

# **Cdk12 Regulates DNA Repair Genes by Suppressing Intronic Polyadenylation**

by

**Sara Jane Dubbury**

B.A. Biological Sciences, Northwestern University (2010)

Submitted to the Department of Biology  
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY  
at the  
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2018

© 2018 Massachusetts Institute of Technology  
All rights reserved

Signature of Author: \_\_\_\_\_

Sara J. Dubbury  
Department of Biology  
January 12, 2018

Certified by: \_\_\_\_\_

Phillip A. Sharp  
Institute Professor of Biology  
Thesis Supervisor

Accepted by: \_\_\_\_\_

Amy E. Keating  
Professor of Biology  
Co-Chair, Biology Graduate Committee



# Cdk12 Regulates DNA Repair Genes by Suppressing Intronic Polyadenylation

by

Sara Jane Dubbury

Submitted to the Department of Biology on January 12, 2018 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Biology

## Abstract

During transcription, cyclin-dependent kinases (CDKs) dynamically phosphorylate the C-terminal domain (CTD) of RNA Polymerase II (RNAPII) to recruit factors that coordinate transcription and mRNA biogenesis. Cdk12 phosphorylates Serine 2 (Ser2) of the RNAPII CTD, a modification associated with the regulation of transcription elongation, splicing, and cleavage/polyadenylation. Unlike other transcriptional CDKs that regulate most expressed genes, Cdk12 depletion abrogates the expression of homologous recombination (HR) genes relatively specifically, suppressing the HR DNA damage repair pathway and sensitizing cells to genotoxic stresses that cause replication fork collapse, such as Parp1 inhibitors. The proposed role for Cdk12 in regulating HR is clinically significant for two reasons. First, Cdk12 loss-of-function mutations populate high-grade serous ovarian carcinoma and castration-resistant prostate tumors raising the possibility that Cdk12 mutational status may predict the effectiveness of chemotherapeutics that target HR-deficient tumors. Second, readily available small molecule inhibitors of Cdk12 induce sensitization of HR-competent tumors to Parp1 inhibitors *in vivo* raising the possibility that inhibitors against Cdk12 could be used as chemotherapeutics. Despite this growing clinical interest, the mechanism behind Cdk12's regulation of HR genes remains unknown.

Here we show that Cdk12 suppresses intronic polyadenylation (IPA) and that this mechanism explains the exquisite sensitivity of HR genes to Cdk12 loss. We find that Cdk12 globally enhances transcription elongation rate to kinetically suppress IPA events. Many HR genes harbor multiple IPA sites per gene, and the cumulative effect of these sites accounts for the increased sensitivity of HR genes to Cdk12. Finally, we find evidence that Cdk12 LOF mutations and deletions cause upregulation of IPA sites in HR genes in human tumors. Our results define the mechanism by which Cdk12 regulates transcription, mRNA biogenesis, and the HR pathway. This work clarifies the biological function of CDK12 and underscores its potential both as a chemotherapeutic target and as a tumor biomarker.

### Thesis Supervisor:

Phillip A. Sharp, Institute Professor of Biology





# Acknowledgements

This thesis is the culmination of a long and winding journey, filled with both highs and lows, during which several people provided critical mentorship, encouragement, and support. Without these people, none of this work would have been possible; I would like to take a moment to thank them here.

First and foremost, I would like to express my deepest gratitude to my advisor, Phil, for his unwavering support of me, both personally and professionally. Phil challenged me to think critically and confidently and to develop my scientific independence, while always providing me with a much-needed encouraging word, scientific suggestion, or metaphorical push at just the right moment. Phil's passion for science, dedication to excellence, and commitment to his people make the Sharp lab a very special place to work, and because of this, the last several years have been a formative and positive experience. I will always be deeply indebted to Phil both for his mentorship and for fostering such a wonderful lab environment. It is truly an honor and a privilege to call myself a "Sharpie."

I would like to thank the members of my thesis committee, Chris Burge and Rick Young, for their helpful feedback. They have been enthusiastic about the Cdk12 work from the beginning and very supportive of my career. This work would never have been successful without technical assistance from the Koch Institute Core Facilities, particularly Stuart Levine from the BioMicro Center and Glenn Paradis from the Flow Cytometry Facility. Both Glenn and Stuart graciously answered my incessant questions and provided key technical expertise. I am especially grateful to Stuart, who went above and beyond to successfully troubleshoot some rather unusual sequencing issues in order to provide me with high quality ChIP data.

To my fellow Sharpies, I am so grateful to have worked with such an amazing group of scientists. I have learned something from each of you, and your constant willingness to share ideas, reagents, and constructive criticism has facilitated my growth in numerous ways. I cannot imagine a better group of people to commiserate with during bad science days and to celebrate with during moments of success. You have made coming to work each day enjoyable and for that I will always be appreciative.

I would particularly like to thank the graduate students of Room 411—Mohini (Mo), Andrew, and Albert. As a new student, you welcomed me into the "party room," showed me the ropes, and provided me with a tremendous amount of scientific advice and personal friendship. To Mo, your sage advice has served me well throughout graduate school. I very much appreciate you always bringing a grounded perspective to the table and our friendship throughout the years. To Andrew, my first baymate, thank you for adding a great deal of levity to my first few years of graduate school and for always being there to pick me up when the chips were down. I would also like to thank Tim, who taught me several experimental techniques early on in grad school, and has become a good friend throughout. Between our mutual love of addicting podcasts and our annual beer bet, you made long hours in lab a bit more fun. Finally, to Margarita, aka "Lab Mom," thank you for always being there with a kind word, an encouraging smile, and a big hug. You definitely took care of me during grad school, made the Sharp Lab a little homier, and always put a smile on my face.

I would especially like to thank my unofficial second mentor, collaborator, and current baymate, Paul. Paul is one of the most thoughtful and creative scientists that I know, and I have learned and grown tremendously as a result of our collaboration. In addition to fostering a great scientific relationship, Paul is also a genuinely wonderful and caring mentor. He has always looked out for my best interest and supported my career in graduate school in every possible way. I would particularly like to thank him for his extreme patience, constant encouragement, and spot-on advice even when I am sure that I was being difficult. I really cannot imagine having done this thesis without him, and all that I can say, inadequately, is “thank you.”

I consider myself exceedingly fortunate to have joined the BioREF program during my second year of graduate school. As a BioREF, I learned a tremendous amount about people, science, and myself. I will carry these lessons with me as I move on from MIT, and they are certainly invaluable to shaping me as a person. The past and current BioREFs are some of the most amazing people and scientists that I will probably ever meet, and I am so lucky that some of them are my closest friends. They have seen me at my best and they have certainly dealt with me at my worst. I will forever be grateful for them occasionally dragging me out of lab and for their support and encouragement over the years. I would especially like to thank Josh and Brian, for recruiting me to the BioREFs and for staying close friends and colleagues throughout the years. The two of you have been constant cheerleaders for me personally and professionally, and your friendship has truly meant a lot. To Lauren, Laura, Emma, Rob, and Sonya- thank you for your friendship, words of encouragement, and support, even when I am at my most unreasonable. It means the world to me. Finally, to Frank Solomon, our BioREF faculty mentor, thank you so much for always taking the time to talk with me about life and science. You have believed in me from the beginning and have always provided me with much needed wisdom and perspective. You have always made yourself available, at any time of the day and on any day of the week, to support me and countless other graduate students. I really don't know what I would have done in graduate school without your wit, wisdom, advice, and encouragement.

I am extremely grateful that I worked at the Muddy Charles Pub during my time in graduate school. My weekly shift always provided a much-needed break from lab. Working with a bunch of PhD candidate bartenders is definitely a unique experience, and I always appreciated that all of the staff implicitly understood the highs and lows of graduate school. After late nights in lab or bad science days, I could always count on the Muddy for both a beer and a friend to commiserate with. To Mike and my fellow bartenders, thank you for everything.

To my friends from before graduate school and at MIT, thank you for your constant faith in me. You have been there for it all and you have supported me in more ways than I can count. I could not have done this without you and for that, I am very grateful. I would especially like to thank my classmates Rini, Kaitlin, Simina, Courtney, Monica, Peter, Rob E., and Daniel. I joined graduate school with a truly incredible group of people and nobody quite understands the highs and lows of grad school like your fellow classmates. We entered the trenches together, and we made it out alive. You have provided invaluable friendship, encouragement, and scientific advice over the years and will continue to be wonderful friends and colleagues of mine in the future. I would also like to thank the ladies of the Lees lab, especially Helen. Having wonderful friends on the fourth floor of the Koch always brightened my day, provided much-needed breaks, and put a smile on my face. Also, thank you for proofreading this thesis.

None of this would have been possible without the unconditional love and support of my family. To Uncle Tom, thank you for coaxing me off ski runs, teaching me how to play Blackjack, proofreading fellowship applications, and for all of the “Uncle Tomisms” that taught me perspective and important life lessons. To my cousin, Jeff, you are like the older brother that I never had. Thank you for always looking out for me, for teaching me how to ski, for all of the fun that we’ve had in Tahoe, and for the many late-night phone calls and Facebook messages filled with love and support. To my younger brother, Eric, thank you for your unwavering enthusiasm and encouragement and for helping Dad organize one awesome surprise graduation party. Finally, most importantly, to my parents, words cannot do justice to how important you have been in my life. You believed in me from the start, always encouraged me to shoot for the stars, and never failed to pick me up when I fell. You taught me to never give up on myself or anything that I set my mind too, and none of this would have been possible without your support. Mom and Dad, this one is for you.

-Sara J. Dubbury



# Table of Contents

<b>Abstract .....</b>	<b>3</b>
<b>Acknowledgements .....</b>	<b>5</b>
<b>Table of Contents .....</b>	<b>9</b>
<b>Chapter 1. Introduction .....</b>	<b>11</b>
The RNA Polymerase II C-Terminal Domain .....	12
Transcription Regulation .....	16
<i>Promoter Proximal Pause Release .....</i>	<i>17</i>
<i>Transcription Elongation Through the Gene Body .....</i>	<i>19</i>
mRNA Cleavage and Polyadenylation .....	22
<i>Cis RNA Elements Define Cleavage and Polyadenylation Sites .....</i>	<i>24</i>
<i>The Cleavage and Polyadenylation Machinery.....</i>	<i>27</i>
<i>Involvement of the RNAPII CTD in Cleavage and Polyadenylation .....</i>	<i>28</i>
Alternative Cleavage and Polyadenylation.....	31
<i>Mechanisms that Regulate Alternative Cleavage and Polyadenylation.....</i>	<i>32</i>
<i>Biological Consequences of IPA Usage .....</i>	<i>36</i>
Biological Consequences of Cdk12 Loss .....	41
References.....	45

<b>Chapter 2. Cdk12 Regulates DNA Repair Genes by Suppressing Intronic Polyadenylation.....</b>	<b>63</b>
Abstract.....	64
Results and Discussion .....	65
Materials and Methods.....	74
References.....	91
Acknowledgements.....	95
Figures.....	96
<b>Chapter 3. Conclusions and Future Directions.....</b>	<b>121</b>
Conclusions.....	122
Future Directions .....	126
<i>Functional Role for Cdk13 in Transcription and mRNA Biogenesis.....</i>	<i>126</i>
<i>Dissecting the Molecular Factors Downstream of Cdk12.....</i>	<i>128</i>
<i>Dissecting the Function of Cdk12 in Ovarian Cancer.....</i>	<i>128</i>
<i>Dissecting the Function of Cdk12 in Breast Cancer .....</i>	<i>129</i>
References.....	131
<b>Biographical Note .....</b>	<b>135</b>

# **Chapter 1**

## **Introduction**

Tightly regulated gene expression establishes appropriate cell identity and allows cells to respond to a variety of intracellular and extracellular signals. Dysregulation of gene expression causes aberrant cellular behavior and promotes many diseases, including cancer. The proper expression of protein-coding genes requires several steps, including: (1) transcription of the nascent message (pre-mRNA); (2) processing of the pre-mRNA into mature, messenger RNA (mRNA); (3) export of the mRNA from the nucleus; and finally, (4) translation of the mRNA into protein. Multiple modes of regulation influence each step in gene expression allowing for exquisite control over individual genes as well as the coordinated expression of many functionally related gene products.

One such step in gene expression, maturation of a pre-mRNA into mRNA, requires several processing steps including: (1) addition of a 7-methylguanosine “cap” to the 5’ end of the transcript, (2) removal of non-coding, intervening RNA sequences (introns) through pre-mRNA splicing, and (3) cleavage and polyadenylation of the 3’ end of the pre-mRNA. These processing steps occur co-transcriptionally, i.e. during transcription of the nascent RNA, allowing for the coordinated regulation of pre-mRNA processing and transcription. This thesis will focus on co-transcriptional mRNA 3’ end processing, and its consequences for a cancer-relevant gene expression program.

### **The RNA Polymerase II C-Terminal Domain**

Three RNA Polymerases (RNA Polymerase I, II, and III) transcribe the eukaryotic genome (Roeder and Rutter, 1969). Each RNA Polymerase is responsible for transcribing a unique set of genomic targets. RNA Polymerase I and III transcribe most of the non-coding, housekeeping RNAs in cells. Meanwhile, RNA Polymerase II (RNAPII) transcribes all protein-



coding mRNAs and several classes of non-coding RNAs. All three eukaryotic RNA Polymerases share significant structural and functional homology with each other and to bacterial and archaeal RNA Polymerases (Cramer et al., 2008). Despite this shared homology, the largest subunit of RNAPII, Rpb1, has a unique C-terminal extension, known as the CTD, consisting of tandem repeats of the heptapeptide (Tyr<sub>1</sub>-Ser<sub>2</sub>-Pro<sub>3</sub>-Thr<sub>4</sub>-Ser<sub>5</sub>-Pro<sub>6</sub>-Ser<sub>7</sub>) (Allison et al., 1985; Corden et al., 1985). The RNAPII CTD is conserved among eukaryotes, and the number of heptapeptide repeats increases with genomic complexity ranging from 26 repeats in *S. cerevisiae* (Allison et al., 1985; Nonet et al., 1987) to 52 repeats in mammals (Corden et al., 1985). The RNAPII CTD is essential for viability in all eukaryotes (Allison et al., 1988; Bartolomei et al., 1988; Nonet et al., 1987; Zehring et al., 1988), however it is not required for *in vitro* transcription (Buratowski and Sharp, 1990; Kim and Dahmus, 1989; Zehring et al., 1988). This observation suggests that the RNAPII CTD is not required for the catalytic activity of RNAPII but that it performs *in vivo*-specific regulatory functions during the transcription of eukaryotic mRNAs.

Early experiments purifying RNAPII suggested that Rpb1 exists in two distinct forms *in vivo* that share identical primary amino acid sequence but differ in electrophoretic mobility (Corden et al., 1985; Schwartz and Roeder, 1975). Shortly after the discovery of the RNAPII CTD, Cadena and Dahmus observed that actively engaged RNAPII was phosphorylated on serine and threonine residues of the RNAPII CTD accounting for the slower running form of RNAPII (Cadena and Dahmus, 1987). We now know that the RNAPII CTD is extensively phosphorylated at the Ser2, Ser5, Ser7, Tyr1, and Thr4 positions of the heptapeptide motif, in addition to several other post-translational modifications (Harlen and Churchman, 2017; Zaborowska et al., 2016). The Ser2 and Ser5 residues of the RNAPII CTD are the most

extensively phosphorylated positions of the heptapeptide repeat, showing almost two orders of magnitude more phosphorylation *in vivo* than the other residues (Suh et al., 2016). Here, we will limit our discussion to Ser2 and Ser5 phosphorylation.

Seminal studies using chromatin immunoprecipitation qPCR showed that Ser2 and Ser5 phosphorylation of the RNAPII CTD showed distinct spatial patterns across actively transcribed genes— an observation consistent with these modifications occurring during specific steps of transcription and pre-mRNA processing (Chonghui and Sharp, 2003; Komarnitsky et al., 2000; Morris et al., 2005). These studies observed that unphosphorylated RNAPII is recruited to the promoter of genes where it is rapidly phosphorylated on Ser5 of the RNAPII CTD during transcription initiation. Ser5 phosphorylated RNAPII levels peak immediately adjacent to the promoter and decrease across the gene body. Conversely, Ser2 phosphorylated RNAPII levels remain low near the promoter and begin to accumulate during productive elongation across the gene, peaking towards the 3' end. More recently, genome-wide studies in yeast (Kim et al., 2010; Mayer et al., 2010) and mammals (Nojima et al., 2015; Rahl et al., 2010) have confirmed these phosphorylation patterns. The preponderance of evidence now suggests that these RNAPII CTD phosphorylation patterns are responsible for the recruitment and activity of factors involved in regulating transcription and co-transcriptional mRNA processing (Buratowski, 2009; Hsin and Manley, 2012).

Several cyclin-dependent kinases (Cdks) and phosphatases are responsible for the dynamic phosphorylation of the RNAPII CTD across expressed genes. Like cell cycle Cdks, transcriptional Cdks are serine/threonine kinases that depend on the binding of cyclin cofactors for activity (Malumbres, 2014). Unlike cell cycle Cdks, whose activities fluctuate with the cell cycle, the activity of transcriptional Cdks is constrained to particular steps during transcription.

A good example of this is the main Ser5 RNAPII CTD kinase, Cdk7. Cdk7 is recruited to promoters in complex with the general transcription factor TFIID, which simultaneously recruits an ATP-dependent helicase (XPB) required for DNA melting and transcription initiation (Kim et al., 2000; Komarnitsky et al., 2000; Lu et al., 1992). The colocalized recruitment of Cdk7 and XPB to promoters allows for the rapid phosphorylation of Ser5 of the RNAPII CTD upon transcription initiation. RNAPII CTD Ser5 phosphorylation subsequently functions to release RNAPII from complexes at the promoter and to recruit and activate the mRNA 5' capping machinery (Buratowski, 2009; Komarnitsky et al., 2000; Rodriguez et al., 2000; Schroeder et al., 2000; Søgaard and Svejstrup, 2007). Cdk7 is therefore a well-studied example of how transcriptional Cdk activity is restricted to particular stages during transcription to spatially and temporally regulate transcription and co-transcriptional mRNA processing events.

RNAPII CTD Ser2 phosphorylation accumulates during RNAPII elongation throughout the gene body and is associated with recruiting factors that regulate transcription elongation, splicing, and cleavage and polyadenylation (Buratowski, 2009; Hsin and Manley, 2012). Until recently, only one RNAPII CTD Ser2 kinase, Cdk9, was known in metazoa. Phylogenetic studies identified two additional mammalian Cdks, Cdk12 and Cdk13, with significant homology to Cdk9 raising the possibility that additional RNAPII CTD Ser2 kinases existed in mammals (Guo and Stiller, 2004; Liu and Kipreos, 2000). Sequence homology suggested that orthologs of Cdk12 are conserved throughout metazoa, while Cdk13 is a mammalian-specific paralog. Cdk12 and Cdk13 are unusually large Cdks, 164 kDa and 165 kDa respectively, that bind the same cyclin co-factor, cyclin K (Bartkowiak et al., 2010; Blazek et al., 2011), and that share an almost identical, central kinase domain (greater than 92% conserved sequence identity across the 300 amino acid domain). Outside of their kinase domains, Cdk12 and Cdk13 are more divergent,

however they both contain arginine/serine-rich and proline-rich motifs in their N-terminus. In addition to these motifs, the N-terminus of Cdk13 contains a highly-repetitive, alanine-rich motif (Bösken et al., 2014; Dixon-Clarke et al., 2015; Greifenberg et al., 2016; Kohoutek and Blazek, 2012). Follow-up biochemical studies confirmed that both Cdk12 and Cdk13 phosphorylated the RNAPII CTD at position 2 *in vitro* and *in vivo* (Bartkowiak et al., 2010; Blazek et al., 2011).

Cdk9, the originally identified mammalian RNAPII CTD Ser2 kinase, has been extensively studied (discussed below). However, the function of Cdk12 and Cdk13, and why there are multiple RNAPII CTD Ser2 kinases in mammals, remains unknown. The remainder of this introduction focuses on the regulation of transcription elongation and mRNA 3' end processing, paying particular attention to the role of RNAPII CTD Ser2 phosphorylation in these processes. Finally, the biological consequences of Cdk12 loss will be reviewed.

## **Transcription Regulation**

Transcription of an mRNA involves several rate-limiting steps, each of which can be regulated to alter the amount of transcript produced (Fuda et al., 2009). Transcription typically begins with the recruitment of RNAPII (in complex with general transcription factors) to the promoter to form the pre-initiation complex. Once RNAPII is recruited to the promoter, the DNA is unwound and RNAPII initiates transcription across the first ~30-60 nucleotides, after which the RNAPII pauses at a promoter-proximal site. Paused RNAPII is subsequently “licensed” for productive elongation, at which point it efficiently transcribes the remainder of the gene. Upon reaching the end of the message, the nascent RNA is cleaved and transcription is rapidly terminated. Original models suggested that RNAPII recruitment to the promoter and transcription initiation were the key rate-limiting steps during RNAPII transcription. We now

know that regulated transcription elongation plays a crucial role in mRNA biogenesis (Jonkers and Lis, 2015). Initial events in the transcription cycle have been heavily reviewed (Fuda et al., 2009; Grünberg and Hahn, 2013; Saunders et al., 2006). Here, we will focus on the regulation of RNAPII elongation. RNAPII elongation is regulated in two main steps: (1) release from the promoter proximal pause and (2) elongation across the gene body (Jonkers and Lis, 2015).

### *Promoter Proximal Pause Release*

The first regulated step in transcription elongation occurs at the promoter proximal pause. During transcription initiation, RNAPII interacts with a set of protein machinery that prevents it from entering productive elongation, including general transcription factors (Kwak et al., 2013; Li and Gilmour, 2013), the negative elongation factor (NELF) (Yamaguchi et al., 1999), and the DRB-sensitivity-inducing factor (DSIF) (Wada et al., 1998a). These protein interactions render RNAPII “elongation incompetent” and the polymerase transcribes a short distance (~30-60 nucleotides) before pausing (Guenther et al., 2007; Kim et al., 2005; Krumm et al., 1992; Rahl et al., 2010; Rasmussen and Lis, 1993; Rougvie and Lis, 1988; Strobl and Eick, 1992). The Cdk9 subunit of the P-TEFb complex promotes release of paused RNAPII by phosphorylating NELF, DSIF, and the Ser2 position of the RNAPII CTD (Fujinaga et al., 2004; Kim and Sharp, 2001; Marshall and Price, 1992; 1995; Marshall et al., 1996; Wada et al., 1998b; Yamada et al., 2006; Zhu et al., 1997). Phosphorylation of these substrates “licenses” RNAPII for transcription elongation by releasing NELF from the RNAPII complex (Fujinaga et al., 2004) and by switching the activity of DSIF from that of an elongation repressor to that of an elongation activator (Yamada et al., 2006). RNAPII CTD Ser2 phosphorylation further aids RNAPII’s transition to proficient elongation by recruiting elongation factors to RNAPII that facilitate its

elongation across a nucleosome template, such as chromatin remodelers, histone chaperones, and histone modifiers (discussed further below).

Studies using small molecule inhibitors of Cdk9 have demonstrated that P-TEFb mediated release of RNAPII from the promoter proximal pause is a universal step during transcription regulation (Henriques et al., 2013; Jonkers et al., 2014; Rahl et al., 2010). Increased RNAPII density is frequently observed adjacent to promoters and a vast majority of this RNAPII density is transcriptionally engaged (Core et al., 2012), suggesting that proximal RNAPII density mostly reflects paused polymerase. Interestingly, peaks of promoter-proximal RNAPII are only observed at approximately 40-70% of expressed genes, indicating that this step is only rate-limiting in a subset of genes (Core et al., 2008; Jonkers et al., 2014; Kwak et al., 2013; Min et al., 2011; Nechaev et al., 2010; Rahl et al., 2010). These genes are often involved in development or signaling pathways that need to be rapidly activated (Adelman et al., 2009; Aida et al., 2006; Krumm et al., 1992; Muse et al., 2007; Zeitlinger et al., 2007) suggesting that these genes may utilize promoter-proximal pausing to maintain genes in a transcriptionally “poised” but inactive state in preparation for rapid induction in response to extracellular or intracellular cues (Adelman and Lis, 2012). A preponderance of evidence now suggests that the degree to which the promoter proximal pause is rate limiting (and therefore regulates transcription) depends upon the balance between repressors and activators of elongation (Jonkers and Lis, 2015). Transcription factors recruited to the promoter by enhancers, such as Myc and Brd4, often recruit P-TEFb to promoters to shift this balance towards elongation competency and to promote gene expression (Jang et al., 2005; Rahl et al., 2010).

*Transcription Elongation Through the Gene Body*

The second step in regulated transcription elongation occurs after the RNAPII is “licensed” for productive elongation by P-TEFb. At this point, RNAPII enters productive elongation across the remainder of the gene body. Historically, RNAPII elongation through the gene body was considered a relatively simple, uniform process across genes. Accumulating evidence now suggests that productive elongation is significantly more complex than originally thought (Jonkers and Lis, 2015). We now know that elongation rates differ between and within genes by as much as three fold (Danko et al., 2013; Fuchs et al., 2014; Jonkers et al., 2014; Veloso et al., 2014). This variation can have significant effects on total mRNA accumulation, and especially, on co-transcriptional mRNA processing (discussed below in models of alternative cleavage and polyadenylation).

P-TEFb “licenses” RNAPII for productive elongation by suppressing the activity of elongation repressors (e.g. NELF and DSIF) and promoting the association of a series of elongation activators. However, the switch from paused polymerase to productively elongating polymerase does not occur instantly downstream of the pause site. Instead, transcription elongation within the first few kilobases is a relatively inefficient process only catalyzing the addition of ~0.5 kilobases per minute. By the time the polymerase transcribes ~15 kilobases downstream, the RNAPII elongation rate typically increases to a much more efficient ~2-5 kilobases per minute (Danko et al., 2013; Fuchs et al., 2014; Jonkers et al., 2014; Veloso et al., 2014). This disparity in elongation rates suggests that perhaps the switch towards an elongation efficient RNAPII is less of an “On/Off” switch and more of a gradual transition as elongation repressors are progressively lost from the complex and positive elongation factors are progressively gained (Jonkers and Lis, 2015). Notably, RNAPII CTD Ser2 phosphorylation continues to accumulate across the gene as RNAPII elongation rate accelerates suggesting a

potential role for RNAPII CTD Ser2 phosphorylation in recruiting elongation factors to mediate rate increases across this region (Harlen and Churchman, 2017).

Downstream of the promoter-proximal pause, RNAPII navigates additional barriers to elongation that can affect its rate. Most notably among these obstacles are nucleosomes, which pose a strong, mechanical barrier to RNAPII *in vivo* (Li et al., 2007). This effect can be observed most strongly at the first nucleosome downstream of the transcription start site (Li and Gilmour, 2013; Weber et al., 2014). Nucleosome-free regions flank promoters and facilitate both the recruitment of general transcription factors and RNAPII as well as transcription initiation (Li et al., 2007). The first nucleosome downstream of the TSS is the strongest nucleosome barrier to productive elongation as evidenced by up to 60% of polymerases backtracking or pausing at this position. Although this first nucleosome can colocalize with the promoter-proximal pause, it often downstream raising the possibility that this “pause” can act as a second rate-limiting step during elongation (Weber et al., 2014). Beyond the first stable nucleosome, transcription through the nucleosome template becomes significantly more efficient. However, density of nucleosomes and histone modifications can still alter how much of a barrier the nucleosomes pose to elongating RNAPII and therefore, how much of an affect they will have on its rate across a gene (Danko et al., 2013; Fuchs et al., 2014; Jonkers and Lis, 2015; Jonkers et al., 2014; Veloso et al., 2014).

Several positive transcription elongation factors complex with the RNAPII after the promoter-proximal pause. Most of these complexes aid RNAPII in transcribing across a DNA-nucleosome template. RNAPII navigates nucleosomes by recruiting nucleosome remodeling complexes, histone chaperones, and histone tail modifiers directly to the elongating RNAPII. This machinery facilitates the removal of nucleosomes downstream of the elongating RNAPII



and the post-translational modification of histone tails that loosen the association of DNA with nucleosomes. Several examples of these complexes have been extensively studied, such as FACT, Spt6, and Paf1 (Jonkers and Lis, 2015; Kwak and Lis, 2013). As RNAPII transcribes across the gene body, Ser2 phosphorylation continues to accumulate on the RNAPII CTD (Harlen and Churchman, 2017). This has led several groups to hypothesize that Ser2 phosphorylation aids in the progressive recruitment of positive elongation factors to the polymerase throughout transcription elongation. Indeed, some of these factors, most notably Spt6, have been shown to associate directly with Ser2 phosphorylation of the RNAPII CTD (Yoh et al., 2007; 2008).

Cdk9 of the P-TEFb complex was originally assumed to be the kinase responsible for increasing RNAPII CTD Ser2 phosphorylation downstream of the promoter proximal pause. However, Cdk12 and Cdk13 are now attractive candidates for this role based upon experiments in *Drosophila*. ChIP-qPCR of Cdk12 across heat shock genes showed that Cdk12 localized downstream of the promoter-proximal pause on actively transcribed genes. Follow-up immunofluorescence demonstrated the Cdk12 and Cdk9 localize independently on polytene chromosomes (Bartkowiak et al., 2010). Given Cdk9's well-established role at the promoter proximal pause, this suggests that perhaps, Cdk12 (and its paralog Cdk13) function downstream of Cdk9 to progressively increase RNAPII CTD Ser2 phosphorylation during transcription.

## mRNA Cleavage and Polyadenylation

Mammalian cell fractionation experiments from the early 1970s purified polysome-associated, translationally-active mRNAs and their nuclear-localized precursors, heterogeneous nuclear (hn)RNAs (pre-mRNAs). Biochemical characterization of these RNAs using RNases with different substrate specificities identified long poly(A) tails attached to the 3' ends of mRNAs (Adesnik et al., 1972; Darnell et al., 1971a; 1971b; Edmonds et al., 1971; Lee et al., 1971; Lim and Canellakis, 1970). Kinetic experiments using the transcriptional inhibitor, Actinomycin D, (Adesnik et al., 1972; Darnell et al., 1971a; Jelinek et al., 1973) as well as experiments comparing the sequence compositions of HeLa cell mRNA and DNA (Birnboim et al., 1973) suggested that mRNA poly(A) tails were untemplated and post-transcriptionally added to pre-mRNA. The importance of these poly(A) tails was evident early on after experiments showed that poly(A) tails were likely required for export of mRNA to the cytoplasm and for association of mRNA with translationally-active polysomes (Adesnik et al., 1972; Darnell et al., 1971a; Jelinek et al., 1973; Lim and Canellakis, 1970). Initial models assumed that the poly(A) tail was post-transcriptionally added to the 3' end of pre-mRNA after the pre-mRNA was released from RNAPII by transcription termination. This model was dismissed after pulse-labeling experiments were used to map nascent transcription at the mouse  $\beta$ -globin (Hofer and Darnell, 1981) and across two viral genomes SV40 (Ford and Hsu, 1978; Lai et al., 1978) and Adenovirus Type-2 (Fraser et al., 1979; Nevins and Darnell, 1978; Nevins et al., 1980). These experiments showed that nascent transcription continued past the mRNA poly(A) site while poly(A)-tailed mRNA kinetically appeared in a manner consistent with co-transcriptional cleavage of the nascent RNA. Moore and Sharp confirmed that 3' end formation of mRNA

requires endonucleolytic cleavage of the nascent RNA using a HeLa nuclear extract system (Moore and Sharp, 1984; 1985).

We now know that all pre-mRNAs, with the exception of canonical, replication-dependent histone genes (Adesnik and Darnell, 1972; Marzluff et al., 2008), are processed by co-transcriptional, endonucleolytic cleavage of the pre-mRNA followed by the post-transcriptional, untemplated addition of a 50-100 nucleotide poly(A) tail. These two independent, yet intrinsically linked, pre-mRNA processing steps are critical for proper gene expression. The first step, co-transcriptional endonucleolytic cleavage of the nascent RNA, defines the 3' end of the pre-mRNA and commits the pre-mRNA to polyadenylation and export from the nucleus. The advent of genome-wide approaches has revealed the extensive use of multiple cleavage and polyadenylation sites within mammalian genes (Derti et al., 2012; Hoque et al., 2013; Ozsolak et al., 2010; Wang et al., 2008). Differential usage of these sites produces alternative isoforms that can vary significantly by coding sequence, stability, translational efficiency, and localization (Tian and Manley, 2017). As the commitment step, regulation of co-transcriptional cleavage determines which isoform will be produced from a gene locus with more than one cleavage and polyadenylation site. The second step in mRNA 3' end formation, the polyadenylation reaction, is equally important. Proper addition of the poly(A) tail is required for release of mRNA from the RNAPII complex, export of the mRNA to the cytoplasm, stability of the mRNA, and translational efficiency (Eckmann et al., 2011; Moore and Proudfoot, 2009; Weill et al., 2012).

Extensive work over the last 35 years has elucidated many salient details about the cleavage and polyadenylation reaction. The development of recombinant DNA technology and cell-free nuclear extracts that recapitulate the cleavage and polyadenylation reaction *in vitro* (Manley, 1983; Moore and Sharp, 1984; 1985; Moore et al., 1986; Skolnik-David et al., 1987)

rapidly accelerated the characterization of the *cis*-acting RNA sequence elements and the *trans*-acting protein machinery involved in cleavage and polyadenylation. More recently, the development of next generation sequencing technologies has allowed researchers to probe the regulation of cleavage and polyadenylation events genome-wide. Here, we review our current understanding of the machinery and mechanisms involved in mammalian mRNA 3' end formation, paying particular attention to what regulates co-transcriptional cleavage at, and therefore usage of, a particular cleavage and polyadenylation site within a given mRNA.

### *Cis RNA Elements Define Cleavage and Polyadenylation Sites*

The cleavage and polyadenylation site on a pre-mRNA is defined by a set of RNA sequencing elements that are recognized by the cleavage and polyadenylation machinery. Three major sequence elements are involved in defining the cleavage and polyadenylation site in mammals: (1) the polyadenylation signal (PAS) located within 40 nucleotides upstream of the cleavage site, (2) the upstream sequence element (USE) that flanks the PAS site upstream of the cleavage site, and (3) the downstream sequence element (DSE) located downstream of the cleavage site (Tian and Graber, 2011). All three of these sequence elements are enriched at cleavage and polyadenylation sites throughout the genome and each sequence element is associated with a set of motifs. However, mammalian cleavage and polyadenylation sites are highly divergent. Many sites lack one or more of these *cis* elements or they utilize sequences at these elements that are highly divergent from the canonical motifs. In general, sites that consist of more sequence elements and that utilize sequences that are closer to the canonical motifs are more efficiently used.

The PAS motif is the most consensus and dominant of the *cis* sequence elements defining mammalian cleavage and polyadenylation sites. The PAS motif was originally identified by purifying six highly-abundant mRNAs from human, rabbit, mouse, and chicken and sequencing their 3' ends (Proudfoot and Brownlee, 1976). This experiment identified a conserved, 6 nucleotide motif (AAUAAA) located 20-30 nucleotides upstream of the cleavage and polyadenylation site in all six mRNAs. Given both its proximity to the poly(A) tail and its high degree of conservation, Proudfoot and Brownlee hypothesized that the PAS motif was involved in defining the cleavage and polyadenylation site of an mRNA. Several subsequent studies confirmed the importance of the PAS site for proper cleavage and polyadenylation by showing that point mutations and deletions of the PAS site in the SV40 late gene (Fitzgerald and Shenk, 1981; Wickens and Stephenson, 1984), the Adenovirus E1A gene (Montell et al., 1983), and the human  $\alpha$ -2 globin gene (Higgs et al., 1983) resulted in abrogated cleavage and polyadenylation of their respective pre-mRNA.

As more cDNAs were cloned and sequenced, a naturally occurring variant of the PAS motif (AUUAAA) was identified. In light of this finding, *in vitro* cleavage and polyadenylation reactions were used to analyze the effect of single nucleotide variants of the canonical (AAUAAA) PAS motif on cleavage and polyadenylation efficiency (Sheets et al., 1990; Wilusz et al., 1989). These experiments showed that most single nucleotide PAS variants abrogated efficient cleavage and polyadenylation activity *in vitro* (Sheets et al., 1990; Wilusz et al., 1989) with the exception of AUUAAA, which retained ~80%, and AGUAAA, which retained ~30%, of the full cleavage and polyadenylation activity associated with the AAUAAA motif (Sheets et al., 1990). Analysis of expressed sequence tags in mammals (Beaudoing et al., 2000; Tian et al., 2005) and 3' end sequencing data (Derti et al., 2012; Hoque et al., 2013) has now provided an

approximate distribution of PAS motifs across the mammalian genome. A majority of used cleavage and polyadenylation sites (~80-90%) have a PAS site within 40 nucleotides upstream of a cleavage site (Beaudoing et al., 2000; Tian et al., 2005). These sites are heavily biased towards the “canonical” PAS motifs (AAUAAA and AUUAAA), which were identified at ~55% and ~15% of cleavage sites, respectively (Beaudoing et al., 2000; Tian et al., 2005). Of the remaining ~30% of cleavage and polyadenylation sites, ~10-20% of sites harbor one of the other less efficient single nucleotide variants of the PAS motif and the remaining ~5-10% of sites have no detectable PAS motif within 40 nucleotides of the cleavage site (Beaudoing et al., 2000; Tian et al., 2005). Further analysis of genomic PAS motif distribution revealed that cleavage and polyadenylation sites that are more conserved and those that are in the 3’ most distal region of a gene are enriched for the stronger, canonical PAS motif (AAUAAA) (Tian et al., 2005; 2007).

Shortly after the identification of the PAS motif, several studies suggested that additional sequence upstream of the PAS motif, the upstream sequence element (USE), (Brackenridge and Proudfoot, 2000; Carswell and Alwine, 1989; DeZazzo et al., 1991; Moreira et al., 1995; Valsamakis et al., 1991) and that sequence downstream of the cleavage site, the downstream sequence element (DSE), (Gil and Proudfoot, 1984; 1987; Levitt et al., 1989; McDevitt et al., 1986; McLauchlan et al., 1985) contributed to the definition and strength of a mammalian cleavage and polyadenylation site. The motifs associated with the USE and DSE are significantly more degenerate than the PAS motif. The USE is typically positioned within 40 nucleotides upstream of the PAS motif and is characterized by U-rich sequences and the motifs, UGUA and UAUU (Hu et al., 2005; Hutchins et al., 2008; Venkataraman et al., 2005). The DSE is located within ~40 nucleotides downstream of the cleavage site and consists of the GU-rich motifs GUGU or UGUG followed by U-rich sequences (>3 sequential uridines) (Gil and Proudfoot,

1987; Hu et al., 2005; Hutchins et al., 2008; Levitt et al., 1989; McDevitt et al., 1986; McLauchlan et al., 1985; Salisbury et al., 2006). The USE and DSE are enriched at mammalian cleavage and polyadenylation sites genome wide (Hu et al., 2005; Hutchins et al., 2008; Salisbury et al., 2006), however, sites without one or both elements exist. In general, inclusion of USE and DSE elements correlates with stronger cleavage and polyadenylation sites (Tian and Graber, 2011).

### *The Cleavage and Polyadenylation Machinery*

The *cis* sequence elements that define mammalian cleavage and polyadenylation sites recruit a complex set of protein machinery that processes the 3' end of pre-mRNA (Colgan and Manley, 1997; Shi and Manley, 2015). Over 85 individual proteins (~15-20 core factors and a host of auxiliary proteins) form the cleavage and polyadenylation machinery (Shi et al., 2009). The core machinery includes 4 multi-subunit complexes (CPSF- cleavage and polyadenylation specificity factor; CstF- cleavage stimulation factor; CFI- cleavage factor I; and CFII- cleavage factor II) in addition to several individual proteins, including Poly(A) Polymerase (PAP) (Shi and Manley, 2015; Takagaki et al., 1989). The protein machinery involved in cleavage and polyadenylation has been extensively characterized and reviewed (Colgan and Manley, 1997; Shi and Manley, 2015). Here, we briefly discuss the role of the core complexes.

Three of the four core complexes directly bind the RNA sequence elements at the cleavage and polyadenylation site. CPSF binds the PAS motif via three (CPSF160, Wdr33, and CPSF30) of its six subunits (CPSF160, Wdr33, CPSF100, CPSF73, Fip1, and CPSF30) (Bienroth et al., 1991; Chan et al., 2014; Keller et al., 1991; Murthy and Manley, 1992; 1995; Schönemann et al., 2014). In addition to recognizing the PAS motif, CPSF catalyzes the

endonucleolytic cleavage of the mRNA via its CPSF73 subunit (Mandel et al., 2006; Ryan et al., 2004), and recruits PAP directly to the cleavage machinery via the Fip1 and CPSF160 subunits (Kaufmann et al., 2004; Murthy and Manley, 1995). The CstF complex (CstF77, CstF50, and CstF64/CstF64 $\tau$ ) and the CFI complex (CFI25 and CFI59/CFI68) bind the DSE and USE respectively (Brown and Gilmartin, 2003; Gilmartin and Nevins, 1991; MacDonald et al., 1994; Rügsegger et al., 1996; Takagaki and Manley, 1997; Takagaki et al., 1990; Yang et al., 2010). Binding of these complexes stabilizes the interaction between CPSF and pre-mRNA, thereby enhancing usage of a given cleavage and polyadenylation site (Gilmartin and Nevins, 1989; 1991; Rügsegger et al., 1996). The function of the fourth core complex, CFII (Pcf11 and Clp1), is the least well understood. CFII does not bind RNA sequence elements directly, instead, CFII complexes with the core machinery through protein-protein interactions and likely serves a scaffolding role that enhances cleavage and polyadenylation (de Vries et al., 2000).

#### *Involvement of the RNAPII CTD in Cleavage and Polyadenylation*

The cleavage and polyadenylation reaction occurs co-transcriptionally, raising the possibility that mRNA 3' end processing is regulated through interactions with the RNAPII complex. Several lines of evidence now suggest that the RNAPII CTD and its post-translational modification are required for efficient mRNA 3' end processing. Initial biochemical evidence in support of such a model was reported by Hirose and Manley, who observed that efficient cleavage of an SV40 substrate required the inclusion of purified RNAPII in a reconstituted *in vitro* reaction (Hirose and Manley, 1998). Interestingly, both purified phosphorylated and unphosphorylated RNAPII CTDs were sufficient to recapitulate the requirement for RNAPII *in vitro* (Hirose and Manley, 1998). Subsequent work demonstrated that the ability of the RNAPII



CTD to facilitate the cleavage reaction was heavily dependent upon the number of heptapeptide repeats included and that increased numbers of consensus repeats promoted more efficient cleavage (Ryan et al., 2002). Concurrently, genetic evidence emerged suggesting a role for the RNAPII CTD in promoting efficient cleavage and polyadenylation. Studies using an  $\alpha$ -amanitin resistant polymerase lacking all but 5 heptapeptide repeats of the RNAPII CTD showed that severe truncation of the RNAPII CTD resulted in decreased cleavage and polyadenylation of mRNAs generated from several reporter constructs including the SV40 gene and the rabbit  $\beta$ -globin gene (McCracken et al., 1997). Similar observations were made in yeast, after a mutant RNAPII, lacking the entire CTD, was expressed in cells using a temperature-sensitive, but otherwise wild type, RNAPII. Growth of the yeast at the non-permissive temperature resulted in transcription by the RNAPII  $\Delta$ CTD mutant, which was shown to reduce cleavage and polyadenylation efficiency at several endogenous yeast genes as well as a reporter plasmid (Licatalosi et al., 2002).

Subsequent biochemical analyses revealed that several components of the cleavage and polyadenylation machinery interact with the RNAPII CTD, including components of the CPSF complex (Dichtl et al., 2002; McCracken et al., 1997), the CstF complex (Barillà et al., 2001; Davidson et al., 2014; Fong and Bentley, 2001; McCracken et al., 1997), and the CFII complex (Barillà et al., 2001; Dichtl et al., 2002; Licatalosi et al., 2002; Meinhart and Cramer, 2004). Intriguingly, one of the components of the CFII complex (Pcf11), binds the RNAPII CTD directly and its binding is dependent upon RNAPII CTD Ser2 phosphorylation (Barillà et al., 2001; Dichtl et al., 2002; Licatalosi et al., 2002; Meinhart and Cramer, 2004). Since RNAPII CTD Ser2 phosphorylation accumulates across gene bodies towards the 3' end of genes, this observation raised the intriguing hypothesis that the RNAPII CTD coordinates the recruitment of

the cleavage and polyadenylation machinery spatially and temporally with transcription elongation. Partial support of this model comes from studies in yeast (Ahn et al., 2004; Kim et al., 2010; 2004; Mayer et al., 2010) and human cells (Davidson et al., 2014) that show that elongation factors are recruited to RNAPII across the 3' ends of genes coincident with Ser2 phosphorylation. Depletion of RNAPII Ser2 phosphorylation by genetic knockout of Ctk1 kinase in yeast (Ahn et al., 2004; Skaar and Greenleaf, 2002), flavipiridol treatment of *Drosophila* cells (Ni et al., 2004), expression of an  $\alpha$ -amanitin resistant mutant RNAPII with all Ser2 residues mutated to alanine in human cells (Gu et al., 2012), or knockdown of Cdk12 in HeLa cells recapitulated the decreased cleavage and polyadenylation efficiency observed upon transcription with a RNAPII  $\Delta$ CTD mutant in yeast (Licatalosi et al., 2002).

Although this data supports a relatively simplistic model whereby Ser2 phosphorylation recruits cleavage and polyadenylation factors to the elongating polymerase, the situation has become significantly more complicated. Several groups have noted that RNAPII CTD Ser2 phosphorylation begins to accumulate over gene bodies shortly downstream of the transcription start site (regardless of the location of the cleavage and polyadenylation site), however cleavage and polyadenylation factors do not crosslink strongly to RNAPII until very close to the polyadenylation site (Ahn et al., 2004; Kim et al., 2010; 2004; Mayer et al., 2010). Additionally, several labs have noted that deletion of cleavage and polyadenylation sites abrogates not only the recruitment of cleavage and polyadenylation machinery to the elongating RNAPII by ChIP but it also decreases Ser2 phosphorylation of the RNAPII surrounding the polyadenylation site (Ahn et al., 2004; Davidson et al., 2014). This data suggests a reciprocal feedback between Ser2 phosphorylation and the cleavage and polyadenylation machinery. Several unanswered questions

remain as to the mechanisms and function of coupling between the cleavage and polyadenylation machinery and the CTD.

### **Alternative Cleavage and Polyadenylation**

It has long been recognized that many genes harbor more than one cleavage and polyadenylation site and that the regulated usage of these sites can generate distinct mRNA isoforms with unique functional behavior. The advent of RNA sequencing approaches to probe transcriptome-wide expression patterns has greatly increased our appreciation of the role that alternative cleavage and polyadenylation (APA) plays in transcriptome diversity and gene regulation (Derti et al., 2012; Hoque et al., 2013; Wang et al., 2008). Indeed, up to 80% of mammalian genes harbor more than one functional cleavage and polyadenylation site and therefore, can be regulated by APA (Derti et al., 2012; Hoque et al., 2013).

APA events can be classified into three main categories based upon the location of the alternative cleavage and polyadenylation site. Tandem 3'UTR-APA occurs when two or more cleavage and polyadenylation sites occur within the same 3' untranslated region (3'UTR). These events generate isoforms that encode the same protein coding sequence attached to different 3' UTRs. mRNA 3'UTR's harbor RNA sequence elements that alter mRNA stability, mRNA localization, translational efficiency, and even protein localization (Berkovits and Mayr, 2015; Elkon et al., 2013; Mayr, 2016; Tian and Manley, 2017). Usage of an upstream (proximal) tandem 3'UTR cleavage and polyadenylation site can produce an isoform with one or more of these sequence elements removed from the 3'UTR, altering its downstream expression and function. Global regulation of tandem 3' UTR-APAs has important functional roles in gene expression networks during T-cell activation (Sandberg et al., 2008), embryonic development (Ji

et al., 2009), and oncogenic transformation (Masamha et al., 2015; Mayr and Bartel, 2009). The biological consequences of tandem 3' UTR-APA and the proposed mechanisms that regulate the choice between these sites has been extensively reviewed and will not be discussed further here (Elkon et al., 2013; Mayr, 2016; Tian and Manley, 2017).

The other two types of APA events (Exonic APA and Intronic APA) occur when a cleavage and polyadenylation site upstream of the last exon is used. Exonic APA events occur when the alternative cleavage and polyadenylation site is located within an upstream exon. These events are extremely rare and typically produce mRNAs without stop codons— substrates for rapid clearance by non-stop decay (Frischmeyer et al., 2002; Tian and Manley, 2017). Intronic APA events utilize cleavage and polyadenylation sites located within introns (intronic polyadenylation sites, IPAs) upstream of the distal most exon. Usage of an IPA site generates an mRNA that can vary in its protein coding potential as well as its 3' UTR (Elkon et al., 2013; Tian and Manley, 2017). Here we review the mechanisms that regulate intronic versus distal polyadenylation site choice and the biological consequences of alternative intronic polyadenylation (IPA-APA).

### *Mechanisms that Regulate Alternative Cleavage and Polyadenylation*

The *cis* RNA sequence elements that define IPA sites are heavily skewed toward weaker cleavage and polyadenylation sites as compared to their distal, 3'-end counterparts (Tian et al., 2007). Presumably, the selective pressure towards stronger cleavage and polyadenylation sites at the 3' ends of genes helps to ensure that processing of the nascent transcript and subsequent transcription termination occur successfully after the distal-most exon of a gene. Conversely, weaker cleavage and polyadenylation sites at IPAs allow for both their suppression as well as

their dynamic upregulation. A preponderance of the literature suggests that three major classes of regulatory pathways exist to increase usage of these weaker IPA sites (Elkon et al., 2013; Tian and Manley, 2017). The first involves regulating the expression level of the cleavage and polyadenylation machinery to increase recognition of weaker polyadenylation sites. The second involves a kinetic competition model between cleavage at an IPA site and splicing of its encompassing intron. Meanwhile, the third involves suppression of intronic polyadenylation sites by the U1 snRNP. These models are not mutually exclusive, and regulation at a given IPA site can depend on contributions from multiple mechanisms.

The first mode of regulation depends upon the expression levels of the cleavage and polyadenylation machinery. This model was originally proposed by work examining regulation at the IgM locus. During B-cell activation, there is a striking shift in IgM isoform expression from cleavage and polyadenylation at a distal polyadenylation site to an IPA site. Early studies examining cleavage and polyadenylation at the IgM locus identified that B-cell activation strongly induced expression of CstF64 (the RNA-binding subunit of the CstF complex), which in turn, activated the weaker IPA site in IgM (Takagaki and Manley, 1998; Takagaki et al., 1996). From this work, Takagaki and Manley hypothesized that stronger cleavage and polyadenylation sites were preferentially bound over weaker cleavage and polyadenylation sites in cells that expressed limiting levels of the cleavage and polyadenylation machinery. Conversely, when the expression levels of cleavage and polyadenylation machinery were high (not limiting), both strong and weak cleavage sites could be recognized and used. IPA sites are transcribed before distal polyadenylation sites, and therefore, they become accessible for cleavage and polyadenylation first. A situation where the levels of cleavage and polyadenylation machinery are abundant would therefore increase the usage of IPA sites over distal polyadenylation sites.

Subsequent work has supported this model at individual IPA sites involving different cleavage and polyadenylation factors. A recent paper from Li, et al. probed this model further by profiling the effect of RNAi knockdown of each of the core cleavage and polyadenylation subunits on IPA-APA sites genome-wide (Li et al., 2015). Unidirectional shifts towards distal polyadenylation sites over IPA sites were observed genome-wide upon knockdown of several of these factors, supporting the model proposed by Takagaki and Manley. Cleavage and polyadenylation machinery expression levels range significantly across different tissue types (Gruber et al., 2014), so it is possible that expression of the cleavage and polyadenylation machinery helps to establish the baseline level of IPA site usage in a given tissue type.

An additional point of regulation for IPA sites involves the kinetic competition between cleavage at the IPA site and splicing of its encompassing intron. A majority of pre-mRNA splicing occurs co-transcriptionally. Once the encompassing intron is removed via splicing, the IPA site can no longer be used to define the 3' end of a transcript. Splicing cannot occur until the downstream 3' splice site is defined, which requires transcription of the full intron and likely the downstream exon due to exon definition. Therefore, the cleavage and polyadenylation machinery temporally has at least until RNAPII finishes transcribing the intron to recognize and cleave the mRNA at the IPA site. Mechanisms that increase the length of time required for RNAPII to transcribe the full intron or that slow splicing of the intron containing the IPA can thus favor usage of the IPA site (Bentley, 2014). Several experiments have implied such a mechanism. Mutations in TFIIS, Spt5, and Rpb2 that slow RNAPII elongation increase IPA site usage in yeast (Cui and Denis, 2003). A recent study in *Drosophila* showed that the expression of a slow-elongation mutant of RNAPII increases IPA site usage across the genome (Liu et al., 2017). Interestingly, in this study, the effect of slow transcription elongation on IPA sites was tissue-

specific, indicating that different mechanisms of IPA regulation may be dominant in different tissues. Finally, a recent study that knocked down the elongation factor PAF in mouse C2C12 cells decreased RNAPII elongation rate (as measured by RNAPII buildup across the gene body) and increased IPA site usage genome-wide (Yang et al., 2016). Also in agreement with this model, IPA sites that are used with higher frequency tend to reside in longer introns, which is consistent with a model in which increasing the time it takes to transcribe the IPA-containing intron results in increasing IPA site usage (Tian et al., 2007).

In addition to modulating the kinetics of transcription to favor IPA site usage, slower or impaired co-transcriptional splicing of the IPA-containing intron also increases IPA site usage. Indeed, knockdown of several key splicing factors, which would cause slower or impaired splicing, resulted in strong global upregulation of IPA sites (Li et al., 2015). In addition to this observation, weak 5' splice sites are strongly correlated with more frequently used IPA sites (Tian et al., 2007), suggesting that highly used IPA sites may select for weaker or slower-splicing introns, consistent with this model. In addition to a direct role in splicing regulation, the U1 snRNP plays a critical role in the suppression of cleavage and polyadenylation sites genome-wide. Several studies have shown that the U1 snRNP binding to RNA protects nascent RNA from downstream premature cleavage and polyadenylation (Almada et al., 2013; Berg et al., 2012; Kaida et al., 2010). Almada *et al.* extended these observations and demonstrated that pervasive transcription in the upstream antisense direction is curtailed by an enrichment of polyadenylation signals and a depletion of U1 snRNP motifs. Conversely, in the sense direction, intronic cleavage and polyadenylation was inhibited by depletion in polyadenylation motifs and an enrichment in U1 snRNP binding sites. Inhibition of U1 snRNPs using an antisense morpholino oligonucleotide dramatically increased cleavage and polyadenylation of transcripts

in the sense direction but had little effect on the cleavage and polyadenylation of transcripts in the antisense direction further supporting this model (Almada et al., 2013). These results suggest that U1 binding upstream of an intronic polyadenylation site and that the strength associated with that binding site may contribute to inhibition of IPAs.

### *Biological Consequences of IPA Usage*

IPA usage results in transcripts with altered coding potential (C-terminal deletions) as well as altered 3' UTRs. The protein products translated from IPA isoforms typically differ from the full-length protein product in one of three ways: (1) the truncated isoform codes for a protein product with distinct biological functions to the full-length isoform, thereby contributing to protein diversification across the genome, (2) the truncated isoform loses protein function and phenocopies a loss-of-function allele, or (3) the truncated protein antagonizes full-length protein function and phenocopies a dominant negative allele (Tian and Manley, 2017).

Two of the earliest and most well characterized examples of regulated IPA site usage (the IgM heavy chain locus and the calcitonin-related polypeptide- $\alpha$  gene) exemplify how two functionally distinct protein isoforms can be produced through regulated IPA-APA. The IgM locus was one of the first described examples of regulated alternative polyadenylation (Alt et al., 1980; Early et al., 1980; Rogers et al., 1980). During B-cell activation, IgM switches from expression of a membrane-bound form to expression of a secreted (soluble) form. Both proteins are expressed from the same genomic locus, however membrane-bound IgM is expressed from the full-length (distal) isoform of the gene, whereas soluble IgM is expressed from a truncated IPA isoform that removes IgM's transmembrane domain through C-terminal deletion. IgM is a key example of signaling-regulated IPA usage. Likewise, calcitonin-related polypeptide- $\alpha$  gene



(CALCA) produces two distinct peptides with independent hormone signaling activities through regulated IPA-APA in a tissue specific manner. Usage of the IPA site predominates in the thyroid producing the protein calcitonin, whereas usage of the distal poly(A) site predominates in the hypothalamus producing calcitonin gene-related peptide 1 (Amara et al., 1982; Rosenfeld et al., 1983).

IPA site usage can also result in the abrogation of protein function, mimicking a loss-of-function allele. One particularly interesting example of this involves the CstF77 gene (Luo et al., 2013). CstF77 forms part of the CstF complex and serves as a scaffolding protein between the RNA binding subunit of the CstF complex (CstF-64 or CstF-64 $\tau$ ) and the rest of the cleavage and polyadenylation machinery (Colgan and Manley, 1997). The CstF77 locus consists of 21 exons with an IPA site downstream of exon 3 that is highly conserved across metazoa (humans through zebrafish). Usage of the IPA site results in a shortened transcript, which produces a severely truncated protein isoform (103 amino acids or 44 amino acids depending upon splicing), in contrast to usage of the distal polyadenylation site, which produces a full-length protein isoform (717 amino acid). Such a large C-terminal truncation from usage of the IPA site can in itself result in protein loss-of-function. Interestingly in this case, however, studies using a dual fluorescent reporter gene recapitulating CstF-77 IPA site usage or western blots with multiple antibodies raised against the N-terminus of CstF-77 suggest that even though both isoforms of the protein are expressed at the mRNA level, only the long version of the protein is expressed at the protein level (Luo et al., 2013). Recent ongoing work has suggested that translation into non-coding regions of the genome (as would occur with the usage of an IPA site) results in the degradation of the protein products by an unknown surveillance mechanism (Arribere et al., 2016; Di Santo et al., 2016). How ubiquitous of a surveillance mechanism this is remains

unclear. However, IPA site usage can abrogate protein function by removing enough functional domains to render a protein product inactive, and potentially also by activating a novel protein surveillance mechanism.

The CstF-77 locus is particularly interesting, because usage of its IPA site is regulated by a negative feedback loop with itself (Luo et al., 2013). As noted above, expression levels of cleavage and polyadenylation factors are one way of regulating alternative cleavage and polyadenylation site choice with high levels of expression correlating with proximal site usage and low levels of expression correlating with distal site usage. High levels of CstF-77 lead to an increased usage of the CstF-77 IPA site, thereby producing the short isoform, which decreases functional CstF-77 levels of protein. As CstF-77 levels decrease, the distal polyadenylation site is used more often and functional CstF-77 protein levels begin to rise. In this way, CstF-77 expression is homeostatically maintained in a tightly-regulated negative feedback loop with itself. Functionally, levels of CstF-77 expression correlate with tandem 3'UTR-APA events genome-wide. Several studies have observed that rapidly proliferating cells use proximal 3'UTR-APA sites (Ji et al., 2009; Mayr and Bartel, 2009; Sandberg et al., 2008). Disrupting this negative feedback loop by lowering full-length CstF-77 levels with RNAi causes loss of cell cycle gene expression and signs of early differentiation through global alterations in 3'UTR-APA site usage. Interestingly, similar observations have been made for RNA binding proteins involved in alternative splicing. These splicing factors, akin to CstF-77's role in cleavage and polyadenylation, establish self-regulated, splicing-based negative feedback loops to maintain their own expression levels and therefore tightly maintain homeostasis over the splicing of their target genes (Jangi and Sharp, 2014; Jangi et al., 2014). Based on the CstF-77 data from Luo et al., it seems likely that alternative cleavage and polyadenylation networks can be regulated

similarly. Regardless of this speculation, CstF-77 is a prime example of the usage of an IPA site abrogating protein function.

The generation of truncated proteins through IPA site usage can also generate isoforms that antagonize the function of their full-length counterparts, mimicking a dominant negative allele. A prime example of this is the family of receptor tyrosine kinases (Vorlová et al., 2011). Receptor tyrosine kinases are cell-surface receptors that transduce extracellular stimuli to regulate key cellular processes, such as proliferation, differentiation, survival, metabolism, migration, and the cell cycle. Fifty-eight receptor tyrosine kinases are expressed in humans, which all have similar protein architecture. The N-terminus of the protein encodes for an extracellular ligand-binding domain, followed by a single transmembrane domain, and finally an intracellular kinase domain. Ligand binding followed by dimerization (or oligomerization) of multiple receptor tyrosine kinases activates downstream signaling. Constitutive and dysregulated signaling by receptor tyrosine kinases is a common feature of disease, particularly cancer (Lemmon and Schlessinger, 2010). Examination of expressed sequence tag databases revealed that receptor tyrosine kinases frequently had IPA sites immediately upstream of the exons encoding either the transmembrane domain or the intracellular kinase domain (Vorlová et al., 2011). These truncated isoforms disrupt receptor tyrosine kinase signaling in several ways. First, usage of the IPA site results in decreased expression of the functional, full-length receptor, and therefore, this truncated receptor acts as a loss-of-function of the protein. This “loss-of-function” characteristic is likely common to all of the receptor tyrosine kinases regulated by these IPA sites. However, expression of truncated, IPA-dependent isoforms, can also generate the expression of dominant negative proteins. If the IPA site is located upstream of the transmembrane domain, a soluble receptor can be expressed and released from the cell. This

soluble receptor can act as a “decoy,” binding ligand and sequestering it away from full-length receptor and thus, preventing activation of signaling. If the IPA site is located upstream of the kinase domain but downstream of the transmembrane domain, truncated receptor can be expressed on the cell surface. This truncated receptor can oligomerize with the full-length receptor preventing proper oligomerization of functional receptors by ligand binding and therefore abrogated signaling. The usage of these sites can be used to decrease receptor tyrosine kinase signaling during normal biology, and likely evolved for this purpose. However, these IPA sites are additionally interesting from a therapeutic perspective. If alternative cleavage and polyadenylation can be controlled to force increased expression of dominant negative (IPA-encoded) isoforms over full-length receptors, this approach could act as a novel therapeutic for receptor tyrosine kinase driven cancers (Vorlová et al., 2011).

It should be noted that in addition to differences in coding sequence, IPA-APA generates isoforms with different 3' UTRs. Alternative 3' UTRs can result in isoforms that differ by stability, translational efficiency, and mRNA localization. Recent work from Taliaferro *et al.* discovered an intriguing spatial relationship between the distribution of alternative isoforms generated by IPA-APA and mRNA localization in the brain (Taliaferro et al., 2016). By isolating mRNA localized to neural projections (neurites) versus those localized to the soma, Taliaferro *et al.* showed that IPA-APA isoforms localized preferentially to the soma, whereas distal polyadenylation isoforms localized preferentially to neurites. This was attributed to the 3' UTRs of the distal-most isoforms, which were enriched in the sequence motifs for muscleblind (Mbnl) RNA binding proteins as well as Mbnl binding itself. Interestingly, these neurite-specific isoforms were induced during neuronal differentiation. This work exemplifies an incredibly important theme in regulated IPA-APA usage. Global networks of regulated IPA-APA events

can be shifted simultaneously in a unidirectional and concerted manner to simultaneously regulate a particular biological process, in this case, neurological development.

### **Biological Consequences of Cdk12 Loss**

Until recently, Cdk9 was the only known RNAPII CTD Ser2 kinase in mammals. Since Cdk12 and Cdk13 were identified as novel RNAPII CTD Ser2 kinases (Bartkowiak et al., 2010; Blazek et al., 2011), their functional roles during mRNA biogenesis have remained unclear. Unlike Cdk9, whose inhibition abrogates transcription at most genes (Henriques et al., 2013; Jonkers et al., 2014; Rahl et al., 2010), several studies observed a relatively specific gene expression signature upon Cdk12 loss (Blazek et al., 2011; Johnson et al., 2016; Zhang et al., 2016). Notably, among the genes whose expression consistently decreased upon Cdk12 depletion were several DNA damage repair genes, particularly those involved in homologous recombination (HR) repair, including: BRCA1, BRCA2, FANCD2, ATM, and ATR (Blazek et al., 2011; Johnson et al., 2016; Zhang et al., 2016). Independently of these studies, a genome-wide shRNA screen provided additional genetic evidence for Cdk12's role in positively regulating the expression of HR repair genes (Bajrami et al., 2014). Of note, Cdk13 knockdown does not show similar effects on gene expression, suggesting that this role may be specific to Cdk12 activity (Liang et al., 2015).

HR repair plays a critical role in maintaining genomic integrity by repairing DNA double strand breaks (DSBs) and stalled replication forks during S-phase, lesions which pose serious threats to genomic integrity and cell viability (Gaillard et al., 2015; Heyer et al., 2010; Prakash et al., 2015). Multiple mechanisms exist in cells to repair DSBs in eukaryotes (Prakash et al., 2015; Symington and Gautier, 2011). The HR repair pathway is considered the most “conservative”

repair mechanism since it repairs the DNA lesion without the introduction of mutations. Loss of HR repair causes cells to depend on alternative repair pathways, most notably non-homologous end joining (NHEJ), which frequently introduces *de novo* mutations (Lieber, 2010). Cells with non-functional HR repair accumulate a distinct DNA mutational profile (Alexandrov et al., 2013) and are at increased risk of accumulating tumorigenic mutations (Lord and Ashworth, 2016). Indeed, patients with germline and somatic loss-of-function mutations in HR repair genes frequently develop some of the most intractable tumor types, including ovarian carcinoma, triple-negative breast cancer, pancreatic cancer, and metastatic castrate-resistant prostate cancer (Lord and Ashworth, 2016).

HR repair-deficient tumors have a unique “Achilles heel” that can be leveraged as a therapeutic target. Compared to HR repair-competent cells, HR repair-deficient cells demonstrate increased sensitivity (synthetic lethality) to small molecule drugs that cause replication fork arrest and lead to an increased number of DSBs (Lord and Ashworth, 2013; 2016). This increased DSB load overstresses the remaining functional DNA damage repair pathways, ultimately causing apoptosis or mitotic catastrophe. Small molecules that induce replication fork arrest ultimately fall into two groups: (1) genotoxic agents that cause DNA intrastrand crosslinks, such as platinum salts and mitomycin C, and (2) small molecule inhibitors of Parp1, which prevent the normal release of Parp1 from DNA lesions after the activation of alternative DNA repair mechanisms (Lord and Ashworth, 2013; 2016). Tumors that display defective HR repair and increased sensitivity to genotoxic agents and Parp1 inhibitors are described as having a “BRCAness” phenotype. Mounting evidence from studies of individual genes, patient sequencing data, and genome-wide screens for Parp1 inhibitor sensitivity have identified 22 genes, whose germline or somatic loss-of-function mutation correlates strongly with the rise of

“BRCAness” tumors (Lord and Ashworth, 2016). Cdk12 was classified as a novel “BRCAness” gene after accumulating evidence suggested that it positively regulates the gene expression of HR repair machinery (Bajrami et al., 2014; Blazek et al., 2011; Ekumi et al., 2015; Johnson et al., 2016; Joshi et al., 2014; Liang et al., 2015; Zhang et al., 2016) and after several studies showed that Cdk12 loss resulted in sensitivity to Parp1 inhibitors *in vitro* and *in vivo* (Bajrami et al., 2014; Johnson et al., 2016; Joshi et al., 2014). Notably, Cdk12 is the only gene among the “BRCAness” set whose function is not directly involved in HR repair (Lord and Ashworth, 2016).

These observations have tremendous clinical implications for tumor treatment. First, known (Ekumi et al., 2015) and predicted Cdk12 loss-of-function mutations occur recurrently in high-grade serous ovarian carcinoma (Carter et al., 2012; The Cancer Genome Atlas Research Network, 2011) and castration-resistant prostate adenocarcinoma (Grasso et al., 2012; Robinson et al., 2015). Cdk12 mutations in these tumors may therefore predict sensitivity to treatment modalities specific to “BRCAness” tumors, such as Parp1 inhibitors and other genotoxic agents. Second, as a broad, positive regulator of HR repair gene expression, Cdk12 inhibition may induce sensitivity of HR repair-competent cells to treatments otherwise reserved for “BRCAness” tumors, such as Parp1 inhibitors. Indeed pharmacological inhibition of Cdk12 has been shown to induce Parp1 inhibitor sensitivity in HR repair-competent tumors (Johnson et al., 2016). Furthermore, Johnson *et al.* showed that Cdk12 inhibition could reverse acquired Parp1 inhibitor resistance (Johnson et al., 2016). Parp1 inhibitor resistance can develop many ways; however, a common mechanism includes the development of secondary mutations in HR repair machinery that re-establish functional HR repair (Lord and Ashworth, 2013). As a broad transcriptional regulator of HR repair, Cdk12 has the potential to simultaneously

transcriptionally abrogate the expression of several members of the HR repair machinery, effectively bypassing the ability of cancer cells to develop Parp1 resistance through such a mechanism.

The mechanisms that underlie Cdk12's apparent specificity for promoting HR repair gene expression and the functional roles of Cdk12 in transcription regulation and mRNA biogenesis remain unclear. This thesis focuses on addressing these outstanding questions.



## References

- Adelman, K., Kennedy, M.A., Nechaev, S., Gilchrist, D.A., Muse, G.W., Chinenov, Y., and Rogatsky, I. (2009). Immediate mediators of the inflammatory response are poised for gene activation through RNA polymerase II stalling. *Proc. Natl. Acad. Sci. U.S.A.* *106*, 18207–18212.
- Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* *13*, 720–731.
- Adesnik, M., Salditt, M., Thomas, W., and Darnell, J.E. (1972). Evidence that all Messenger RNA Molecules (except Histone Messenger RNA) contain Poly(A) Sequences and that the Poly(A) has a Nuclear Function. *J. Mol. Biol.* *71*, 21–30.
- Adesnik, M., and Darnell, J.E. (1972). Biogenesis and Characterization of Histone Messenger RNA in HeLa Cells. *J. Mol. Biol.* *67*, 397–406.
- Ahn, S.H., Kim, M., and Buratowski, S. (2004). Phosphorylation of Serine 2 within the RNA Polymerase II C-Terminal Domain Couples Transcription and 3' End Processing. *Mol. Cell* *13*, 1–10.
- Aida, M., Chen, Y., Nakajima, K., Yamaguchi, Y., Wada, T., and Handa, H. (2006). Transcriptional pausing caused by NELF plays a dual role in regulating immediate-early expression of the junB gene. *Mol. Cell. Biol.* *26*, 6094–6104.
- Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational processes in human cancer. *Nature* *500*, 415–421.
- Allison, L.A., Wong, J.K., Fitzpatrick, V.D., Moyle, M., and Ingles, C.J. (1988). The C-terminal domain of the largest subunit of RNA polymerase II of *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and mammals: a conserved structure with an essential function. *Mol. Cell. Biol.* *8*, 321–329.
- Allison, L.A., Moyle, M., Shales, M., and Ingles, C.J. (1985). Extensive Homology among the Largest Subunits of Eukaryotic and Prokaryotic RNA Polymerases. *Cell* *42*, 599–610.
- 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* *499*, 360–363.
- Alt, F.W., Bothwell, A.L.M., Knapp, M., Siden, E., Mather, E., Koshland, M., and Baltimore, D. (1980). Synthesis of Secreted and Membrane-Bound Immunoglobulin Mu Heavy Chains Is Directed by mRNAs That Differ at Their 3' Ends. *Cell* *20*, 293–301.
- Amara, S.G., Jonas, V., Rosenfeld, M.G., Ong, E.S., and Evans, R.M. (1982). Alternative RNA processing in calcitonin gene expression generates mRNAs encoding different polypeptide products. *Nature* *298*, 240–244.

- Arribere, J.A., Cenik, E.S., Jain, N., Hess, G.T., Lee, C.H., Bassik, M.C., and Fire, A.Z. (2016). Translation readthrough mitigation. *Nature* 534, 719–723.
- Bajrami, I., Frankum, J.R., Konde, A., Miller, R.E., Rehman, F.L., Brough, R., Campbell, J., Sims, D., Rafiq, R., Hooper, S., et al. (2014). Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. *Cancer Res.* 74, 287–297.
- Barrilà, D., Lee, B.A., and Proudfoot, N.J. (2001). Cleavage/polyadenylation factor IA associates with the carboxyl-terminal domain of RNA polymerase II in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.* 98, 445–450.
- Bartkowiak, B., Liu, P., Phatnani, H.P., Fuda, N.J., Cooper, J.J., Price, D.H., Adelman, K., Lis, J.T., and Greenleaf, A.L. (2010). CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes Dev.* 24, 2303–2316.
- Bartolomei, M.S., Halden, N.F., Cullen, C.R., and Corden, J.L. (1988). Genetic Analysis of the Repetitive Carboxyl-Terminal Domain of the Largest Subunit of Mouse RNA Polymerase II. *Mol. Cell. Biol.* 8, 330–339.
- Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.-M., and Gautheret, D. (2000). Patterns of Variant Polyadenylation Signal Usage in Human Genes. *Genome Res.* 10, 1001–1010.
- Bentley, D.L. (2014). Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* 15, 163–175.
- Berg, M.G., Singh, L.N., Younis, I., Liu, Q., Pinto, A.M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L., et al. (2012). U1 snRNP determines mRNA length and regulates isoform expression. *Cell* 150, 53–64.
- Berkovits, B.D., and Mayr, C. (2015). Alternative 3' UTRs act as scaffolds to regulate membrane protein localization. *Nature* 522, 363–367.
- Bienroth, S., Wahle, E., Suter-Crazzolara, C., and Keller, W. (1991). Purification of the cleavage and polyadenylation factor involved in the 3'-processing of messenger RNA precursors. *J. Biol. Chem.* 266, 19768–19776.
- Birnboim, H.C., Mitchel, R., and Straus, N.A. (1973). Analysis of Long Pyrimidine Polynucleotides in HeLa Cell Nuclear DNA: Absence of Polydeoxythymidylate. *Proc. Natl. Acad. Sci. U.S.A.* 70, 2189–2192.
- Blazek, D., Kohoutek, J., Bartholomeeusen, K., Johansen, E., Hulinkova, P., Luo, Z., Cimermanic, P., Ule, J., and Peterlin, B.M. (2011). The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev.* 25, 2158–2172.

- Bösken, C.A., Farnung, L., Hintermair, C., Merzel Schachter, M., Vogel-Bachmayr, K., Blazek, D., Anand, K., Fisher, R.P., Eick, D., and Geyer, M. (2014). The structure and substrate specificity of human Cdk12/Cyclin K. *Nat. Commun.* *5*, 1-14.
- Brackenridge, S., and Proudfoot, N.J. (2000). Recruitment of a basal polyadenylation factor by the upstream sequence element of the human lamin B2 polyadenylation signal. *Mol. Cell. Biol.* *20*, 2660–2669.
- Brown, K.M., and Gilmartin, G.M. (2003). A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor Im. *Mol. Cell* *12*, 1467–1476.
- Buratowski, S., and Sharp, P.A. (1990). Transcription initiation complexes and upstream activation with RNA polymerase II lacking the C-terminal domain of the largest subunit. *Mol. Cell. Biol.* *10*, 5562–5564.
- Buratowski, S. (2009). Progression through the RNA Polymerase II CTD Cycle. *Mol. Cell* *36*, 541–546.
- Cadena, D.L., and Dahmus, M.E. (1987). Messenger RNA synthesis in mammalian cells is catalyzed by the phosphorylated form of RNA polymerase II. *J. Biol. Chem.* *262*, 12468–12474.
- Carswell, S., and Alwine, J.C. (1989). Efficiency of utilization of the simian virus 40 late polyadenylation site: effects of upstream sequences. *Mol. Cell. Biol.* *9*, 4248–4258.
- Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* *30*, 413–421.
- Chan, S.L., Huppertz, I., Yao, C., Weng, L., Moresco, J.J., Yates, J.R., Ule, J., Manley, J.L., and Shi, Y. (2014). CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing. *Genes Dev.* *28*, 2370–2380.
- Chonghui, C., and Sharp, P.A. (2003). RNA polymerase II accumulation in the promoter-proximal region of the dihydrofolate reductase and gamma-actin genes. *Mol. Cell. Biol.* *23*, 1961–1967.
- Colgan, D.F., and Manley, J.L. (1997). Mechanism and regulation of mRNA polyadenylation. *Genes Dev.* *11*, 2755–2766.
- Corden, J.L., Cadena, D.L., Joseph M Ahearn, J., and Dahmus, M.E. (1985). A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. *Proc. Natl. Acad. Sci. U.S.A.* *82*, 7934–7938.
- Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* *322*, 1845–1848.
- 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 Rep.* *2*, 1025–1035.

- Cramer, P., Armache, K.J., Baumli, S., Benkert, S., Brueckner, F., Buchen, C., Damsma, G.E., Dengl, S., Geiger, S.R., Jasiak, A.J., et al. (2008). Structure of Eukaryotic RNA Polymerases. *Annu. Rev. Biophys.* 37, 337–352.
- Cui, Y., and Denis, C.L. (2003). In Vivo Evidence that Defects in the Transcriptional Elongation Factors RPB2, TFIIIS, and SPT5 Enhance Upstream Poly(A) Site Utilization. *Mol. Cell. Biol.* 23, 7887–7901.
- Danko, C.G., Hah, N., Luo, X., Martins, A.L., Core, L., Lis, J.T., Siepel, A., and Kraus, W.L. (2013). Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol. Cell* 50, 212–222.
- Darnell, J.E., Philipson, L., Wall, R., and Adesnik, M. (1971a). Polyadenylic Acid Sequences: Role in Conversion of Nuclear RNA into Messenger RNA. *Science* 174, 507–510.
- Darnell, J.E., Wall, R., and Tushinski, R.J. (1971b). An Adenylic Acid-Rich Sequence in Messenger RNA of HeLa Cells and Its Possible Relationship to Reiterated Sites in DNA. *Proc. Natl. Acad. Sci. U.S.A.* 68, 1321–1325.
- Davidson, L., Muniz, L., and West, S. (2014). 3' end formation of pre-mRNA and phosphorylation of Ser2 on the RNA polymerase II CTD are reciprocally coupled in human cells. *Genes Dev.* 28, 342–356.
- de Vries, H., Rügsegger, U., Hübner, W., Friedlein, A., Langen, H., and Keller, W. (2000). Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. *EMBO J.* 19, 5895–5904.
- Derti, A., Garrett-Engle, P., MacIsaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M., and Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. *Genome Res.* 22, 1173–1183.
- DeZazzo, J.D., Kilpatrick, J.E., and Imperiale, M.J. (1991). Involvement of long terminal repeat U3 sequences overlapping the transcription control region in human immunodeficiency virus type 1 mRNA 3' end formation. *Mol. Cell. Biol.* 11, 1624–1630.
- Di Santo, R., Aboulhoda, S., and Weinberg, D.E. (2016). The fail-safe mechanism of post-transcriptional silencing of unspliced HAC1 mRNA. *eLife* 5, 1-23.
- Dichtl, B., Blank, D., Sadowski, M., Hübner, W., Weiser, S., and Keller, W. (2002). Yhh1p/Cft1p directly links poly(A) site recognition and RNA polymerase II transcription termination. *EMBO J.* 21, 4125–4135.
- Dixon-Clarke, S.E., Elkins, J.M., Cheng, S.-W.G., Morin, G.B., and Bullock, A.N. (2015). Structures of the CDK12/CycK complex with AMP-PNP reveal a flexible C-terminal kinase extension important for ATP binding. *Sci. Rep.* 5, 1-13.

- Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R., and Hood, L. (1980). Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell* 20, 313–319.
- Eckmann, C.R., Rammelt, C., and Wahle, E. (2011). Control of poly(A) tail length. *Wiley Interdiscip. Rev. RNA* 2, 348–361.
- Edmonds, M., Vaughan JR, M.H., and Nakazato, H. (1971). Polyadenylic Acid Sequences in the Heterogeneous Nuclear RNA and Rapidly-Labeled Polyribosomal RNA of HeLa Cells: Possible Evidence for a Precursor Relationship. *Proc. Natl. Acad. Sci. U.S.A.* 68, 1336–1340.
- Ekumi, K.M., Paculova, H., Lenasi, T., Pospichalova, V., Böskén, C.A., Rybarikova, J., Bryja, V., Geyer, M., Blazek, D., and Barboric, M. (2015). Ovarian carcinoma CDK12 mutations misregulate expression of DNA repair genes via deficient formation and function of the Cdk12/CycK complex. *Nucleic Acids Res.* 43, 2575–2589.
- Elkon, R., Ugalde, A.P., and Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.* 14, 496–506.
- Fitzgerald, M., and Shenk, T. (1981). The sequence 5′-AAUAAA-3′ Forms Parts of the Recognition Site for Polyadenylation of Late SV40 mRNAs. *Cell* 24, 251–260.
- Fong, N., and Bentley, D.L. (2001). Capping, splicing, and 3′ processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. *Genes Dev.* 15, 1783–1795.
- Ford, J.P., and Hsu, M.T. (1978). Transcription pattern of In Vivo-Labeled Late Simian Virus 40 RNA: Equimolar Transcription Beyond the mRNA 3′ Terminus. *J. Virol.* 28, 795–801.
- Fraser, N.W., Nevins, J.R., Ziff, E., and Darnell, J.E. (1979). The Major Late Adenovirus type-2 Transcription Unit: Termination is Downstream From the Last Poly(A) Site. *J. Mol. Biol.* 129, 643–656.
- Frischmeyer, P.A., van Hoof, A., ODonnell, K., Guerrierio, A.L., Parker, R., and Dietz, H.C. (2002). An mRNA Surveillance Mechanism That Eliminates Transcripts Lacking Termination Codons. *Science* 295, 2258–2261.
- Fuchs, G., Voichek, Y., Benjamin, S., Gilad, S., Amit, I., and Oren, M. (2014). 4sUDRB-seq: measuring genome-wide transcriptional elongation rates and initiation frequencies within cells. *Genome Biol.* 15, R69.
- Fuda, N.J., Ardehali, M.B., and Lis, J.T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* 461, 186–192.
- 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. *Mol. Cell. Biol.* 24, 787–795.

- Gaillard, H., García-Muse, T., and Aguilera, A. (2015). Replication stress and cancer. *Nat. Rev. Cancer* *15*, 276–289.
- Gil, A., and Proudfoot, N.J. (1984). A sequence downstream of AAUAAA is required for rabbit Beta-globin mRNA 3'-end formation. *Nature* *312*, 473–474.
- Gil, A., and Proudfoot, N.J. (1987). Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation. *Cell* *49*, 399–406.
- Gilmartin, G.M., and Nevins, J.R. (1989). An ordered pathway of assembly of components required for polyadenylation site recognition and processing. *Genes Dev.* *3*, 2180–2190.
- Gilmartin, G.M., and Nevins, J.R. (1991). Molecular analyses of two poly(A) site-processing factors that determine the recognition and efficiency of cleavage of the pre-mRNA. *Mol. Cell. Biol.* *11*, 2432–2438.
- Grasso, C.S., Wu, Y., Robinson, D.R., Cao, X., Dhanasekaran, S.M., Khan, A.P., Quist, M.J., Jing, X., Lonigro, R.J., Brenner, J.C., et al. (2012). The mutational landscape of lethal castration-resistant prostate cancer. *Nature* *487*, 239–243.
- Greifenberg, A.K., Hönig, D., Pilarova, K., Düster, R., Bartholomeeusen, K., Böskén, C.A., Anand, K., Blazek, D., and Geyer, M. (2016). Structural and Functional Analysis of the Cdk13/Cyclin K Complex. *Cell Rep.* *14*, 320–331.
- Gruber, A.R., Martin, G., Keller, W., and Zavolan, M. (2014). Means to an end: mechanisms of alternative polyadenylation of messenger RNA precursors. *Wiley Interdiscip. Rev. RNA* *5*, 183–196.
- Grünberg, S., and Hahn, S. (2013). Structural insights into transcription initiation by RNA polymerase II. *Trends Biochem. Sci.* *38*, 603–611.
- Gu, B., Eick, D., and Bensaude, O. (2012). CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo. *Nucleic Acids Res.* *41*, 1591–1603.
- Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* *130*, 77–88.
- Guo, Z., and Stiller, J.W. (2004). Comparative genomics of cyclin-dependent kinases suggest co-evolution of the RNAPII C-terminal domain and CTD-directed CDKs. *BMC Genomics* *5*, 69–81.
- Harlen, K.M., and Churchman, L.S. (2017). The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nat. Rev. Mol. Cell Biol.* *18*, 263–273.
- 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. *Mol. Cell* *52*, 517–528.

- Heyer, W.-D., Ehmsen, K.T., and Liu, J. (2010). Regulation of homologous recombination in eukaryotes. *Annu. Rev. Genet.* *44*, 113–139.
- Higgs, D.R., Goodbourn, S.E.Y., Lamb, J., Clegg, J.B., and Weatherall, D.J. (1983). Alpha-Thalassaemia caused by a polyadenylation signal mutation. *Nature* *306*, 398–400.
- Hirose, Y., and Manley, J.L. (1998). RNA polymerase II is an essential mRNA polyadenylation factor. *Nature* *395*, 93–96.
- Hofer, E., and Darnell, J.E. (1981). The primary transcription unit of the mouse beta-major globin gene. *Cell* *23*, 585–593.
- Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J.Y., Yehia, G., and Tian, B. (2013). Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat. Meth.* *10*, 133–139.
- Hsin, J.P., and Manley, J.L. (2012). The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev.* *26*, 2119–2137.
- Hu, J., Lutz, C.S., Wilusz, J., and Tian, B. (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA* *11*, 1485–1493.
- Hutchins, L.N., Murphy, S.M., Singh, P., and Graber, J.H. (2008). Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics* *24*, 2684–2690.
- Jang, M.K., Mochizuki, K., Zhou, M., Jeong, H.-S., Brady, J.N., and Ozato, K. (2005). The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Mol. Cell* *19*, 523–534.
- Jangi, M., and Sharp, P.A. (2014). Building robust transcriptomes with master splicing factors. *Cell* *159*, 487–498.
- Jangi, M., Boutz, P.L., Paul, P., and Sharp, P.A. (2014). Rbfox2 controls autoregulation in RNA-binding protein networks. *Genes Dev.* *28*, 637–651.
- Jelinek, W., Adesnik, M., Salditt, M., Sheiness, D., Wall, R., Molloy, G., Philipson, L., and Darnell, J.E. (1973). Further Evidence on the Nuclear Origin and Transfer to the Cytoplasm of Polyadenylic Acid Sequences in Mammalian Cell RNA. *J. Mol. Biol.* *75*, 515–532.
- Ji, Z., Lee, J.Y., Pan, Z., Jiang, B., and Tian, B. (2009). Progressive lengthening of 3' untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development. *Proc. Natl. Acad. Sci. U.S.A.* *106*, 7028–7033.
- Johnson, S.F., Cruz, C., Greifenberg, A.K., Dust, S., Stover, D.G., Chi, D., Primack, B., Cao, S., Bernhardt, A.J., Coulson, R., et al. (2016). CDK12 Inhibition Reverses De Novo and Acquired PARP Inhibitor Resistance in BRCA Wild-Type and Mutated Models of Triple-Negative Breast Cancer. *Cell Rep.* *17*, 2367–2381.

- Jonkers, I., and Lis, J.T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* *16*, 167–177.
- 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* *3*, 1-25.
- Joshi, P.M., Sutor, S.L., Huntoon, C.J., and Karnitz, L.M. (2014). Ovarian cancer-associated mutations disable catalytic activity of CDK12, a kinase that promotes homologous recombination repair and resistance to cisplatin and poly(ADP-ribose) polymerase inhibitors. *J. Biol. Chem.* *289*, 9247–9253.
- Kaida, D., Berg, M.G., Younis, I., Kasim, M., Singh, L.N., Wan, L., and Dreyfuss, G. (2010). U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* *468*, 664–668.
- Kaufmann, I., Martin, G., Friedlein, A., Langen, H., and Keller, W. (2004). Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J.* *23*, 616–626.
- Keller, W., Bienroth, S., Lang, K.M., and Christofori, G. (1991). Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA. *EMBO J.* *10*, 4241–4249.
- Kim, H., Erickson, B., Luo, W., Seward, D., Graber, J.H., Pollock, D.D., Megee, P.C., and Bentley, D.L. (2010). Gene-specific RNA polymerase II phosphorylation and the CTD code. *Nat. Struct. Mol. Biol.* *17*, 1279–1286.
- 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. *J. Biol. Chem.* *276*, 12317–12323.
- Kim, M., Ahn, S.H., Krogan, N.J., Greenblatt, J.F., and Buratowski, S. (2004). Transitions in RNA polymerase II elongation complexes at the 3' ends of genes. *EMBO J.* *23*, 354–364.
- Kim, T.K., Ebright, R.H., and Reinberg, D. (2000). Mechanism of ATP-dependent promoter melting by transcription factor IIIH. *Science* *288*, 1418–1422.
- 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*, 876–880.
- Kim, W.Y., and Dahmus, M.E. (1989). The major late promoter of adenovirus-2 is accurately transcribed by RNA polymerases IIO, IIA, and IIB. *J. Biol. Chem.* *264*, 3169–3176.
- Kohoutek, J., and Blazek, D. (2012). Cyclin K goes with Cdk12 and Cdk13. *Cell Div.* *7*, 12.



- Komarnitsky, P., Cho, E.-J., and Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev.* *14*, 2452–2460.
- Krumm, A., Meulia, T., Brunvand, M., and Groudine, M. (1992). The block to transcriptional elongation within the human c-myc gene is determined in the promoter-proximal region. *Genes Dev.* *6*, 2201–2213.
- Kwak, H., and Lis, J.T. (2013). Control of transcriptional elongation. *Annu. Rev. Genet.* *47*, 483–508.
- 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* *339*, 950–953.
- Lai, C.J., Dhar, R., and Khoury, G. (1978). Mapping the spliced and unspliced late lytic SV40 RNAs. *Cell* *14*, 971–982.
- Lee, S.Y., Mendecki, J., and Brawerman, G. (1971). A Polynucleotide Segment Rich in Adenylic Acid in Rapidly-Labeled Polyribosomal RNA Component of Mouse Sarcoma 180 Ascites Cells. *Proc. Natl. Acad. Sci. U.S.A.* *68*, 1331–1335.
- Lemmon, M.A., and Schlessinger, J. (2010). Cell signaling by receptor tyrosine kinases. *Cell* *141*, 1117–1134.
- Levitt, N., Briggs, D., Gil, A., and Proudfoot, N.J. (1989). Definition of an efficient synthetic poly(A) site. *Genes Dev.* *3*, 1019–1025.
- Li, B., Carey, M., and Workman, J.L. (2007). The Role of Chromatin during Transcription. *Cell* *128*, 707–719.
- Li, J., and Gilmour, D.S. (2013). Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor. *EMBO J.* *32*, 1829–1841.
- Li, W., You, B., Hoque, M., Zheng, D., Luo, W., Ji, Z., Park, J.Y., Gunderson, S.I., Kalsotra, A., Manley, J.L., et al. (2015). Systematic Profiling of Poly(A)<sup>+</sup> Transcripts Modulated by Core 3' End Processing and Splicing Factors Reveals Regulatory Rules of Alternative Cleavage and Polyadenylation. *PLoS Genet.* *11*, 1–28.
- Liang, K., Gao, X., Gilmore, J.M., Florens, L., Washburn, M.P., Smith, E., and Shilatifard, A. (2015). Characterization of Human Cyclin-Dependent Kinase 12 (CDK12) and CDK13 Complexes in C-Terminal Domain Phosphorylation, Gene Transcription, and RNA Processing. *Mol. Cell. Biol.* *35*, 928–938.
- Licatalosi, D.D., Geiger, G., Minet, M., Schroeder, S., Cilli, K., McNeil, J.B., and Bentley, D.L. (2002). Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II. *Mol. Cell* *9*, 1101–1111.

- Lieber, M.R. (2010). The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem.* *79*, 181–211.
- Lim, L., and Canellakis, E.S. (1970). Adenine-rich Polymer associated with Rabbit Reticulocyte Messenger RNA. *Nature* *227*, 710–712.
- Liu, J., and Kipreos, E.T. (2000). Evolution of Cyclin-Dependent Kinases (CDKs) and CDK-Activating Kinases (CAKs): Differential Conservation of CAKs in Yeast and Metazoa. *Mol. Biol. Evol.* *17*, 1–14.
- Liu, X., Freitas, J., Zheng, D., Hoque, M., Oliveira, M.S., Martins, T., Henriques, T., Bin Tian, and Moreira, A. (2017). Transcription elongation rate has a tissue-specific impact on alternative cleavage and polyadenylation in *Drosophila melanogaster*. *RNA* *23*, 1807–1816.
- Lord, C.J., and Ashworth, A. (2013). Mechanisms of resistance to therapies targeting BRCA-mutant cancers. *Nat. Med.* *19*, 1381–1388.
- Lord, C.J., and Ashworth, A. (2016). BRCAness revisited. *Nat. Rev. Cancer* *16*, 110–120.
- Lu, H., Zawel, L., Fisher, L., Egly, J.M., and Reinberg, D. (1992). Human general transcription factor IIIH phosphorylates the C-terminal domain of RNA polymerase II. *Nature* *358*, 641–645.
- Luo, W., Ji, Z., Pan, Z., You, B., Hoque, M., Li, W., Gunderson, S.I., and Tian, B. (2013). The conserved intronic cleavage and polyadenylation site of *CstF-77* gene imparts control of 3' end processing activity through feedback autoregulation and by U1 snRNP. *PLoS Genet.* *9*, 1-14.
- MacDonald, C.C., Wilusz, J., and Shenk, T. (1994). The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. *Mol. Cell. Biol.* *14*, 6647–6654.
- Malumbres, M. (2014). Cyclin-dependent kinases. *Genome Biol.* *15*, 1-10.
- Mandel, C.R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J.L., and Tong, L. (2006). Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature* *444*, 953–956.
- Manley, J.L. (1983). Accurate and specific polyadenylation of mRNA precursors in a soluble whole-cell lysate. *Cell* *33*, 595–605.
- Marshall, N.F., and Price, D.H. (1992). Control and formation of two distinct classes of RNA Polymerase II Elongation Complexes. *Mol. Cell. Biol.* *12*, 2078–2090.
- Marshall, N.F., and Price, D.H. (1995). Purification of P-TEFb, a Transcription Factor Required for the Transition into Productive Elongation. *J. Biol. Chem.* *270*, 12335–12338.
- Marshall, N.F., Peng, J., Xie, Z., and Price, D.H. (1996). Control of RNA Polymerase II Elongation Potential by a Novel Carboxyl-terminal Domain Kinase. *J. Biol. Chem.* *271*, 27176–27183.

- Marzluff, W.F., Wagner, E.J., and Duronio, R.J. (2008). Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nat. Rev. Genet.* *9*, 843–854.
- Masamha, C.P., Xia, Z., Yang, J., Albrecht, T.R., Li, M., Shyu, A.-B., Li, W., and Wagner, E.J. (2015). CFIm25 links alternative polyadenylation to glioblastoma tumour suppression. *Nature* *510*, 412–416.
- Mayer, A., Lidschreiber, M., Siebert, M., Leike, K., Söding, J., and Cramer, P. (2010). Uniform transitions of the general RNA polymerase II transcription complex. *Nat. Struct. Mol. Biol.* *17*, 1272–1278.
- Mayr, C. (2016). Evolution and Biological Roles of Alternative 3'UTRs. *Trends Cell Biol.* *26*, 227–237.
- Mayr, C., and Bartel, D.P. (2009). Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* *138*, 673–684.
- McCracken, S., Fong, N., Yankulov, K., Ballantyne, S., Pan, G., Greenblat, J., Patterson, S.D., Wickens, M., and Bentley, D. (1997). The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* *385*, 357–361.
- McDevitt, M.A., Hart, R.P., Wong, W.W., and Nevins, J.R. (1986). Sequences capable of restoring poly(A) site function define two distinct downstream elements. *EMBO J.* *5*, 2907–2913.
- McLauchlan, J., Gaffney, D., Whitton, J.L., and Clements, J.B. (1985). The consensus sequence YGTGTTY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini. *Nucleic Acids Res.* *13*, 1347–1368.
- Meinhart, A., and Cramer, P. (2004). Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. *Nature* *430*, 223–226.
- Min, I.M., Waterfall, J.J., Core, L.J., Munroe, R.J., Schimenti, J., and Lis, J.T. (2011). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev.* *25*, 742–754.
- Montell, C., Fisher, E.F., Caruthers, M.H., and Berk, A.J. (1983). Inhibition of RNA cleavage but not polyadenylation by a point mutation in mRNA 3' consensus sequence AAUAAA. *Nature* *305*, 600–605.
- Moore, C.L., and Sharp, P.A. (1984). Site-Specific Polyadenylation in a Cell-Free Reaction. *Cell* *36*, 581–591.
- Moore, C.L., and Sharp, P.A. (1985). Accurate Cleavage and Polyadenylation of Exogenous RNA Substrate. *Cell* *41*, 845–855.
- Moore, C.L., Skolnik-David, H., and Sharp, P.A. (1986). Analysis of RNA cleavage at the adenovirus-2 L3 polyadenylation site. *EMBO J.* *5*, 1929–1938.

- Moore, M.J., and Proudfoot, N.J. (2009). Pre-mRNA Processing Reaches Back to Transcription and Ahead to Translation. *Cell* *136*, 688–700.
- Moreira, A., Wollerton, M., Monks, J., and Proudfoot, N.J. (1995). Upstream sequence elements enhance poly(A) site efficiency of the C2 complement gene and are phylogenetically conserved. *EMBO J.* *14*, 3809–3819.
- Morris, D.P., Michelotti, G.A., and Schwinn, D.A. (2005). Evidence That Phosphorylation of the RNA Polymerase II Carboxyl-terminal Repeats Is Similar in Yeast and Humans. *J. Biol. Chem.* *280*, 31368–31377.
- Murthy, K.G., and Manley, J.L. (1992). Characterization of the multisubunit cleavage-polyadenylation specificity factor from calf thymus. *J. Biol. Chem.* *267*, 14804–14811.
- Murthy, K.G., and Manley, J.L. (1995). The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. *Genes Dev.* *9*, 2672–2683.
- 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. *Nat. Genet.* *39*, 1507–1511.
- Nechaev, S., Fargo, D.C., Santos, dos, 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* *327*, 335–338.
- Nevins, J.R., and Darnell, J.E. (1978). Steps in the Processing of Ad2 mRNA: Poly(A)<sup>+</sup> Nuclear Sequences Are Conserved and Poly(A) Addition Precedes Splicing. *Cell* *15*, 1477–1493.
- Nevins, J.R., Blanchard, J.M., and Darnell, J.E. (1980). Transcription units of adenovirus type 2. Termination of transcription beyond the poly(A) addition site in early regions 2 and 4. *J. Mol. Biol.* *144*, 377–386.
- Ni, Z., Schwartz, B.E., Werner, J., Suarez, J.-R., and Lis, J.T. (2004). Coordination of transcription, RNA processing, and surveillance by P-TEFb kinase on heat shock genes. *Mol. Cell* *13*, 55–65.
- 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*, 526–540.
- Nonet, M., Sweester, D., and Young, R.A. (1987). Functional Redundancy and Structural Polymorphism in the Large Subunit of RNA Polymerase II. *Cell* *50*, 909–915.
- Ozsolak, F., Kapranov, P., Foissac, S., Kim, S.W., Fishilevich, E., Monaghan, A.P., John, B., and Milos, P.M. (2010). Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation. *Cell* *143*, 1018–1029.

- Prakash, R., Zhang, Y., Feng, W., and Jasin, M. (2015). Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins. *Cold Spring Harb. Perspect. Biol.* *7*, 1–27.
- Proudfoot, N.J., and Brownlee, G.G. (1976). 3' Non-coding region sequences in eukaryotic messenger RNA. *Nature* *263*, 211–214.
- 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* *141*, 432–445.
- Rasmussen, E.B., and Lis, J.T. (1993). In vivo transcriptional pausing and cap formation on three *Drosophila* heat shock genes. *Proc. Natl. Acad. Sci. U.S.A.* *90*, 7923–7927.
- Robinson, D., Van Allen, E.M., Wu, Y., Schultz, N., Lonigro, R.J., Mosquera, J.-M., Montgomery, B., Taplin, M.-E., Pritchard, C.C., Attard, G., et al. (2015). Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* *162*, 1215–1228.
- Rodriguez, C.R., Cho, E.-J., Keogh, M.C., Moore, C.L., Greenleaf, A.L., and Buratowski, S. (2000). Kin28, the TFIIF-Associated Carboxy-Terminal Domain Kinase, Facilitates the Recruitment of mRNA Processing Machinery to RNA Polymerase II. *Mol. Cell. Biol.* *20*, 104–112.
- Roeder, R., and Rutter, W. (1969). Multiple Forms of DNA-dependent RNA Polymerase in Eukaryotic Organisms. *Nature* *224*, 234–237.
- Rogers, J., Early, P., Carter, C., Calame, K., Bond, M., Hood, L., and Wall, R. (1980). Two mRNAs with different 3' ends encode membrane-bound and secreted forms of immunoglobulin mu chain. *Cell* *20*, 303–312.
- Rosenfeld, M.G., Mermod, J., Amara, S.G., Swanson, L.W., Sawchenko, P.E., River, J., Vale, W.W., and Evans, R.M. (1983). Production of a novel neuropeptide encoded by the calcitonin gene via tissue-specific RNA processing. *Nature* *304*, 129–135.
- 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*, 795–804.
- Rüegsegger, U., Beyer, K., and Keller, W. (1996). Purification and characterization of human cleavage factor Im involved in the 3' end processing of messenger RNA precursors. *J. Biol. Chem.* *271*, 6107–6113.
- Ryan, K., Calvo, O., and Manley, J.L. (2004). Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. *RNA* *10*, 565–573.
- Ryan, K., Murthy, K.G.K., Kaneko, S., and Manley, J.L. (2002). Requirements of the RNA polymerase II C-terminal domain for reconstituting pre-mRNA 3' cleavage. *Mol. Cell. Biol.* *22*, 1684–1692.

- Salisbury, J., Hutchison, K.W., and Graber, J.H. (2006). A multispecies comparison of the metazoan 3'-processing downstream elements and the CstF-43 RNA recognition motif. *BMC Genomics* 7, 55–13.
- Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A., and Burge, C.B. (2008). Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science* 320, 1643–1647.
- Saunders, A., Core, L.J., and Lis, J.T. (2006). Breaking barriers to transcription elongation. *Nat. Rev. Mol. Cell Biol.* 7, 557–567.
- Schönemann, L., Kühn, U., Martin, G., Schäfer, P., Gruber, A.R., Keller, W., Zavolan, M., and Wahle, E. (2014). Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Genes Dev.* 28, 2381–2393.
- Schroeder, S.C., Schwer, B., Shuman, S., and Bentley, D. (2000). Dynamic association of capping enzymes with transcribing RNA polymerase II. *Genes Dev.* 14, 2435–2440.
- Schwartz, L.B., and Roeder, R.G. (1975). Purification and subunit structure of deoxyribonucleic acid-dependent ribonucleic acid polymerase II from the mouse plasmacytoma, MOPC 315. *J. Biol. Chem.* 250, 3221–3228.
- Sheets, M.D., Ogg, S.C., and Wickens, M.P. (1990). Point mutations in AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res.* 18, 5799–5805.
- Shi, Y., and Manley, J.L. (2015). The end of the message: multiple protein–RNA interactions define the mRNA polyadenylation site. *Genes Dev.* 29, 889–897.
- Shi, Y., Di Giammartino, D.C., Taylor, D., Sarkeshik, A., Rice, W.J., Yates, J.R., III, Frank, J., and Manley, J.L. (2009). Molecular Architecture of the Human Pre-mRNA 3' Processing Complex. *Mol. Cell* 33, 365–376.
- Skaar, D.A., and Greenleaf, A.L. (2002). The RNA Polymerase II CTD Kinase CTDK-1 Affects Pre-mRNA 3' Cleavage/Polyadenylation through the Processing Component Pti1p. *Mol. Cell* 10, 1429–1439.
- Skolnik-David, H., Moore, C.L., and Sharp, P.A. (1987). Electrophoretic separation of polyadenylation-specific complexes. *Genes Dev.* 1, 672–682.
- Strobl, L.J., and Eick, D. (1992). Hold back of RNA polymerase II at the transcription start site mediates down-regulation of c-myc in vivo. *EMBO J.* 11, 3307–3314.
- Suh, H., Ficarro, S.B., Kang, U.-B., Chun, Y., Marto, J.A., and Buratowski, S. (2016). Direct Analysis of Phosphorylation Sites on the Rpb1 C-Terminal Domain of RNA Polymerase II. *Mol. Cell* 61, 297–304.

- Symington, L.S., and Gautier, J. (2011). Double-Strand Break End Resection and Repair Pathway Choice. *Annu. Rev. Genet.* 45, 247–271.
- Søgaard, T.M.M., and Svejstrup, J.Q. (2007). Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator. *J. Biol. Chem.* 282, 14113–14120.
- Takagaki, Y., and Manley, J.L. (1997). RNA recognition by the human polyadenylation factor CstF. *Mol. Cell. Biol.* 17, 3907–3914.
- Takagaki, Y., and Manley, J.L. (1998). Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. *Mol. Cell* 2, 761–771.
- Takagaki, Y., Manley, J.L., MacDonald, C.C., Wilusz, J., and Shenk, T. (1990). A multisubunit factor, CstF, is required for polyadenylation of mammalian pre-mRNAs. *Genes Dev.* 4, 2112–2120.
- Takagaki, Y., Ryner, L.C., and Manley, J.L. (1989). Four factors are required for 3'-end cleavage of pre-mRNAs. *Genes Dev.* 3, 1711–1724.
- Takagaki, Y., Seipelt, R.L., Peterson, M.L., and Manley, J.L. (1996). The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. *Cell* 87, 941–952.
- Taliaferro, J.M., Vidaki, M., Oliveira, R., Olson, S., Zhan, L., Saxena, T., Wang, E.T., Graveley, B.R., Gertler, F.B., Swanson, M.S., et al. (2016). Distal Alternative Last Exons Localize mRNAs to Neural Projections. *Mol. Cell* 61, 821–833.
- The Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615.
- Tian, B., Hu, J., Zhang, H., and Lutz, C.S. (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* 33, 201–212.
- Tian, B., and Graber, J.H. (2011). Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip. Rev. RNA* 3, 385–396.
- Tian, B., and Manley, J.L. (2017). Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* 18, 18–30.
- Tian, B., Pan, Z., and Lee, J.Y. (2007). Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Res.* 17, 156–165.
- Valsamakis, A., Zeichner, S., Carswell, S., and Alwine, J.C. (1991). The human immunodeficiency virus type 1 polyadenylation signal: a 3' long terminal repeat element upstream of the AAUAAA necessary for efficient polyadenylation. *Proc. Natl. Acad. Sci. U.S.A.* 88, 2108–2112.

- Veloso, A., Kirkconnell, K.S., Magnuson, B., Biewen, B., Paulsen, M.T., Wilson, T.E., and Ljungman, M. (2014). Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.* *24*, 896–905.
- Venkataraman, K., Brown, K.M., and Gilmartin, G.M. (2005). Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes Dev.* *19*, 1315–1327.
- Vorlová, S., Rocco, G., Lefave, C.V., Jodelka, F.M., Hess, K., Hastings, M.L., Henke, E., and Cartegni, L. (2011). Induction of antagonistic soluble decoy receptor tyrosine kinases by intronic polyA activation. *Mol. Cell* *43*, 927–939.
- Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G.A., Winston, F., et al. (1998a). DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev.* *12*, 343–356.
- Wada, T., Takagi, T., Yamaguchi, Y., Watanabe, D., and Handa, H. (1998b). Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro. *EMBO J.* *17*, 7395–7403.
- Wang, E.T., Sandberg, R., Luo, S., Khrebtkova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* *456*, 470–476.
- Weber, C.M., Ramachandran, S., and Henikoff, S. (2014). Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol. Cell* *53*, 819–830.
- Weill, L., Belloc, E., Bava, F.-A., and Méndez, R. (2012). Translational control by changes in poly(A) tail length: recycling mRNAs. *Nat. Struct. Mol. Biol.* *19*, 577–585.
- Wickens, M., and Stephenson, P. (1984). Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA 3' end formation. *Science* *226*, 1045–1051.
- Wilusz, J., Pettine, S.M., and Shenk, T. (1989). Functional analysis of point mutations in the AAUAAA motif of the SV40 late polyadenylation signal. *Nucleic Acids Res.* *17*, 3899–3908.
- 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. *Mol. Cell* *21*, 227–237.
- 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*, 41–51.
- Yang, Q., Gilmartin, G.M., and Doublé, S. (2010). Structural basis of UGUA recognition by the Nudix protein CFI(m)25 and implications for a regulatory role in mRNA 3' processing. *Proc. Natl. Acad. Sci. U.S.A.* *107*, 10062–10067.



- Yang, Y., Li, W., Hoque, M., Hou, L., Shen, S., Tian, B., and Dynlacht, B.D. (2016). PAF Complex Plays Novel Subunit-Specific Roles in Alternative Cleavage and Polyadenylation. *PLoS Genet.* *12*, 1–28.
- Yoh, S.M., Cho, H., Pickle, L., Evans, R.M., and Jones, K.A. (2007). The Spt6 SH2 domain binds Ser2-P RNAPII to direct Iws1-dependent mRNA splicing and export. *Genes Dev.* *21*, 160–174.
- Yoh, S.M., Lucas, J.S., and Jones, K.A. (2008). The Iws1:Spt6:CTD complex controls cotranscriptional mRNA biosynthesis and HYPB/Setd2-mediated histone H3K36 methylation. *Genes Dev.* *22*, 3422–3434.
- Zaborowska, J., Egloff, S., and Murphy, S. (2016). The pol II CTD: new twists in the tail. *Nat. Struct. Mol. Biol.* *23*, 771–777.
- Zehring, W.A., Lee, J.M., Weeks, J.R., Jokerst, R.S., and Greenleaf, A.L. (1988). The C-terminal repeat domain of RNA polymerase II largest subunit is essential in vivo but is not required for accurate transcription initiation in vitro. *Proc. Natl. Acad. Sci. U.S.A.* *85*, 3698–3702.
- Zeitlinger, J., Stark, A., Kellis, M., Hong, J.-W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nat. Genet.* *39*, 1512–1516.
- Zhang, T., Kwiatkowski, N., Olson, C.M., Dixon-Clarke, S.E., Abraham, B.J., Greifenberg, A.K., Ficarro, S.B., Elkins, J.M., Liang, Y., Hannett, N.M., et al. (2016). Covalent targeting of remote cysteine residues to develop CDK12 and CDK13 inhibitors. *Nat. Chem. Biol.* *12*, 876–884.
- Zhu, Y., Pe'ery, T., Peng, J., Ramanathan, Y., Marshall, N., Marshall, T., Amendt, B., Mathews, M.B., and Price, D.H. (1997). Transcription elongation factor P-TEFb is required for HIV-1 tat transactivation in vitro. *Genes Dev.* *11*, 2622–2632.



## Chapter 2

# **Cdk12 regulates DNA repair genes by suppressing intronic polyadenylation**

The work in this chapter is adapted with permission from the following manuscript:

Sara J. Dubbury\*, Paul L. Boutz\*, Phillip A. Sharp. Cdk12 regulates DNA repair genes by suppressing intronic polyadenylation. (In submission)

\*Authors contributed equally

Sara Dubbury, Paul Boutz, and Phillip Sharp conceived and designed the research. Sara Dubbury performed all experiments. Paul Boutz performed all computational analysis. Sara Dubbury, Paul Boutz, and Phillip Sharp analyzed the data and wrote the manuscript.

## Abstract

Mutations that attenuate DNA repair by homologous recombination (HR) promote tumorigenesis and sensitize cells to chemotherapeutic agents that cause replication fork collapse, a phenotype known as “BRCAness.”<sup>1</sup> BRCAness tumors arise from loss-of-function mutations in 22 genes.<sup>1</sup> Of these genes, all but one (Cdk12) directly function in the HR repair pathway.<sup>1</sup> Cdk12 phosphorylates Serine 2 of the RNA Polymerase II (RNAPII) C-terminal domain (CTD) heptapeptide repeat,<sup>2-6</sup> a modification that regulates transcription elongation, splicing, and cleavage/polyadenylation.<sup>7-10</sup> Genome-wide expression studies suggest that Cdk12 depletion abrogates the expression of several HR genes relatively specifically, blunting HR repair.<sup>3-6,11-13</sup> This observation suggests that Cdk12 mutational status may predict tumor sensitivity to targeted treatments against BRCAness, such as Parp 1 inhibitors, and that small-molecule inhibitors of Cdk12 may induce sensitization of otherwise HR-competent tumors to these treatments.<sup>6,11,13</sup> Despite this growing clinical interest, the mechanism behind the apparent specificity of Cdk12 in regulating HR genes remains unknown. Here we find that Cdk12 globally suppresses intronic polyadenylation events, enabling the production of full-length gene products. Many HR genes harbor significantly more intronic polyadenylation sites compared to all expressed genes, and the cumulative effect of these sites accounts for the increased sensitivity of HR gene expression to Cdk12 loss. Finally, we find evidence that Cdk12 loss-of-function mutations cause increased intronic polyadenylation within HR genes in human tumors, suggesting that this mechanism is conserved. This work clarifies the biological function of CDK12 and underscores its potential both as a chemotherapeutic target and as a tumor biomarker.

## Results and Discussion

Cdk12 regulates the expression of HR genes by an unknown mechanism. In order to dissect Cdk12's molecular function, we sought to establish a genetic knockout of Cdk12 in wild type mouse embryonic stem cells (mESCs). mESCs spend most of their cell cycle in S-phase and fail to activate a G1/S checkpoint in response to DNA damage, making them heavily dependent upon replication-coupled HR repair and, therefore, highly sensitive to HR defects.<sup>14-16</sup> We generated Cdk12 knockout clones (Cdk12 $\Delta$ ) in mESCs that express a complementing, doxycycline (Dox)-inducible HA-epitope tagged Cdk12 transgene in the presence of continuous Dox treatment (Extended Data Fig. 1A,B). To acutely ablate Cdk12, we removed Dox from the growth media and observed significant loss of Cdk12 after 24 hours and undetectable levels after 48 hours (Fig. 1A, Extended Data Fig. 1C). Cdk12 loss yielded a substantial, progressive viability defect beginning at 72 hours of Dox depletion, which was grossly quantified by measuring cell doublings (Fig. 1B). Importantly, the first 48 hours of Cdk12 depletion had minimal consequences on viability, providing a 48-hour window of time in which to genetically probe the function of Cdk12. Notably, the gross cellular defects were largely reversible upon Cdk12 re-expression (Fig. 1B).

The viability defect observed upon Cdk12 loss could be caused by decreased proliferation and/or increased cell death. To characterize defects in proliferation, we profiled the cell cycle upon Cdk12 loss (Fig. 1C). We observed two major cell-cycle defects. First, we noted a striking decrease in nucleotide incorporation during S-phase (as measured by EdU intensity after a short EdU pulse). Second, we noted a decreased percentage of cells in active S-phase with an increasing proportion of cells in G1. The reversibility of the proliferation defect upon restoration of Cdk12 expression correlated with a return to normal levels of EdU incorporation (Fig. 1C). To

determine if increased cell death also contributed to the observed viability defects, we measured the percentage of cells undergoing active apoptosis (Fig. 1D). We noted increased apoptosis upon Cdk12 depletion.

Failure to repair DNA damage during S-phase results in replication fork stalling/collapse and inhibited DNA replication,<sup>17,18</sup> which is consistent with the decrease in EdU signal that we observed upon Cdk12 depletion. Persistent, unrepaired DNA damage and incomplete DNA replication drive mESCs to differentiate or initiate programmed cell death.<sup>19,20</sup> The accumulation of cells in G1 that we observe upon Cdk12 loss is consistent with differentiating cells that have longer G1-stages and competent DNA damage-activated checkpoints.<sup>21</sup> Furthermore, the increased percentage of cells undergoing apoptosis is consistent with programmed cell death in response to unrepaired DNA damage. Because DNA damage induces p53, we looked for evidence of a p53 response upon Cdk12 depletion. Indeed, total p53 and Ser15-phosphorylated (activated) p53<sup>22-24</sup> were both upregulated upon Cdk12 loss (Fig. 1E). Taken in sum, Cdk12 loss in mESCs results in phenotypes consistent with defective HR repair.

We next addressed the molecular consequences of Cdk12 loss on gene expression by sequencing polyA-selected RNA after 24 and 48 hours of Cdk12 depletion. We observed that 140 genes after 24 hours and 814 genes after 48 hours changed significantly in expression at the total mRNA level upon Cdk12 depletion (Extended Data Fig. 2A). In line with the observed p53 activation upon Cdk12 loss, we detected a robust p53-dependent gene expression signature among the set of affected genes. Approximately 33% of the affected genes are known p53 target genes in mESCs<sup>25</sup> and changed in the direction consistent with DNA-damage induced p53-activation. (Extended Data Fig. 2B). Since p53 activation induces differentiation in mESCs, we also looked for enrichment of genes that have bivalent chromatin modifications (H3K4me3 and

H3K27me3) at their promoters,<sup>26</sup> a characteristic of lineage-specific early differentiation genes in mESCs. Indeed, an additional 12% of genes that changed in expression were such lineage-specific early differentiation genes. These two gene signatures accounted for approximately 70% of the genes whose expression increased and ~20% of the genes whose expression decreased upon Cdk12 depletion. Excluding these genes as likely secondary effects of p53 activation and early differentiation, Cdk12 depletion after 48 hours modestly affected the total expression level of 428 genes, or 3% of the 13,594 expressed genes, and a majority of those decreased in expression upon Cdk12 loss (Extended Data Fig. 2B).

Surprisingly, we did not observe by analysis of total RNA-seq a statistically significant decrease in gene expression of the HR machinery. We therefore asked whether Cdk12 might influence isoform composition. Although Cdk12 has been implicated in alternative splicing regulation,<sup>27</sup> we observed few changes in alternative splicing genome-wide (Extended Data Fig. 2C), suggesting little direct role for Cdk12 in splicing regulation. We then examined isoform differences caused by alternative cleavage and polyadenylation. In addition to the polyadenylation site located after the 3' most exon of a gene (the distal polyadenylation site), intronic polyadenylation sites (IPAs) frequently occur within the internal introns of a gene. Usage of these IPAs results in truncated mRNA isoforms that can vary in coding potential, stability, translational efficiency, and localization.<sup>28-30</sup> We used previously-published 3' end sequencing data from mESCs<sup>31</sup> to identify IPA and distal isoforms genome-wide; when we quantified the use of these sites in Cdk12-depleted cells we observed a striking increase in IPA termination events at the expense of distal sites (Fig. 2A,B). Among 33,115 IPA sites we identified in 13,594 expressed genes, 2009 individual IPA isoforms (~6.4% of all identified IPA isoforms) were differentially expressed upon Cdk12 loss at the level of statistical significance

(Extended Data Fig. 2D). Notably, a vast majority of these IPA isoforms (1824, 91%) increased in expression upon Cdk12 depletion.

We also quantified differential expression of isoforms resulting from usage of the distal polyadenylation site by measuring the fold change in the expression of the distal-most exon as normalized to the rest of the transcript (Fig. 2A). Opposite to the trend in IPA isoforms, the majority (1848, 75%) of all significantly changing distal isoforms decreased in expression upon Cdk12 loss (Fig. 2A, Extended Data Fig. 2D). In a subset of these genes (571), we could also detect a corresponding, statistically significant increase in at least one IPA isoform (Extended Data Fig. 2E). A majority (56%) of the remaining genes with a statistically significant decrease in distal isoform expression contained at least one IPA site that increased in usage even though they did not reach the threshold for statistical significance. Thus, we consider these 1,848 genes with significantly decreasing distal exons to also be altered by Cdk12-dependent IPA usage.

A single intron may contain multiple IPA sites (i.e. tandem polyadenylation sites), and a gene may contain multiple IPA isoforms (i.e. IPAs in multiple introns). In order to determine the extent of this regulatory mechanism on expressed genes, we collapsed the data to single genes. Approximately 22% of all expressed genes had at least one significantly increasing IPA isoform, a significantly decreasing distal isoform, or both (Fig. 2C). These 2,948 genes are the genes most strongly affected by a shift in isoform usage from the distal (full-length) isoform to the proximal IPA isoform(s) after 48 hours of Cdk12 loss. We chose to focus on this set of genes for the later mechanistic studies presented here; however, it is notable that IPA isoform usage increased and distal isoform usage decreased globally whether or not they reached statistical significance (Fig. 2D). Therefore, we conclude that the primary effect of Cdk12 on gene expression is to suppress IPA sites genome-wide and thus to promote distal (full-length) isoform expression.



Since cleavage and polyadenylation occur co-transcriptionally and Cdk12 is a RNAPII CTD Ser2 kinase, we sought to determine if Cdk12 depletion resulted in changes to RNAPII or its Ser2 phosphorylation (Ser2p) status that might explain the increased IPA site usage. We used a high-resolution chromatin immunoprecipitation sequencing method<sup>32,33</sup> to map RNAPII and Ser2p RNAPII density genome-wide in the presence (+Dox) or absence (-Dox 48 hours) of Cdk12. To increase robustness in our ChIP sequencing data, we isolated ChIP samples for both RNAPII and Ser2p RNAPII using two independent antibodies for each target and repeated the ChIP in biological duplicate for each one of these antibodies and conditions (16 libraries in total). Comparing metagene profiles of the average read densities of all ChIPs across biological replicates indicates that ChIP profiles from the two independent antibodies are highly reproducible and similar (Extended Data Fig. 3). We therefore aggregated data from both antibodies and both biological replicates throughout our metagene analyses in order to effectively incorporate four independent biological replicates per condition (Extended Data Fig. 4A). Furthermore, we developed a statistical framework (Extended Data Fig. 4B-D) to measure the significance of the differences observed in ChIP signal in the presence and absence of Cdk12 in order to achieve a high level of confidence that we can reliably and reproducibly quantify even subtle differences in ChIP read density.

To analyze the effects of Cdk12 loss on RNAPII elongation, we plotted metagene profiles of RNAPII density across genes with statistically significant Cdk12-sensitive IPA or distal isoform changes from the transcriptional start site (TSS) to the distal polyadenylation site (Distal PAS) (Fig. 3A). Because RNAPII density correlates strongly with gene expression levels and inversely with gene length, we conservatively removed the shortest and longest length quartiles and focused on the middle two quartiles of affected genes (Extended Data Fig. 5A,B). Cdk12

loss resulted in decreased RNAPII density at the 5' end of genes, transitioning to increased RNAPII density toward the 3' end (Fig. 3B, Extended Data Fig. 5B,C). Notably, since increased IPA usage upon Cdk12 depletion would result in the dissociation of RNAPII prior to the 3' end of the metagene profiles, the magnitude of the increase in RNAPII signal the 3' end of the metagene profile is probably underestimated. This pattern was not specific to genes with statistically-significant Cdk12-sensitive IPA or distal polyA isoforms and was observed in two different length- and expression-matched control gene sets (Extended Data Fig. 6).

In order to determine whether the decreased RNAPII density at the 5' end of genes could be caused by less RNAPII entering productive elongation upon Cdk12 loss, we aligned metagene profiles on the first stable nucleosome downstream of the promoter, a spatial barrier associated with promoter-proximal pausing prior to the polymerase entering productive elongation (Fig. 3A).<sup>34</sup> Indeed, upon Cdk12 loss, RNAPII density increased just upstream of, and decreased immediately downstream of the first stable nucleosome, indicating that less RNAPII entered productive elongation upon Cdk12 depletion and accounting for the decrease in RNAPII density at the 5' end of genes (Fig. 3C, Extended Data Fig. 7). Based on this observation, the increased RNAPII density over the 3' ends of genes observed upon Cdk12 loss is most parsimoniously explained by a decrease in the transcription elongation rate of the RNAPII that results in a progressive accumulation of RNAPII density across the gene body in the Cdk12 depleted condition. Consistent with its putative role as a RNAPII Ser2 kinase, Cdk12 loss resulted in a substantial decrease in Ser2p RNAPII across the entire gene body (Fig. 3D, Extended Data Fig. 8). Ser2p has been shown to recruit a variety of transcription elongation factors to RNAPII.<sup>35-37</sup> Therefore, our data suggest that Cdk12-mediated phosphorylation of the RNAPII CTD is required to maintain efficient transcription elongation across all or most expressed genes.

The observed effects on RNAPII elongation provide a likely mechanistic explanation for the global increase in IPA site usage observed upon Cdk12 loss. IPA usage is in kinetic competition with the splicing of its encompassing intron, which must be transcribed completely in order to be excised. Reducing the transcription elongation rate would therefore alter the kinetic balance to favor IPA usage by allowing more time between transcription of the IPA site and transcription of the downstream exon (Extended Data Fig. 9). This model is consistent with previous studies that demonstrated that slower RNAPII elongation rates, due to mutant polymerases or alterations in transcription elongation factors, increased IPA usage over that of distal sites.<sup>8,38-40</sup> Our data suggest that Cdk12 activity globally increases elongation rates across gene bodies to suppress IPA usage. Although the reduction in elongation efficiency when Cdk12 is depleted is probably small enough in magnitude so as not to affect the expression of the majority of genes at the total transcript level, those with active IPA sites undergo enhanced IPA site usage and decreased production of full-length isoforms.

Given that we did not observe expression differences in the HR genes at the total gene level, we asked whether changes in IPA usage could explain the sensitivity of HR genes to Cdk12 depletion. Gene ontology analysis revealed that homologous recombination (as well as other DNA damage repair) genes are enriched in statistically significant Cdk12-sensitive IPA sites, including over half of the genes that contribute to the “BRCAness” phenotype.<sup>1</sup> However, this observation did not explain the apparent specificity of the HR machinery to Cdk12 loss since many other gene ontology categories were also enriched. This observation raised the possibility that HR genes were more profoundly affected by IPA usage as a group than genes in other functional categories. To test this hypothesis, we compared the total number of IPA sites per

gene and found that the Cdk12-sensitive HR genes were significantly enriched in the frequency of IPA sites per gene compared to all expressed genes (Fig. 4A).

In genes where multiple IPA sites are used, the negative effect of terminating at each individual IPA site on the amount of full-length isoform is cumulative. Indeed, we observed a strong correlation between increasing numbers of IPA sites per gene and the effect of Cdk12 depletion on the amount of full-length isoform produced for that gene (Fig. 4B). Consistent with their tendency toward higher numbers of IPA sites, HR genes were among those exhibiting the most profound decrease in full-length isoform expression. To confirm that the changes observed at the mRNA level impact protein expression, we performed immunoblotting on several HR family members over a time course of Cdk12 depletion, and observed significant decreases in the expression of full-length protein (Fig. 4C,D). Since Cdk12 activity maintains the full-length expression of over half of the identified BRCAness genes, we propose that the combined effect of strong downregulation of multiple gene products within the same functional pathway contributes to the HR-deficient phenotypes observed upon Cdk12 loss and is consistent with the cellular phenotypes that we observe in mESCs. Therefore, Cdk12 appears to be a “master regulator” of HR genes.

Predicted and validated<sup>5</sup> loss-of-function (LOF) point mutations and genomic deletions in Cdk12 have been identified recurrently in genomic sequencing of prostate<sup>41-43</sup> and ovarian<sup>44,45</sup> tumors. We analyzed RNA sequencing data from ovarian serous adenocarcinoma<sup>44</sup> and prostate adenocarcinoma<sup>42</sup> tumors to determine whether CDK12 LOF affects IPA usage in human tumor samples. We found that tumors harboring predicted loss-of-function point mutations in CDK12 and/or gene deletions across the CDK12 locus consistently showed evidence of upregulated IPA site usage within several key BRCAness genes, including ATM, ATR, WRN, FANCA and

FANCD2, compared to tumors that were diploid or amplified in copy number and wild type in CDK12 gene expression (Fig. 4E,F). Notably, the only CDK12-mutated tumor that had no apparent effect on an ATM IPA site harbored a missense mutation (K975E) that was previously shown to not cause loss-of-function of CDK12 activity.<sup>5</sup> These data suggest that the gene expression mechanism we describe here is conserved in humans and functions in human cancer to downregulate expression of HR genes. Differential IPA usage may therefore function as a biomarker for functional loss of CDK12 (and thus HR) in human tumors and could potentially be used to identify patients with tumors that would respond to targeted treatments against “BRCAness” phenotypes, e.g. PARP1 inhibitor treatment.

## Materials and Methods

To ensure reproducibility, all experiments were repeated at least three times.

### Cell Culture, Cell Line Generation, Drug Conditions

V6.5 (C57Bl/6-129) mESCs and derived cell lines were cultured on 0.2% gelatin-coated tissue culture plates in ES media (HEPES-buffered DMEM (Thermo Fisher) supplemented with 15% FBS (Hyclone), 1000U/mL LIF (Millipore), non-essential amino acids, L-glutamine, BME, penicillin, and streptomycin). Cdk12 $\Delta$  clones were maintained in 1 $\mu$ g/mL doxycycline (dox) (Sigma) in ES media (changed daily) to sustain complementing levels of Cdk12. To investigate Cdk12 loss, cells were washed at time zero with HBS and switched to ES media without dox.

Cdk12 $\Delta$  clones were generated as follows (see Extended Data Fig. 1A). A Cdk12-Flox clone was isolated using CRISPR/Cas9 genome editing technology.<sup>46</sup> Two sgRNA sequences targeting introns 3 and 4 of the endogenous Cdk12 locus were cloned into pX330 (a gift from Feng Zhang, Addgene# 42230).<sup>47</sup> Lipofectamine®2000 (Thermo Fisher) was used to co-transfect wild type V6.5 cells with the sgRNA plasmids, along with single-stranded oligodeoxynucleotides (ssODNs) from Integrated DNA Technologies (IDT) containing a LoxP sequence adjacent to an NcoI restriction site flanked on either side by 60-nucleotide homology arms complementary to intron 3 or intron 4 surrounding the sgRNA cut site, and the pLKO.1 plasmid harboring a puromycin-resistance gene. Cells were selected 24 hours after transfection with 1 $\mu$ g/mL puromycin (Sigma) for 48 hours and single-cell cloned. Clones were screened for homozygous LoxP site insertion into both introns by PCR followed by NcoI digest. Positive clones were confirmed by Sanger sequencing. A dox-inducible Cdk12 transgene was stably introduced into the Cdk12-Flox cell line using a piggybac retrotransposon system. N-terminal

Flag- HA- tandem epitope tagged Cdk12 (NM\_001109626.1) was cloned from polyA-selected mouse cDNA into pCR8/GW/TOPO (Thermo Fisher) followed by transfer into the doxycycline-inducible piggybac expression vector, PBNeoTetO-Dest (a gift from A.W. Cheng), using standard TOPO and Gateway cloning kits (Thermo Fisher). This expression vector was cotransfected with pAC4 (constitutively expressing M2rtTA, the dox-inducible transactivator, flanked by piggybac recombination sites, A.W. Cheng) and mPBase (piggybac transposase expression plasmid, A.W. Cheng) using Lipofectamine®2000. 24 hours after transfection, cells were selected with 150 ug/mL Hygromycin (Thermo Fisher) and 200 ug/mL G418 (Sigma) to select stable transformants. Subsequently, a constitutive Cre expression plasmid, pPGK-Cre-bpA (a gift from Klaus Rajewsky, Addgene plasmid # 11543), and a constitutive mCherry expression vector, pCAGGS-mCherry<sup>48</sup> were co-transfected using Lipofectamine®2000 into the Cdk12 Flox cells with stably integrated doxycycline-inducible Cdk12. 48 hours after transfection, the cell population was single-cell FACS sorted for mCherry positive cells. Beginning 4-6 hours after Cre transfection, the cells were treated daily with 1ug/mL dox (Sigma) supplemented ES media to express rescuing levels of Cdk12 protein. PCR was used to select clones harboring homozygous deletions of exon 4 and exon 4 deletion was confirmed by Sanger sequencing across the locus. Several homozygous knockout clones were isolated and two clones were picked for subsequent analysis.

Endogenous N-terminal, V5-epitope tagged *ATM*, *BRCA2*, and *FANCD2* cell lines were made as follows. sgRNAs targeting genomic loci near the start codon of *ATM*, *BRCA2*, and *FANCD2* were cloned into pX330. gBlocks (IDT) containing a V5 epitope tag positioned in-frame, immediately adjacent to the start codon flanked by 64 to 354 nucleotides of homology were TOPO cloned and sequenced verified. For *ATM* where homologous insertion of the V5 tag

disrupted sgRNA/Cas9 cutting, we added a restriction enzyme site in frame to the end of the V5 tag to facilitate screening. For *FANCD2* and *BRCA2* where homologous insertion of the V5 epitope tag did not inhibit sgRNA/Cas9 re-cleavage, we also engineered point mutations into the gBlock construct that would introduce a novel restriction enzyme site adjacent to the sgRNA PAM motif to disrupt sgRNA/Cas9 re-cutting and facilitate screening. Cdk12 $\Delta$  cells were co-transfected with the appropriate sgRNA, TOPO-cloned gBlock, and pLKO.1 with blasticidin resistance. Cells were selected with 2 $\mu$ g/mL Blasticidin and single cell cloned. Cells were screened by PCR followed by restriction enzyme digest. Heterozygous and homozygous insertions of the V5 tag were isolated and confirmed by Sanger sequencing across the locus. Two independent clones with homozygous or heterozygous insertions of the V5 tag were isolated and experiments were replicated in at least two independent clones.

### **FACS Assays**

FACS analyses were performed using BD FACS machines: FACS Celesta, LSRII, FACS Canto II, FACS LSR Fortessa, and FACS Aria IIIu. Data was collected using FACS Diva Version 8.0.1 and data was analyzed using FlowJo version 1.0.1.

### *Growth Curve Analysis*

24 hours prior to starting the time course, cells were plated at the same cell number in biological triplicate for each of the first 3 time points of the experiment (0, 24, and 48 hours) in +dox media such that they would reach no more than 90% confluency after 72 hours of growth in the +dox condition. Since mESCs do not tolerate growth on cell cultures dishes for longer than 72 hours, in order to observe cell growth out to 96 hours, additional cells were grown in parallel



cultures. After 24 hours, we split these parallel cultures into biological triplicates per condition for the final three time points of the experiment (48, 72, and 96 hours). Starting at time 0, cells received daily media changes with +dox or -dox media as appropriate. At each time point, triplicate cell cultures were washed with HBS and harvested by trypsinization. Each biological replicate was resuspended in 450uL of ES media followed by addition of 1uL of 50uM calcein-AM in DMSO and 2uL of 2mM ethidium homodimer-1 (Thermo Fisher). Samples were incubated for 15-20 minutes at room temperature protected from light. After staining, 50uL of CountBright Absolute Counting Beads (Thermo Fisher) was added to each sample. Samples were analyzed by flow cytometry such that at least ~5000 CountBright Absolute Counting Beads (~100uL) were recorded per sample, during which samples were vortexed every minute to prevent counting beads from settling out of solution. The number of live cells per replicate was quantified in each sample by counting the number of live (Calcein-AM positive and Ethidium Homodimer-1 negative) cells and comparing it to the number of CountBright Absolute Counting Beads (with known concentration). The number of live cells was averaged across biological replicates. To calculate the number of viable cell doublings over time, the number of live cells at each time point was compared to the previous time point to give the ratio of live cells after 24 hours. This number was converted into number of cell doublings by taking the log (ratio of viable cells)/log(2).

### *Cell Cycle Profiling*

Cells were plated at approximately equal cell density 24 hours prior to profiling, such that cells were 50-80% confluent at the time of harvest. Cells were pulsed with 10uM 5-ethynyl-2'-deoxyuridine (EdU) for 1 hour under standard growth conditions, then harvested by

trypsinization. Collected cell pellets were fixed, permeabilized, and stained for EdU incorporation with Alexa Fluor 647 using Click-iT EdU Flow Cytometry Assay Kit (Thermo Fisher) according to manufacturer's instructions. After EdU staining, cells were resuspended in 1x Click-iT saponin-based permeabilization and wash reagent (Thermo Fisher) with 50ug/mL propidium iodide to label total DNA content and 100ug/mL RNase A. Cells were incubated at room temperature in the dark for 30 minutes, and at least 50,000 cells were analyzed by flow cytometry for EdU content (AlexaFluor 647) and total DNA content (propidium iodide).

### *Apoptosis*

Cells were plated at approximately equal cell density at least 24 hours prior to harvest. Cells were harvested by trypsinization. The growth media and HBS wash prior to trypsinization were collected and centrifuged with the trypsinized cell population to collect any apoptosing cells with decreased adherence to the plate. Cell pellets were washed twice with cold PBS. Cells were fixed, permeabilized, and stained for cleaved Caspase-3 (apoptosis) using the FITC Active Caspase-3 Apoptosis Kit (BD Pharmingen) and the recommended protocol. At least 50,000 stained cells were analyzed by FACS.

### **Western Blotting**

Whole cell extract was harvested from cells by washing the cells in cold phosphate-buffered saline (PBS) and lysing in RIPA (10mM Tris pH7.4, 150mM NaCl, 1% TritonX-100, 0.1% SDS, 0.5% Sodium Deoxycholate, and 1mM EDTA) supplemented with 1x cOmplete, EDTA-free Protease Inhibitors (Roche), 2uL/mL Benzonase Nuclease (Sigma), and if needed, 1x Halt Phosphatase Inhibitor Cocktail (Thermo Fisher). Lysates were incubated on ice for at least

30min, centrifuged for 10 min at 4°C and max speed, and the cleared lysate was quantified using a standard BCA assay (Thermo Fisher). Lysates were normalized in 1x Loading Dye (62.5mM Tris pH6.8, 5% glycerol, 2% SDS, 16.67% BME, and 0.083% bromophenol blue) or 1x NuPAGE LDS Sample Buffer (Thermo Fisher) with 1x NuPAGE Reducing Agent (Thermo Fisher). Normalized lysates were boiled for 5 min or incubated at 70°C for 10 min and run on one of the following types of precast gels: NuPAGE 4-12% Bis-Tris Gels (Thermo Fisher), NuPAGE 3-8% Tris-Acetate Protein Gels (Thermo Fisher), Novex 4-20% Tris Glycine Gels (Thermo Fisher), Novex 10-20% Tris Glycine gels. Gels were transferred overnight (30 V) to PVDF in 10% methanol supplemented 1x NuPAGE Transfer Buffer (NuPAGE Bis-Tris Gels) for Bis-Tris and Tris-Acetate gels or 20% methanol supplemented 1x Novex Tris-Glycine Transfer Buffer (Thermo Fisher) for Tris-Glycine gels. Primary antibodies used for blotting: Anti-HA High Affinity Antibody (Roche 11867423001), Enolase I (CST 3810S), Vinculin (Sigma V9131), p53 (1C12) (CST2524S), P-p53 Serine15 (CST 9284S), ATR (CST 13934S), FANCD2 (Abcam ab108928), and V5 (Life Technologies R96025). Secondary antibodies used all blots except the V5 epitope tag: ECL Anti-Rat IgG (GE Healthcare NA935V), ECL Anti-Mouse IgG (GE Healthcare NA931V), and ECL Anti-Rabbit IgG (GE Healthcare NA934V). For blots with the V5 epitope tag, we used the Anti-Mouse IgG, HRP-linked antibody (CST 7076S). Blots were exposed with Western Lightning Plus-ECL (Perkin Elmer) or SuperSignal West Dura Extended Duration Substrate (Thermo Fisher).

## **RNA Sequencing**

Total RNA was harvested using Trizol Reagent (Thermo Fisher) from two independent Cdk12Δ clones each in biological duplicate from cells maintained in dox (+dox) or withdrawn

from dox for 24 hours (-dox 24 hours) or 48 hours (-dox 48 hours). In parallel, total RNA was harvested from Cdk12 floxed cells (without integrated dox-inducible transgene) that had been pre-treated with 1 $\mu$ g/mL dox daily for 17 days and subjected to the same dox conditions (+dox, -dox 24 hours, -dox 48 hours) in biological duplicate to serve as a control for gene expression effects of long-term dox treatment followed by short-term withdrawal. RNA was extracted following the standard Trizol protocol and subsequently DNase treated with Turbo DNase (Thermo Fisher) under standard reaction conditions. RNA quality was assessed by Agilent 2100 Bioanalyzer and only samples with a RIN value  $\geq 9$  were used for library preparation and sequencing. PolyA-selected libraries were made from 1 $\mu$ g of total RNA input using the TruSeq Stranded mRNA Library Prep Kit (Illumina RS-122-2102) with multiplexing barcodes, following the standard protocol with the following specifications: (1) 5 min RNA fragmentation time, (2) Superscript III (Thermo Fisher) was used for reverse transcription, (3) 15 cycles of PCR were used during the library amplification step, and (4) AMPure beads (Beckman Coulter) were used to size select/purify the library post PCR amplification instead of gel size selection. The 18 libraries were pooled and sequenced on two lanes of an Illumina NextSeq500.

### **Chromatin Immunoprecipitation Sequencing**

We modified a high resolution, micrococcal nuclease (MNase) digestion-based ChIP methodology.<sup>32,33</sup> We performed ChIP on Cdk12 $\Delta$  cells in +dox or -dox 48 hour conditions for total RNAPII density and Ser2p RNAPII density. Two independent antibodies were used for each antigen as follows: 8WG16 (Abcam ab817) and Rpb3 (Bethyl A303-771A) for total RNAPII density; and H5 Clone (Abcam ab24758) and 3E10 Clone (Millipore 04-1571) for Ser2p RNAPII density. Two biological replicates were processed for each antibody and dox

condition (for example of the experimental setup for one protein target, e.g. RNAPII, see Extended Data Fig. 4). For each dox condition, we also processed 4 negative control libraries: one whole-cell extract (WCE) sample and 3 mock IP samples with the following antibodies: goat anti-mouse IgM (Thermo Fisher 31172), goat anti-rat IgG (Thermo Fisher 31226), mouse IgG2a [MOPC-173] Isotype Control (Abcam ab18413).

Chromatin immunoprecipitation was performed as follows. Briefly, 48 hours before cells were harvested, 10e6 CDK12Δ cells were plated in 15 cm dishes either in +dox or -dox ES media. Cells were crosslinked directly in media on the plate in 1% methanol-free formaldehyde (Thermo Fisher) for 10 minutes at room temperature. Crosslinking was quenched with 250mM glycine. Cells were washed 3x in 10mL chilled PBS and harvested by scraping. Cells were pelleted, washed in 10mL chilled PBS, and pelleted again. PBS was aspirated off the cells and the pellets were flash frozen in liquid nitrogen and stored at -80°C. Pellets were thawed on ice and resuspended in 0.9mL of CHIP Lysis Buffer (1% SDS, 10mM EDTA, 50mM Tris-HCL pH7.5) supplemented with 1x cOmplete Protease Inhibitors (Roche) and 1x Halt Phosphatase Inhibitors (Thermo Fisher) and lysed for 10 minutes at room temperature. Pellets were diluted in 8.1mL of CHIP Dilution Buffer (1% TritonX-100, 2mM EDTA, 150mM NaCl, 20mM Tris-HCl pH7.5) supplemented with 1x cOmplete Protease Inhibitors (Roche) and 1x Halt Phosphatase Inhibitors (Thermo Fisher) and 3mM CaCl<sub>2</sub>. Each tube was pre-warmed at 37°C for 2 minutes. 12 Units of micrococcal nuclease (MNase, Sigma) were added per tube and samples were digested for 30 min at 37°C with rotation. The digestion reaction was quenched with the addition of 180uL 500mM EDTA and 360uL 500mM EGTA per tube. Samples were sonicated in 15mL polystyrene tubes in a BioRupter (Diagenode) for 20 cycles on high (1 cycle = 30 seconds ON/30 seconds OFF). Samples were cleared by centrifugation (max speed, 4°C, 10min).

Soluble material was transferred to a new tube and each sample (one pellet) was split into four 1.8mL parts. Lysate was incubated overnight at 4°C plus rotation with 10ug or 10uL of acites fluid of the following antibodies: Total RNAPII (8WG16 and Rpb3), and RNAPII CTD Ser2p (H5 and 3E10). After the overnight incubation, the IPs were incubated for 2 hours at 4°C as follows: The 8WG16 and Rpb3 IPs were incubated with 100uL of Protein G Dynabeads (Thermo Fisher). The H5 IPs were incubated with 100uL of Protein G Dynabeads pre-conjugated overnight with 20ug goat anti-mouse IgM (Thermo Fisher 31172) in Bead Preparation Solution (9 ChIP Dilution Buffer: 1 ChIP Lysis Buffer). The 3E10 IP was incubated with 100uL of Protein G Dynabeads pre-conjugated with 20ug goat anti-rat IgG (Thermo Fisher 31226). IPs were washed as follows 2 x 2mL LB3 (20mM Tris-HCl pH7.5, 150mM NaCl, 2mM EDTA, 0.1% SDS, 1% Triton X-100); 1x 2mL LB3+ (20mM Tris-HCl pH7.5, 500mM NaCl, 2mM EDTA, 0.1% SDS, 1% Triton X-100); 1x 2mL Lithium Chloride Buffer (10mM Tris-HCl, pH 7.5, 250mM LiCl, 1mM EDTA, 1% NP-40), and 1x 2mL TE+50mM NaCl (10mM Tris-HCl, pH 7.5, 1mM EDTA, 50mM NaCl). After washing, beads were resuspended in 200uL of Extraction Buffer (50 mM Tris pH 8, 10mM EDTA, 5mM EGTA, 1% SDS). For WCE control, 100uL of soluble input material was added to 100uL of Extraction buffer. Samples were incubated overnight at 65°C +1000rpm shaking to elute IP and reverse crosslinks. 200uL of eluted samples was cleared of beads and added to 200uL TE pH 8. Samples were digested with 0.2mg/mL RNaseA for 1 hour at 37°C. 7 uL of CaCl<sub>2</sub> solution (10mM Tris-HCl, pH 8 and 300mM CaCl<sub>2</sub>) and 4uL of 20mg/mL proteinase K were added to each sample and incubated for 1 hour at 55°C. Samples were extracted 1 x phenol:chloroform followed by 1 x chloroform and precipitated overnight at -80°C with a standard NaCl, ethanol, and glycogen DNA precipitation. Pellets were washed 2x 1mL 70% ethanol, dried, and resuspended in 70uL 0.1x TE pH 8.

All of the ChIP material (70 uL resuspended) or 200ng of input material (WCE) was used to prepare libraries. Sample DNA was end-repaired for 30 min at 20°C in a 100uL reaction: 1x T4 DNA Ligase Buffer (New England Biolabs, NEB), 0.4 mM dNTPs (NEB), 15 U of T4 DNA Polymerase (NEB), 5 U of Klenow enzyme (NEB), 50 U of T4 PNK (NEB). Samples were purified by Invitrogen Purelink Kit (Thermo Fisher) following standard conditions and eluted in 33uL. 32uL of purified product was A-tailed for 37°C for 30 min in a 50uL reaction with 1x NEB Buffer 2, 2 uM dATP (NEB), and 15 U of Klenow exo- (NEB). Samples were purified by Invitrogen Purelink Kit and eluted in 11uL. Illumina genomic adapters were ligated onto 10uL of the purified product for 15min at 20°C in the following 50uL reaction with 1x Quick Ligase Buffer (NEB), 400nM of Y-shaped Adapter Oligo, and 5uL of Quick Ligase. The reaction was cleaned up by a double size selection with AMPure beads as described by the manufacturer with the initial size selection using a 0.9x AMPure ratio and keeping the supernatant (to select against large products – mononucleosomes and larger) followed by a 1.8x AMPure selection keeping the bead bound material (to select against adapter dimers and free adapters). Note: When switching between first and second size selection, we used the following formula provided by the manufacturer to calculate the amount of additional beads to add:  $(\text{second ratio} - \text{first ratio}) * \text{volume transferred}$ . Illumina adapter oligos were added to the size selected product in a 50 uL PCR reaction with 200uM dNTPs (NEB), 1x High Fidelity Phusion Buffer (NEB), 1uL Phusion Polymerase (NEB), 0.5uM forward Illumina oligo adapter, and 0.5uM reverse Illumina oligo adapter with 16x cycles of standard Phusion PCR conditions (annealing: 65°C 30 sec and extension: 72°C 30 sec). PCR products were AMPure purified (1.8x ratio) according to manufacturer guidelines. Libraries were run on the BioAnalyzer. Libraries with adapter dimer contamination were repurified with extra AMPure selections until adapter dimer contamination

disappeared. The libraries were sequenced with 40bp paired-end reads on 4 lanes of an Illumina HiSeq 2000. Sequencing results from the four lanes were pooled and analyzed.

## **Bioinformatics and statistical analyses**

### *Mapping reads and junctions*

Raw RNA-seq reads were mapped using STAR aligner version 2.4.1d<sup>49</sup> with the parameters:

```
STAR --runMode alignReads --runThreadN 2 --genomeDir UCSC_mm9 --twopassMode Basic -  
-sjdbOverhang 74 --outSAMtype BAM SortedByCoordinate --outFilterMultimapNmax 20 --  
outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.04 --alignIntronMin 70 --  
alignIntronMax 500000 --alignMatesGapMax 500000 --alignSJoverhangMin 8 --  
alignSJDBoverhangMin 1 --outSAMstrandField intronMotif --outFilterType BySJout
```

### *Gene expression quantitation*

Gene-level quantification was performed using Rsem version 1.2.26<sup>50</sup> and EBSeq version 1.10.1<sup>51</sup> using the UCSC mm9 genome annotation with four independent replicates each (2x technical replicates of 2x independent clones) for Cdk12-expressing and Cdk12-depleted samples at 24 and 48 hours post dox-withdrawal. Parameters:

```
rsem-calculate-expression --forward-prob 0.5 --output-genome-bam -p 2 --paired-end  
UCSC_mm9
```

Rsem analysis was followed by read-count matrix assembly and standard EBSeq differential expression analysis.



Gene-level and quantification was also performed using DESeq<sup>52</sup> to obtain independent validation as well as to derive an adjusted p value for use in gene expression volcano plots. Gene expression differences at the total gene level were considered significant if the posterior probability of differential expression (PPDE) as determined by EBSeq > 0.95. Genes that exhibited statistically significant changes in response to doxycycline alone were excluded from the set of Cdk12-affected genes.

#### *Annotation of p53-target genes and bivalent promoters*

p53 target genes were derived from previously published data.<sup>25</sup> In order to be considered a direct p53 transcriptional target gene, the gene was required to have a p53 ChIP peak within ~5.5 kb upstream and 2.5 kb downstream of the TSS, and exhibit a p53-dependent transcriptional response to DNA-damage induced by Adriamycin (doxorubicin) dosing in mouse ES cells. Directionality of transcriptional responses was determined by microarray-assayed gene expression changes.

Classification of bivalent-promoter genes was based on previously published data.<sup>26</sup> Genome wide ChIP experiments were used as the basis for classifying genes based on histone modifications within the promoter region. Genes with H3K4-trimethylation overlapping H3K27-trimethylation in mouse ES cells were considered to belong to the bivalent class.

#### *Determination and quantitation of alternative splicing events*

Mapped splice junctions detected in all samples were combined and processed using custom Python scripts to filter out junctions representing < ~1% of transcripts. Alternative and constitutive intron classifications were performed using custom Python scripts, and are agnostic

with regard to existing annotations other than known gene boundaries. If no overlapping introns exist for a given intron, it is assigned to the constitutive class. The subgroups containing overlapping introns are assigned a splicing classification if the start and end coordinates of all of the constituent introns fall into a pattern representing a known splice type (cassette, mutually exclusive, alternative 5' splice site, alternative 3' splice site).

To quantify differences in alternative splicing events between Cdk12-expressing and Cdk12-depleted cells, splicing events determined from mapped junctions as described above were converted into event-specific gff3 annotation files compatible with MISO.<sup>53</sup> We then performed two separate MISO analyses (one for each clone, comparing Cdk12-expressing versus Cdk12-depleted samples) per alternative splicing type. To be considered significant, both independent clones were required to change in percent-spliced in (delta-PSI) in the same direction, both with a Bayes factor of  $\geq 5$ .

#### *Annotation of intronic polyadenylation (IPA) sites*

We used previously published 3' end sequencing data<sup>31</sup> to identify genome-wide polyadenylation sites in the same strain of mouse ES cells used to generate the Cdk12 knockout lines. These data were generated from poly(A)-selected RNA that was oligo-dT primed and reverse transcribed. cDNAs were circularized, PCR amplified, and sequenced. After mapping the reads were filtered to remove genomically-encoded poly(A) tracts and polyadenylation sites associated with B2-SINE retrotransposons. Putative cleavage sites were then required to have one of 36 polyadenylation signal sequences within 40 nucleotides upstream. In the data analysis performed here, we first combined all cleavage sites from both control and U1-snRNA antisense morpholino oligonucleotide treated cells, two replicates each, and cleavage sites within 40

nucleotides of contiguous sequence were combined into clusters. Clusters were then required to contain a minimum of 20 reads to be included.

#### *Genome-wide IPA transcript annotation and quantification*

We used custom Python scripts to derive a transcriptome annotation based on Mm9 (Mus\_musculus\_NCBI\_build37.1) gene start and end boundaries, the location of polyadenylation sites as determined by 2-p seq (described above), and the genomic locations of all mapped splice junctions from the RNA-seq data. The annotation the distal polyadenylation site isoform for each gene was set to identify the consensus isoform, i.e. the junctions defining each exon are the most frequently used junction detected within all of the combined samples. Each IPA site within a gene was then assigned to an additional transcript based on the consensus isoform but terminating at the IPA cleavage site. These gene annotations were then converted to DEXseq exon parts using DEXseq-associated script `dexseq_prepare_annotation.py` and the reads mapping to each exon part in each sample were counted using `dexseq_count.py`. Using the counts matrix thus derived for all biological replicates in each condition, DEXseq was used to identify changes in the relative abundance of each exon part as normalized to all exon parts within the gene, including those representing the IPAs and distal polyadenylation site isoform 3' terminal exon. This gave us a log<sub>2</sub>-fold change and FDR adjusted p value for each exon part as it differed between the Cdk12-expressing and Cdk12-depleted samples. IPA sites or distal polyadenylation site isoform 3' terminal exons whose exon part exhibited an adjusted p value < 0.05 were considered statistically significant.

#### *Determination of first stable nucleosome dyad positions*

Genomic coordinates of nucleosome dyads in mESCs were download from previously published data.<sup>54</sup> To identify regions with stable nucleosomes, five different mESC MNase-seq datasets were analyzed with NucTools.<sup>55</sup> Stable nucleosome regions were determined using `stable_nucs_replicates.pl` with a sliding window of 50 bases and a relative error based on five replicates  $< 0.5$ . The NucTools-defined stable region dyad downstream and most proximal to the transcription start site was regarded as the +1 dyad (first stable nucleosome).

#### *Chromatin Immunoprecipitation sequencing analysis*

Custom Python and R scripts were used to calculate normalized read densities of ChIP data, bin for metagene analysis, and perform statistical tests. Briefly, genomic coordinates for full genes or TSS-flanking regions for specific gene sets were first compiled. Mapped reads from ChIP data sets were counted within the specified regions using the Bedtools coverageBed tool,<sup>56</sup> combining both replicates of both antibodies for the ChIP under analysis. Each gene was divided into 100 equal-length bins and the total read counts per nucleotide for each gene within each bin were summed. Summed read counts were normalized by total mapped reads for the sample under consideration and the average of these count values across all genes was plotted as the normalized read density for that bin. Bin-wise p values were obtained using a one-sided Kolmogorov-Smirnov test comparing the distributions of normalized reads across all genes in the group between Cdk12 +/- samples.

#### *Determination of index of Cdk12 sensitivity*

In order to determine the cumulative effect of IPA site usage on full-length gene expression, the  $\log_2$ -fold change between Cdk12-expressing and Cdk12-depleted samples in the ratios of the previously annotated (see above) first and last (distal polyA isoform) exons of each

gene was calculated. We refer to this change in ratio as the “index of Cdk12 sensitivity”. The index for each gene was plotted as part of a distribution containing all other genes sharing the same total number of detected IPA sites as determined by the IPA annotation (see above).

*Identification and quantitation of IPA sites in The Cancer Genome Atlas (TCGA) tumor samples*

Patient samples from the prostate adenocarcinoma<sup>42</sup> and ovarian serous cystadenocarcinoma<sup>44</sup> TCGA cohorts were assessed for the presence of missense or truncating point mutations as well as copy-number variations (amplifications or deletions) in *CDK12* using cBioPortal ([www.cbioportal.org](http://www.cbioportal.org)).<sup>57,58</sup> Additionally, normalized *CDK12* mRNA expression levels were considered. A set of patients from each cohort with wild type, diploid *CDK12* loci were selected along with all available *CDK12*-mutation carrying tumors and a subset of tumors carrying *CDK12* gene deletions, as well as one to two samples from each tumor type carrying an amplified locus. Aligned reads (STAR-mapped .bam files) covering relevant gene loci were downloaded from the Genomic Data Commons (<https://gdc.cancer.gov/>) using the bam-slicing tool. Reads were visually inspected to identify regions with clear IPA events that aligned with a canonical PAS motif. Regions spanning from the upstream 5' splice site of the preceding exon to the PAS site were added to gtf annotations of the gene of interest, and DEXseq tools were used for annotating and counting reads mapping to each exon part as described above; these were used to generate a matrix of counts per exon part in each individual tumor sample. All exonic reads aligning within the gene were summed for each sample, and then read counts for each exon part were normalized to exon length and total exon counts in the gene for each sample to obtain a normalized IPA site usage metric for that sample. GraphPad PRISM 7 software was used to generate plots. Significance levels of the differences in IPA site usage between wild type diploid

versus the deletion and mutation group were determined using the Mann-Whitney U test (one-tailed).

*Data and software availability*

RNAseq and CHIP-seq data are deposited in GEO. Custom scripts used for analysis will be made available upon request.

## References

1. Lord, C. J. & Ashworth, A. BRCAness revisited. *Nat. Rev. Cancer* **16**, 110–120 (2016).
2. Bartkowiak, B. *et al.* CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes Dev.* **24**, 2303–2316 (2010).
3. Blazek, D. *et al.* The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev.* **25**, 2158–2172 (2011).
4. Zhang, T. *et al.* Covalent targeting of remote cysteine residues to develop CDK12 and CDK13 inhibitors. *Nat. Chem. Biol.* **12**, 876–884 (2016).
5. Ekumi, K. M. *et al.* Ovarian carcinoma CDK12 mutations misregulate expression of DNA repair genes via deficient formation and function of the Cdk12/CycK complex. *Nucleic Acids Res.* **43**, 2575–2589 (2015).
6. Johnson, S. F. *et al.* CDK12 Inhibition Reverses De Novo and Acquired PARP Inhibitor Resistance in BRCA Wild-Type and Mutated Models of Triple-Negative Breast Cancer. *Cell Rep.* **17**, 2367–2381 (2016).
7. Zaborowska, J., Egloff, S. & Murphy, S. The pol II CTD: new twists in the tail. *Nat. Struct. Mol. Biol.* **23**, 771–777 (2016).
8. Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* **15**, 163–175 (2014).
9. Hsin, J. P. & Manley, J. L. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev.* **26**, 2119–2137 (2012).
10. Buratowski, S. Progression through the RNA Polymerase II CTD Cycle. *Mol. Cell* **36**, 541–546 (2009).
11. Bajrami, I. *et al.* Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. *Cancer Res.* **74**, 287–297 (2014).
12. Liang, K. *et al.* Characterization of Human Cyclin-Dependent Kinase 12 (CDK12) and CDK13 Complexes in C-Terminal Domain Phosphorylation, Gene Transcription, and RNA Processing. *Mol. Cell. Biol.* **35**, 928–938 (2015).
13. Joshi, P. M., Sutor, S. L., Huntoon, C. J. & Karnitz, L. M. Ovarian cancer-associated mutations disable catalytic activity of CDK12, a kinase that promotes homologous recombination repair and resistance to cisplatin and poly(ADP-ribose) polymerase inhibitors. *J. Biol. Chem.* **289**, 9247–9253 (2014).
14. Hong & Stambrook, P. J. Restoration of an absent G1 arrest and protection from apoptosis in embryonic stem cells after ionizing radiation. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 14443–14448 (2004).
15. Aladjem, M. I. *et al.* ES cells do not activate p53-dependent stress responses and undergo p53-independent apoptosis in response to DNA damage. *Curr. Biol.* **8**, 145–155 (1998).
16. Tichy, E. D. *et al.* Mouse Embryonic Stem Cells, but Not Somatic Cells, Predominantly Use Homologous Recombination to Repair Double-Strand DNA Breaks. *Stem Cells Dev.* **19**, 1699–1711 (2010).
17. Bartek, J., Lukas, C. & Lukas, J. Checking on DNA damage in S phase. *Nat. Rev. Mol. Cell Biol.* **5**, 792–804 (2004).

18. Zeman, M. K. & Cimprich, K. A. Causes and consequences of replication stress. *Nat. Cell Biol.* **16**, 2–9 (2013).
19. Lin, T. *et al.* p53 induces differentiation of mouse embryonic stem cells by suppressing Nanog expression. *Nat. Cell Biol.* **7**, 165–171 (2005).
20. Liu, J. C. *et al.* High mitochondrial priming sensitizes hESCs to DNA-damage-induced apoptosis. *Cell Stem Cell* **13**, 483–491 (2013).
21. van der Laan, S., Tsanov, N., Crozet, C. & Maiorano, D. High Dub3 expression in mouse ESCs couples the G1/S checkpoint to pluripotency. *Mol. Cell* **52**, 366–379 (2013).
22. Shieh, S. Y., Ikeda, M., Taya, Y. & Prives, C. DNA damage-induced phosphorylation of p53 alleviates inhibition by MDM2. *Cell* **91**, 325–334 (1997).
23. Siliciano, J. D. *et al.* DNA damage induces phosphorylation of the amino terminus of p53. *Genes Dev.* **11**, 3471–3481 (1997).
24. Bode, A. M. & Dong, Z. Post-translational modification of p53 in tumorigenesis. *Nat. Rev. Cancer* **4**, 793–805 (2004).
25. Lee, K.-H. *et al.* A genomewide study identifies the Wnt signaling pathway as a major target of p53 in murine embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 69–74 (2010).
26. Ku, M. *et al.* Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet.* **4**, 1–14 (2008).
27. Tien, J. F. *et al.* CDK12 regulates alternative last exon mRNA splicing and promotes breast cancer cell invasion. *Nucleic Acids Res.* **45**, 6698–6716 (2017).
28. Elkon, R., Ugalde, A. P. & Agami, R. Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.* **14**, 496–506 (2013).
29. Tian, B. & Manley, J. L. Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* **18**, 18–30 (2017).
30. Tian, B., Pan, Z. & Lee, J. Y. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Research* **17**, 156–165 (2007).
31. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).
32. Skene, P. J. & Henikoff, S. A simple method for generating high-resolution maps of genome-wide protein binding. *eLife* **4**, 1–9 (2015).
33. Skene, P. J., Hernandez, A. E., Groudine, M. & Henikoff, S. The nucleosomal barrier to promoter escape by RNA polymerase II is overcome by the chromatin remodeler Chd1. *eLife* **3**, 1–19 (2014).
34. Weber, C. M., Ramachandran, S. & Henikoff, S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol. Cell* **53**, 819–830 (2014).
35. Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 167–177 (2015).
36. Yoh, S. M., Lucas, J. S. & Jones, K. A. The Iws1:Spt6:CTD complex controls cotranscriptional mRNA biosynthesis and HYPB/Setd2-mediated histone H3K36 methylation. *Genes Dev.* **22**, 3422–3434 (2008).
37. Yoh, S. M., Cho, H., Pickle, L., Evans, R. M. & Jones, K. A. The Spt6 SH2 domain binds Ser2-P RNAPII to direct Iws1-dependent mRNA splicing and export. *Genes Dev.* **21**, 160–174 (2007).

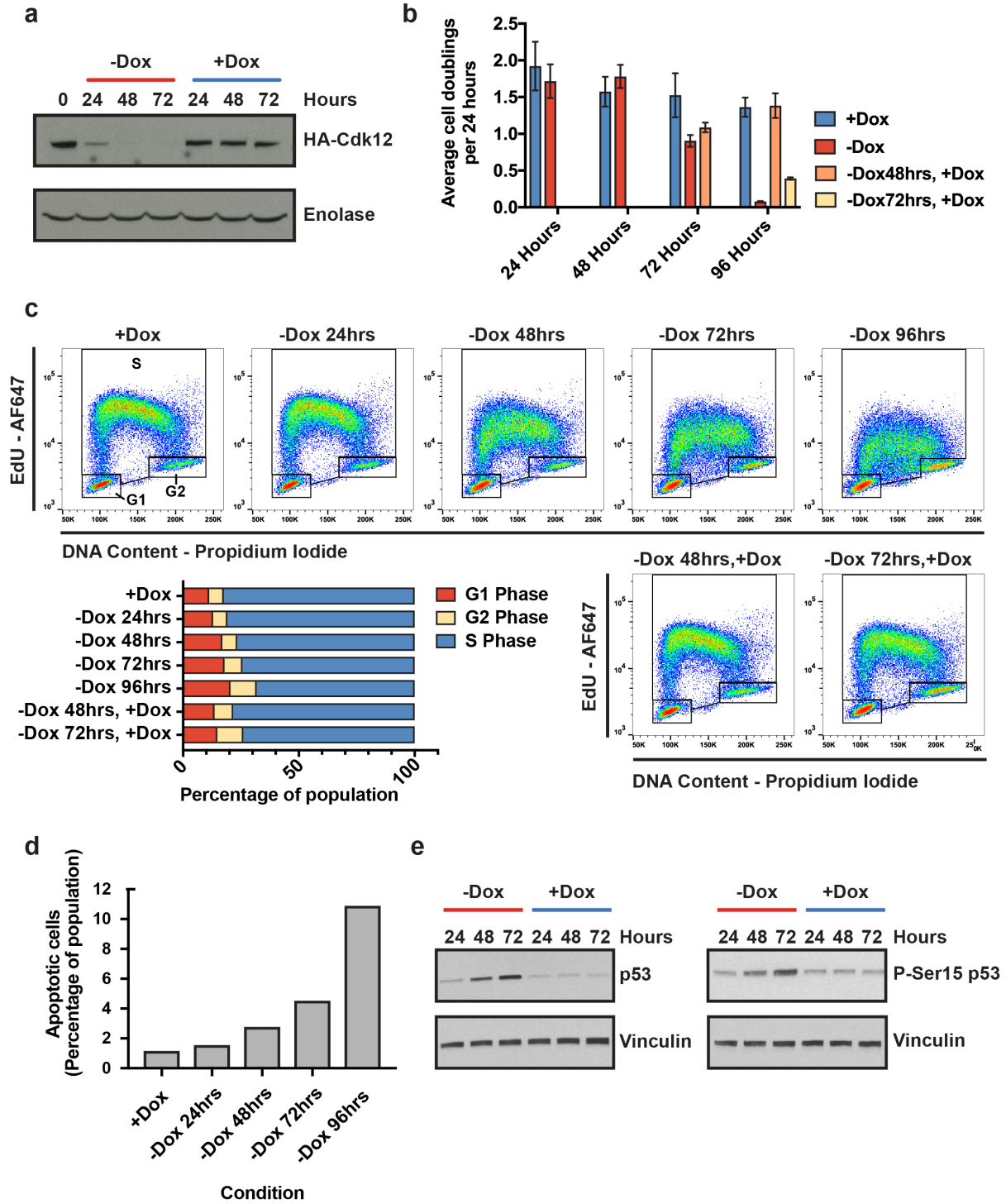


38. Cui, Y. & Denis, C. L. In Vivo Evidence that Defects in the Transcriptional Elongation Factors RPB2, TFIIIS, and SPT5 Enhance Upstream Poly(A) Site Utilization. *Mol. Cell Biol.* **23**, 7887–7901 (2003).
39. Yang, Y. *et al.* PAF Complex Plays Novel Subunit-Specific Roles in Alternative Cleavage and Polyadenylation. *PLoS Genet.* **12**, 1–28 (2016).
40. Liu, X. *et al.* Transcription elongation rate has a tissue-specific impact on alternative cleavage and polyadenylation in *Drosophila melanogaster*. *RNA* **23**, 1807–1816 (2017).
41. Grasso, C. S. *et al.* The mutational landscape of lethal castration-resistant prostate cancer. *Nature* **487**, 239–243 (2012).
42. The Cancer Genome Atlas Research Network *et al.* The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025 (2015).
43. Robinson, D. *et al.* Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* **162**, 1215–1228 (2015).
44. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
45. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
46. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
47. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
48. Gurtan, A. M., Lu, V., Bhutkar, A. & Sharp, P. A. In vivo structure-function analysis of human Dicer reveals directional processing of precursor miRNAs. *RNA* **18**, 1116–1122 (2012).
49. Dobin, A. *et al.* STAR: an ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
50. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* **12**, 323 (2011).
51. Leng, N. *et al.* EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* **29**, 1035–1043 (2013).
52. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, 1–12 (2010).
53. Katz, Y., Wang, E. T., Airoidi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Meth.* **7**, 1009–1015 (2010).
54. Voong, L. N. *et al.* Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. *Cell* **167**, 1555–1570 (2016).
55. Vainshtein, Y., Rippe, K. & Teif, V. B. NucTools: analysis of chromatin feature occupancy profiles from high-throughput sequencing data. *BMC Genomics* **18**, 1–13 (2017).
56. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
57. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, 1–19 (2013).

58. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data *Cancer Discov.* **2**, 401–404 (2012).

## Acknowledgements

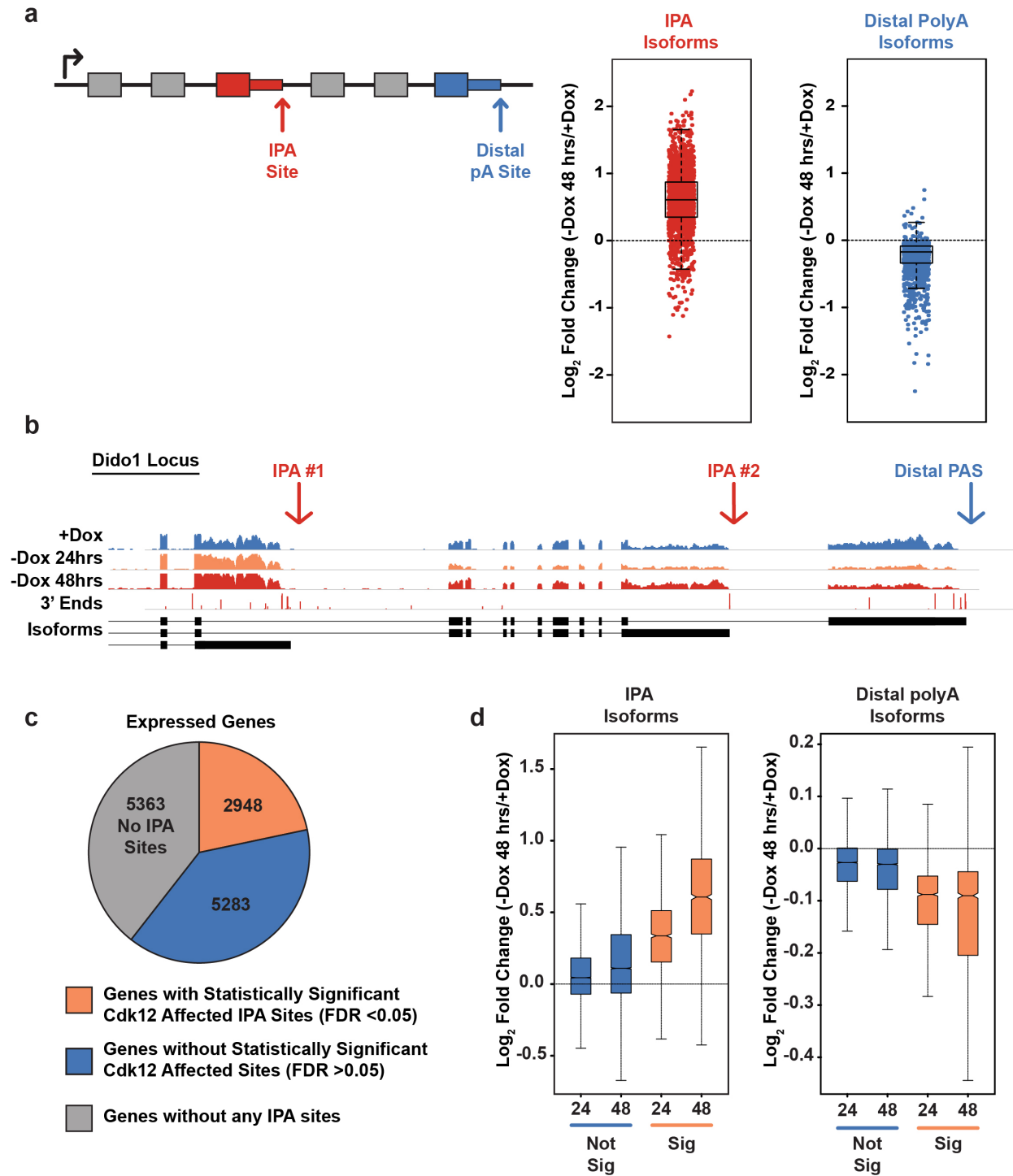
We would like to thank members of the Sharp laboratory for insightful discussions and critical reading of the manuscript. pAC4 and PBNeoTetO-Dest were generous gifts from Albert W. Cheng (The Jackson Laboratory). We thank Hiroshi Suzuki for the first stable nucleosome coordinates. The results shown here are in whole or part based upon data generated by the TCGA Research Network: <http://cancergenome.nih.gov/>. We thank the Koch Institute's Robert A. Swanson (1969) Biotechnology Center at the Massachusetts Institute of Technology for technical support, specifically Glenn Paradis, Michael Jennings, and Michele Griffin of the Flow Cytometry Core and Stuart Levine and Noelani Kamelamela of the MIT BioMicro Center. The research described here was supported by Program Project Grant P01CA042063 from the National Cancer Institute, by United States Public Health Service grants R01GM34277 and R01CA133404 from the National Institutes of Health, and partially by the Koch Institute Support (core) grant P30-CA14051 from the National Cancer Institute. Sara Dubbury was also supported by a David H. Koch Fellowship from MIT's School of Science (2014-2015) and by the NIH Pre-Doctoral Training Grant T32GM007287 to the MIT Department of Biology.



**Figure 1. Loss of Cdk12 leads to profound, but reversible, viability defects in mESCs**  
(legend on next page)

**Figure 1. Loss of Cdk12 leads to profound, but reversible, viability defects in mESCs**

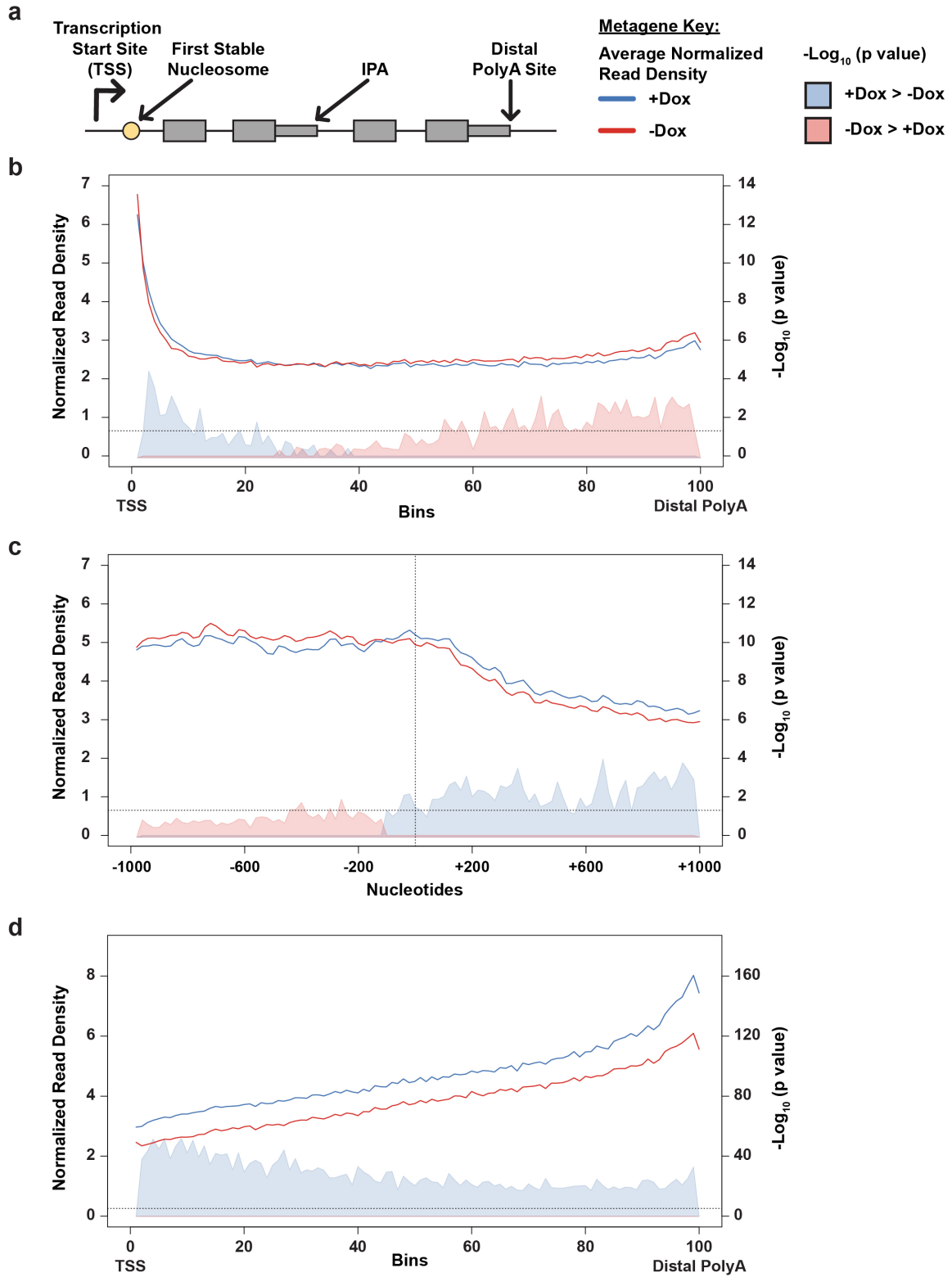
**a**, Immunoblot for Cdk12 transgene (HA epitope) in one representative Cdk12 $\Delta$  clone. Enolase serves as a loading control. **b**, Average number of cell doublings for two independent clones quantified by FACS every 24 hours. +Dox (blue bars) and -Dox (red bars) conditions maintained throughout time-course. Orange bars (-Dox 48hrs) and yellow bars (-Dox 72hrs) followed by re-addition of Dox for 48 or 24 hours, respectively. Error bars represent standard error of the mean. **c**, FACS-based cell cycle profiling of one representative clone and biological replicate for the same conditions as in **(b)** (top) and quantification (bottom). **d**, FACS-based quantification of cleaved Caspase 3 positive (apoptotic) cells. One representative clone and biological replicate shown. **e**, Immunoblot for total (left) and phosphorylated Ser15 (P-Ser15) p53 (right) upon Cdk12 loss for the indicated times. Vinculin serves as a loading control.



**Figure 2. Cdk12 loss increases intronic polyadenylation and decreases distal polyadenylation** (legend on next page)

**Figure 2. Cdk12 loss increases intronic polyadenylation and decreases distal polyadenylation**

**a**, Schematic of a gene with IPA site and distal polyadenylation site indicated. Right: boxplot of all statistically significantly changing IPA isoforms (red) and distal polyadenylation isoforms (blue) indicating the  $\log_2$  fold change in normalized read densities in cells 48 hours after Dox removal compared to cells grown in the presence of Dox. **b**, IGV browser image of RNA sequencing read density at the 3' end of one example gene (*Dido1*) with two significantly increasing IPA sites (red) and a significantly decreasing terminal exon (blue) grown in Dox or after 24 and 48 hours post-Dox removal. Density histograms for sequenced reads representing 3' ends are shown below. A schematic of the gene structure is indicated, note there are additional exons on the 5' end not shown. **c**, Pie chart showing proportions of expressed genes with at least one significantly increasing IPA and/or a significantly decreasing distal isoform (orange), with at least one identified IPA isoform that does not change significantly nor does its terminal isoform (blue), or genes without any identified IPA site (grey). **d**,  $\log_2$  fold changes of all IPA sites (left) and all terminal sites (right) in expressed genes that change statistically significantly upon Dox depletion for 24 or 48 hours (orange) or do not change statistically significantly upon Dox depletion after 24 or 48 hours (blue).

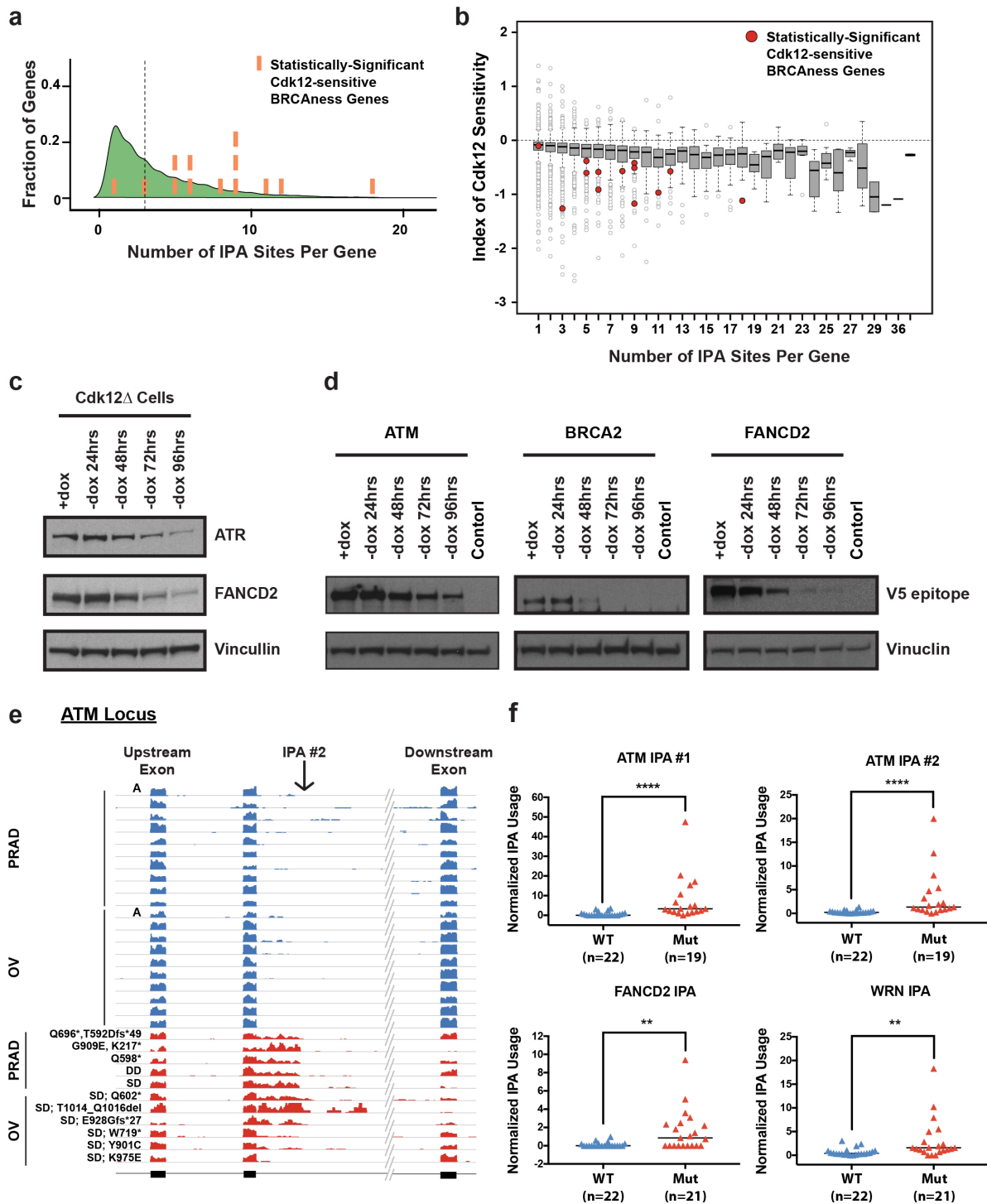


**Figure 3. Cdk12 loss results in reduced RNAPII elongation rate and a corresponding decrease in RNAPII-CTD Ser2 phosphorylation (legend on next page)**



**Figure 3. Cdk12 loss results in reduced RNAPII elongation rate and a corresponding decrease in RNAPII-CTD Ser2 phosphorylation**

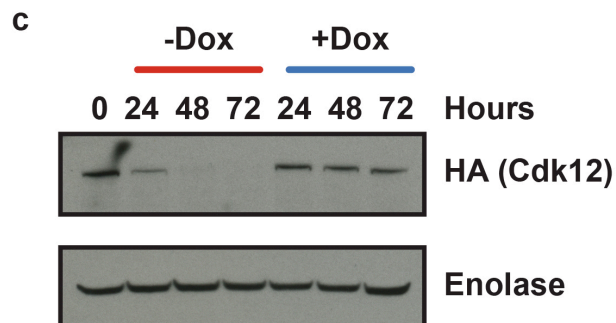
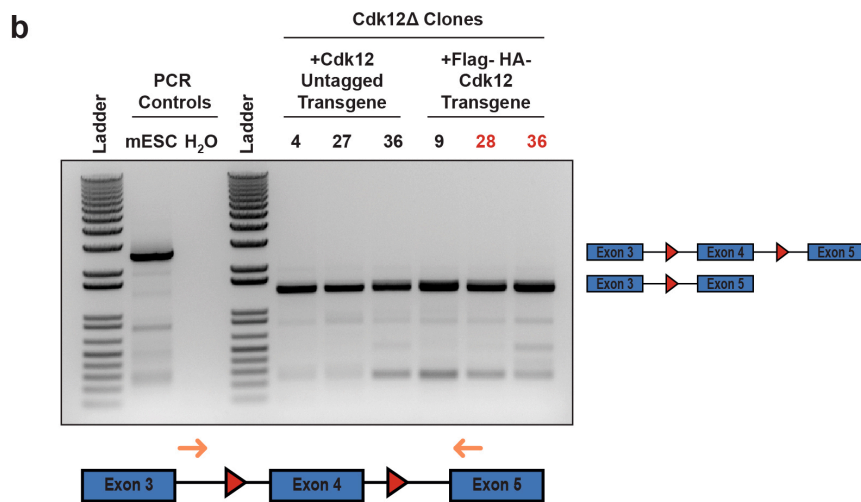
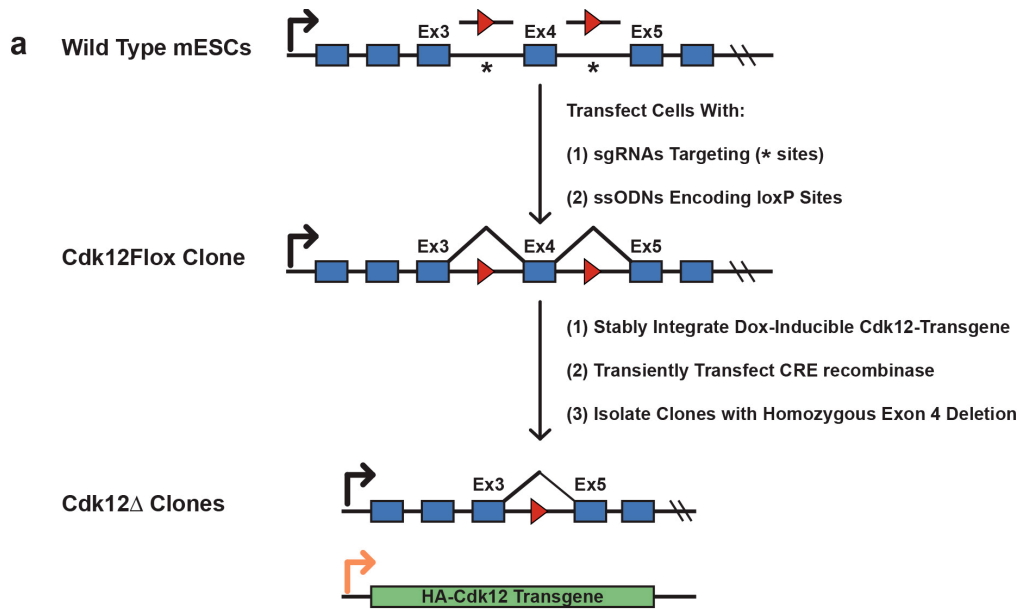
**a**, Schematic of gene elements and key to metagene plots. **b**, Metagene profiles of total RNAPII density across the gene body from the transcription start site (TSS) to the distal polyadenylation site for the significantly changing IPA/distal isoforms. Shortest and longest gene length quartiles are excluded (see Extended Data Fig. 5). In **(b)**, **(c)**, and **(d)** solid lines indicate normalized read depth with (blue) or without (red) Cdk12, and shaded areas indicate the negative  $\log_{10}$  of the bin-wise p values (Kolmogorov-Smirnov one-sided test) of the difference in read depth, with blue shading indicating the plus Cdk12 signal is significantly greater, and pink shading indicating the minus Cdk12 signal is significantly greater. The negative  $\log_{10}$  of the p value is shown in the axis on the right, and the horizontal dashed line is  $p = 0.05$ . **c**, Total RNAPII metagene density 1 kb upstream and 1kb downstream of the first stable nucleosome for the significantly changing IPA/distal isoforms. Vertical dashed line indicates the first stable nucleosome dyad. **d**, RNAPII CTD Ser2p metagene density across the entire gene body for the significantly changing IPA/distal isoforms. As in **(b)**, shortest and longest length quartiles are excluded.



**Figure 4.** Many BRCAness HR genes contain multiple IPAs, making them particularly sensitive to Cdk12 loss; human tumors with CDK12 LOF upregulate IPAs (legend on next page)

**Figure 4. Many BRCAness HR genes contain multiple IPAs, making them particularly sensitive to Cdk12 loss; human tumors with CDK12 LOF upregulate IPAs**

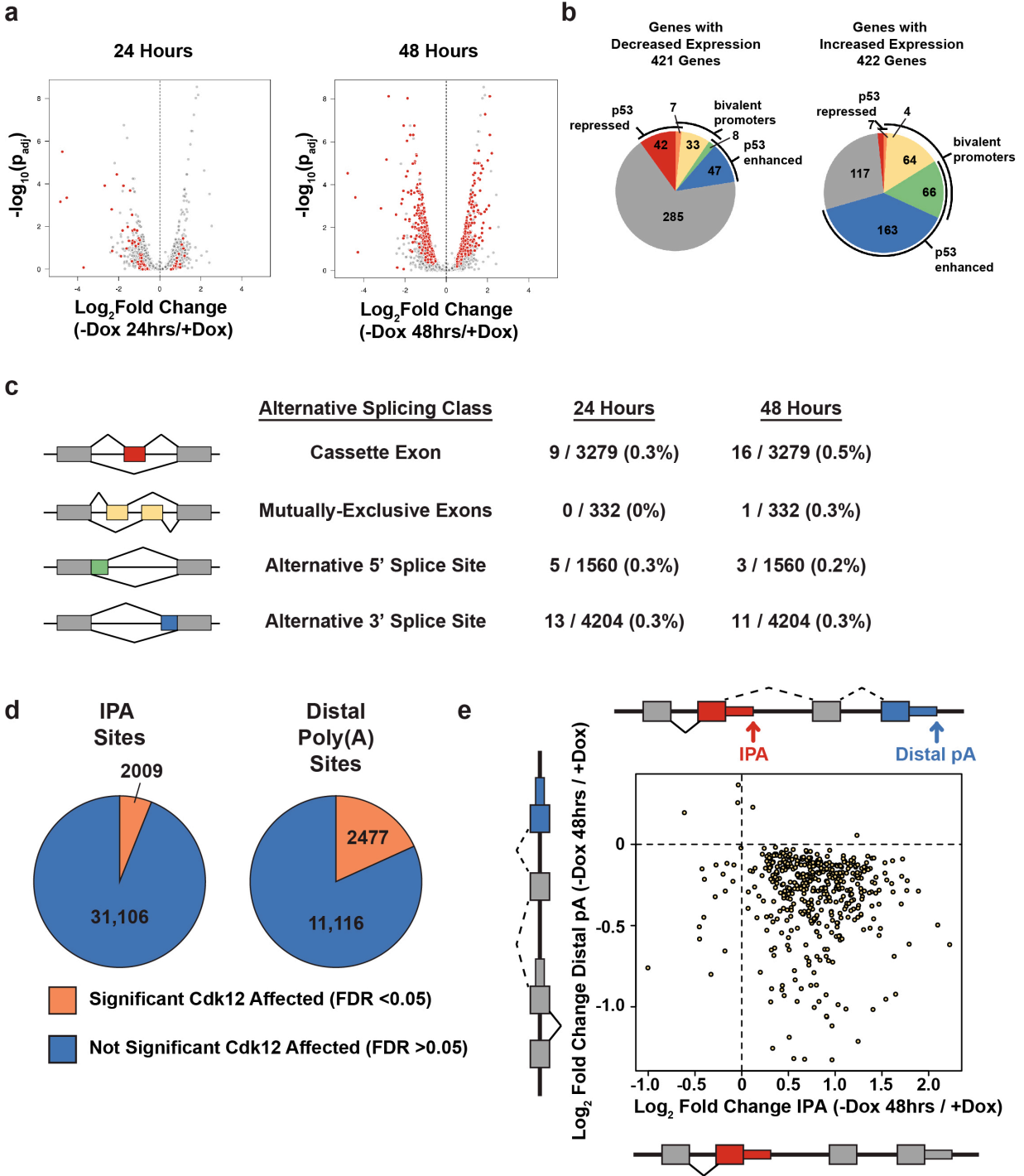
**a**, Kernel density plot of the distribution of genes with different numbers of IPA sites (x-axis). Statistically-significantly Cdk12 sensitive BRCAness HR genes are indicated as orange bars. **b**, Boxplots showing the distributions of Cdk12-loss dependent changes in the ratio of full-length isoforms to IPA usage (as quantified by comparing the expression of the distal-most exon to the expression of the first exon, “Index of Cdk12 Sensitivity”, y-axis), compared to the number of IPA sites per gene (x-axis). BRCAness genes with statistically significant Cdk12-sensitivity are highlighted in red. **c**, Immunoblots of Cdk12 $\Delta$  after Dox removal for the indicated times using endogenous antibodies against ATR and FANCD2. Vinculin serves as a loading control. **d**, ATM, BRCA2, and FANCD2 were endogenously, N-terminally V5 epitope tagged in Cdk12 $\Delta$  cells using CRISPR/Cas9. Immunoblots of V5 tag for one representative clone each of ATM, BRCA2, and FANCD2 tagged cells are shown after Dox removal for the indicated times. Lysate from untagged Cdk12 $\Delta$  cells (+Dox) are shown as an antibody control. Vinculin serves as a loading control. **e**, IGV browser shot of RNA sequencing read density from TCGA tumors of prostate adenocarcinoma (PRAD) and ovarian cystadenocarcinoma (OV) with the indicated mutational status at a CDK12-sensitive IPA site in the human ATM locus (ATM IPA #2). Tumors shown in blue are wild type for CDK12 and diploid unless marked as amplified (A). Tumors shown in red carry missense or truncating mutations and/or shallow (SD) or deep (DD) gene deletions at the CDK12 locus. **f**, Quantification of usage of two different IPA sites in human ATM and at IPA sites in FANCD2 and WRN. Tumors with wild type or amplified *CDK12* are shown in blue (WT), those with *CDK12* deletions, missense mutations, or truncating mutations in red (Mut). Medians are indicated by horizontal black bars, sample size indicated below (2 CDK12-deletion tumors showed no reads mapping to ATM, possibly due to a deletion at this locus), and p-values determined by one-sided Mann-Whitney U test (\*\* p < 0.01, \*\*\*\* p < 0.0001).



Extended Data Figure 1. Generation of Cdk12 genetic knockouts in mESCs (legend on next page)

**Extended Data Figure 1. Generation of Cdk12 genetic knockouts in mESCs**

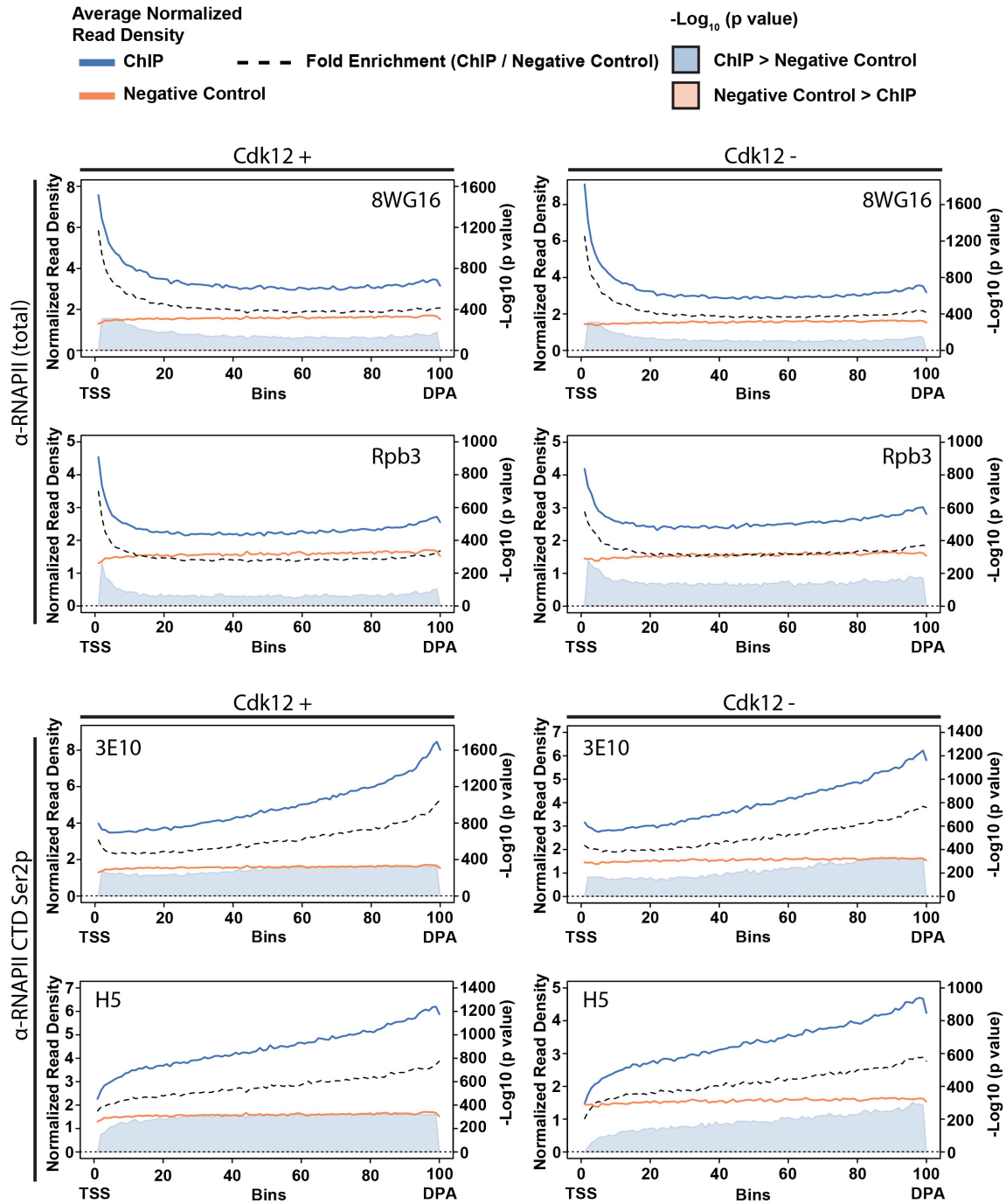
**a**, Schematic of Cdk12 $\Delta$  cell line generation, LoxP sites (red triangles), sgRNA cut sites (\*), endogenous promoter (black arrows), Doxycycline-inducible promoter (orange arrow). **b**, PCR products across the Cdk12 locus flanking exon 4 (primers shown as orange arrows) for wild type mESCs and Cdk12 $\Delta$  clones. Clones 28 and 36 used in this study are indicated in red. **c**, Immunoblot for Cdk12 transgene (HA epitope) in second independent Cdk12 $\Delta$  clone used throughout the paper.



**Extended Data Figure 2. Gene expression changes in Cdk12-depleted mESCs are dominated by increased IPA usage (legend on next page)**

**Extended Data Figure 2. Gene expression changes in Cdk12-depleted mESCs are dominated by increased IPA usage**

**a**, Volcano plots of significant gene expression events (colored dots) at the total gene level after 24 hours (left) or 48 hours (right) of Dox withdrawal. **b**, Pie chart indicating genes that decrease (left) or increase (right) in total gene expression at a statistically significant level after 24 or 48 hours of Dox withdrawal (combined). Likely secondary effects are indicated: p53 repressed genes (red), p53 enhanced genes (blue), bivalent promoter genes (yellow), p53 repressed and bivalent promoter genes (orange), and p53 enhanced and bivalent promoter genes (green). Genes belonging to none of the above groups are indicated in gray. **c**, Table summarizing significant alternative splicing events observed after 24 and 48 hours of Dox depletion in Cdk12 $\Delta$  cells. **d**, Left: IPA sites exhibiting a statistically significant change (orange) or not (blue) in expressed genes after 24 or 48 hours of Dox depletion. Right: Expressed genes with terminal polyadenylation sites that are significantly changed (orange) or not statistically significant (blue) as normalized to the rest of the transcript. **e**, Scatterplot showing log<sub>2</sub> fold changes upon Cdk12 loss in distal exons (y-axis) versus IPA sites (x-axis) in genes that have both a statistically significant IPA site and a statistically significant distal polyadenylation change.

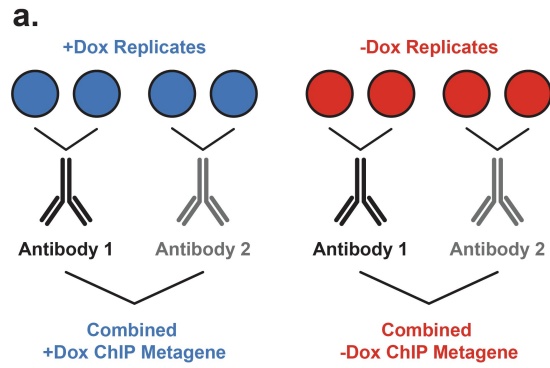


**Extended Data Figure 3. ChIP antibodies recognizing the same target protein exhibit strongly overlapping metagene patterns (legend on next page)**

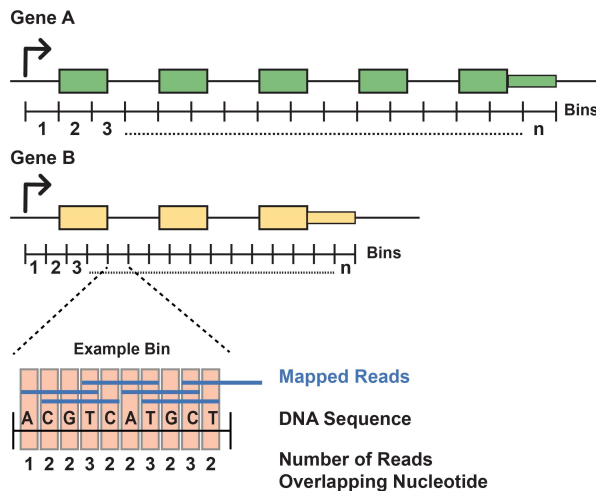


**Extended Data Figure 3. ChIP antibodies recognizing the same target protein exhibit strongly overlapping metagene patterns**

Metagene profiles broken down by individual antibodies used. Blue lines indicate normalized read density for the specific ChIP in the two combined biological replicates. Orange lines indicate the negative control (combined whole cell extract and all antibody negative controls combined for both replicates). Black dashed lines indicate the fold-enrichment (specific ChIP / negative control). Shaded areas indicate the negative  $\log_{10}$  of the bin-wise p values (Kolmogorov-Smirnov one-sided test) of the difference in read depth, with blue shading indicating the plus Cdk12 signal is significantly greater. The negative  $\log_{10}$  of the p value is shown in the axis on the right, and the horizontal dashed line is  $p = 0.05$ . TSS=Transcription Start Site; DPA=Distal PolyAdenylation site. Both total RNAPII antibodies (8WG16 and Rpb3) and both Ser2 RNAPII antibodies (3E10 and H5) show similar density patterns across genes. The patterns are consistent, though the magnitudes differ, between Cdk12+ (left panels) and Cdk12-depleted (right panels)



**b. Step 1: Divide each gene into  $n$  equal length bins**

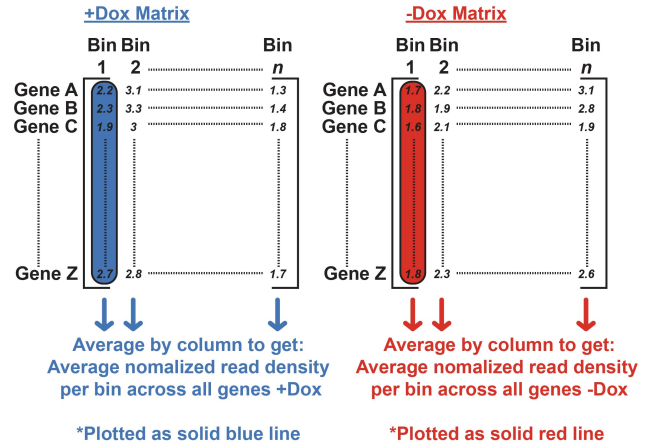


**C. Step 2: Determine average read density per nucleotide per bin in each gene and condition**

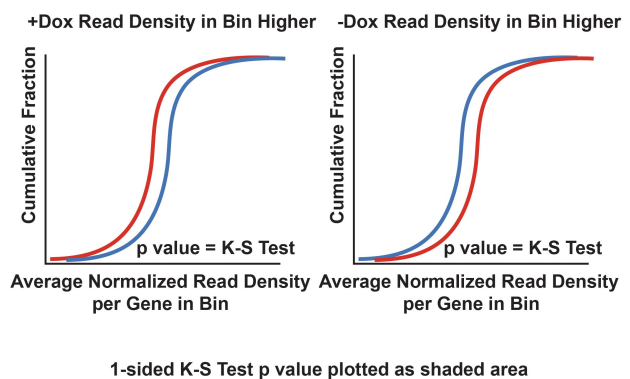
For Example Bin from Step 1:

$$\frac{(1+2+2+3+2+2+3+2+3+2) \text{ reads}}{10 \text{ nucleotides per bin}} = 2.2 \text{ average reads per nt per bin}^*$$

\* Normalize value by mapped reads in library



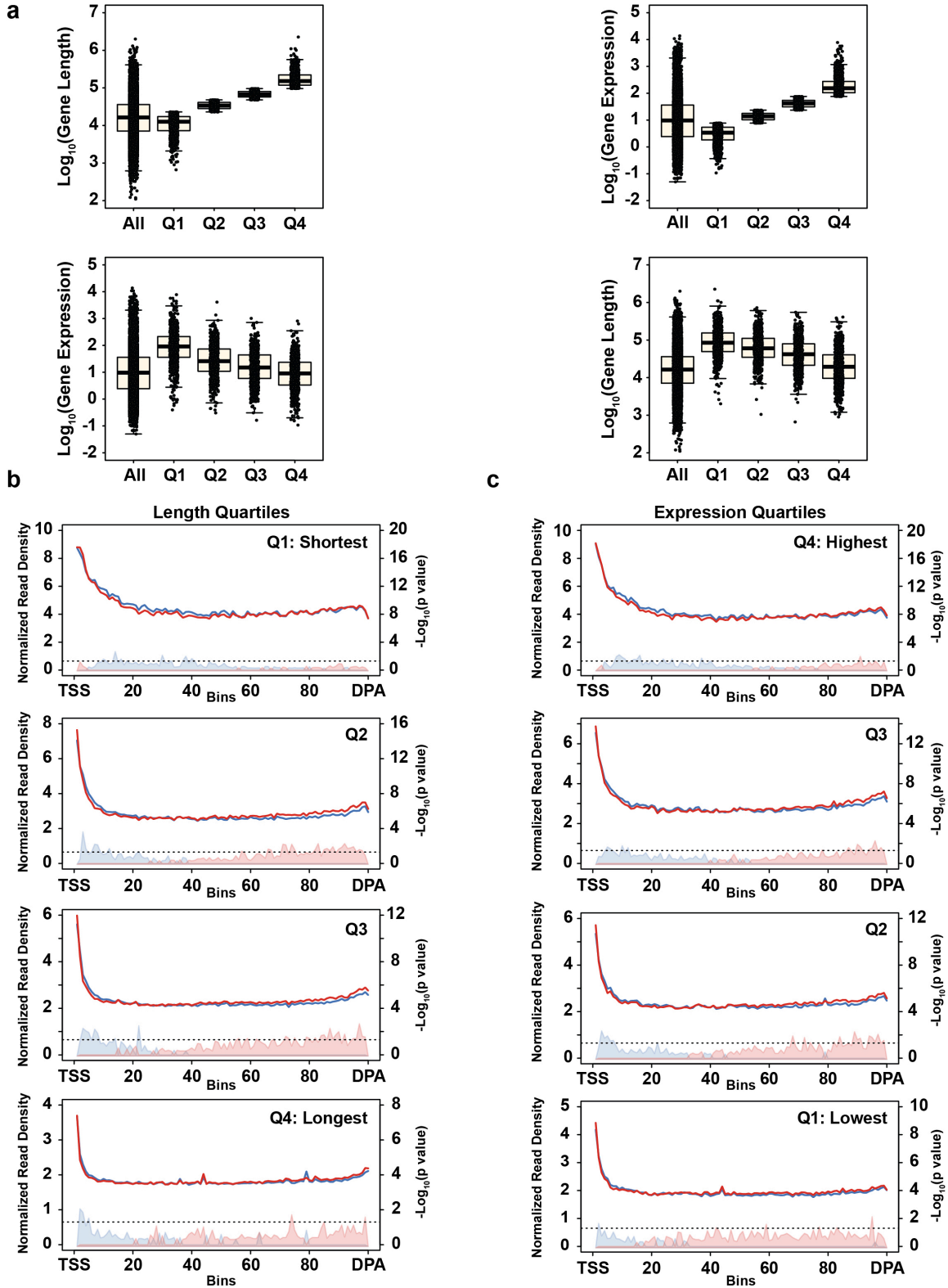
**d. Step 3: Compare distribution of normalized read density in each bin across all genes by condition (+Dox or -Dox) and determine statistical significance via K-S Test: CDF plots representative of shaded columns in Step 2 matrices**



### Extended Data Figure 4. Schematic of ChIP experiments and data analysis

**a,** Schematic of biological replicate/antibody replicate experimental design. Each ChIP set (RNAPII and Ser2p, in Cdk12<sup>+/-</sup>) consists of 2 biological replicates each ChIP'ed with two different antibodies recognizing the same protein. These four replicates were then combined for the ChIP metagene analyses. **b-d,** Schematic of the steps used to determine average read densities for the ChIP assays, and the statistical test used to determine significant differences in the read density dependent on Cdk12 expression.

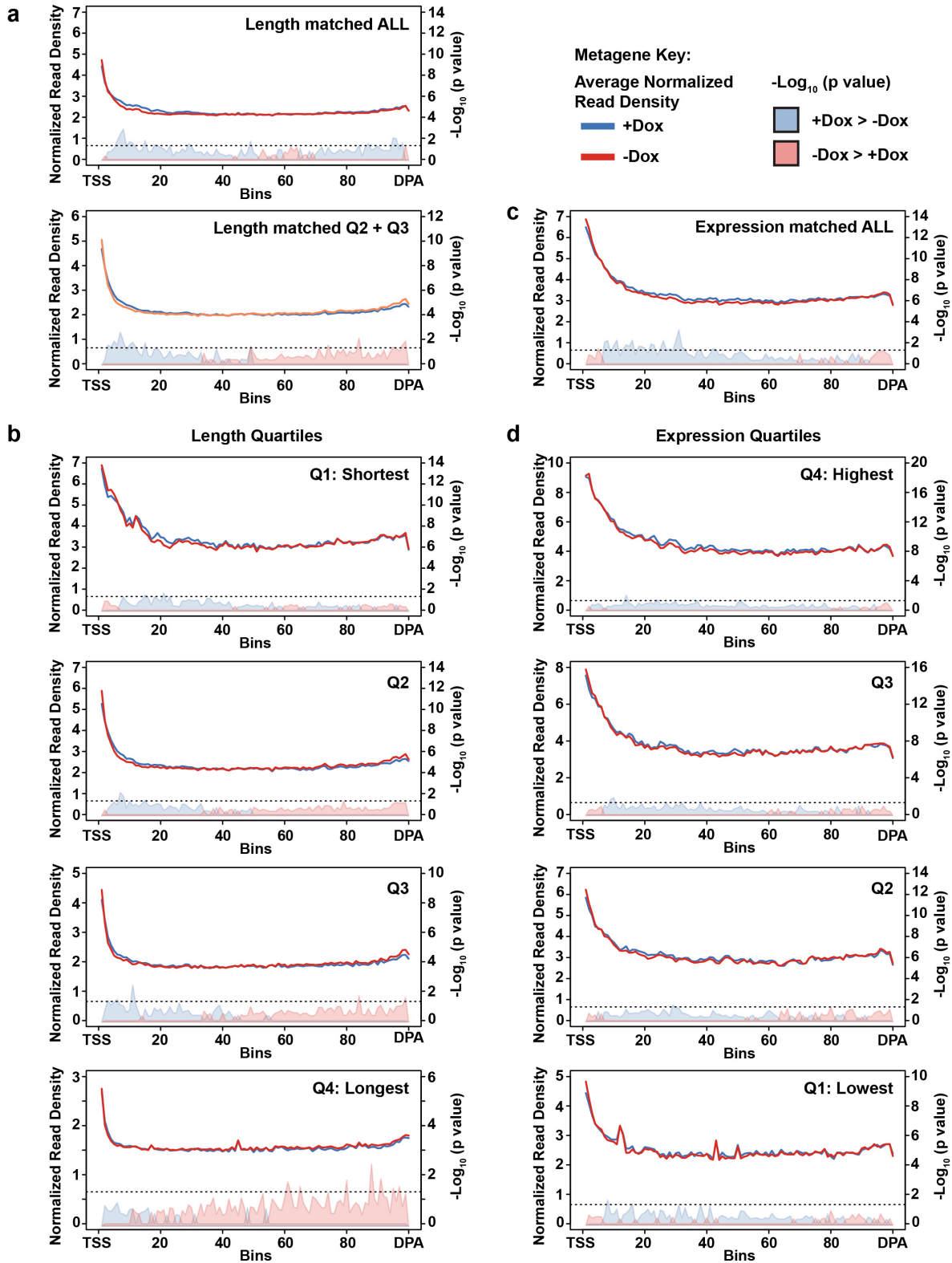




**Extended Data Figure 5. RNAPII metagene patterns are influenced by gene length and expression** (legend on next page).

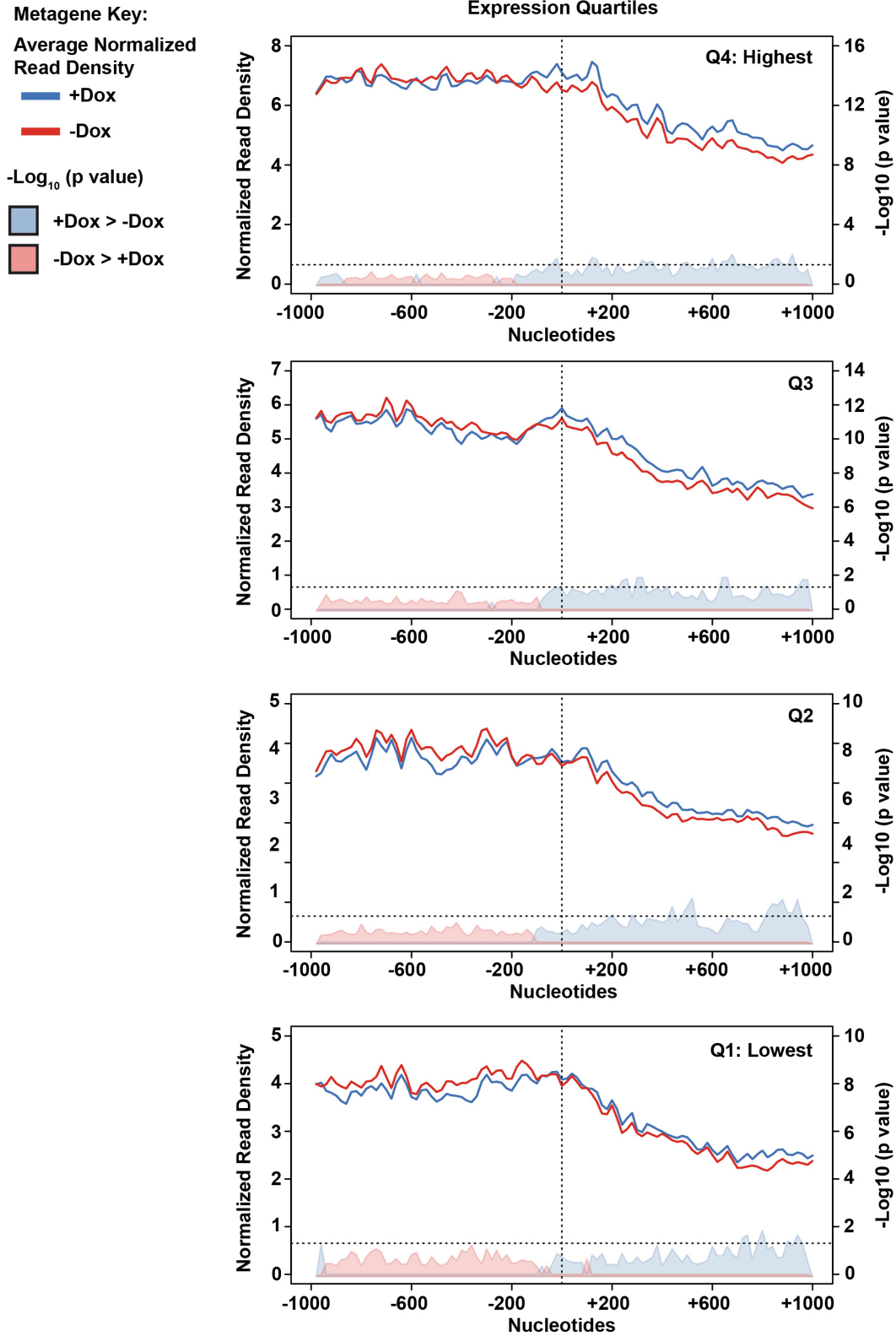
**Extended Data Figure 5. RNAPII metagene patterns are influenced by gene length and expression**

**a**, Boxplots of length and expression quartiles of the significantly changing IPA/distal isoform genes. Top panels: size distributions ( $\log_{10}$  of length in nucleotides) of each length quartile (left) and gene expression distributions ( $\log_{10}$  of transcripts per million) of each expression quartile (right) compared to the respective distributions of all expressed genes. Bottom panels: expression distributions for each length quartile (left) and length distributions for each expression quartile (right). Note that gene length is generally inversely correlated with expression level, but median expression of all quartiles of the significantly changing IPA/distal isoforms is higher than the median for all expressed genes. Additionally, the median length of all expression quartiles of the significantly changing IPA/distal isoforms is longer than the median for all expressed genes. Thus, the genes comprising the significantly changing IPA/distal isoform set are longer and more highly expressed for their length than the broader gene population. **b**, Metagene profiles of RNAPII density in genes with a statistically-significant Cdk12-sensitive IPA or terminal site divided into length-based quartiles. TSS=Transcription Start Site; DPA=Distal PolyAdenylation site. Note that in the shortest quartile, the Cdk12-depleted cells show a trend toward increased density at the 3' end, but the shortest genes terminate before the polymerase can reach a higher density than the Cdk12 competent cells. Conversely, the longest genes are expressed at a lower level (see **a**) resulting in lower RNAPII ChIP signal. For these reasons, the shortest and longest length quartiles were excluded in Fig. 3**b,d**. **c**, Metagene profiles of RNAPII density in genes with a statistically-significant Cdk12-sensitive IPA or terminal site divided into expression-based quartiles.



Extended Data Figure 6. RNAPII ChIP pattern is not specific to Cdk12 IPA-affected genes (legend on next page)

**Extended Data Figure 6. RNAPII ChIP pattern is not specific to Cdk12 IPA-affected genes**  
**a**, Metagene profile of RNAPII density in a set of control genes length-matched to the significantly changing IPA/distal isoforms gene set. Top: all control genes. Bottom: shortest and longest quartiles removed (as in Fig. 3B). TSS=Transcription Start Site; DPA=Distal PolyAdenylation site. **b**, Metagene profile of RNAPII density in a set of control genes length-matched to the significantly changing IPA/distal isoforms gene set divided into length quartiles. **c**, Metagene profile of RNAPII density in a set of control genes expression-matched to the significantly changing IPA/distal isoforms gene set. **d**, Metagene profile of RNAPII density in a set of control genes expression-matched to the significantly changing IPA/distal isoforms gene set divided into expression quartiles.

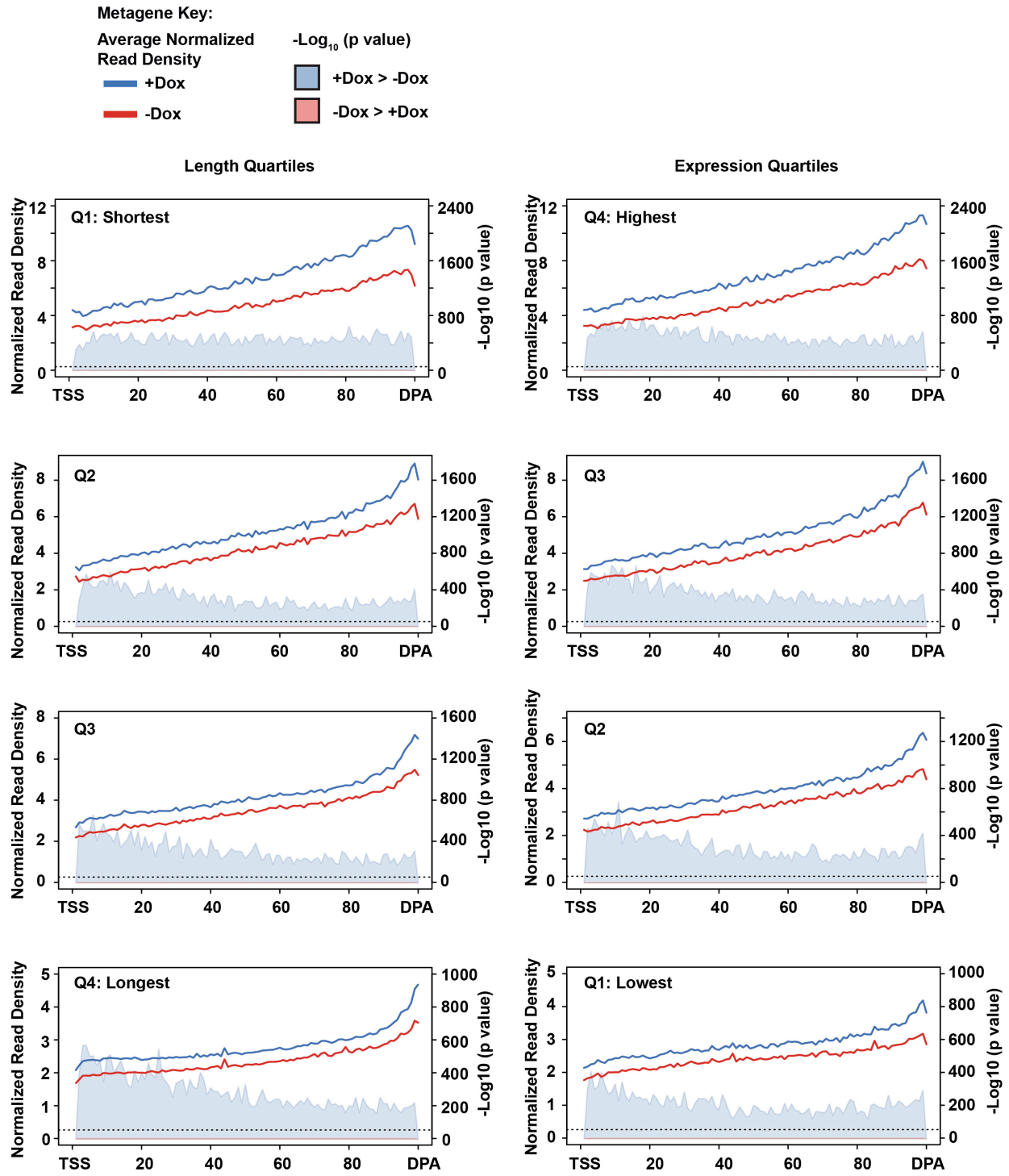


**Extended Data Figure 7. Increased RNAPII upstream and decreased RNAPII downstream of first stable nucleosome occurs in all gene expression quartiles (legend on next page)**



**Extended Data Figure 7. Increased RNAPII upstream and decreased RNAPII downstream of first stable nucleosome occurs in all gene expression quartiles**

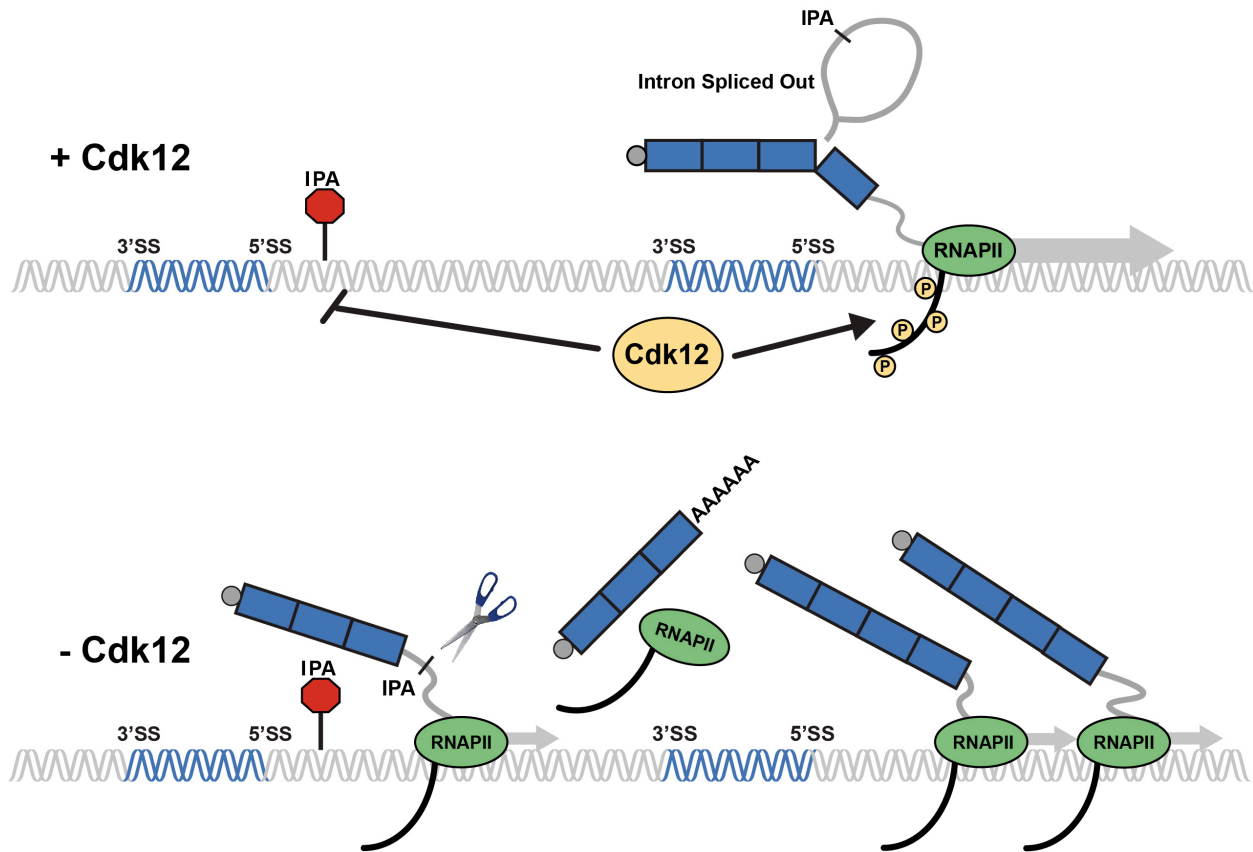
Total RNAPII metagene density 1 kb upstream and 1kb downstream of the first stable nucleosome for the significantly changing IPA/distal isoforms divided into gene expression quartiles. As in Fig 3C, solid lines indicate normalized read depth with (blue) or without (red) Cdk12, and shaded areas indicate the negative  $\log_{10}$  of the bin-wise p values (Kolmogorov-Smirnov one-sided test) of the difference in read depth, with blue shading indicating the plus Cdk12 signal is significantly greater, and pink shading indicating the minus Cdk12 signal is significantly greater. The negative  $\log_{10}$  of the p value is shown in the axis on the right, and the horizontal dashed line is  $p = 0.05$ . Vertical dashed line indicates the position of the first stable nucleosome dyad.



**Extended Data Figure 8. Ser2p is depleted by Cdk12 loss and metagene patterns are influenced by gene expression and length (legend on next page)**

**Extended Data Figure 8. Ser2p is depleted by Cdk12 loss and metagene patterns are influenced by gene expression and length**

**a**, Metagene profiles of Ser2p density in genes with a statistically-significant Cdk12-sensitive IPA or terminal site divided into length-based quartiles. As in Fig. 3D, solid blue lines indicate average normalized read density in Cdk12+ cells, red solid lines are the average normalized read density in Cdk12-depleted samples. Light blue shading indicates the plus Cdk12 signal is significantly greater. TSS=Transcription Start Site; DPA=Distal PolyAdenylation site. Note that in all quartiles, Ser2p signal increases from TSS to TES and is profoundly depleted in Cdk12 depleted cells. However, shorter genes have a steeper, linear increase while longer genes begin with a shallower slope that increases in slope toward the TSS; in shorter genes the difference between Cdk12 + and – cells is similar across the gene body, while the difference is greatest in the 5' end of longer genes and becomes less pronounced toward the 3' ends. **b**, Metagene profiles of Ser2p density in genes with a statistically-significant Cdk12-sensitive IPA or terminal site divided into expression-based quartiles.



**Extended Data Figure 9. Model for Cdk12-dependent effect on elongation rate affecting kinetics of IPA site usage.**

**Upper panel:** As RNAPII transcribes through a region of a gene (exonic regions shown in blue with 5' and 3' splice sites (SS) indicated, introns in gray) containing an IPA site (red octagon), Cdk12-dependent RNAPII-CTD Ser2 phosphorylation suppresses IPA site usage by enhancing the rate of transcription elongation (thick arrow), thereby increasing the frequency at which the polymerase transcribes through the downstream exon to enable splicing of the IPA-containing intron before the IPA site can be cleaved for termination and polyadenylation.

**Lower panel:** In the absence of Cdk12, RNAPII-CTD Ser2 phosphorylation is decreased, and the rate of elongation slows, resulting in an increase in the use of the IPA site for cleavage and polyadenylation. Polymerase that transcribes through the downstream exon, enabling removal of the intron prior to cleavage and polyadenylation, accumulates with increasing density toward the 3' end of the gene.

## **Chapter 3**

### **Conclusions and Future Directions**

Until recently, Cdk9 was the only known RNAPII CTD Ser2 kinase in mammals. Since Cdk12 and Cdk13 were identified as novel RNAPII CTD Ser2 kinases, their functional roles during mRNA biogenesis have remained unclear. Mounting evidence that Cdk12 functions as a “master regulator” of homologous recombination (HR) repair machinery piqued substantial clinical interest in Cdk12 as a potential biomarker for “BRCAness” tumors and in small-molecule inhibitors of Cdk12 as potential chemotherapeutics to induce Parp1 inhibitor sensitivity in otherwise HR-competent tumor cells. Despite this growing clinical interest, the mechanism behind Cdk12’s apparent “specificity” for regulating the HR repair machinery remained unknown.

## **Conclusions**

In this thesis, we established a genetic system in mouse embryonic stem cells (mESCs) to conditionally ablate Cdk12 activity and probe the molecular function of Cdk12. Cdk12 ablation in mESCs caused striking cellular phenotypes suggestive of impaired HR repair. Using this system, we found that Cdk12 loss resulted in few gene expression changes at the total gene level, consistent with several previous studies (Blazek et al., 2011; Johnson et al., 2016; Liang et al., 2015; Zhang et al., 2016). In contrast, we observed profound changes in mRNA isoform composition caused by increased usage of intronic cleavage and polyadenylation sites (IPAs) genome-wide. This data suggests that Cdk12 primarily affects mRNA processing by suppressing IPA sites across the genome.

High-resolution ChIP sequencing data allowed us to probe the transcriptional consequences of Cdk12 loss. Using this method, we showed that Cdk12 loss resulted in global

increases in total RNAPII density across the 3' end of genes irrespective of whether those genes contained Cdk12-sensitive IPAs. This pattern occurred in conjunction with subtly decreased RNAPII density entering productive elongation immediately downstream of the first stable nucleosome. These total RNAPII density patterns coincided with significantly decreased RNAPII CTD Ser2 phosphorylation across the entire gene. This data is consistent with CDK12 knockout causing a decrease in RNAPII elongation rate across the gene body, presumably through the loss of RNAPII CTD Ser2 phosphorylation. As described in the introduction to this thesis, RNAPII CTD Ser2 phosphorylation is known to recruit several factors to elongation-competent RNAPII. Future experiments are needed to decipher the molecular factors mediating this effect.

mRNA splicing and cleavage/polyadenylation occur co-transcriptionally and this temporal coupling allows for the regulation of IPA via transcription elongation rate (Bentley, 2014). Usage of IPA sites is in kinetic competition with splicing of the encompassing intron, which cannot occur until the downstream 3' splice site is transcribed. Once the encompassing intron is spliced out, which often occurs rapidly after transcription of the downstream exon, the IPA can no longer be used. Therefore, the cleavage and polyadenylation machinery has a temporal window between transcription of the IPA site and transcription of the downstream 3' splice site to recognize the IPA site and cleave the RNA. Modulation of transcription elongation rate would thus be expected to alter this kinetic balance. A decrease in the transcription elongation rate would result in longer amounts of time between transcription of the IPA site and transcription of the downstream exon, and therefore increased likelihood that the IPA site will be recognized and cleaved. Alternatively, an increased transcription elongation rate would mean less time between transcription of these two sites and a decreased likelihood that the IPA site

would be used. Little is known about the mechanisms that regulate RNAPII elongation rate. However, several studies utilizing mutant polymerases or knockdown of elongation factors that slow RNAPII elongation rates have shown that alterations in transcription elongation rates do indeed alter IPA usage, supporting the kinetic model proposed above (Cui and Denis, 2003; Liu et al., 2017; Yang et al., 2016). Although the alteration in RNAPII elongation rate that we observed upon Cdk12 knockout is not enough to affect the expression of a majority of genes, those with active IPA sites showed enhanced IPA site usage and increased production of truncated isoforms suggesting that this mechanism is potentially extremely sensitive to even low magnitude changes in elongation rate.

Usage of IPA sites results in the expression of truncated RNA isoforms, which can alter protein-coding sequence and/or RNA stability, frequently resulting in the loss-of-function of the protein product (Tian and Manley, 2017). Our data suggests that Cdk12's role in suppressing IPA site usage accounts for its phenotypic specificity and proposed role as a "master regulator" of HR repair machinery. Although Cdk12 depletion resulted in increased IPA usage genome-wide, we found that as a group, many HR repair genes have significantly more IPA sites per gene. These include some of the most functionally essential genes in the HR repair pathway, including ATM, ATR, BRCA2, and FANCD2. As expected, genes with more IPA sites showed enhanced repression of their full-length isoforms upon Cdk12 loss. Over half of the BRCAness genes are enriched in the number of IPA sites per gene and consequently, their full-length isoforms show greater repression upon Cdk12 loss. We propose that the combinatorial affect of attenuating full-length isoform expression of multiple genes within the same functional pathway makes HR repair exquisitely sensitive to Cdk12 loss.



Finally, our work suggests several important implications for the diagnosis and treatment of ovarian and prostate cancer. Previous studies suggested that recurrent CDK12 loss-of-function (LOF) mutations in high-grade serous ovarian carcinoma (Carter et al., 2012; Ekumi et al., 2015; The Cancer Genome Atlas Research Network, 2011) and castration-resistant prostate cancer (Grasso et al., 2012; Robinson et al., 2015) may predict ovarian or prostate tumors with “BRCAness” phenotypes and therefore, sensitivity to particular drug regimens. Our data suggests that using RNA sequencing data to analyze IPA site usage in HR repair machinery in conjunction with Cdk12 mutational status may generate a more accurate set of diagnostic biomarkers. Although a number of Cdk12 mutations have been validated *in vitro* for their effects on Cdk12 activity (Ekumi et al., 2015), not all Cdk12 mutations, especially novel missense mutations, have known functional consequences. These mutations could be significant events or passenger mutations, and therefore these mutations would have unknown prognostic potential in the clinic. To circumvent this problem, IPA site usage in key HR repair genes (such as ATM) could be quantified from RNA sequencing data isolated in parallel from the patient’s tumor as a functional readout for Cdk12 activity. Our analysis of RNA sequencing data from TCGA patient cohorts provided proof of concept for this idea with a tumor harboring the Cdk12 E928Gfs\*27 mutation (confirmed LOF mutation *in vitro*) showing dramatic upregulation of an ATM IPA site while a tumor harboring the Cdk12 K975E (a mutation with near wild-type levels of Cdk12 activity *in vitro*) showed no usage of the ATM IPA site (Ekumi et al., 2015).

## Future Directions

### *Functional Role for Cdk13 in Transcription and mRNA Biogenesis*

Our work has begun to characterize Cdk12's role in transcription, mRNA biogenesis, and tumor biology. However, the function of Cdk12's mammalian specific paralog, Cdk13, remains almost completely unexplored. These two RNAPII CTD Ser2 kinases share the same cyclin co-factor (cyclin K) and almost identical kinase domains (92% conserved sequence identity across the 302 amino acid domain). Given these shared features, several groups have speculated that the two kinases may perform similar molecular functions. Several pieces of evidence from our work and others now suggest that this is likely not the case. First, Cdk13 is expressed in our Cdk12 knockout clones. However, this level of expression is incapable of rescuing the fitness defects observed upon Cdk12 knockout. Second, although Cdk12 is essential in mESCs, Cdk13 knockouts are phenotypically normal (unpublished data). Finally, accumulating evidence suggests that Cdk12 and Cdk13 LOF mutations cause distinct physiological defects in humans. While Cdk12 LOF mutations are strongly implicated in the development of "BRCAness" tumors, mounting evidence suggests that Cdk13 LOF mutations cause developmental defects including congenital heart abnormalities, dysmorphic facial features, and intellectual developmental delays (Bostwick et al., 2017; Deciphering Developmental Disorder Study, 2017; Sifrim et al., 2016). Taken together, these observations suggest that Cdk12 and Cdk13 may perform independent functions in mammals.

How Cdk13 regulates gene expression and how its activity relates to Cdk12 (and the other transcriptional Cdks) still remain unanswered. Answers to these questions would not only further our understanding of gene expression mechanisms but would also provide key insight into the development of Cdk12 inhibitors as chemotherapeutics. Current inhibitors of Cdk12

interact directly with the kinase domain (Johnson et al., 2016; Kwiatkowski et al., 2014; Zhang et al., 2016). Due to the strong sequence conservation between the kinase domains of Cdk12 and Cdk13, even the most specific inhibitors of Cdk12 show little distinction between their affinity for Cdk12 and Cdk13 (Kwiatkowski et al., 2014; Zhang et al., 2016). Recently, Cdk12 inhibitors have shown exciting results *in vitro* and *in vivo* for inducing Parp1 inhibitor sensitivity in cells that either have *de novo* or acquired Parp1 inhibitor resistance (Johnson et al., 2016). These effects have largely been attributed to Cdk12 based upon literature precedence that Cdk12 is a positive regulator of HR repair genes (Bajrami et al., 2014; Blazek et al., 2011; Ekumi et al., 2015; Johnson et al., 2016; Joshi et al., 2014; Liang et al., 2015); our own work supports this model. However, whether or not inhibition of Cdk13 (or other cross-targeted Cdks) contributes to these effects have not been formally tested.

To probe the function of Cdk13 in gene expression, we would mimic our experimental design used here for analyzing Cdk12 function in Cdk13 knockout clones. This would allow us to compare the data from our single-kinase knockout cell lines, which would ultimately allow us to tease apart independent effects of the two kinases from potentially redundant roles. Furthermore, by establishing single knockouts of each kinase in mESCs, we could use these cell lines to characterize specific and non-specific effects of Cdk12/Cdk13 inhibitors. To do this, we would treat each single knockout with or without inhibitor and identify which gene expression and cellular phenotypes (e.g. Parp1 inhibitor sensitivity) were caused by Cdk12 inhibition versus Cdk13 inhibition. Formally testing the molecular specificity of these cellular phenotypes could have important implications for next-generation small-molecule inhibitors and their desired specificity.

### *Dissecting the Molecular Factors Downstream of Cdk12*

As mentioned above, our experiments do not suggest what molecular players mediate the effect of Cdk12 on elongation rate. RNAPII CTD Ser2 phosphorylation has been suggested to recruit elongation factors to the transcribing polymerase. Follow-up studies could attempt to use a targeted approach to test individual elongation factors known to interact with Ser2 phosphorylation on RNAPII CTD, such as Spt6 (Yoh et al., 2007; 2008), to see if overexpression partially (or fully) rescues the Cdk12 knockout phenotype. These experiments are complicated, however, by our relatively naive understanding of the factors involved in regulating transcription elongation. Instead, we could take an unbiased approach and perform parallel knockout and activating genome-wide CRISPR/Cas9 screens to identify factors that suppress the viability phenotype observed upon Cdk12 loss. Identification of factors from these screens can then be followed up biochemically.

### *Dissecting the Function of Cdk12 in Ovarian Cancer*

Accumulating evidence now suggests that inactivating mutations in Cdk12 promote tumorigenesis in ovarian carcinomas, likely by fostering a “BRCAness” phenotype and promoting the accumulation of tumorigenic mutations. Interestingly, despite ovarian tumor cells selecting for Cdk12 loss-of-function mutations, Cdk12 is absolutely essential for viability in mESCs. This suggests that either: (1) Cdk12 is essential only in certain cell types, or (2) ovarian tumors acquire additional mutations that permit cell viability following Cdk12 LOF. Distinguishing between these two scenarios may have profound implications for the treatment of ovarian carcinoma. If Cdk12 LOF mutations require a secondary mutation for viability, identification of these mutations could identify additional (Parp1 inhibitor-independent)

druggable targets that would induce synthetic lethality in Cdk12 LOF cells. This type of treatment would be especially useful in Cdk12 knockout tumors that develop Parp1 inhibitor resistance. To test this, we would create inducible Cdk12 knockout clones (akin to how the Cdk12 $\Delta$  mESCs were made) in hTERT-immortalized ovarian epithelial cells that are otherwise untransformed and non-tumorigenic (Merritt et al., 2013). If Cdk12 LOF is not compatible with viability in this cell line, this would imply the second scenario. We could then perform parallel CRISPR knockout and CRISPR activation screens to identify necessary secondary mutations required for Cdk12 LOF viability in ovarian epithelial cells. If Cdk12 loss is compatible with viability in this cell line, then we could use the same system with and without Parp1 inhibitor treatment to identify mutations that cause resistance to Parp1 inhibitors in Cdk12 LOF ovarian epithelial cells.

#### *Dissecting the Function of Cdk12 in Breast Cancer*

Much of this thesis has been dedicated to how Cdk12 LOF promotes tumorigenesis by fostering a “BRCAness” phenotype in ovarian and prostate tumors. Interestingly, however, mutations in Cdk12 are also frequently identified in Her2-amplified breast cancer tumors. Surprisingly, instead of harboring LOF mutations in Cdk12, these Her2-amplified breast cancers are highly enriched for Cdk12 amplification at the DNA, RNA, and protein levels (Mertins et al., 2016). Complicating this molecular profile, Cdk12 is located adjacent to the Her2 gene (ERBB2) in the genome. As with many tumorigenic amplifications, selection for Her2 amplifications drives amplification of many of the surrounding genes as passenger mutations. These observations raise two potential hypotheses, either: (1) Cdk12 is amplified as a passenger mutation during ERBB2 genomic amplification; or (2) intriguingly, Cdk12 can act as a tumor

suppressor in certain contexts (for example ovarian tumors), and as an oncogene in other contexts (for example Her2-amplified breast cancer).

Given its role in suppressing IPA sites, several mechanisms could account for Cdk12 switching between the role of a tumor suppressor and that of an oncogene depending upon the cellular context. Our work and others have shed significant light on the “tumor-suppressive” role for Cdk12 in a “BRCAness” type context. Any potential “oncogenic” roles for Cdk12 are based solely on conjecture. However, literature precedent suggests one very intriguing hypothesis: perhaps, amplified expression of Her2 requires high levels of Cdk12 activity. Her2 is a receptor tyrosine kinase (RTK). As mentioned in Chapter 1, RTKs are critical for a host of cellular processes that drive tumorigenesis, and as a family, RTKs are enriched for strong IPA sites upstream of their transmembrane domain. Usage of these IPA sites frequently results in dominant negative versions of the RTK that are endogenously used to decrease the activity of these receptors. Vorlova *et al* recently screened for all RTKs regulated by IPA in this manner and identified 19 IPA-sensitive RTKs, one of which was Her2 (Vorlová et al., 2011). Perhaps, amplified levels of Cdk12 activity are required to suppress IPA site usage and effectively increase full-length (functional) Her2 expression in tumors dependent upon Her2-amplified signaling. Several experiments are required to begin testing if Cdk12 activity can act as an oncogene in addition to a tumor suppressor. An initial test to distinguish between a “passenger mutation” and an “oncogenic” role for Cdk12 in Her2-amplified tumors, would be to knockdown Cdk12 in breast cancers cells that are amplified for both Her2 and Cdk12 and identify if there is a proliferative disadvantage or decreased sensitivity to Her2 ligand binding.

## References

- Bajrami, I., Frankum, J.R., Konde, A., Miller, R.E., Rehman, F.L., Brough, R., Campbell, J., Sims, D., Rafiq, R., Hooper, S., et al. (2014). Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. *Cancer Res.* *74*, 287–297.
- Bentley, D.L. (2014). Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* *15*, 163–175.
- Blazek, D., Kohoutek, J., Bartholomeeusen, K., Johansen, E., Hulinkova, P., Luo, Z., Cimermancic, P., Ule, J., and Peterlin, B.M. (2011). The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev.* *25*, 2158–2172.
- Bostwick, B.L., McLean, S., Posey, J.E., Streff, H.E., Gripp, K.W., Blesson, A., Powell-Hamilton, N., Tusi, J., Stevenson, D.A., Farrelly, E., et al. (2017). Phenotypic and molecular characterisation of CDK13-related congenital heart defects, dysmorphic facial features and intellectual developmental disorders. *Genome Med.* *9*, 1–9.
- Carter, S.L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P.W., Onofrio, R.C., Winckler, W., Weir, B.A., et al. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* *30*, 413–421.
- Cui, Y., and Denis, C.L. (2003). In Vivo Evidence that Defects in the Transcriptional Elongation Factors RPB2, TFIIS, and SPT5 Enhance Upstream Poly(A) Site Utilization. *Mol. Cell. Biol.* *23*, 7887–7901.
- Deciphering Developmental Disorder Study (2017). Prevalence and architecture of de novo mutations in developmental disorders. *Nature* *542*, 433–438.
- Ekumi, K.M., Paculova, H., Lenasi, T., Pospichalova, V., Bösken, C.A., Rybarikova, J., Bryja, V., Geyer, M., Blazek, D., and Barboric, M. (2015). Ovarian carcinoma CDK12 mutations misregulate expression of DNA repair genes via deficient formation and function of the Cdk12/CycK complex. *Nucleic Acids Res.* *43*, 2575–2589.
- Grasso, C.S., Wu, Y., Robinson, D.R., Cao, X., Dhanasekaran, S.M., Khan, A.P., Quist, M.J., Jing, X., Lonigro, R.J., Brenner, J.C., et al. (2012). The mutational landscape of lethal castration-resistant prostate cancer. *Nature* *487*, 239–243.
- Johnson, S.F., Cruz, C., Greifenberg, A.K., Dust, S., Stover, D.G., Chi, D., Primack, B., Cao, S., Bernhardt, A.J., Coulson, R., et al. (2016). CDK12 Inhibition Reverses De Novo and Acquired PARP Inhibitor Resistance in BRCA Wild-Type and Mutated Models of Triple-Negative Breast Cancer. *Cell Rep.* *17*, 2367–2381.

- Joshi, P.M., Sutor, S.L., Huntoon, C.J., and Karnitz, L.M. (2014). Ovarian cancer-associated mutations disable catalytic activity of CDK12, a kinase that promotes homologous recombination repair and resistance to cisplatin and poly(ADP-ribose) polymerase inhibitors. *J. Biol. Chem.* *289*, 9247–9253.
- Kwiatkowski, N., Zhang, T., Rahl, P.B., Abraham, B.J., Reddy, J., Ficarro, S.B., Dastur, A., Amzallag, A., Ramaswamy, S., Tesar, B., et al. (2014). Targeting transcription regulation in cancer with a covalent CDK7 inhibitor. *Nature* *511*, 616–620.
- Liang, K., Gao, X., Gilmore, J.M., Florens, L., Washburn, M.P., Smith, E., and Shilatifard, A. (2015). Characterization of Human Cyclin-Dependent Kinase 12 (CDK12) and CDK13 Complexes in C-Terminal Domain Phosphorylation, Gene Transcription, and RNA Processing. *Mol. Cell. Biol.* *35*, 928–938.
- Liu, X., Freitas, J., Zheng, D., Hoque, M., Oliveira, M.S., Martins, T., Henriques, T., Bin Tian, and Moreira, A. (2017). Transcription elongation rate has a tissue-specific impact on alternative cleavage and polyadenylation in *Drosophila melanogaster*. *RNA* *23*, 1807–1816.
- Merritt, M.A., Bentink, S., Schwede, M., Iwanicki, M.P., Quackenbush, J., Woo, T., Agoston, E.S., Reinhardt, F., Crum, C.P., Berkowitz, R.S., et al. (2013). Gene expression signature of normal cell-of-origin predicts ovarian tumor outcomes. *PLoS ONE* *8*, 1–10.
- Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., et al. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* *534*, 55–62.
- Robinson, D., Van Allen, E.M., Wu, Y., Schultz, N., Lonigro, R.J., Mosquera, J.-M., Montgomery, B., Taplin, M.-E., Pritchard, C.C., Attard, G., et al. (2015). Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* *162*, 1215–1228.
- Sifrim, A., Hitz, M.-P., Wilsdon, A., Breckpot, J., Turki, A.I., S.H., Thienpont, B., McRae, J., Fitzgerald, T.W., Singh, T., Swaminathan, G.J., et al. (2016). Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing. *Nat. Genet.* *48*, 1060–1065.
- The Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* *474*, 609–615.
- Tian, B., and Manley, J.L. (2017). Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.* *18*, 18–30.
- Vorlová, S., Rocco, G., Lefave, C.V., Jodelka, F.M., Hess, K., Hastings, M.L., Henke, E., and Cartegni, L. (2011). Induction of antagonistic soluble decoy receptor tyrosine kinases by intronic polyA activation. *Mol. Cell* *43*, 927–939.
- Yang, Y., Li, W., Hoque, M., Hou, L., Shen, S., Tian, B., and Dynlacht, B.D. (2016). PAF Complex Plays Novel Subunit-Specific Roles in Alternative Cleavage and Polyadenylation. *PLoS Genet.* *12*, 1–28.



Yoh, S.M., Cho, H., Pickle, L., Evans, R.M., and Jones, K.A. (2007). The Spt6 SH2 domain binds Ser2-P RNAPII to direct Iws1-dependent mRNA splicing and export. *Genes Dev.* *21*, 160–174.

Yoh, S.M., Lucas, J.S., and Jones, K.A. (2008). The Iws1:Spt6:CTD complex controls cotranscriptional mRNA biosynthesis and HYPB/Setd2-mediated histone H3K36 methylation. *Genes Dev.* *22*, 3422–3434.

Zhang, T., Kwiatkowski, N., Olson, C.M., Dixon-Clarke, S.E., Abraham, B.J., Greifenberg, A.K., Ficarro, S.B., Elkins, J.M., Liang, Y., Hannett, N.M., et al. (2016). Covalent targeting of remote cysteine residues to develop CDK12 and CDK13 inhibitors. *Nat. Chem. Biol.* *12*, 876–884.



# Sara Jane Dubbury

Tel: 425-501-0761

Email: sara.dubbury@gmail.com

---

## Education

**Massachusetts Institute of Technology**, Cambridge, MA

*Ph.D., Department of Biology*, February 2018

Thesis Advisor: Professor Phillip Sharp, Koch Institute for Integrative Cancer Research

**Northwestern University**, Evanston, IL

*BA, Biological Sciences (Molecular Biology & Genetics), Minor in Chemistry*, June 2010

Magna Cum Laude; Departmental Honors in Biology; Phi Beta Kappa

GPA: 3.850/4.000

## Research Experience

**PhD Candidate**, Massachusetts Institute of Technology

Department of Biology, June 2011-February 2018

Thesis Advisor: Professor Phillip Sharp, Koch Institute for Integrative Cancer Research

**Undergraduate Honors Thesis Research**, Northwestern University

Biological & Chemical Engineering Department, September 2008-June 2010

Advisor: Professor Lonnie Shea

- Developed the use of lentiviral vectors as the gene delivery mechanism in 3D live cell arrays— an innovative, high-throughput tool (developed in the Shea lab) that simultaneously quantifies the functional activity of transcription factors from a wide range of intracellular signaling pathways in 2D and 3D cell culture. These live cell arrays were aimed at analyzing developmental and disease-causing phenomena at the molecular level by measuring the output (modification in transcription factor activity) of key intracellular signaling pathways in parallel over extended time periods either in response to *in vitro* tissue development or to various extracellular treatments (stresses, drugs, etc).

**Summer Cancer Research Training Fellow**, National Cancer Institute, National Institutes of Health

Laboratory of Cellular Biology, Chemical Immunology Section, Summer 2008

Advisor: Dr. Ettore Appella

- Quantitative, temporal analysis of the wip1 and mdm2 negative feedback inhibition of the p53 tumor suppressor gene

**Undergraduate Chemistry Research Intern**, Northwestern University

Chemistry Department, Marks Research Group, January 2007-June 2008

Advisor: Professor Tobin Marks

- Designed, synthesized, and characterized a bimetallic lanthanide catalyst for use in hydroalkoxylation polymerization synthesis of small alkenols

**Undergraduate Chemistry Research Intern**, University of Washington  
Chemistry Department, Dalton Research Group, June 2005-September 2006, June 2007-  
September 2007

Advisors: Professor Larry Dalton and Dr. Scott Hammond

- Designed multi-step organic synthesis and compound characterization for two projects aimed at synthesizing modified peripheries onto standard non-linear optic chromophores with the purpose of increasing noncentrosymmetric ordering within a thin film chromophore material

### **Publications**

**Dubbury, S.J.\***, Boutz, P.L.\*, Sharp, P.A., Cdk12 regulates DNA repair genes by suppressing intronic polyadenylation. *In submission*. \*Co-first Authors

Weiss, M.S., Peñalver Bernabé, B., Shin, S., Asztalos, S., **Dubbury, S.J.**, Mui, M.D., Bellis, A.D., Bluver, D., Tonetti, D.A., Rodriguez, J.S., Broadbelt, L.J., Jeruss, J.S., Shea, L.D. Dynamic transcription factor activity and networks during ErbB2 breast oncogenesis and targeted therapy. *Integrative Biology* **6**, 1170-1182 (2014).

Hammond, S.R., Sinness, J., **Dubbury, S.**, Firestone, K.A., Wawrzak, Z., Bennedict, J., Clot, O., Reid, P., Dalton, L.R. Molecular Engineering of Nanoscale Order in Dipolar Organic Glasses. *Journal of Materials Chemistry* **22**, 6752-6764 (2012).

### **Poster Presentations**

Dubbury, S.J., Boutz, P.L., and Sharp, P.A. (August 22-August 26, 2017). “Cdk12 Suppresses Early Transcript Cleavage and Polyadenylation.” 2017 Cold Spring Harbor Lab Eukaryotic mRNA Processing Meeting — Cold Spring Harbor, NY

Dubbury, S.J., Boutz, P.L., and Sharp, P. A. (July 10-July 15, 2016). “Loss of CDK12 causes intronic polyadenylation and p53 activation.” 2016 Gordon Conference on Post-Transcriptional Gene Regulation — Stowe, VT

Dubbury, S.J., Boutz, P.L., and Sharp, P. A. (August 18-August 22, 2015). “Characterizing CDK12 and CDK13 Activity in Transcription and mRNA Biogenesis in Mouse Embryonic Stem Cells.” 2015 Cold Spring Harbor Lab Eukaryotic mRNA Processing Meeting — Cold Spring Harbor, NY

Dubbury, S.J., Boutz, P.L., and Sharp, P. A. (March 28-April 1, 2015). “Characterizing CDK12 and CDK13 Activity in Transcription and mRNA Biogenesis in Mouse Embryonic Stem Cells.” Experimental Biology 2015 Meeting, American Society of Biochemistry and Molecular Biology – Boston, MA

Dubbury, S.J. and Sharp, P.A. (June 13-15, 2014). “Characterizing CDK12 and CDK13 activity in transcription and mRNA biogenesis in mouse embryonic stem cells.” 2014 Sharp Lab Reunion Symposium on RNA Biology – Koch Institute, Cambridge, MA

## Research Grant Awards and Fellowships

- Koch Institute Marlena Felter Bradford Research Travel Fellowship- Travel award to attend the 2015 CSHL Eukaryotic mRNA Processing Meeting
- David H. Koch Fellowship in Cancer Research Award- Full tuition, stipend, and health insurance funding for the 2014-2015 academic year
- Katherine L. Kriehbaum Scholarship Award- 1 of 5 students in the College of Arts and Sciences Class of 2010 to receive specialized research funds for proposed honors thesis research, Northwestern University, 2009-2010
- Undergraduate Research Grant Summer Award- \$3000 to support summer undergraduate research project, Northwestern University Office of the Provost, Summer 2009
- Undergraduate Research Grant Academic Year Award- \$1000 to support winter quarter undergraduate research project, Northwestern University Office of the Provost, Winter 2009
- Summer Cancer Research Training Award, National Cancer Institute, NIH, 2008

## Teaching Experience

### MIT Graduate Teaching Assistant, MIT, September 2013-December 2013

- 7.03 Genetics, Professors: Gerald Fink and Peter Reddien
- Responsible for two weekly recitations (discussion sections), making problem sets, revising exams, holding office hours, grading, and organizing course logistics

### MIT Graduate Teaching Assistant, MIT, September 2011-December 2011

- 7.06 Cell Biology, Professors: Angelika Amon and Harvey Lodish
- Responsible for running a weekly recitation (discussion section), making problem sets, revising exams, holding office hours, grading, and organizing course logistics

### MIT Principles of Biochemical Analysis (7.51) Tutor, MIT, September 2011-December 2011

- Tutored Graduate Biochemistry (7.51) to first-year graduate students

### MIT Office of Minority Education, MIT, October 2010-December 2010

- Led an academic enrichment seminar for freshmen focused on problem solving skills and mastery of core subject matter in General Chemistry
- Facilitated weekly drop-in tutor hours for Introductory Biology and General Chemistry students

### Mill Creek Education, Mill Creek, WA, June 2010-August 2010

- Tutored individual and small groups of high school students in all levels of math, chemistry, and biology (including AP courses)

### N'CAT Tutoring Program, Northwestern University, September 2009-June 2010

- Tutored individual student-athletes in General Chemistry, Organic Chemistry, Physics, General Biology, Single-Variable Calculus, and Statistics
- Facilitated group drop-in tutoring sessions for 3-12 student-athletes in General Chemistry

**NU Tutors**, Northwestern University, March 2009-March 2010

- Tutored individual high school students in all levels of Biology, Chemistry, Physics, and Math
- All students increased their performance by at least one letter grade

### **Academic Honors and Scholarships**

- Phi Beta Kappa Honors Society, 2010
- Graduated Magna Cum Laude from Northwestern University, 2010
- Graduated with Departmental Honors in Biology from Northwestern University, 2010
- F. Sheppard Shanley Fund Endowed Tuition Scholarship, Northwestern University, 2008-2010
- J.G. Nolan Endowed Tuition Scholarship, Northwestern University, 2007-2010
- Women's Board Endowed Tuition Scholarship, Northwestern University, 2007-2008
- Alpha Lambda Delta Honors Society, 2007-2010
- 10 Quarters on the Dean's List, 2006-2010

### **Service Activities**

**MIT BioREFS**, Massachusetts Institute of Technology, June 2011-February 2018

- Member of the MIT BioREFS (Resource for Easing Friction and Stress)- We are a group of MIT biology graduate students and postdocs that serve as an independent resource and support system to the biology community to encourage, diffuse, and mediate to fellow graduate students and postdocs.
- 40 hours of mediation training through MIT Division of Student Life