as in Figure 3. 12. To calculate this, we normalize TSNs so that the sum of all reads from 20 to 60 is 1. The cumulative sum at each nt within the pause is calculated by scanning from 20 to 60. The sum of this cumulative density function (CDF), or rather the area under the CDF curve, quantifies how skewed the distribution is toward the beginning (high area) or end (low area) of the pause window.

Metaplots and heatmaps

All metaplots in this work show a bootstrapped estimate of average signal from the sites being summarized, along with 87.5% and 12.5% confidence intervals. Briefly, this is done by taking 1000 random samples of 10% of the data, and calculating the median and confidence intervals from the averages of each of these 1000 samples.

Heatmaps in this work summarize sorted data into 200 lines averaging every $N/200$ rows to produce a representative heatmap. This was initially developed for coPRO data, as individual TSNs' stochastic pause profiles are sparse, and was subsequently used for all other data for which heatmaps are shown. The rationale here is that the resolution of the files used for figures, and the screen or printer used to display them often would only allocate at best a few hundred pixels to the heatmap, so it is best to intentionally bin data rather than allowing binning to occur by default.

Chapter 5: General Conclusions and Future Directions

General Conclusions

My work in as a graduate student all served the goal of understanding how gene regulation is carried out, with a focus on the mechanism of promoter proximal pausing. In the pursuit of this, I carried out both extremely targeted studies of the role of NELF and extremely broad analyses of transcription with coPRO. Thus, all of this work fits somewhere along the spectrum of functional levels of gene regulation outlined in the Introduction. HiTS-RAP is a highly targeted technique: the goal of a HiTS-RAP experiment is to understand how an RNA binding protein interacts with RNA. This quantitative measure of a single set of interactions constitutes one small piece of the broader regulatory puzzle. Expressing NELFapt to explore the importance of interactions between NELF-E and nascent RNA in pausing is a broader study. Unlike HiTS-RAP, where one feature is interrogated essentially without other moderating influences, this necessitated the inclusion of many other factors to understand the effects on pausing. Finally, the coPRO study spanned virtually all of the functional levels of gene regulation. Some targeted analyses were highly informative, such as asking what the sequence determinants of pausing are and where transcripts become capped. The interplay between pausing and capping involved several variables. Finally, analyses of TIDs was a systems level question: we asked what is the fundamental unit of sites of transcription initiation and saw that, from the perspective of chromatin, it is a large clusters of TSSes that are largely grouped into divergent pairs. Biology makes the most sense when a result is put into this framework. Is this telling me about a specific interaction? What does that mean for chromatin? How does that feed back to the promoter? How does this affect other nearby loci? The general strategy of

observation, perturbation, re-observation is most effective when the question being addressed is contextualized in this way.

Future Directions

HiTS-RAP

Interactions between RNA and protein are a pervasive avenue of gene regulation, and noncoding RNAs are postulated to play important roles in gene expression(Lee, 2012; Kim and Shiekhattar, 2016). Thus, large-scale characterizations of interactions between key regulatory proteins and RNA could provide valuable biological insights. As a new graduate student, I was fascinated by these potential modes of regulation. The noncoding RNA field was at its zenith: enhancer RNAs had recently been discovered, and new roles were being found for noncoding RNAs. We were particularly excited about a study showing that eRNAs may help activate their target loci through interactions with Mediator(Lai *et al.*, 2013). Enhancer RNAs would later be shown to have functional interactions with other factors critical for the regulation of transcription, such as NELF(Schaukowitch *et al.*, 2014) and the histone acetyltransferase P300(Bose *et al.*, 2017). However, for every study that found a function for eRNAs, others find that they are not functional as RNAs but are rather a byproduct of activity of enhancers(Hah *et al.*, 2013; Young *et al.*, 2016). Furthermore, many long noncoding RNA promoters act as potent enhancers(Engreitz *et al.*, 2016), raising the possibility that many lncRNAs are simply enhancers that have evolved stability and slicing, and are not functional as RNAs. In fact, there is now considerable debate about whether HOTAIR, one of the first discovered and most well-known lncRNAs(Rinn *et al.*, 2007), is actually functional(Amândio *et al.*, 2016). Further complicating matters is the fact that the rules for RNA binding specificity are different than DNA binding

specificity, so much of the logic for understanding interactions with DNA do not translate well to RNA. RNA-protein interactions often have a strong structural component, and thus cannot be neatly summarized as a short motif as DNA binding interactions can, as was the case with both the GFP and NELF aptamers. Furthermore, many RNA binding proteins bind with less specificity than we are accustomed to from DNA-protein interactions. In fact, the most specific interactions with RNA that come to mind are microRNA interactions, which are mediated by basepairing with other RNAs(Agarwal *et al.*, 2015). Another example is riboswitches, which are natural RNAs that have high affinity for small molecules, and can induce structural changes that help regulate metabolism related to the target(Serganov and Nudler, 2013). Defining specificity is sometimes problematic as well(Jankowsky and Harris, 2015): for example, the tRNA processing factor C5 from *E. coli* binds RNA with little sequence preference, but its catalytic activity does have strong sequence preference(Guenther *et al.*, 2013). In summary, while there absolutely are functional RNAs in gene regulation that have specific and important interactions with proteins, I decided that it was unlikely that I would discover these interactions with HiTS-RAP.

Despite these problems in interpreting the relevance of biological RNA-protein interactions, there are some cases where a HiTS-RAP experiment measuring the affinity of genome-derived RNAs to a protein would be very illuminating. However, the expectation is different than what I was hoping for eRNAs and lncRNAs near the beginning of my PhD. I initially hoped that such measurements of binding specificity, similar to the high-throughput measurements of transcription factor binding specificity mentioned in the Introduction(Jolma *et al.*, 2013, 2015), could be used to quantitatively model gene regulation. If the strength of each interaction in a system is known with high precision, then we could predict the outcome given a

few starting parameters. I do not believe, however, that there are many RNA-protein interactions where this would be possible. I think that it is unlikely that one of the protein factors relevant to transcriptional regulation, like NELF or Mediator, has highly specific interactions with a set of RNAs that play an interesting role in gene regulation. But, I do think that factors like NELF that bind RNA as they function normally could have preferences for certain RNAs that would be informative in understanding pausing. One feature of HiTS-RAP is that any Illumina sequencing library could be adapted for it (such as ATAC, CHIP, PRO-seq, RNA-seq, etc). I once had plans to use a PRO-seq (or better yet, PRO-cap) library for HiTS-RAP against NELF. This would allow us to determine whether NELF interacts with nascent RNA with some preference. Furthermore, I could ask whether NELF has higher affinity to RNA before or after pause escape by comparing PRO-RAP (affinity to the paused RNA... it already has an acronym) and GRO-RAP (longer run-on simulates pause escape, in the RNA length). While I do not think we would see orders of magnitude specificity, I think that there would be slight differences that would be very informative. Such measurements have already been successfully carried out by other groups (She *et al.*, 2017), and have been highly informative. In addition to HiTS-RAP, we could carry out measurements of DNA binding affinity as well with the infrastructure that I have set up.

Another useful application of HiTS-RAP, and indeed our original goal with it, is the characterization of SELEX libraries. In an unpublished series of experiments, I carried out HiTS-RAP for the NELF-E SELEX libraries and Fabiana's HSF1 SELEX libraries. With NELF, I found that aptamers with perfect matches to the NBE all bind similarly, while aptamers with no NBE do not bind well. And in the HSF pool, most aptamers' K_d is within the same order of magnitude, though with the large number of measurements that I have with HiTS-RAP in a run

of 20 million reads, I am highly confident in measuring small differences in affinity. Therefore, we need the right kind of library for HiTS-RAP to be successful and useful. It must be highly enriched, as we need at least 5 reads in to make a measurement and we would like to make at least a few thousand high confidence measurements. It needs to have been an ok but not stellar selection: HSF1 aptamers bind with similar affinity so not much was gained by HiTS-RAP. And it would need to be worthwhile: with NELF, we already knew from a motif search that the best aptamers have a 7-mer NBE. I was hoping to find an aptamer to another surface of NELF (that would thus lack the NBE), but did not find any other candidates that looked promising. If we had such a library, and stood to learn a lot with a good aptamer to the target, then a HiTS-RAP could be extremely valuable.

As our HiTS-RAP paper was under review, the Greenleaf lab published a paper on RNA-MaP, a very similar technology (Buenrostro *et al.*, 2014). Their method also makes RNA binding affinity measurements on an Illumina sequencing instrument, but has many fundamental differences. As it is, HiTS-RAP uses an unmodified sequencer, and Illumina's software, so I am confident that any competent graduate student could do it. RNA-MaP, however, uses the optics of the GAIIX, but with many modifications enabling imaging of other fluors and greater control of reagent delivery and imaging times. With this, they are able to measure rate constants as well as equilibrium binding. We should consider borrowing from this. Rate constants could be very informative in some cases: for example, in the HSF1 aptamer SELEX library, we see no correlation between enrichment in SELEX and equilibrium binding affinity, but may see differences with rate constants. I can envision an assay where we follow association by using the timestamps on images of different tiles to infer association dynamics for abundant sequences that are represented many times in each tile. But, as it is, their setup is far superior for measuring

kinetics. A used GAIIX costs less than 1/10th what it did when I started. If we had the right expertise in the lab, someone could buy a few extras and scavenge parts to set up some custom workstations.

Aptamers as Inhibitors

Strategies for improving aptamer inhibition studies are listed in the conclusion of Chapter 3. In the specific case of NELFapt, I wanted to try another non-specific control to try to understand the scrambled aptamer result explained in Chapter 3. The Scrambled aptamer had the same effect as NELFapt. This is either because one of the scrambled aptamer units binds NELF specifically, or because any RNA overexpressed in the nucleus would produce the same result. If I were carrying out another aptamer inhibition experiment, I would want to understand what went wrong with the NELFapt experiment. To this end, I made a stably transfected S2 line with a scrambled aptamer to DSIF that was made by Judhajeet Ray in the lab. The PRO-seq for this experiment failed, but there are more nuclei in the freezer.

Given my experience with using NELFapt as an inhibitor, I would be very cautious before attempting to use an aptamer *in vivo* again. It would have to be for a case where I knew the specific interaction that the aptamer is disrupting, and only with an experimental design that is well suited to a genome-wide readout, as discussed at the end of Chapter 3. Thus, given the time and uncertainty involved in selecting a suitable aptamer, and the difficulties of using it in a way that meaningful biological insights stand to be gained, I do not think that this is the most efficient way of using resources in the lab when the goal is to gain biological insights. Future efforts should focus on developing technology for aptamer inhibition that address some of the issues seen in the NELF experiment. Our broader group has vastly improved the process of

selecting and characterizing aptamers: now it is time to focus our efforts on ways to use aptamers. For now, an alternative approach has a much higher likelihood of success for deciphering the biological role of factors when this is the goal. Knockdowns have repeatedly been used successfully in the past(Core *et al.*, 2012; Duarte *et al.*, 2016). Several new strategies for perturbation are extremely promising. Auxin induced degrons are used to rapidly and specifically degrade proteins with a small peptide tag inserted in the endogenous gene upon addition of a drug, and often function more completely than knockdowns, and take effect rapidly(Nishimura *et al.*, 2009), thus negating aptamer's promise of rapid induction. Analogue sensitive mutants can be made to render specific enzymes sensitive to a bulky substrate analogue molecule(Gregan *et al.*, 2007). This strategy has been used successfully in our lab to rapidly and specifically inhibit Cdk7 and Cdk9(Booth *et al.*, 2017). These types of experiments target catalytic activities of factors rapidly and specifically, and thus have virtually all of the features that we were attempting to gain with aptamers. This strategy is limited to biological questions concerning the catalytic activity of the target. Similarly, because aptamers to non-nucleic acid binding domains of factors are exceedingly difficult to select and characterize(Shui *et al.*, 2012), I would argue that aptamer inhibition is somewhat limited to perturbations of nucleic acid interactions. In total, I believe that our efforts are much more likely to be rewarded if we focus on some of these other strategies rather than RNA aptamers for the time being. Investment in strategies for aptamer inhibition, as a purely technology development oriented project, could bring us to a point where aptamers can routinely be used as a tool to answer biological questions in the future.

Termination, Capping and Transcription Initiation Domains

The greatest untapped potential of coPRO lies in a study of termination. Terminating Pol II has undergone cleavage and polyadenylation and is being chased down by the exonuclease Xrn2(Proudfoot, 2016). This means that the nascent transcript associated with terminating Pol II has a 5' monophosphate end, and is short enough to make a good insert for an Illumina library, and is thus incorporated into my uncapped coPRO library. This significantly hurt our ability to look at capping with the experiment as is: though there is a strong peak of signal at TSNs corresponding to 5' triphosphate transcripts (pre-capped transcripts), most of the reads in the library map to the termination region of genes (Figure 5. 1A). This significantly reduced our coverage of pre-capped RNA. With an experiment designed intentionally to interrogate monophosphate 5' ends, we could draw important conclusions about termination. This monophosphate specific library would be made by proceeding immediately to 5' adapter ligation after the run-on (thus excluding capped and 5' triphosphate pre-capped RNA). As it is, the uncapped library will also incorporate degradation products of RNA hydrolysis during handling (PNK treatment rescues 5' hydroxyl RNA), so a monophosphate specific library could be analyzed with greater confidence. With such an analysis, we would gain new mechanistic insights into termination. As it is, the data already show a stair shaped pattern in paired coPRO Uncapped plots similar to Figure 4. 1 (see Figure 5. 1B). Termination occurs at sites where the DNA template sequence stalls polymerase: this stair pattern supports a model where an individual polymerase stalls repeatedly in the termination region, while Xrn2 uniformly chases it down. Sites of fixed 5' ends, similar to initiation sites, that are unique to the uncapped library indicate that multiple rounds of cleavage and polyadenylation could occur. Such uncapped, polyadenylated chasedown RNA products would likely be highly unstable. We could

specifically sequence these to validate this model of repeated cleavage and polyadenylation. Other groups detect rapidly degraded RNAs, like upstream divergent and enhancer RNAs, by depleting the RNA exosome and mapping sites of initiation with CAGE(Andersson *et al.*, 2014). We could map chasedown products by depleting exosome and sequencing uncapped polyadenylated RNAs specifically, similar to coPRO uncapped.

A thoughtfully designed coPRO experiment could ask how close Xrn2 is as it chases polymerase, where termination occurs, and whether multiple rounds of cleavage and polyadenylation occur. An Xrn2 degron could compliment this by validating the patterns observed as exonucleolytic cleavage. We could also measure, quantitatively, the amount of termination near promoters by careful comparison between capped and uncapped nascent RNA there. This would give new insight into the prevalence of termination from the pause region, without pause release. Someone doing this needs to think very carefully about normalizing for the length of RNA: inserts beyond 100 bp are subject to an exponential decay in their ability to be sequenced. This would be another study where clarity of what is being measured is paramount, so while it is an easy experiment, the analysis is not simple. All genome-wide assays are a snapshot of the average of millions of cells. Deriving rates of movement, or frequencies of termination from such data requires careful thought.

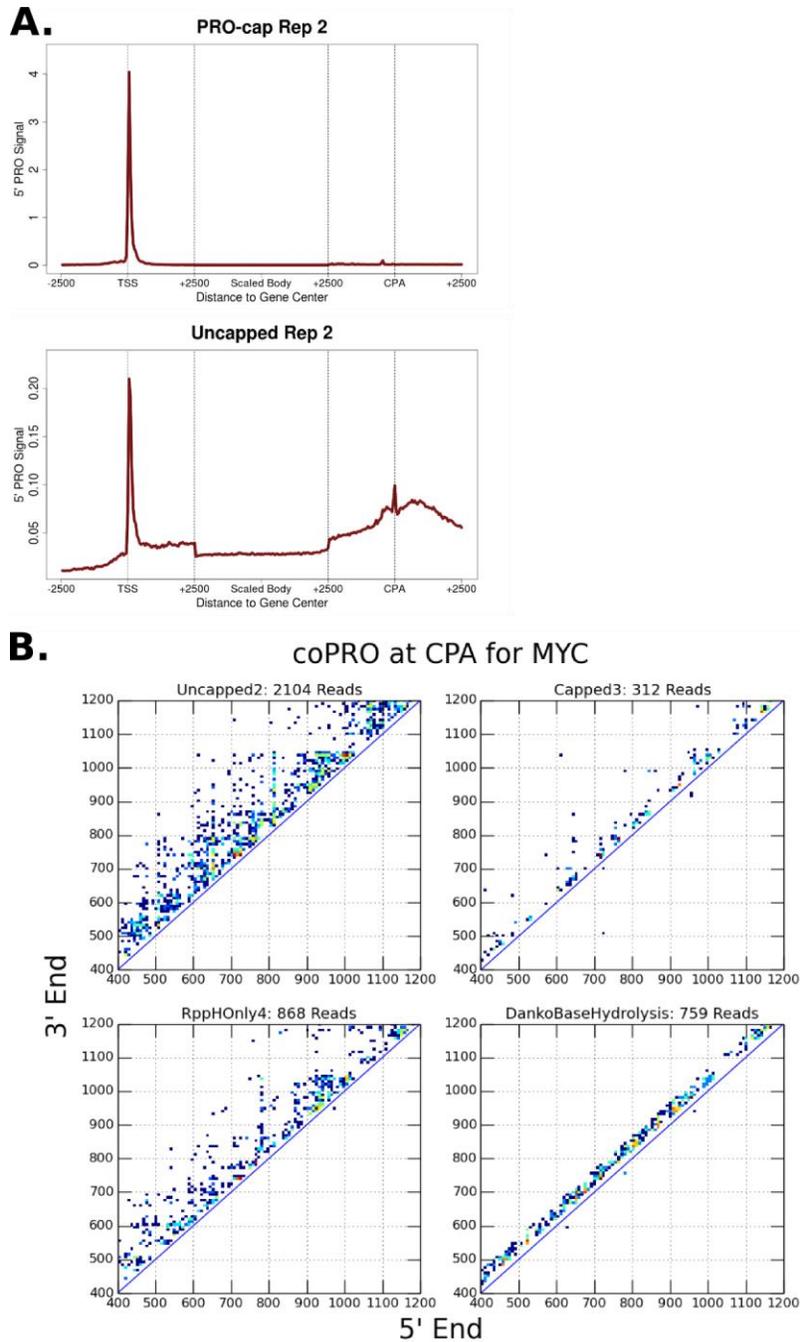


Figure 5. 1: Observing Pol II termination with coPRO Uncapped.

A.) Metaplot of coPRO at RefSeq annotated genes. Signal is plotted in fixed bins ± 2500 bp to the annotated TSS and CPA signal, and scaled in the middle (the jump in bin size when scaling creates some discontinuity). In coPRO capped, most signal is at the TSS. In coPRO Uncapped, much of the signal is from 5' pRNA in the termination region. **B.)** Paired plots of the termination region of *Myc* similar to Figure 4. 1A, for different coPRO libraries (and a regular PRO-seq library with the subset of reads spanned with 75 bp reads). 0 corresponds to the annotated CPA. Note the strong enrichment of uncapped over capped (coPRO capped here is likely uncapped background that is captured at a low frequency). The stair pattern of termination is apparent in RppH only (the Capped + Uncapped library), and Uncapped.

In Chapter 4, I describe functional relationships between capping and pausing that were illuminated by coPRO. These results suggest that capping is not always complete before pausing. Some simple experiments could provide additional insight into their interplay. Using the Cdk7 inhibitor THZ1, the Price group showed that Cdk7 mediated phosphorylation of the CTD affects capping enzyme activity (Nilson *et al.*, 2015). We could explore this genome-wide with THZ1 or, better yet, analogue sensitive Cdk7 mutants made by the Fisher group, with coPRO for capped and uncapped RNA as readout. It would be interesting to see if capping is impaired uniformly, or if different pause classes responded differently. I suspect that the Late pause class will be more sensitive to this inhibition, as I believe that capping there is more dependent upon DSIF mediated recruitment of capping enzyme. Thus, these promoters seem to adhere to a more rigid ordered assembly of factors. We could further explore the interplay between capping and pausing by perturbing the capping enzyme, for example with an auxin induced degron. This strategy is more direct than trying to assess the importance of capping by inhibiting Cdk7. With coPRO as readout, we could assess whether impaired capping inhibits normal pausing and pause release, and increases the rate of premature termination after pause release. In such experiments, we could design an enzymatic selection that would be more specific to pre-capped RNA than what we have already used, thus greatly reducing the background from termination as we assess capping. To do this, I would first use Terminator exonuclease to degrade monophosphate RNA. I have confirmed that pre-capped RNA is much less susceptible to Terminator degradation than monophosphate (Figure 4. 23), as advertised by the supplier, Lucigen. Apyrase, an enzyme from NEB, could then be used to specifically reduce 5' triphosphate RNA to monophosphate, without decapping capped transcripts. I suspect that this enrichment scheme would allow us to make a pre-capped library where we could sequence

capping with much higher coverage from the same depth, and therefore provide greater insight into any perturbations of capping and pausing.

Another rich avenue for further experimentation is tests of the functional significance of minor TSSes within TIDs, and their relationship to chromatin environment. One potential way that we could assess the relationship could be by perturbing the chromatin environment in TIDS. As a part of their work showing that broad active chromatin regions at promoters are important in regulating cell to cell consistency (Benayoun *et al.*, 2014), the Burnet group modulated the width of H3K4me3 peaks by knocking down erasers and writers of this mark. Knockdown of JARID2, an eraser of this mark, resulted in a broadening of H3K4me3 domains. Similarly, they found that knockdown of WDR5, a core subunit of the H3K4me3 methyltransferase complex COMPASS, reduced the breadth of H3K4me3 peaks. By using coPRO after such a perturbation, we could ask whether the peripheral TSSes of TIDs are specifically affected or not. With this, we would gain insight into the interplay between chromatin environment and initiation: I suspect that the peripheral TSSes of TIDs are involved in the fine tuning of the boundaries of chromatin domains and are thus sensitive to changes in spreading of H3K4me3. Stronger peripheral TSSes would therefore be less sensitive. I also suspect that H3K27ac would have a much stronger effect, if a similar strategy were devised for it.

Bibliography

- Adelman, K. and Lis, J. T. (2012) 'Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans.', *Nature reviews. Genetics*. Nature Publishing Group, 13(10), pp. 720–31. doi: 10.1038/nrg3293.
- Agarwal, V., Bell, G. W., Nam, J.-W. and Bartel, D. P. (2015) 'Predicting effective microRNA target sites in mammalian mRNAs', *eLife*. eLife Sciences Publications Limited, 4, p. e05005. doi: 10.7554/eLife.05005.
- Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. and Sharp, P. A. (2013) 'Promoter directionality is controlled by U1 snRNP and polyadenylation signals.', *Nature*. Nature Publishing Group, 499(7458), pp. 360–3. doi: 10.1038/nature12349.
- Amândio, A. R., Necsulea, A., Joye, E., Mascrez, B. and Duboule, D. (2016) 'Hotair Is Dispensable for Mouse Development', *PLOS Genetics*. Edited by G. S. Barsh. Public Library of Science, 12(12), p. e1006232. doi: 10.1371/journal.pgen.1006232.
- Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., Bagger, F. O., Jørgensen, M., Andersen, P. R., Bertin, N., Rackham, O., Burroughs, A. M., Baillie, J. K., Ishizu, Y., Shimizu, Y., Furuhashi, E., Maeda, S., Negishi, Y., Mungall, C. J., Meehan, T. F., Lassmann, T., Itoh, M., Kawaji, H., Kondo, N., Kawai, J., Lennartsson, A., Daub, C. O., Heutink, P., Hume, D. A., Jensen, T. H., Suzuki, H., Hayashizaki, Y., Müller, F., Consortium, T. F., Forrest, A. R. R., Carninci, P., Rehli, M. and Sandelin, A. (2014) 'An atlas of active enhancers across human cell types and tissues', *Nature*, 507(7493), pp. 455–461. doi: 10.1038/nature12787.
- Andrulis, E. D. (2000) 'High-resolution localization of Drosophila Spt5 and Spt6 at heat shock genes in vivo: roles in promoter proximal pausing and transcription elongation', *Genes & Development*, 14(20), pp. 2635–2649. doi: 10.1101/gad.844200.
- Arensbergen, J. Van, Fitzpatrick, V. D., Haas, M. De, Pagie, L., Sluimer, J., Bussemaker, H. J. and Steensel, B. Van (2016) 'Genome-wide mapping of autonomous promoter activity in human cells', *Nature Publishing Group*. Nature Publishing Group, 35(2). doi: 10.1038/nbt.3754.
- Arnold, C. D., Gerlach, D., Stelzer, C., Boryn, L. M., Rath, M., Stark, A., Boryn, L. M., Rath, M., Stark, A., Boryn, L. M., Rath, M. and Stark, A. (2013) 'Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq', *Science*, 339(6123), pp. 1074–1077. doi: 10.1126/science.1232542.
- Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. and Noble, W. S. (2009) 'MEME Suite: Tools for motif discovery and searching', *Nucleic Acids Research*, 37(SUPPL. 2), pp. 202–208. doi: 10.1093/nar/gkp335.
- Bannister, A. J. and Kouzarides, T. (2011) 'Regulation of chromatin by histone modifications', *Cell Research*. Nature Publishing Group, 21(3), pp. 381–395. doi: 10.1038/cr.2011.22.
- Baranello, L., Wojtowicz, D., Cui, K., Devaiah, B. N., Chung, H.-J., Chan-Salis, K. Y., Guha, R., Wilson, K., Zhang, X., Zhang, H., Piotrowski, J., Thomas, C. J., Singer, D. S., Pugh, B. F., Pommier, Y., Przytycka, T. M., Kouzine, F., Lewis, B. A., Zhao, K. and Levens, D. (2016) 'RNA Polymerase II Regulates Topoisomerase 1 Activity to Favor Efficient Transcription', *Cell*, 165(2), pp. 357–371. doi: 10.1016/j.cell.2016.02.036.
- Barlow, R. and Blake, J. F. (1989) 'Hill coefficients and the logistic equation.', *Trends in*

- pharmacological sciences*, 10(11), pp. 440–1. doi: 10.1016/S0165-6147(89)80006-9.
- Beagan, J. A., Duong, M. T., Titus, K. R., Zhou, L., Cao, Z., Ma, J., Lachanski, C. V., Gillis, D. R. and Phillips-Cremins, J. E. (2017) ‘YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment’, *Genome Research*, 27(7), pp. 1139–1152. doi: 10.1101/gr.215160.116.
- Benayoun, B. A., Pollina, E. A., Ucar, D., Mahmoudi, S., Karra, K., Wong, E. D., Devarajan, K., Daugherty, A. C., Kundaje, A. B., Mancini, E., Hitz, B. C., Gupta, R., Rando, T. A., Baker, J. C., Snyder, M. P., Cherry, J. M. and Brunet, A. (2014) ‘H3K4me3 breadth is linked to cell identity and transcriptional consistency’, *Cell*. Cell Press, 158(3), pp. 673–688. doi: 10.1016/j.cell.2014.06.027.
- Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. a, Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M. J., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M. D., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. a, Benoit, V. a, Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. a, Brown, R. C., Brown, A. a, Buermann, D. H., Bundu, A. a, Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. a, Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V, Johnson, M. Q., James, T., Huw Jones, T. a, Kang, G.-D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. a, Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. a, Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O’Neill, M. J., Osborne, M. a, Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. a, Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovskiy, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurler, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R. and Smith, A. J. (2008) ‘Accurate whole human genome sequencing using reversible terminator chemistry.’, *Nature*, 456(7218), pp. 53–9. doi: 10.1038/nature07517.
- Berger, S. L. (2002) ‘Histone modifications in transcriptional regulation’, *Current Opinion in*

- Genetics & Development*, 12(2), pp. 142–148. doi: 10.1016/S0959-437X(02)00279-4.
- Booth, G. T., Parua, P. K., Sansó, M., Fisher, R. P. and Lis, T. (2017) ‘Cdk9 regulates a promoter-proximal checkpoint to modulate RNA Polymerase II elongation rate’, *bioRxiv*, (607), pp. 1–43.
- Booth, G. T., Wang, I. X., Cheung, V. G. and Lis, J. T. (2016) ‘Divergence of a conserved elongation factor and transcription regulation in budding and fission yeast’, *Genome Research*, p. gr.204578.116-. doi: 10.1101/gr.204578.116.
- Bose, D. A., Donahue, G., Reinberg, D., Shiekhatar, R., Bonasio, R. and Berger, S. L. (2017) ‘RNA Binding to CBP Stimulates Histone Acetylation and Transcription’, *Cell*, 168(1–2), p. 135–149.e22. doi: 10.1016/j.cell.2016.12.020.
- Brannan, K., Kim, H., Erickson, B., Glover-Cutter, K., Kim, S., Fong, N., Kiemele, L., Hansen, K., Davis, R., Lykke-Andersen, J. and Bentley, D. L. (2012) ‘mRNA Decapping Factors and the Exonuclease Xrn2 Function in Widespread Premature Termination of RNA Polymerase II Transcription’, *Molecular Cell*, 46(3), pp. 311–324. doi: 10.1016/j.molcel.2012.03.006.
- Buckley, M. S., Kwak, H., Zipfel, W. R. and Lis, J. T. (2014) ‘Kinetics of promoter Pol II on Hsp70 reveal stable pausing and key insights into its regulation.’, *Genes & development*, 28(1), pp. 14–9. doi: 10.1101/gad.231886.113.
- Buenrostro, J. D., Araya, C. L., Chircus, L. M., Layton, C. J., Chang, H. Y., Snyder, M. P. and Greenleaf, W. J. (2014) ‘Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes’, *Nature Biotechnology*, (April). doi: 10.1038/nbt.2880.
- Bunch, T. A., Grinblat, Y. and Goldstein, L. S. (1988) ‘Characterization and use of the *Drosophila* metallothionein promoter in cultured *Drosophila melanogaster* cells.’, *Nucleic acids research*, 16(3), pp. 1043–61. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=334736&tool=pmcentrez&rendertype=abstract> (Accessed: 20 August 2014).
- Busslinger, G. A., Stocsits, R. R., Van Der Lelij, P., Axelsson, E., Tedeschi, A., Galjart, N. and Peters, J. M. (2017) ‘Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl’, *Nature*. Nature Publishing Group, 544(7651), pp. 503–507. doi: 10.1038/nature22063.
- Calo, E. and Wysocka, J. (2013) ‘Modification of Enhancer Chromatin: What, How, and Why?’, *Molecular Cell*. Elsevier Inc., 49(5), pp. 825–837. doi: 10.1016/j.molcel.2013.01.038.
- Campbell, Z. T., Bhimsaria, D., Valley, C. T., Rodriguez-Martinez, J. A., Menichelli, E., Williamson, J. R., Ansari, A. Z. and Wickens, M. (2012) ‘Cooperativity in RNA-protein interactions: global analysis of RNA binding specificity.’, *Cell reports*, 1(5), pp. 570–81. doi: 10.1016/j.celrep.2012.04.003.
- Canny, M. D., Jucker, F. M. and Pardi, A. (2007) ‘Efficient ligation of the *Schistosoma* hammerhead ribozyme.’, *Biochemistry*, 46(12), pp. 3826–34. doi: 10.1021/bi062077r.
- Carninci, P., Sandelin, T. A., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C. A., Taylor, M. S., Engstrom, P., Frith, M., Alkema, W. B., Tan, S. L., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori, M., Kawaji, H., Kai, C., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., Persichetti, F., Suzuki, H., Grimmond, S. M., Wells, C., Orlando, V., Wahlestedt, C., Liu, E. T., Harbers, M., Kawai, J., Bajic, V. B., Hume, D. and Hayashizaki, Y. (2006) ‘Genome-wide analysis of mammalian promoter architecture and evolution’, *Nat Genet*, 38(6), pp. 626–635. doi: 10.1038/ng1789.

- Chao, S. H. and Price, D. H. (2001) 'Flavopiridol inactivates P-TEFb and blocks most RNA polymerase II transcription in vivo.', *The Journal of biological chemistry*, 276(34), pp. 31793–9. doi: 10.1074/jbc.M102306200.
- Chen, F., Gao, X. and Shilatifard, A. (2015) 'Stably paused genes revealed through inhibition of transcription initiation by the TFIIH inhibitor triptolide.', *Genes & development*, 29(1), pp. 39–47. doi: 10.1101/gad.246173.114.
- Chen, F. X., Xie, P., Collings, C. K., Cao, K., Aoi, Y., Marshall, S. A., Rendleman, E. J., Ugarenko, M., Ozark, P. A., Zhang, A., Shiekhatar, R., Smith, E. R., Zhang, M. Q. and Shilatifard, A. (2017) 'PAF1 regulation of promoter-proximal pause release via enhancer activation', *Science*, 3269(August), pp. 1–11. doi: 10.1126/science.aan3269.
- Chen, K., Chen, Z., Wu, D., Zhang, L., Lin, X., Su, J., Rodriguez, B., Xi, Y., Xia, Z., Chen, X., Shi, X., Wang, Q. and Li, W. (2015) 'Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes', *Nature Genetics*. Nature Publishing Group, 47(10), pp. 1149–1157. doi: 10.1038/ng.3385.
- Chen, Y., Jorgensen, M., Kolde, R., Zhao, X., Parker, B., Valen, E., Wen, J. and Sandelin, A. (2011) 'Prediction of RNA Polymerase II recruitment, elongation and stalling from histone modification data.', *BMC Genomics*, 12(1), p. 544. doi: 10.1186/1471-2164-12-544.
- Chen, Y., Pai, A. A., Herudek, J., Lubas, M., Meola, N., Järvelin, A. I., Andersson, R., Pelechano, V., Steinmetz, L. M., Jensen, T. H. and Sandelin, A. (2016) 'Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters', *Nature Genetics*. Nature Publishing Group, 48(9), pp. 984–994. doi: 10.1038/ng.3616.
- Cherbas, L. and Cherbas, P. (2007) 'Drosophila cell culture and transformation.', *CSH protocols*. Cold Spring Harbor Laboratory Press, 2007(8), p. pdb.top6. doi: 10.1101/PDB.TOP6.
- Cho, M., Soo Oh, S., Nie, J., Stewart, R., Eisenstein, M., Chambers, J., Marth, J. D., Walker, F., Thomson, J. A. and Soh, H. T. (2013) 'Quantitative selection and parallel characterization of aptamers.', *Proceedings of the National Academy of Sciences of the United States of America*, 110(46), pp. 18460–5. doi: 10.1073/pnas.1315866110.
- Cinghu, S., Yang, P., Kosak, J. P., Conway, A. E., Kumar, D., Oldfield, A. J., Adelman, K. and Jothi, R. (2017) 'Intragenic Enhancers Attenuate Host Gene Expression', *Molecular Cell*. Elsevier Inc., 68(1), p. 104–117.e6. doi: 10.1016/j.molcel.2017.09.010.
- Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A. and Lis, J. T. (2014) 'Analysis of transcription start sites from nascent RNA supports a unified architecture of mammalian promoters and enhancers', *Nature genetics*. Nature Publishing Group, 46(12), pp. 1311–1320. doi: 10.1038/ng.3142.
- Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A. and Lis, J. T. (2014) 'Supplement: Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers', *Nature Genetics*, 46(12), pp. 1311–20. doi: 10.1038/ng.3142.
- Core, L. J., Waterfall, J. J., Gilchrist, D. A., Fargo, D. C., Kwak, H., Adelman, K. and Lis, J. T. (2012) 'Defining the Status of RNA Polymerase at Promoters', *Cell Reports*. The Authors, 2(4), pp. 1025–1035. doi: 10.1016/j.celrep.2012.08.034.
- Core, L. J., Waterfall, J. J. and Lis, J. T. (2008) 'Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters.', *Science (New York, N.Y.)*, 322(5909), pp. 1845–8. doi: 10.1126/science.1162228.

- Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K. C., Huang, H., Liu, T., Marina, R. J., Jung, I., Shen, Y., Guan, K. and Ren, B. (2017) ‘A tiling-deletion-based genetic screen for cis - regulatory element identification in mammalian cells’, *Nature Methods*. Nature Publishing Group, (April), pp. 1–11. doi: 10.1038/nmeth.4264.
- Diao, Y., Li, B., Meng, Z., Jung, I., Lee, A. Y., Dixon, J., Maliskova, L., Guan, K.-L., Shen, Y. and Ren, B. (2016) ‘A new class of temporarily phenotypic enhancers identified by CRISPR/Cas9-mediated genetic screening.’, *Genome research*. Cold Spring Harbor Laboratory Press, 26(3), pp. 397–405. doi: 10.1101/gr.197152.115.
- Dixit, A., Parnas, O., Li, B., Weissman, J. S., Friedman, N., Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-aron, L., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N. and Regev, A. (2016) ‘Perturb-Seq : Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Resource Perturb-Seq : Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens’, *Cell*. Elsevier Inc., 167(7), p. 1853–1857.e17. doi: 10.1016/j.cell.2016.11.038.
- Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. and Ren, B. (2012) ‘Topological domains in mammalian genomes identified by analysis of chromatin interactions’, *Nature*. Nature Publishing Group, 485(7398), pp. 376–380. doi: 10.1038/nature11082.
- Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakraborty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R. and Gingeras, T. R. (2012) ‘Landscape of transcription in human cells’, *Nature*. Nature Research, 489(7414), pp. 101–108. doi: 10.1038/nature11233.
- Dorigi, K. M., Swigut, T., Henriques, T., Garcia, B. A., Adelman, K., Dorigi, K. M., Swigut, T., Henriques, T., Bhanu, N. V., Scruggs, B. S. and Nady, N. (2017) ‘Mll3 and Mll4 Facilitate EnhancerRNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation’, *Molecular Cell*. Elsevier Inc., 66(4), p. 568–576.e4. doi: 10.1016/j.molcel.2017.04.018.
- Duarte, F. M., Fuda, N. J., Mahat, D. B., Core, L. J., Guertin, M. J. and Lis, J. T. (2016) ‘Transcription factors GAF and HSF act at distinct regulatory steps to modulate stress-induced gene activation.’, *Genes & development*. Cold Spring Harbor Laboratory Press, 30(15), pp. 1731–46. doi: 10.1101/gad.284430.116.
- Duttke, S. H. C., Lacadie, S. A., Ibrahim, M. M., Glass, C. K., Corcoran, D. L., Benner, C., Heinz, S., Kadonaga, J. T. and Ohler, U. (2015) ‘Human Promoters Are Intrinsically Directional’, *Molecular Cell*. Elsevier, 57(4), pp. 674–684. doi: 10.1016/j.molcel.2014.12.029.

- Ellington, A. D. and Szostak, J. W. (1990) 'In vitro selection of RNA molecules that bind specific ligands', *Nature*. Nature Publishing Group, 346(6287), pp. 818–822. doi: 10.1016/0021-9797(80)90501-9.
- Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M. and Lander, E. S. (2016) 'Local regulation of gene expression by lncRNA promoters, transcription and splicing', *Nature*. Nature Research. doi: 10.1038/nature20149.
- Epshtein, V., Toulmé, F., Rahmouni, A. R., Borukhov, S. and Nudler, E. (2003) 'Transcription through the roadblocks: the role of RNA polymerase cooperation.', *The EMBO journal*, 22(18), pp. 4719–27. doi: 10.1093/emboj/cdg452.
- Ernst, J. and Kellis, M. (2012) 'ChromHMM: Automating chromatin-state discovery and characterization', *Nature Methods*. Nature Publishing Group, 9(3), pp. 215–216. doi: 10.1038/nmeth.1906.
- Esteller, M. (2011) 'Non-coding RNAs in human disease.', *Nature reviews. Genetics*. Nature Publishing Group, 12(12), pp. 861–74. doi: 10.1038/nrg3074.
- Evanko, D. (2011) 'Next-generation protein binding', *Nature Methods*. Nature Publishing Group, 8(8), pp. 619–619. doi: 10.1038/nmeth0811-619.
- Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L. and Quake, S. R. (2010) 'Analysis of the size distributions of fetal and maternal cell-free DNA by paired-end sequencing', *Clinical Chemistry*, 56(8), pp. 1279–1286. doi: 10.1373/clinchem.2010.144188.
- Ferrai, C., Torlai Triglia, E., Risner-Janiczek, J. R., Rito, T., Rackham, O. J., de Santiago, I., Kukalev, A., Nicodemi, M., Akalin, A., Li, M., Ungless, M. A. and Pombo, A. (2017) 'RNA polymerase II primes Polycomb-repressed developmental genes throughout terminal neuronal differentiation', *Molecular Systems Biology*, 13(10), p. 946. doi: 10.15252/msb.20177754.
- Fishburn, J., Galburt, E. and Hahn, S. (2016) 'Transcription Start Site Scanning and the Requirement for ATP during Transcription Initiation by RNA Polymerase II *', *Journal of Biological Chemistry*, 291(25), pp. 13040–13047. doi: 10.1074/jbc.M116.724583.
- Fuda, N. J., Ardehali, M. B. and Lis, J. T. (2009) 'Defining mechanisms that regulate RNA polymerase II transcription in vivo.', *Nature*, 461(7261), pp. 186–92. doi: 10.1038/nature08449.
- Fuda, N. J., Guertin, M. J., Sharma, S., Danko, C. G., Martins, A. L., Siepel, A. and Lis, J. T. (2015) 'GAGA Factor Maintains Nucleosome-Free Regions and Has a Role in RNA Polymerase II Recruitment to Promoters.', *PLoS genetics*, 11(3), p. e1005108. doi: 10.1371/journal.pgen.1005108.
- Fujinaga, K., Irwin, D., Huang, Y., Taube, R., Kurosu, T. and Peterlin, B. M. (2004) 'Dynamics of Human Immunodeficiency Virus Transcription : P-TEFb Phosphorylates RD and Dissociates Negative Effectors from the Transactivation Response Element', 24(2), pp. 787–795. doi: 10.1128/MCB.24.2.787.
- Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S. and Engreitz, J. M. (2016) 'Systematic mapping of functional enhancer-promoter connections with CRISPR interference', *Science*, 6056(September), pp. 1–8.
- Galbraith, M. D., Allen, M. A., Bensard, C. L., Wang, X., Schwinn, M. K., Qin, B., Long, H. W., Daniels, D. L., Hahn, W. C., Dowell, R. D. and Espinosa, J. M. (2013) 'HIF1A Employs CDK8-Mediator to Stimulate RNAPII Elongation in Response to Hypoxia', *Cell*, 153(6),

- pp. 1327–1339. doi: 10.1016/j.cell.2013.04.048.
- Gasparini, M., Findlay, G. M., McKenna, A., Milbank, J. H., Lee, C., Zhang, M. D., Cusanovich, D. A. and Shendure, J. (2016) ‘Paired CRISPR/Cas9 guide-RNAs enable high-throughput deletion scanning (ScanDel) of a Mendelian disease locus for functionally critical non-coding elements’, *bioRxiv*. Cold Spring Harbor Labs Journals, p. 92445. doi: 10.1101/092445.
- Germer, K., Leonard, M. and Zhang, X. (2013) ‘RNA aptamers and their therapeutic and diagnostic applications.’, *International journal of biochemistry and molecular biology*, 4(1), pp. 27–40. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3627066&tool=pmcentrez&rendertype=abstract> (Accessed: 24 March 2014).
- Gibson, B. A., Zhang, Y., Jiang, H., Hussey, K. M., Shrimp, J. H., Lin, H., Schwede, F., Yu, Y., Kraus, W. L., Gibson, B. A., Kraus, W. L., Vyas, S., Matic, I., Uchima, L., Rood, J., Zaja, R., Hay, R. T., Ahel, I., Chang, P., Hottiger, M. O., Daniels, C. M., Ong, S.-E., Leung, A. K. L., Carter-O’Connell, I., Jin, H., Morgan, R. K., David, L. L., Cohen, M. S., Carter-O’Connell, I., Jin, H., Morgan, R. K., Zaja, R., David, L. L., Ahel, I., Cohen, M. S., Jiang, H., Kim, J. H., Frizzell, K. M., Kraus, W. L., Lin, H., Specht, K. M., Shokat, K. M., Zhang, Y., Wang, J., Ding, M., Yu, Y., Petesch, S. J., Lis, J. T., Tulin, A., Spradling, A., Adelman, K., Lis, J. T., Yamaguchi, Y., Inukai, N., Narita, T., Wada, T., Handa, H., Guo, J., Price, D. H., Bartolomei, G., Leutert, M., Manzo, M., Baubec, T., Hottiger, M. O., Core, L. J., Waterfall, J. J., Lis, J. T., Yamaguchi, Y., Filipovska, J., Yano, K., Furuya, A., Inukai, N., Narita, T., Wada, T., Sugimoto, S., Konarska, M. M., Handa, H., Krishnakumar, R., Gamble, M. J., Frizzell, K. M., Berrocal, J. G., Kininis, M., Kraus, W. L., Curtin, N. J., Szabo, C., Kim, M. Y., Mauro, S., Gévry, N., Lis, J. T., Kraus, W. L., Colowick, S. P., Kaplan, N. O., Ciotti, M. M., Kumar, A., Colman, R. F., Liang, G., He, J., Zhang, Y., Wessel, D., Flügge, U. I., Dignam, J. D., Lebovitz, R. M., Roeder, R. G., Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., Stephens, R., Baseler, M. W., Lane, H. C., Lempicki, R. A., Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H. C., Lempicki, R. A., Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., Sullivan, M., Hendriks, I. A., D’Souza, R. C., Yang, B., Vries, M. V., Mann, M., Vertegaal, A. C., Chou, M. F., Schwartz, D., Schwartz, D., Gygi, S. P., Sun, M., Gadad, S. S., Kim, D.-S., Kraus, W. L., Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., Pachter, L., Luo, X., Chae, M., Krishnakumar, R., Danko, C. G., Kraus, W. L., Hah, N., Danko, C. G., Core, L., Waterfall, J. J., Siepel, A., Lis, J. T., Kraus, W. L., Langmead, B., Trapnell, C., Pop, M., Salzberg, S. L., Quinlan, A. R., Hall, I. M., Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., Mesirov, J. P., Thorvaldsdóttir, H., Robinson, J. T., Mesirov, J. P., Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenko, V. V., Ren, B., Sun, J., Pan, H., Lei, C., Yuan, B., Nair, S. J., April, C., Parameswaran, B., Klotzle, B., Fan, J. B., Ruan, J., Li, R., Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K., Peng, W., Saldanha, A. J., Martin, C., Danko, C. G., Hah, N., Luo, X., Martins, A. L., Core, L., Lis, J. T., Siepel, A., Kraus, W. L., Robinson, M. D., McCarthy, D. J., Smyth, G. K., Jungmichel, S., Rosenthal, F., Altmeyer, M., Lukas, J., Hottiger, M. O., Nielsen, M. L., Hah, N., Murakami, S., Nagari, A., Danko, C. G. and

- Kraus, W. L. (2016) ‘Chemical genetic discovery of PARP targets reveals a role for PARP-1 in transcription elongation.’, *Science (New York, N.Y.)*. American Association for the Advancement of Science, 13(7), pp. 411–424. doi: 10.1126/science.aaf7865.
- Gilchrist, D. a, Nechaev, S., Lee, C., Ghosh, S. K. B., Collins, J. B., Li, L., Gilmour, D. S. and Adelman, K. (2008) ‘NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly.’, *Genes & development*, 22(14), pp. 1921–33. doi: 10.1101/gad.1643208.
- Gilchrist, D. a, Dos Santos, G., Fargo, D. C., Xie, B., Gao, Y., Li, L. and Adelman, K. (2010) ‘Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation.’, *Cell*. Elsevier Inc., 143(4), pp. 540–51. doi: 10.1016/j.cell.2010.10.004.
- Gilmour, D. S. and Lis, J. T. (1986) ‘RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in *Drosophila melanogaster* cells.’, *Molecular and cellular biology*. American Society for Microbiology, 6(11), pp. 3984–9. doi: 10.1128/MCB.6.11.3984.
- Grant, C. E., Bailey, T. L. and Noble, W. S. (2011) ‘FIMO: scanning for occurrences of a given motif.’, *Bioinformatics (Oxford, England)*, 27(7), pp. 1017–8. doi: 10.1093/bioinformatics/btr064.
- Gravina, M. T., Lin, J. H. and Levine, S. S. (2013) ‘Lane-by-lane sequencing using Illumina’s Genome Analyzer II’, *BioTechniques*, 54(5), pp. 265–269.
- Gregan, J., Zhang, C., Rumpf, C., Cipak, L., Li, Z., Uluocak, P., Nasmyth, K. and Shokat, K. M. (2007) ‘Construction of conditional analog-sensitive kinase alleles in the fission yeast *Schizosaccharomyces pombe*’, *Nature Protocols*. Nature Publishing Group, 2(11), pp. 2996–3000. doi: 10.1038/nprot.2007.447.
- Gressel, S., Schwalb, B., Decker, T. M., Qin, W., Leonhardt, H., Eick, D. and Cramer, P. (2017) ‘CDK9-dependent RNA polymerase II pausing controls transcription initiation’, *eLife*, 6, pp. 1–24. doi: 10.7554/eLife.29736.
- Grohmann, K., Amairic, F., Crews, S. and Attardi, G. (1978) ‘Failure to detect "cap" structures in mitochondrial DNA-coded poly(A)-containing RNA from HeLa cells.’, *Nucleic acids research*. Oxford University Press, 5(3), pp. 637–51. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/643614> (Accessed: 5 February 2017).
- Guajardo, R. and Sousa, R. (1999) ‘Characterization of the effects of *Escherichia coli* replication terminator protein (Tus) on transcription reveals dynamic nature of the Tus block to transcription complex progression’, *Nucleic Acids Research*, 27(13), pp. 2814–2824. doi: 10.1093/nar/27.13.2814.
- Guenther, U.-P., Yandek, L. E., Niland, C. N., Campbell, F. E., Anderson, D., Anderson, V. E., Harris, M. E. and Jankowsky, E. (2013) ‘Hidden specificity in an apparently nonspecific RNA-binding protein’, *Nature*. Nature Publishing Group, 502(7471), pp. 385–388. doi: 10.1038/nature12543.
- Guertin, M. J. and Lis, J. T. (2010) ‘Chromatin landscape dictates HSF binding to target DNA elements.’, *PLoS genetics*. Edited by M. Lichten. Public Library of Science, 6(9), p. e1001114. doi: 10.1371/journal.pgen.1001114.
- Guertin, M. J., Martins, A. L., Siepel, A. and Lis, J. T. (2012) ‘Accurate prediction of inducible transcription factor binding intensities in vivo.’, *PLoS genetics*. Edited by M. Snyder. Public Library of Science, 8(3), p. e1002610. doi: 10.1371/journal.pgen.1002610.
- Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A.,

- Ascano, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M. and Tuschl, T. (2010) ‘Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.’, *Cell*, 141(1), pp. 129–41. doi: 10.1016/j.cell.2010.03.009.
- Hah, N., Murakami, S., Nagari, A., Danko, C. and Kraus, W. L. (2013) ‘Enhancer Transcripts Mark Active Estrogen Receptor Binding Sites.’, *Genome research*. doi: 10.1101/gr.152306.112.
- Hall, M. A., Shundrovsky, A., Bai, L., Fulbright, R. M., Lis, J. T. and Wang, M. D. (2009) ‘High-resolution dynamic mapping of histone-DNA interactions in a nucleosome’, *Nature Structural and Molecular Biology*, 16(2), pp. 124–129. doi: 10.1038/nsmb.1526.
- Harlen, K. M. and Churchman, L. S. (2017) ‘The code and beyond: Transcription regulation by the RNA polymerase II carboxy-terminal domain’, *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 18(4), pp. 263–273. doi: 10.1038/nrm.2017.10.
- Harlen, K. M., Trotta, K. L., Smith, E. E., Mosaheb, M. M., Fuchs, S. M. and Churchman, L. S. (2016) ‘Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue’, *Cell Reports*. doi: 10.1016/j.celrep.2016.05.010.
- Hartzog, G. a, Wada, T., Handa, H. and Winston, F. (1998) ‘Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in *Saccharomyces cerevisiae*.’, *Genes & development*, 12(3), pp. 357–69. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=316481&tool=pmcentrez&rendertype=abstract>.
- Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Calcar, S. Van, Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E. and Ren, B. (2007) ‘Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome’, 39(3), pp. 311–318. doi: 10.1038/ng1966.
- Henriques, T., Gilchrist, D. A., Nechaev, S., Bern, M., Muse, G. W., Burkholder, A., Fargo, D. C. and Adelman, K. (2013) ‘Stable Pausing by RNA Polymerase II Provides an Opportunity to Target and Integrate Regulatory Signals’, *Molecular Cell*. Available at: <http://www.sciencedirect.com/science/article/pii/S109727651300717X> (Accessed: 10 November 2013).
- Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., Young, R. A., Miyagoe-Suzuki, Y., Takeda, S., Kawakami, K. and al., et (2013) ‘Super-Enhancers in the Control of Cell Identity and Disease’, *Cell*. Elsevier, 155(4), pp. 934–947. doi: 10.1016/j.cell.2013.09.053.
- Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. and Sharp, P. A. (2017) ‘A Phase Separation Model for Transcriptional Control’, *Cell*. Elsevier Inc., 169(1), pp. 13–23. doi: 10.1016/j.cell.2017.02.007.
- Holeman, L. A., Robinson, S. L., Szostak, J. W. and Wilson, C. (1998) ‘Isolation and characterization of fluorophore-binding RNA aptamers.’, *Folding & design*, 3(6), pp. 423–31. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/9889155> (Accessed: 11 October 2013).
- Jacob, F. and Monod, J. (1961) ‘Genetic regulatory mechanisms in the synthesis of proteins’, *Journal of Molecular Biology*, 3(3), pp. 318–356. doi: 10.1016/S0022-2836(61)80072-7.
- Jankowsky, E. and Harris, M. E. (2015) ‘Specificity and nonspecificity in RNA-protein interactions’, *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 16(9),

- pp. 533–544. doi: 10.1038/nrm4032.
- Jimeno-González, S., Ceballos-Chávez, M. and Reyes, J. C. (2015) ‘A positioned +1 nucleosome enhances promoter-proximal pausing.’, *Nucleic acids research*, pp. 1–11. doi: 10.1093/nar/gkv149.
- Jolma, A., Yan, J., Whittington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T. and Taipale, J. (2013) ‘DNA-binding specificities of human transcription factors’, *Cell*. Elsevier Inc., 152(1–2), pp. 327–339. doi: 10.1016/j.cell.2012.12.009.
- Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E. and Taipale, J. (2015) ‘DNA-dependent formation of transcription factor pairs alters their binding specificity’, *Nature*. Nature Publishing Group, 527(7578), pp. 384–388. doi: 10.1038/nature15518.
- Jonkers, I., Kwak, H. and Lis, J. T. (2014) ‘Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons.’, *eLife*. eLife Sciences Publications Limited, 3, p. e02407. doi: 10.7554/eLife.02407.
- Jonkers, I. and Lis, J. T. (2015) ‘Getting up to speed with transcription elongation by RNA polymerase II.’, *Nature reviews. Molecular cell biology*. Nature Publishing Group, 16(3), pp. 167–177. doi: 10.1038/nrm3953.
- Juven-Gershon, T., Cheng, S. and Kadonaga, J. T. (2006) ‘Rational design of a super core promoter that enhances gene expression’, *Nature Methods*, 3(11), pp. 917–922. doi: 10.1038/nmeth937.
- Kadonaga, J. T. (2012) ‘Perspectives on the RNA polymerase II core promoter’, *Wiley Interdisciplinary Reviews: Developmental Biology*, 1(1), pp. 40–51. doi: 10.1002/wdev.21.
- Kamada, K., Horiuchi, T., Ohsumi, K., Shimamoto, N. and Morikawa, K. (1996) ‘Structure of a replication-terminator protein complexed with DNA.’, *Nature*, 383(6601), pp. 598–603. doi: 10.1038/383598a0.
- Karlic, R., Chung, H.-R., Lasserre, J., Vlahovicek, K. and Vingron, M. (2010) ‘Histone modification levels are predictive for gene expression’, *Proceedings of the National Academy of Sciences*, 107(7), pp. 2926–2931. doi: 10.1073/pnas.0909344107.
- Katsamba, P. S., Park, S. and Laird-Offringa, I. A. (2002) ‘Kinetic studies of RNA-protein interactions using surface plasmon resonance.’, *Methods (San Diego, Calif.)*, 26(2), pp. 95–104. doi: 10.1016/S1046-2023(02)00012-9.
- Kim, J. B. and Sharp, P. A. (2001) ‘Positive transcription elongation factor B phosphorylates hSPT5 and RNA polymerase II carboxyl-terminal domain independently of cyclin-dependent kinase-activating kinase.’, *The Journal of biological chemistry*, 276(15), pp. 12317–23. doi: 10.1074/jbc.M010908200.
- Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. a, Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G. and Greenberg, M. E. (2010) ‘Widespread transcription at neuronal activity-regulated enhancers.’, *Nature*. Nature Publishing Group, 465(7295), pp. 182–7. doi: 10.1038/nature09033.
- Kim, T.-K. and Shiekhattar, R. (2016) ‘Diverse regulatory interactions of long noncoding RNAs.’, *Current opinion in genetics & development*, 36, pp. 73–82. doi: 10.1016/j.gde.2016.03.014.

- Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. a, Richmond, T. a, Wu, Y., Green, R. D. and Ren, B. (2005) ‘A high-resolution map of active promoters in the human genome.’, *Nature*, 436(7052), pp. 876–80. doi: 10.1038/nature03877.
- König, J., Zarnack, K., Luscombe, N. M. and Ule, J. (2011) ‘Protein-RNA interactions: new genomic technologies and perspectives.’, *Nature reviews. Genetics*. Nature Publishing Group, 13(2), pp. 77–83. doi: 10.1038/nrg3141.
- Krebs, A. R., Imanci, D., Hoerner, L., Gaidatzis, D., Burger, L. and Schübeler, D. (2017) ‘Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters’, *Molecular Cell*, pp. 1–12. doi: 10.1016/j.molcel.2017.06.027.
- Kruesi, W. S., Core, L. J., Waters, C. T., Lis, J. T. and Meyer, B. J. (2013) ‘Condensin controls recruitment of RNA polymerase ii to achieve nematode X-chromosome dosage compensation’, *eLife*, 2013(2), pp. 1–31. doi: 10.7554/eLife.00808.
- Kwak, H., Fuda, N. J., Core, L. J. and Lis, J. T. (2013) ‘Precise maps of RNA polymerase reveal how promoters direct initiation and pausing.’, *Science (New York, N.Y.)*, 339(6122), pp. 950–3. doi: 10.1126/science.1229386.
- Lai, F., Orom, U. a, Cesaroni, M., Beringer, M., Taatjes, D. J., Blobel, G. a and Shiekhattar, R. (2013) ‘Activating RNAs associate with Mediator to enhance chromatin architecture and transcription.’, *Nature*. Nature Publishing Group, 494(7438), pp. 497–501. doi: 10.1038/nature11884.
- Latulippe, D. R., Szeto, K., Ozer, A., Duarte, F. M., Kelly, C. V., Pagano, J. M., White, B. S., Shalloway, D., Lis, J. T. and Craighead, H. G. (2013) ‘Multiplexed microcolumn-based process for efficient selection of RNA aptamers.’, *Analytical chemistry*. American Chemical Society, 85(6), pp. 3417–24. doi: 10.1021/ac400105e.
- Lavender, C. A., Cannady, K. R., Hoffman, J. A., Trotter, K. W., Gilchrist, D. A., Bennett, B. D., Burkholder, A. B., Burd, C. J., Fargo, D. C., Archer, T. K., Butler, J., Kadonaga, J., Boyle, A., Davis, S., Shulha, H., Meltzer, P., Margulies, E., Lenhard, B., Sandelin, A., Carninci, P., Core, L., Waterfall, J., Lis, J., Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M., Seila, A., Calabrese, J., Levine, S., Yeo, G., Rahl, P., Core, L., Martins, A., Danko, C., Waters, C., Siepel, A., Flynn, R., Almada, A., Zamudio, J., Sharp, P., Ntini, E., Jarvelin, A., Bornholdt, J., Chen, Y., Boyd, M., Duttke, S., Lacadie, S., Ibrahim, M., Glass, C., Corcoran, D., Scruggs, B., Gilchrist, D., Nechaev, S., Muse, G., Burkholder, A., Minchiotti, G., Nocera, P. Di, Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Djebali, S., Davis, C., Merkel, A., Dobin, A., Lassmann, T., Mayer, A., Iulio, J. di, Maleri, S., Eser, U., Vierstra, J., Nechaev, S., Fargo, D., Santos, G. dos, Liu, L., Gao, Y., Gardiner-Garden, M., Frommer, M., Mathelier, A., Zhao, X., Zhang, A., Parcy, F., Worsley-Hunt, R., Siepel, A., Bejerano, G., Pedersen, J., Hinrichs, A., Hou, M., Bucher, P., Gilchrist, D., Adelman, K., Takaku, M., Grimm, S., Shimbo, T., Perera, L., Menafra, R., Burd, C., Ward, J., Crusselle-Davis, V., Kissling, G., Phadke, D., Goodman, R., Smolik, S., Kim, T., Xu, Z., Clauder-Munster, S., Steinmetz, L., Buratowski, S., Shearwin, K., Callen, B., Egan, J., Carissimi, C., Laudadio, I., Cipolletta, E., Gioiosa, S., Mihailovich, M., Fryer, C., Archer, T., Martin, M., Langmead, B., Trapnell, C., Pop, M., Salzberg, S., Pruitt, K., Brown, G., Hiatt, S., Thibaud-Nissen, F., Astashyn, A., Jothi, R., Cuddapah, S., Barski, A., Cui, K., Zhao, K., Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y., Crooks, G., Hon, G., Chandonia, J., Brenner, S., Grant, C., Bailey, T., Noble, W., Bailey, T., Boden, M., Buske, F., Frith, M., Grant, C., Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Trapnell, C.,

- Roberts, A., Goff, L., Pertea, G., Kim, D., Ramirez, F., Dundar, F., Diehl, S., Gruning, B. and Manke, T. (2016) 'Downstream Antisense Transcription Predicts Genomic Features That Define the Specific Chromatin Environment at Mammalian Promoters', *PLOS Genetics*. Edited by B. van Steensel. Public Library of Science, 12(8), p. e1006224. doi: 10.1371/journal.pgen.1006224.
- Lee, C., Li, X., Hechmer, A., Eisen, M., Biggin, M. D., Venters, B. J., Jiang, C., Li, J., Pugh, B. F. and Gilmour, D. S. (2008) 'NELF and GAGA factor are linked to promoter-proximal pausing at many genes in *Drosophila*.' *Molecular and cellular biology*, 28(10), pp. 3290–300. doi: 10.1128/MCB.02224-07.
- Lee, J. T. (2012) 'Epigenetic regulation by long noncoding RNAs.' *Science (New York, N.Y.)*, 338(6113), pp. 1435–9. doi: 10.1126/science.1231776.
- Lee, T. I. and Young, R. A. (2013) 'Transcriptional regulation and its misregulation in disease', *Cell*. Elsevier Inc., 152(6), pp. 1237–1251. doi: 10.1016/j.cell.2013.02.014.
- Lenhard, B., Sandelin, A. and Carninci, P. (2012) 'Metazoan promoters: Emerging characteristics and insights into transcriptional regulation', *Nature Reviews Genetics*. Nature Publishing Group, 13(4), pp. 233–245. doi: 10.1038/nrg3163.
- Leontis, N. B., Stombaugh, J. and Westhof, E. (2002) 'The non-Watson-Crick base pairs and their associated isostericity matrices.' *Nucleic acids research*, 30(16), pp. 3497–531. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=134247&tool=pmcentrez&rendertype=abstract> (Accessed: 11 January 2014).
- Levine, M. (2011) 'Paused RNA polymerase II as a developmental checkpoint.' *Cell*. Elsevier, 145(4), pp. 502–11. doi: 10.1016/j.cell.2011.04.021.
- Li, J. and Gilmour, D. S. (2013) 'Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor.' *The EMBO journal*, 32(13), pp. 1829–41. doi: 10.1038/emboj.2013.111.
- Li, J., Liu, Y., Rhee, H. S., Ghosh, S. K. B., Bai, L., Pugh, B. F. and Gilmour, D. S. (2013) 'Kinetic Competition between Elongation Rate and Binding of NELF Controls Promoter-Proximal Pausing', *Molecular Cell*. Elsevier Inc., 50(5), pp. 711–722. doi: 10.1016/j.molcel.2013.05.016.
- Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C. and Darnell, R. B. (2008) 'HITS-CLIP yields genome-wide insights into brain alternative RNA processing.' *Nature*, 456(7221), pp. 464–9. doi: 10.1038/nature07488.
- Lieberman-aiden, E., Berkum, N. L. Van, Williams, L., Imakaev, M., Ragozy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S. and Dekker, J. (2009) 'Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome', *Science*, 33292(October), pp. 289–294.
- Limpert, E. and Stahel, W. a (2011) 'Problems with using the normal distribution--and ways to improve quality and efficiency of data analysis.' *PloS one*, 6(7), p. e21403. doi: 10.1371/journal.pone.0021403.
- Lis, J. T., Mason, P., Peng, J., Price, D. H. and Werner, J. (2000) 'P-TEFb kinase recruitment and function at heat shock loci P-TEFb kinase recruitment and function at heat shock loci', pp. 792–803. doi: 10.1101/gad.14.7.792.

- Love, M. I., Anders, S. and Huber, W. (2014) *Differential analysis of count data - the DESeq2 package*. doi: 10.1101/002832.
- Lukong, K. E., Chang, K., Khandjian, E. W. and Richard, S. (2008) 'RNA-binding proteins in human genetic disease.', *Trends in genetics : TIG*, 24(8), pp. 416–25. doi: 10.1016/j.tig.2008.05.004.
- Luo, Z., Lin, C. and Shilatifard, A. (2012) 'The super elongation complex (SEC) family in transcriptional control', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 13(9), pp. 543–547. doi: 10.1038/nrm3417.
- Mahat, D. B., Kwak, H., Booth, G. T., Jonkers, I. H., Danko, C. G., Patel, R. K., Waters, C. T., Munson, K., Core, L. J. and Lis, J. T. (2016) 'Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq)', *Nature Protocols*. Nature Research, 11(8), pp. 1455–1476. doi: 10.1038/nprot.2016.086.
- Mahat, D. B., Salamanca, H. H., Duarte, F. M., Danko, C. G. and Lis, J. T. (2016) 'Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation', *Molecular Cell*, 62(1), pp. 63–78. doi: 10.1016/j.molcel.2016.02.025.
- Mahat, D. B., Salamanca, H. H., Duarte, F. M., Danko, C. G. and Lis, J. T. (2016) 'Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation', *Molecular Cell*. Elsevier Inc., 62(1), pp. 63–78. doi: 10.1016/j.molcel.2016.02.025.
- Mandal, S. S., Chu, C., Wada, T., Handa, H., Shatkin, A. J. and Reinberg, D. (2004) 'Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 101(20), pp. 7572–7. doi: 10.1073/pnas.0401493101.
- Marshall, N. F. and Price, D. H. (1992) 'Control of formation of two distinct classes of RNA polymerase II elongation complexes.', *Molecular and cellular biology*, 12(5), pp. 2078–90. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=364379&tool=pmcentrez&rendertype=abstract>.
- Martin, L., Meier, M., Lyons, S. M., Sit, R. V, Marzluff, W. F., Quake, S. R. and Chang, H. Y. (2012) 'Systematic reconstruction of RNA functional motifs with high-throughput microfluidics.', *Nature methods*, 9(12), pp. 1192–4. doi: 10.1038/nmeth.2225.
- Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J. A. and Churchman, L. S. (2015a) 'Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution', *Cell*. Elsevier Inc., 161(3), pp. 541–554. doi: 10.1016/j.cell.2015.03.010.
- Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J. A. and Churchman, L. S. (2015b) 'Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution', *Cell*, 161(3), pp. 541–554. doi: 10.1016/j.cell.2015.03.010.
- McCall, M. J., Hendry, P. and Jennings, P. A. (1992) 'Minimal sequence requirements for ribozyme activity.', *Proceedings of the National Academy of Sciences of the United States of America*, 89(13), pp. 5710–4. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=49366&tool=pmcentrez&rendertype=abstract> (Accessed: 24 November 2014).
- Mccullum, E. O., Williams, B. A. R., Zhang, J. and Chaput, J. C. (2010) 'In Vitro Mutagenesis

- Protocols'. Edited by J. Braman. Totowa, NJ: Humana Press (Methods in Molecular Biology), 634(3), pp. 103–109. doi: 10.1007/978-1-60761-652-8.
- Miller, E. L., Hargreaves, D. C., Kadoch, C., Chang, C.-Y., Calarco, J. P., Hodges, C., Buenrostro, J. D., Cui, K., Greenleaf, W. J., Zhao, K. and Crabtree, G. R. (2017) 'TOP2 synergizes with BAF chromatin remodeling for both resolution and formation of facultative heterochromatin', *Nature Structural & Molecular Biology*. Nature Publishing Group, 24(4), pp. 344–352. doi: 10.1038/nsmb.3384.
- Missra, A. and Gilmour, D. S. (2010) 'Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex.', *Proceedings of the National Academy of Sciences of the United States of America*, 107(25), pp. 11301–6. doi: 10.1073/pnas.1000681107.
- Mohanty, B. K. (1998) 'Mechanistic Studies on the Impact of Transcription on Sequence-specific Termination of DNA Replication and Vice Versa', *Journal of Biological Chemistry*, 273(5), pp. 3051–3059. doi: 10.1074/jbc.273.5.3051.
- Mohanty, B. K., Sahoo, T. and Bastia, D. (1996) 'The relationship between sequence-specific termination of DNA replication and transcription.', *The EMBO journal*, 15(10), pp. 2530–9. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=450185&tool=pmcentrez&rendertype=abstract> (Accessed: 21 July 2013).
- Mulcair, M. D., Schaeffer, P. M., Oakley, A. J., Cross, H. F., Neylon, C., Hill, T. M. and Dixon, N. E. (2006) 'A molecular mousetrap determines polarity of termination of DNA replication in E. coli.', *Cell*. Elsevier, 125(7), pp. 1309–19. doi: 10.1016/j.cell.2006.04.040.
- Mulugu, S., Potnis, a, Shamsuzzaman, Taylor, J., Alexander, K. and Bastia, D. (2001) 'Mechanism of termination of DNA replication of Escherichia coli involves helicase-contrahelicase interaction.', *Proceedings of the National Academy of Sciences of the United States of America*, 98(17), pp. 9569–74. doi: 10.1073/pnas.171065898.
- Murakami, K., Elmlund, H., Kalisman, N., Bushnell, D. A., Adams, C. M., Azubel, M., Elmlund, D., Levi-Kalisman, Y., Liu, X., Gibbons, B. J., Levitt, M. and Kornberg, R. D. (2013) 'Architecture of an RNA polymerase II transcription pre-initiation complex', *Science*, 342(6159). doi: 10.1126/science.1238724.
- Muse, G. W., Gilchrist, D. a, Nechaev, S., Shah, R., Parker, J. S., Grissom, S. F., Zeitlinger, J. and Adelman, K. (2007) 'RNA polymerase is poised for activation across the genome.', *Nature genetics*, 39(12), pp. 1507–11. doi: 10.1038/ng.2007.21.
- Nechaev, S., Fargo, D. C., dos Santos, G., Liu, L., Gao, Y. and Adelman, K. (2010) 'Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila.', *Science (New York, N.Y.)*, 327(5963), pp. 335–8. doi: 10.1126/science.1181421.
- Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., Maurano, M. T., Humbert, R., Rynes, E., Wang, H., Vong, S., Lee, K., Bates, D., Diegel, M., Roach, V., Dunn, D., Neri, J., Schafer, A., Hansen, R. S., Kutayavin, T., Giste, E., Weaver, M., Canfield, T., Sabo, P., Zhang, M., Balasundaram, G., Byron, R., MacCoss, M. J., Akey, J. M., Bender, M. a, Groudine, M., Kaul, R. and Stamatoyannopoulos, J. a (2012) 'An expansive human regulatory lexicon encoded in transcription factor footprints.', *Nature*. Nature Publishing Group, 489(7414), pp. 83–90. doi: 10.1038/nature11212.

- Nguyen, V. T., Kiss, T., Michels, A. A. and Bensaude, O. (2001) '7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes.', *Nature*, 414(6861), pp. 322–5. doi: 10.1038/35104581.
- Ni, X., Castanares, M., Mukherjee, A. and Lupold, S. E. (2011) 'Nucleic acid aptamers: clinical applications and promising new horizons.', *Current medicinal chemistry*, 18(27), pp. 4206–14. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3260938&tool=pmcentrez&rendertype=abstract> (Accessed: 5 February 2014).
- Nilson, K. A., Guo, J., Turek, M. E., Brogie, J. E., Delaney, E., Luse, D. S. and Price, D. H. (2015) 'THZ1 Reveals Roles for Cdk7 in Co-transcriptional Capping and Pausing', *Molecular Cell*. Elsevier Inc., 59(4), pp. 576–587. doi: 10.1016/j.molcel.2015.06.032.
- Nilson, K. A., Lawson, C. K., Mullen, N. J., Ball, C. B., Spector, B. M., Meier, J. L. and Price, D. H. (2017) 'Oxidative stress rapidly stabilizes promoter-proximal paused Pol II across the human genome', *Nucleic Acids Research*. Oxford University Press, 45(19), pp. 11088–11105. doi: 10.1093/nar/gkx724.
- Nishimura, K., Fukagawa, T., Takisawa, H., Kakimoto, T. and Kanemaki, M. (2009) 'An auxin-based degron system for the rapid depletion of proteins in nonplant cells', *Nature Methods*. Nature Publishing Group, 6(12), pp. 917–922. doi: 10.1038/nmeth.1401.
- Nojima, T., Gomes, T., Grosso, A. R. F., Kimura, H., Dye, M. J., Dhir, S., Carmo-Fonseca, M. and Proudfoot, N. J. (2015) 'Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing', *Cell*, 161(3), pp. 526–540. doi: 10.1016/j.cell.2015.03.027.
- Nutiu, R., Friedman, R. C., Luo, S., Khrebtukova, I., Silva, D., Li, R., Zhang, L., Schroth, G. P. and Burge, C. B. (2011) 'Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument.', *Nature biotechnology*, 29(7), pp. 659–64. doi: 10.1038/nbt.1882.
- Ohler, U., Liao, G. C., Niemann, H. and Rubin, G. M. (2002) 'Computational analysis of core promoters in the Drosophila genome', *Genome Biol*, 3(12), pp. 1–12.
- Ozer, A., Pagano, J. M. and Lis, J. T. (2014) 'New Technologies Provide Quantum Changes in the Scale, Speed, and Success of SELEX Methods and Aptamer Characterization.', *Molecular therapy. Nucleic acids*, 3(August), p. e183. doi: 10.1038/mtna.2014.34.
- Ozer, A., Tome, J. M., Friedman, R. C., Gheba, D., Schroth, G. P. and Lis, J. T. (2015) 'Quantitative assessment of RNA-protein interactions with high-throughput sequencing-RNA affinity profiling.', *Nature protocols*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved., 10(8), pp. 1212–33. doi: 10.1038/nprot.2015.074.
- Ozer, A., White, B. S., Lis, J. T. and Shalloway, D. (2013) 'Density-dependent cooperative non-specific binding in solid-phase SELEX affinity selection', *Nucleic Acids Research*, 41(14), pp. 7167–7175. doi: 10.1093/nar/gkt477.
- Pagano, J. M., Clingman, C. C. and Ryder, S. P. (2011) 'Quantitative approaches to monitor protein-nucleic acid interactions using fluorescent probes.', *RNA (New York, N.Y.)*, 17(1), pp. 14–20. doi: 10.1261/rna.2428111.
- Pagano, J. M., Kwak, H., Waters, C. T., Sprouse, R. O., White, B. S., Ozer, A., Szeto, K., Shalloway, D., Craighead, H. G. and Lis, J. T. (2014) 'Defining NELF-E RNA Binding in HIV-1 and Promoter-Proximal Pause Regions', *PLoS Genetics*. Edited by D. Schübeler. Public Library of Science, 10(1), p. e1004090. doi:

- 10.1371/journal.pgen.1004090.
- Peng, J. (1996) 'Control of RNA Polymerase II Elongation Potential by a Novel Carboxyl-terminal Domain Kinase', *Journal of Biological Chemistry*, 271(43), pp. 27176–27183. doi: 10.1074/jbc.271.43.27176.
- Peng, J. (1998) 'Identification of a Cyclin Subunit Required for the Function of Drosophila P-TEFb', *Journal of Biological Chemistry*, 273(22), pp. 13855–13860. doi: 10.1074/jbc.273.22.13855.
- Peng, J., Zhu, Y., Milton, J. T. and Price, D. H. (1998) 'Identification of multiple cyclin subunits of human P-TEFb', pp. 755–762.
- Perkel, J. M. (2015) *BioTechniques - Hacking the Sequencer*, *BioTechniques*. Available at: http://www.biotechniques.com/news/Hacking-the-Sequencer/biotechniques-356823.html?autnID=336695#.VOaCMfnd_CI (Accessed: 20 February 2015).
- Peterlin, B. M. and Price, D. H. (2006) 'Controlling the elongation phase of transcription with P-TEFb.', *Molecular cell*, 23(3), pp. 297–305. doi: 10.1016/j.molcel.2006.06.014.
- Pradeepa, M. M., Grimes, G. R., Kumar, Y., Olley, G., Taylor, G. C. A., Schneider, R. and Bickmore, W. A. (2016) 'Histone H3 globular domain acetylation identifies a new class of enhancers', *Nature Genetics*. Nature Publishing Group, 48(6), pp. 681–686. doi: 10.1038/ng.3550.
- Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M. S., Mapendano, C. K., Schierup, M. H. and Jensen, T. H. (2008) 'RNA exosome depletion reveals transcription upstream of active human promoters.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 322(5909), pp. 1851–4. doi: 10.1126/science.1164096.
- Prelich, G. (2002) 'RNA Polymerase II Carboxy-Terminal Domain Kinases: Emerging Clues to Their Function', *Eukaryotic Cell*, 1(2), pp. 153–162. doi: 10.1128/EC.1.2.153-162.2002.
- Proudfoot, N. J. (2016) 'Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut', *Science*, 352(6291), pp. 715–718. doi: 10.1126/science.
- Pugh, B. F. and Venters, B. J. (2016) 'Genomic Organization of Human Transcription Initiation Complexes.', *PloS one*. Public Library of Science, 11(2), p. e0149339. doi: 10.1371/journal.pone.0149339.
- Pundhir, S., Bagger, F. O., Lauridsen, F. B., Rapin, N. and Porse, B. T. (2016) 'Peak-valley-peak pattern of histone modifications delineates active regulatory elements and their directionality', *Nucleic Acids Research*, 44(9), pp. 4037–4051. doi: 10.1093/nar/gkw250.
- Rahl, P. B., Lin, C. Y., Seila, A. C., Flynn, R. a, McCuine, S., Burge, C. B., Sharp, P. a and Young, R. a (2010) 'c-Myc regulates transcriptional pause release.', *Cell*. Elsevier Ltd, 141(3), pp. 432–45. doi: 10.1016/j.cell.2010.03.030.
- Ramanathan, A., Robb, G. B., Chan, S. and Biolabs, N. E. (2016) 'mRNA capping : biological functions and applications', pp. 1–16. doi: 10.1093/nar/gkw551.
- Rao, J. N., Neumann, L., Wenzel, S., Schweimer, K., Rösch, P. and Wöhrle, B. M. (2006) 'Structural studies on the RNA-recognition motif of NELF E, a cellular negative transcription elongation factor involved in the regulation of HIV transcription.', *The Biochemical journal*, 400(3), pp. 449–56. doi: 10.1042/BJ20060421.
- Rao, J. N., Schweimer, K., Wenzel, S., Wöhrle, B. M. and Rösch, P. (2008) 'NELF-E RRM undergoes major structural changes in flexible protein regions on target RNA binding.', *Biochemistry*, 47(12), pp. 3756–61. doi: 10.1021/bi702429m.
- Rasmussen, E. B. and Lis, J. T. (1993) 'In vivo transcriptional pausing and cap formation on

- three *Drosophila* heat shock genes.’, *Proceedings of the National Academy of Sciences of the United States of America*, 90(17), pp. 7923–7. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=47259&tool=pmcentrez&rendertype=abstract>.
- Reinholt, S. J., Ozer, A., Lis, J. T. and Craighead, H. G. (2016) ‘Highly Multiplexed RNA Aptamer Selection using a Microplate-based Microcolumn Device’, *Scientific Reports*. Nature Publishing Group, 6(1), p. 29771. doi: 10.1038/srep29771.
- Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Bruggmann, S. a, Goodnough, L. H., Helms, J. a, Farnham, P. J., Segal, E. and Chang, H. Y. (2007) ‘Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs.’, *Cell*, 129(7), pp. 1311–23. doi: 10.1016/j.cell.2007.05.022.
- Roberts, D. B. (David B. . (1998) *Drosophila : a practical approach*. IRL Press at Oxford University Press. Available at: <https://books.google.com/books/about/Drosophila.html?id=uoLwAAAAMAAJ> (Accessed: 17 January 2018).
- Roeder, R. G. and Rutter, W. J. (1969) ‘Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms’, *Nature*, 224(5216), pp. 234–237. doi: 10.1038/224234a0.
- Rougvie, a E. and Lis, J. T. (1988) ‘The RNA polymerase II molecule at the 5’ end of the uninduced hsp70 gene of *D. melanogaster* is transcriptionally engaged.’, *Cell*, 54(6), pp. 795–804. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/3136931>.
- Ryder, S. P., Recht, M. I. and Williamson, J. R. (2008) ‘Quantitative analysis of protein-RNA interactions by gel mobility shift.’, *Methods in molecular biology (Clifton, N.J.)*, 488, pp. 99–115. doi: 10.1007/978-1-60327-475-3_7.
- Sainsbury, S., Bernecky, C. and Cramer, P. (2015) ‘Structural basis of transcription initiation by RNA polymerase II’, *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 16(3), pp. 129–143. doi: 10.1038/nrm3952.
- Salamanca, H. H., Antonyak, M. A., Cerione, R. A., Shi, H. and Lis, J. T. (2014) ‘Inhibiting heat shock factor 1 in human cancer cells with a potent RNA aptamer.’, *PloS one*. Edited by S. D. Westerheide. Public Library of Science, 9(5), p. e96330. doi: 10.1371/journal.pone.0096330.
- Salamanca, H. H., Fuda, N., Shi, H. and Lis, J. T. (2011) ‘An RNA aptamer perturbs heat shock transcription factor activity in *Drosophila melanogaster*.’, *Nucleic acids research*, 39(15), pp. 6729–40. doi: 10.1093/nar/gkr206.
- Salim, N. N. and Feig, A. L. (2009) ‘Isothermal titration calorimetry of RNA.’, *Methods (San Diego, Calif.)*, 47(3), pp. 198–205. doi: 10.1016/j.ymeth.2008.09.003.
- Sanger, F., Nicklen, S. and Coulson, A. R. (1977) ‘DNA sequencing with chain-terminating inhibitors’, *Proceedings of the National Academy of Sciences*, 74(12), pp. 5463–5467. doi: 10.1073/pnas.74.12.5463.
- Satya Prakash Avva, S. V. and Hart, C. M. (2016) ‘Characterization of the *drosophila* BEAF-32A and BEAF-32B insulator proteins’, *PLoS ONE*, 11(9), pp. 1–17. doi: 10.1371/journal.pone.0162906.
- Saunders, A., Core, L. J. and Lis, J. T. (2006) ‘Breaking barriers to transcription elongation.’, *Nature reviews. Molecular cell biology*, 7(8), pp. 557–67. doi: 10.1038/nrm1981.
- Schaukowitz, K., Joo, J.-Y., Liu, X., Watts, J. K., Martinez, C. and Kim, T.-K. (2014) ‘Enhancer RNA Facilitates NELF Release from Immediate Early Genes’, *Molecular Cell*. Elsevier Inc., pp. 1–14. doi: 10.1016/j.molcel.2014.08.023.

- Schuller, R., Forne, I., Straub, T., Schriebeck, A., Texier, Y., Shah, N., Decker, T. M., Cramer, P., Imhof, A. and Eick, D. (2016) ‘Heptad-Specific Phosphorylation of RNA Polymerase II CTD’, *Molecular Cell*, 61(2), pp. 305–314. doi: 10.1016/j.molcel.2015.12.003.
- Schütze, T., Wilhelm, B., Greiner, N., Braun, H., Peter, F., Mörl, M., Erdmann, V. A., Lehrach, H., Konthur, Z., Menger, M., Arndt, P. F. and Glökler, J. (2011) ‘Probing the SELEX process with next-generation sequencing.’, *PloS one*. Edited by J. D. Hoheisel. Public Library of Science, 6(12), p. e29604. doi: 10.1371/journal.pone.0029604.
- Schwalb, B., Michel, M., Zacher, B., Frühauf, K., Demel, C., Tresch, A., Gagneur, J. and Cramer, P. (2016) ‘TT-seq maps the human transient transcriptome’, *Science (New York, N.Y.)*. American Association for the Advancement of Science, 352(6290), pp. 1225–8. doi: 10.1126/science.aad9841.
- Scott, W. G., Finch, J. T. and Klug, A. (1995) ‘The crystal structure of an AII-RNAhammerhead ribozyme: A proposed mechanism for RNA catalytic cleavage’, *Cell*, 81(7), pp. 991–1002. doi: 10.1016/S0092-8674(05)80004-2.
- Scruggs, B. S., Gilchrist, D. A., Nechaev, S., Muse, G. W., Burkholder, A., Fargo, D. C. and Adelman, K. (2015) ‘Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin.’, *Molecular cell*. doi: 10.1016/j.molcel.2015.04.006.
- Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., Young, R. A. and Sharp, P. A. (2008) ‘Divergent transcription from active promoters’, *Science*, 322(5909), pp. 1849–1851. doi: 10.1126/science.1162253.
- Serfling, E., Jasin, M. and Schaffner, W. (1985) ‘Enhancers and eukaryotic gene transcription’, *Trends in Genetics*, 1(C), pp. 224–230. doi: 10.1016/0168-9525(85)90088-5.
- Serganov, A. and Nudler, E. (2013) ‘A decade of riboswitches.’, *Cell*. Elsevier, 152(1–2), pp. 17–24. doi: 10.1016/j.cell.2012.12.024.
- Sevilimedu, A., Shi, H. and Lis, J. T. (2008) ‘TFIIB aptamers inhibit transcription by perturbing PIC formation at distinct stages’, *Nucleic Acids Research*, 36(9), pp. 3118–3127. doi: 10.1093/nar/gkn163.
- Shaner, N. C., Campbell, R. E., Steinbach, P. a, Giepmans, B. N. G., Palmer, A. E. and Tsien, R. Y. (2004) ‘Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein.’, *Nature biotechnology*, 22(12), pp. 1567–72. doi: 10.1038/nbt1037.
- Shao, W. and Zeitlinger, J. (2017) ‘Paused RNA polymerase II inhibits new transcriptional initiation’, *Nature Publishing Group*. Nature Publishing Group, (May). doi: 10.1038/ng.3867.
- She, R., Chakravarty, A. K., Layton, C. J., Chircus, L. M., Andreasson, J. O. L., Damaraju, N., McMahon, P. L., Buenrostro, J. D., Jarosz, D. F. and Greenleaf, W. J. (2017) ‘Comprehensive and quantitative mapping of RNA-protein interactions across a transcribed eukaryotic genome.’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 114(14), pp. 3619–3624. doi: 10.1073/pnas.1618370114.
- Sheffield, P., Garrard, S. and Derewenda, Z. (1999) ‘Overcoming Expression and Purification Problems of RhoGDI Using a Family of “ Parallel ” Expression Vectors’, 39, pp. 34–39.
- Shetty, A., Kallgren, S. P., Demel, C., Maier, K. C., Spatt, D., Alver, B. H., Cramer, P., Park, P. J. and Winston, F. (2017) ‘Spt5 Plays Vital Roles in the Control of Sense and Antisense Transcription Elongation’, *Molecular Cell*. Elsevier Inc., 66(1), p. 77–88.e5. doi:

- 10.1016/j.molcel.2017.02.023.
- Shi, H., Hoffman, B. E. and Lis, J. T. (1999) 'RNA aptamers as effective protein antagonists in a multicellular organism', *Proceedings of the National Academy of Sciences*, 96(18), pp. 10033–10038. doi: 10.1073/pnas.96.18.10033.
- Shui, B., Ozer, A., Zipfel, W., Sahu, N., Singh, A., Lis, J. T., Shi, H. and Kotlikoff, M. I. (2012) 'RNA aptamers that functionally interact with green fluorescent protein and its derivatives.', *Nucleic acids research*, 40(5), p. e39. doi: 10.1093/nar/gkr1264.
- Siebert, M. and Söding, J. (2014) 'Universality of core promoter elements?', *Nature*, 511(7510), pp. E11–E12. doi: 10.1038/nature13587.
- Smith, E. and Shilatifard, A. (2014) 'Enhancer biology and enhanceropathies', *Nature Structural & Molecular Biology*. Nature Publishing Group, 21(3), pp. 210–219. doi: 10.1038/nsmb.2784.
- Sundaram, P., Kurniawan, H., Byrne, M. E. and Wower, J. (2013) 'Therapeutic RNA aptamers in clinical trials.', *European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences*, 48(1–2), pp. 259–71. doi: 10.1016/j.ejps.2012.10.014.
- Szeto, K., Latulippe, D. R., Ozer, A., Pagano, J. M., White, B. S., Shalloway, D., Lis, J. T. and Craighead, H. G. (2013) 'RAPID-SELEX for RNA aptamers.', *PloS one*. Public Library of Science, 8(12), p. e82667. doi: 10.1371/journal.pone.0082667.
- Szeto, K., Reinholt, S. J., Duarte, F. M., Pagano, J. M., Ozer, A., Yao, L., Lis, J. T. and Craighead, H. G. (2014) 'High-throughput binding characterization of RNA aptamer selections using a microplate-based multiplex microcolumn device.', *Analytical and bioanalytical chemistry*. Springer, 406(11), pp. 2727–32. doi: 10.1007/s00216-014-7661-7.
- Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S. Z., Penrad-Mobayed, M., Sachs, L. M., Ruan, X., Wei, C.-L., Liu, E. T., Wilczynski, G. M., Plewczynski, D., Li, G. and Ruan, Y. (2015) 'CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription.', *Cell*. NIH Public Access, 163(7), pp. 1611–27. doi: 10.1016/j.cell.2015.11.024.
- Tenenbaum, S. A., Carson, C. C., Lager, P. J. and Keene, J. D. (2000) 'Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays.', *Proceedings of the National Academy of Sciences of the United States of America*, 97(26), pp. 14085–90. doi: 10.1073/pnas.97.26.14085.
- Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kuttyavin, T., Lajoie, B., Lee, B. K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E. and Stamatoyannopoulos, J. A. (2012) 'The accessible chromatin landscape of the human genome', *Nature*. Nature Publishing Group, 489(7414), pp. 75–82. doi: 10.1038/nature11232.

- Tome, J. M., Ozer, A., Pagano, J. M., Gheba, D., Schroth, G. P. and Lis, J. T. (2014) 'Comprehensive analysis of RNA-protein interactions by high-throughput sequencing–RNA affinity profiling', *Nature Methods*. Nature Publishing Group. doi: 10.1038/nmeth.2970.
- Traut, T. W. (1994) 'Physiological concentrations of purines and pyrimidines i', pp. 1–22.
- Tuerk, C. and Gold, L. (1990) 'Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.', *Science (New York, N.Y.)*, 249(4968), pp. 505–510. doi: 10.1126/science.2200121.
- Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. and Luscombe, N. M. (2009) 'A census of human transcription factors: Function, expression and evolution', *Nature Reviews Genetics*, 10(4), pp. 252–263. doi: 10.1038/nrg2538.
- Vihervaara, A., Mahat, D. B., Guertin, M. J., Chu, T., Danko, C. G., Lis, J. T. and Sistonon, L. (2017) 'Transcriptional response to stress is pre-wired by promoter and enhancer architecture', *Nature Communications*. Springer US, 8(1), pp. 1–15. doi: 10.1038/s41467-017-00151-0.
- Vo Ngoc, L., Wang, Y.-L., Kassavetis, G. A. and Kadonaga, J. T. (2017) 'The punctilious RNA polymerase II core promoter.', *Genes & development*. Cold Spring Harbor Laboratory Press, 31(13), pp. 1289–1301. doi: 10.1101/gad.303149.117.GENES.
- Vos, S. M., Pöllmann, D., Caizzi, L., Hofmann, K. B., Rombaut, P., Zimniak, T., Herzog, F. and Cramer, P. (2016) 'Architecture and RNA binding of the human negative elongation factor', *eLife*, 5, p. e14981. doi: 10.7554/eLife.14981.
- Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, a, Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G. a, Winston, F., Buratowski, S. and Handa, H. (1998) 'DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs.', *Genes & development*, 12(3), pp. 343–56. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=316480&tool=pmcentrez&rendertype=abstract>.
- Wada, T., Takagi, T., Yamaguchi, Y., Watanabe, D. and Handa, H. (1998) 'Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro.', *The EMBO journal*, 17(24), pp. 7395–403. doi: 10.1093/emboj/17.24.7395.
- Weber, C. M., Ramachandran, S. and Henikoff, S. (2014) 'Nucleosomes Are Context-Specific, H2A.Z-Modulated Barriers to RNA Polymerase', *Molecular Cell*, 53(5), pp. 819–830. doi: 10.1016/j.molcel.2014.02.014.
- Weber, C. M., Ramachandran, S. and Henikoff, S. (2014) 'Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase.', *Molecular cell*. Elsevier Inc., 53(5), pp. 819–30. doi: 10.1016/j.molcel.2014.02.014.
- Weber, J., Jelinek, W. and Darnell, J. E. (1977) 'The Definition of a Large Viral Transcription Unit Late in Ad2 Infection of HeLa Cells: Mapping of Nascent RNA Molecules Labeled in Isolated Nuclei', *Cell*, 10(0), pp. 611–616. Available at: https://ac-els-cdn-com.proxy.library.cornell.edu/0092867477900939/1-s2.0-0092867477900939-main.pdf?_tid=88ceee86-fb35-11e7-a23a-00000aacb35e&acdnat=1516159481_75292c7fb4e39f51386b4f8d7aaf2f3d (Accessed: 16 January 2018).
- Wei, B., Jolma, A., Sahu, B., Orre, L. M., Zhong, F., Zhu, F., Kivioja, T., Sur, I. K., Lehtio, J.,

- Taipale, M. and Taipale, J. (2017) ‘Strong binding activity of few transcription factors is a major determinant of open chromatin’, *bioRxiv*, p. 204743. doi: 10.1101/204743.
- Weintraub, A. S., Li, C. H., Zamudio, A. V., Sigova, A. A., Hanne, N. M., Day, D. S., Abraham, B. J., Cohen, M. A., Nabet, B., Buckley, D. L., Guo, Y. E., Hnisz, D., Jaenisch, R., Bradner, J. E., Gray, N. S. and Young, R. A. (2018) ‘YY1 Is a Structural Regulator of Enhancer-Promoter Loops’, *Cell*. Elsevier Inc., (172), pp. 1–16. doi: 10.1016/j.cell.2017.11.008.
- Wen, Y. and Shatkin, A. J. (1999) ‘Transcription elongation factor hSPT5 stimulates mRNA capping.’, *Genes & development*. Cold Spring Harbor Laboratory Press, 13(14), pp. 1774–9. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10421630> (Accessed: 19 October 2016).
- Whyte, W. A., Bilodeau, S., Orlando, D. A., Hoke, H. A., Frampton, G. M., Foster, C. T., Cowley, S. M. and Young, R. A. (2012) ‘Enhancer decommissioning by LSD1 during embryonic stem cell differentiation’, *Nature*. Nature Publishing Group, 482(7384), pp. 221–225. doi: 10.1038/nature10805.
- Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., Young, R. A., Lander, E. S., Mesirov, J. P. and al., et (2013) ‘Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes’, *Cell*. Elsevier, 153(2), pp. 307–319. doi: 10.1016/j.cell.2013.03.035.
- Wong, I. and Lohman, T. M. (1993) ‘A double-filter method for nitrocellulose-filter binding: application to protein-nucleic acid interactions.’, *Proceedings of the National Academy of Sciences of the United States of America*, 90(12), pp. 5428–32. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=46733&tool=pmcentrez&rendertype=abstract> (Accessed: 23 October 2013).
- Wu, C.-H., Lee, C., Fan, R., Smith, M. J., Yamaguchi, Y., Handa, H. and Gilmour, D. S. (2005) ‘Molecular characterization of Drosophila NELF.’, *Nucleic acids research*, 33(4), pp. 1269–79. doi: 10.1093/nar/gki274.
- Wu, C.-H., Yamaguchi, Y., Benjamin, L. R., Horvat-Gordon, M., Washinsky, J., Enerly, E., Larsson, J., Lambertsson, A., Handa, H. and Gilmour, D. (2003) ‘NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila.’, *Genes & development*, 17(11), pp. 1402–14. doi: 10.1101/gad.1091403.
- Xayaphoummine, A., Bucher, T. and Isambert, H. (2005) ‘Kinofold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots.’, *Nucleic acids research*, 33(Web Server issue), pp. W605–10. doi: 10.1093/nar/gki447.
- Yamada, T., Yamaguchi, Y., Inukai, N., Okamoto, S., Mura, T. and Handa, H. (2006) ‘P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation.’, *Molecular cell*, 21(2), pp. 227–37. doi: 10.1016/j.molcel.2005.11.024.
- Yamaguchi, Y., Inukai, N., Narita, T., Wada, T. and Handa, H. (2002) ‘Evidence that Negative Elongation Factor Represses Transcription Elongation through Binding to a DRB Sensitivity-Inducing Factor / RNA Polymerase II Complex and RNA’, 22(9), pp. 2918–2927. doi: 10.1128/MCB.22.9.2918.
- Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, a, Sugimoto, S., Hasegawa, J. and Handa, H. (1999) ‘NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation.’, *Cell*, 97(1), pp. 41–51. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/10199401>.

- Yang, J. P., Ramos, E. and Corces, V. G. (2012) ‘The BEAF-32 insulator coordinates genome organization and function during the evolution of *Drosophila* species’, *Genome Research*, 22(11), pp. 2199–2207. doi: 10.1101/gr.142125.112.
- Yang, Z., Yik, J. H. N., Chen, R., He, N., Jang, M. K., Ozato, K. and Zhou, Q. (2005) ‘Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4.’, *Molecular cell*, 19(4), pp. 535–45. doi: 10.1016/j.molcel.2005.06.029.
- Young, L., Sung, J., Stacey, G. and Masters, J. R. (2010) ‘Detection of Mycoplasma in cell cultures’, *Nature Protocols*, 5(5), pp. 929–934. doi: 10.1038/nprot.2010.43.
- Young, R. S., Kumar, Y., Bickmore, W. A. and Taylor, M. S. (2016) ‘Bidirectional transcription marks accessible chromatin and is not specific to enhancers’, *bioRxiv*, p. 48629. doi: 10.1101/048629.
- Zabidi, M. A. and Stark, A. (2016) ‘Regulatory Enhancer–Core–Promoter Communication via Transcription Factors and Cofactors’, *Trends in Genetics*. Elsevier Ltd, 32(12), pp. 801–814. doi: 10.1016/j.tig.2016.10.003.
- Zaret, K. S. and Carroll, J. S. (2011) ‘Pioneer transcription factors : establishing competence for gene expression Parameters affecting transcription factor access to target sites in chromatin Initiating events in chromatin : pioneer factors bind first’, *Genes and Development*, pp. 2227–2241. doi: 10.1101/gad.176826.111.GENES.
- Zhao, X., Shi, H., Sevilimedu, A., Liachko, N., Nelson, H. C. M. and Lis, J. T. (2006) ‘An RNA aptamer that interferes with the DNA binding of the HSF transcription activator’, *Nucleic Acids Research*, 34(13), pp. 3755–3761. doi: 10.1093/nar/gkl470.
- Zhou, R., Mohr, S., Hannon, G. J. and Perrimon, N. (2013) ‘Inducing RNAi in *Drosophila* cells by transfection with dsRNA.’, *Cold Spring Harbor protocols*, 2013(5), pp. 461–3. doi: 10.1101/pdb.prot074351.
- Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S., Cramer, P. and Taipale, J. (2017) ‘The interaction landscape between transcription factors and the nucleosome’, *bioRxiv*. doi: 10.1101/240598.
- Zobeck, K. L., Buckley, M. S., Zipfel, W. R. and Lis, J. T. (2010) ‘Recruitment timing and dynamics of transcription factors at the Hsp70 loci in living cells.’, *Molecular cell*. Elsevier, 40(6), pp. 965–75. doi: 10.1016/j.molcel.2010.11.022.
- Zuker, M. (2003) ‘Mfold web server for nucleic acid folding and hybridization prediction.’, *Nucleic acids research*, 31(13), pp. 3406–15. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=169194&tool=pmcentrez&rendertype=abstract> (Accessed: 2 January 2014).
- Zykovich, A., Korf, I. and Segal, D. J. (2009) ‘Bind-n-Seq: high-throughput analysis of in vitro protein-DNA interactions using massively parallel sequencing.’, *Nucleic acids research*, 37(22), p. e151. doi: 10.1093/nar/gkp802.