Bioinformatic analysis of multi-omics data elucidates transcriptional cyclin dependent kinase mediated transcriptional regulation

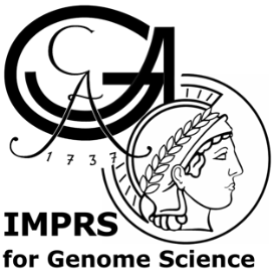
Doctoral Dissertation

for the award of the degree

"Doctor of philosophy" (Ph.D.)

of the Georg-August-Universität Göttingen

within the doctoral program



International Max Planck Research School for Genome Science

of the Göttingen Graduate School for Neurosciences, Biophysics, and Molecular Biosciences (GGNB)

submitted by

Eusra Mohammad

from Dhaka, Bangladesh

Göttingen 2022

Members of Thesis Advisory Committee (TAC)

Prof. Dr. Patrick Cramer
Department of Molecular Biology
Max Planck Institute for Multidisciplinary Sciences (MPI-NAT), Göttingen, Germany

Prof. Dr. Axel Munk
Institute for Mathematical Stochastics
Georg-August-University Göttingen, Göttingen, Germany
Research Group Statistical Inverse Problems in Biophysics Institute of Cell Biochemistry
Max Planck Institute for Multidisciplinary Sciences (MPI-NAT), Göttingen, Germany

Dr. Melina Schuh
Department of Meiosis
Max Planck Institute for Multidisciplinary Sciences (MPI-NAT), Göttingen, Germany

Members of the examination board

First reviewer: Prof. Dr. Patrick Cramer
Department of Molecular Biology
Max Planck Institute for Multidisciplinary Sciences (MPI-NAT), Göttingen, Germany

Second reviewer: Prof. Dr. Axel Munk
Institute for Mathematical Stochastics
Georg-August-University Göttingen, Göttingen, Germany
Research Group Statistical Inverse Problems in Biophysics Institute of Cell Biochemistry
Max Planck Institute for Multidisciplinary Sciences (MPI-NAT), Göttingen, Germany

Other members of the examination board

Dr. Melina Schuh
Department of Meiosis
Max Planck Institute for Multidisciplinary Sciences (MPI-NAT), Göttingen, Germany

Prof. Dr. Gregor Eichele
Emeritus Group Genes and Behavior
Max Planck Institute for Multidisciplinary Sciences (MPI-NAT), Göttingen, Germany

Prof. Dr. Matthias Dobbelstein
Department of Molecular Oncology
University Medical Center Göttingen, Germany

Dr. Ufuk Günesdogan
Department of Developmental Biology
Göttingen Center for Molecular Biology , Göttingen, Germany

Date of thesis submission:	March 31, 2022
Date of the oral examination:	June 3, 2022

ACKNOWLEDGEMENTS

With heart full of gratitude, I thank the omnipresent, omniscient and omnipotent for blessing me with courage and patience, for enabling me to accomplish the work, protecting me and providing me with endurance and wisdom in all spheres of life.

I would like to express my sincere gratitude to all the people who have motivated me throughout the journey. Their direct and indirect inspirations aided the completion of the work successfully.

Prof. Dr. Patrick Cramer provided me with me the opportunity to work in a scientifically enriched environment in Max Planck Society. I feel blessed to be a part of the dynamic community while pursuing the PhD, which shaped my perspective on science and research, entirely. I would also like to thank him for displaying patience, reassurance and support during difficult times which enabled the successful completion of the PhD.

The thesis would not have been possible without the guidance, support and assistance from the two mentors I had during the course of the PhD - *Dr. Björn Schwalb* and *Dr. Michael Lidschreiber*. Their combined contribution at different stages of the PhD life formed the foundation of the thesis. During the early stage of the PhD, *Dr Björn Schwalb* guided me through the fascinating world of transcription kinetics and data analysis. In the later stages of the PhD life, *Dr. Michael Lidschreiber* assisted me to navigate through grueling times. I am indebted to him for encouraging me to develop and explore my own ideas along with his insightful remarks for my betterment, always. He guided me immensely throughout the learning process. I sincerely appreciate his timely response to all of the questions. I must acknowledge both their mentoring, guidance, motivation and most of all, limitless patience.

I would like to thank the thesis committee members *Prof. Dr. Axel Munk* and *Dr. Melina Schuh* for keeping track of my progress and their support.

I would like to express my gratitude to *Prof. Dr. Gregor Eichele*, *Prof. Dr. Matthias Dobbelstein* and *Dr. Ufuk Güneşdogan* for their interest in my work and agreeing to be a part of the PhD examination board.

The work presented in the thesis are outcomes of great internal and external collaborations. I would like to first express my gratitude to all the laboratory internal collaborators. A large part of the thesis is culminated with the collaboration of an excellent PhD student in the laboratory *Taras Velychko*. He enriched the project with conducting the experiments, his knowledge and comprehension of the dataset. Other internal collaborators include *Dr. Kristina Žumer* who taught me to think critically, *Dr. Livia Caizzi* taught me optimism and *Dr. Sara Osman* taught me the virtue of patience during our fruitful collaborations all of which

culminated into published manuscripts. A special mention to two laboratory rotation students, *Artem Babych* and *Marcel Werner* who assisted with the data analysis and experiments during their rotation.

I would like to thank the external collaborators, *Dr. Edith Heard* and her lab members, *Dr. Shona Murphy* and her lab members for the excellent collaborations and their valuable insight.

I would like to thank *Dr. Kerstin Meyer* and *Dr. Petra Rus* for their great support in the sequencing facility, which is the basis of all the bioinformatic data analysis.

I am thankful to IMPRS-GS and the present coordinators *Dr. Henriette Irmer, Frauke Bergmann* and past coordinator *Dr. Katja Lidschreiber* for their support during the PhD study.

I am thankful to *Janine Blümel*, *Kirsten Backs* and *Almuth Burgdorf* for taking care of all the administrative issues as well as making my life easier by helping me with adapting to a new environment with new language in a new city, country and continent. I would also like to mention *Mario Klein* for his skilled technical support in managing the resources whenever needed.

Thanks to all present and past members of the @CramerLaboratory, especially the experimental and computational functional genomics people with whom I have an acquaintance with. They enriched me with perspectives of different aspects of life, a gesture I genuinely appreciate.

On a personal note, I would like to express my immense gratitude to *Dr. Henriette Irmer*. Her unconditional support during difficult times and unmatchable kindness and compassion were instrumental for the completion of the PhD. I owe a special mention of thanks to *Dr. Katja Lidschreiber and Dr. Kerstin Meyer* for the encouragement, affection and positivity that helped me overcome my difficulties to a greater extent.

I would like to acknowledge all my past mentors who contributed to my ever-growing passion for science. A special acknowledgement to my friends and well-wishers across the world who happen to be one of my strongest support systems. Thanks to my friends for being there for me, for all the timely support and enthusiasm and for always trying to keep my spirits up.

Last but not the least, I wish to thank my family, and in particular my parents, for raising me to be what I am today, with all my characteristics and peculiarities, and for fully supporting and guiding me through the ongoing process of growing-up. Thanks to all my particularly splendid sibling's. Words cannot truly describe my appreciation for the prayers, blessings, love and support my family has given me over the years.

SUMMARY

Transcriptional regulation is a highly dynamic biological process which is governed by a complex network of proteins for precise expression of a gene. One of the most important regulators of gene expression are cyclin dependent kinases (CDKs), which transduces signal by phosphorylation of its substrate. CDKs are broadly classified into two groups – cell cycle CDKs and transcriptional CDKs(tCDKs). Due to limited structural and functional knowledge of tCDKs, till date it remains to be a fascinating field of research. One of the early functions discovered for tCDKs is the phosphorylation of the C-terminal domain (CTD) of RNA polymerase II (Pol II), but the knowledge remains poorly understood *in vivo* till date. This phosphorylation cycle of Pol II CTD is crucial to co-ordinate between different transcriptional processes namely initiation, pause release, elongation, as well as termination of transcription. Additionally, tCDKs can also interact with other proteins in transcription cycle for orchestrating regulation intertwined with chromatin accessibility and epigenetic mechanisms. Since the discovery of tCDKs, scientists have been trying to unravel the functional mystery of the individual kinases in transcriptional cycle regulation. tCDKs are structurally and functionally conserved, therefore dissecting the role of individual kinases is a challenging task. This thesis work presents a collaborative effort to piece together the transcriptional kinase puzzle by defining the primary role of individual tCDKs in transcription cycle. This will eventually lead to uncover the detailed molecular mechanism of tCDKs and their impact on transcriptional regulation. Research focused on tCDKs for functional interpretation are mostly done with chemical inhibitors e.g., THZ1 which has off target effects as well as longer inhibition duration introduces secondary effects. To circumvent these limitations, the studies presented in the thesis uses an analog sensitive kinase approach as a powerful tool to study the function of individual kinases with specific inhibition *in vivo*. Restricting the inhibition duration to a shorter time span allows to investigate the primary function of the individual kinases. Precisely, we investigated the function of three different kinases, CDK7, CDK12 in human cell line and CDK8 in yeast. To obtain novel insights into the underlying mechanism, we combined complementary functional genomic methods, for example, Transient Transcriptome Sequencing (TT-seq) and mammalian Native Elongating Transcript Sequencing (mNET-seq) with proteomics and structural studies which gives us promising results to determine the exact functionality of the individual tCDKs. This broadens our knowledge in the field of tCDK mediated transcriptional regulation and its involvement in all of the relevant biological processes. More careful analysis needs to be done to interpret how kinases interact under physiological conditions, but the work presented in this thesis significantly contributed to the understanding of individual tCDKs function *in vivo*. In recent times, tCDKs have been implicated in cancer and their inhibition in cancer therapy and presence as biomarkers are being explored extensively. The knowledge of individual tCDK function will help us to manipulate the activity of these kinases to substantially increase the therapeutic potential.

PUBLICATIONS

Part of this work has been published:

2021 **The Cdk8 kinase module regulates Mediator-RNA polymerase II interaction**

Sara Osman, **Eusra Mohammad**, Michael Lidschreiber, Alexandra Stuetzer, Fanni Bazsó, Kerstin Maier, Henning Urlaub, Patrick Cramer

Journal of Biological Chemistry (JBC)
Volume 296
January-June 2021
Section: Gene Regulation

DOI: https://doi.org/10.1016/j.jbc.2021.100734

Author contributions: S.O. designed and performed experiments and evaluated data unless otherwise stated. A.S., F.L.B., and H.U. performed MS measurements. K.C.M. performed the 4tU-seq experiments. **E.M.** and M.L. performed the bioinformatics analysis of the 4tU-seq data. P.C. supervised research. S.O. and P.C. wrote the article with input from all authors.

2020 **CDK12 globally stimulates RNA polymerase II transcription elongation and carboxyl-terminal domain phosphorylation**

Michael Tellier, Justyna Zaborowska, Livia Caizzi, **Eusra Mohammad**, Taras Velychko, Björn Schwalb, Ivan Ferrer-Vicens, Daniel Blears, Takayuki Nojima, Patrick Cramer and Shona Murphy

Nucleic Acids Research
Volume 48, Issue 14, 20 August 2020, Pages 7712–7727
Section: GENE REGULATION, CHROMATIN AND EPIGENETICS

DOI: https://doi.org/10.1093/nar/gkaa514

Author contributions: M.T. and J.Z. carried out most of the experimental analyses, M.T. carried out all the bioinformatic analysis of the mNET-seq and ChIP-seq data, L.C. and T.V. performed TT-seq, **E.M.** and B.S. carried out bioinformatic analysis of TT-seq data. I.F-V. and D.B. aided some experimental analyses and T.N. provided support with mNET-seq. S.M. produced materials, carried out experimental analyses and supervised most of the research. P.C. supervised the TT-seq analyses and helped with data interpretation. S.M., M.T. and J.Z. wrote the paper with contributions from all the authors.

✓ If applicable, a list of contributions can be found at the beginning of each subsection or paragraph in Chapter: Methods and Chapter: Results and Discussions.

Contribution to other publication:

2019 **The Implication of Early Chromatin Changes in X Chromosome Inactivation**

Jan Jakub Żylicz, Aurélie Bousard, Kristina Žumer, Francois Dossin, **Eusra Mohammad**, Simão Teixeira da Rocha, Björn Schwalb, Laurène Syx, Florent Dingli, Damarys Loew, Patrick Cramer, Edith Heard

Cell.
Volume 176, Issues 1-2, 10 January 2019, Pages 182-197.e23
Section: Articles

DOI: 10.1016/j.cell.2018.11.041

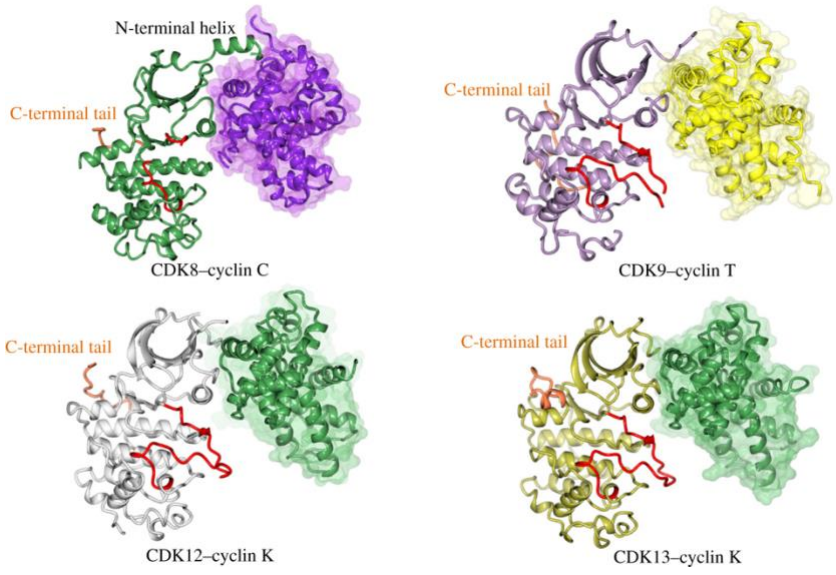
Author contributions: Conceptualization, J.J.Z. and E.H.; Methodology, J.J.Z. and F. Dossin; Software, A.B., B.S., **E.M.**, and L.S.; Validation, J.J.Z.; Formal Analysis, A.B., B.S., **E.M.**, L.S., F. Dingli, and D.L.; Investigation, J.J.Z., K.Z., S.T.R., and F. Dossin.; Data Curation, A.B., B.S., **E.M.**, L.S., F. Dingli, and D.L.; Writing – Original Draft, J.J.Z. and E.H.; Writing – Review & Editing, J.J.Z., E.H., A.B., B.S., **E.M.**, K.Z., S.T.R., and P.C.; Visualization, J.J.Z., A.B., and **E.M.**; Supervision, E.H. and P.C.; Project Administration, J.J.Z. and E.H.; Funding Acquisition, E.H., J.J.Z., and P.C.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

The aim of this chapter is to provide the reader with the relevant biological concepts for comprehension of the thesis. This chapter is written with sources from published review articles. Each subsection is wrapped up with the relevant literature information.



1.1 Central dogma of molecular biology

The eukaryotic cell nucleus contains the genetic blueprint of an organism in the form of deoxyribonucleic acid (DNA) organized in chromosomes (Watson and Crick, 1953). The genetic information encoded in the double helix DNA sequence is transferred into proteins through the ribonucleic acid (RNA) intermediate. This flow of genetic information is referred to as the central dogma of molecular biology **(Figure 1.1)**, a phrase first enunciated by Francis Crick in 1958 (Crick, 1970). The underlying biological processes, namely DNA replication, transcription into RNA, and translation into proteins ensure the transfer in a deterministic manner and are crucial for the development and function of all living organisms (Crick, 1970).

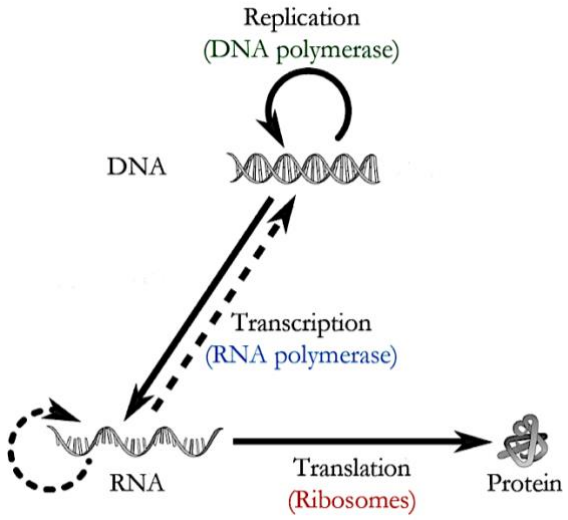


Figure 1.1 | The central dogma of molecular biology. The flow of genetic information along the three important classes of biopolymers: DNA, RNA and protein. DNA is transcribed (copied) to mRNA which is translated to proteins using amino acids. Adapted from (Crick, 1970).

RNA, one of the three important biopolymers, is a highly versatile compound. It functions in the process of translation as rRNA and tRNA, contributes to the gene expression regulation as eRNA, miRNA, siRNA and other non-coding RNA, or the mRNA acts as a pure messenger. The underlying mechanisms, with the extensive and crucial roles they play in a biological system, are of high interest for their association with diseases (Li and Liu, 2019; Statello *et al.*, 2020; Bhatti *et al.*, 2021).

1.2 Gene transcription in eukaryotes

The first step in the flow of the genetic information pathway is transcription of genes from DNA template to produce complementary single-stranded RNA (Crick, 1970). The enzyme that carries out transcription is the DNA dependent RNA polymerase (Pol) (Hurwitz, 2005). While bacteria and archaea have only one type of RNA polymerase, eukaryotic cells have

five different nuclear RNA polymerases that perform transcription of the nuclear genome: Pol I, II, III, IV and V (Hurwitz, 2005). Different RNA polymerases synthesize functionally distinct transcripts. Pol I synthesize ribosomal RNAs (rRNAs). Pol II produces protein-coding RNAs (mRNA) as well as small nuclear (snRNAs) and small nucleolar RNAs (snoRNAs). Pol III is responsible for the synthesis of transfer RNAs (tRNAs), the 5S rRNA and other small RNA molecules (Carter and Drouin, 2009). The plant-specific enzymes Pol IV and Pol V produce small interfering RNAs (siRNAs) that are involved in gene silencing (Zhou and Law, 2015). All polymerases are structurally related and consist of a conserved 10-subunit core and specific additional subunits (Cramer, 2019a; Barba-Aliaga, Alepuz and Pérez-Ortín, 2021).

1.3 RNA polymerase II (Pol II)

The synthesis of eukaryotic mRNA is carried out by RNA polymerase II (Pol II) (Kornberg, 1999; Cramer, Bushnell and Kornberg, 2001). Pol II, is a multi-subunit complex composed of 12 subunits, termed Rpb1-12 and is the only polymerase containing a flexible C-terminal domain (CTD) on its largest subunit Rpb1 (Corden *et al.*, 1985). CTD consists of multiple, highly conserved heptapeptide tandem repeats. While the CTD of the yeast *S. cerevisiae* contains 26 heptapeptide repeats, the human *Homo sapiens* CTD is composed of 52 repeats **(Figure 1.2)** (Stiller and Hall, 2002). Control of Pol II activity is specifically regulated at individual genes and is absolutely crucial for homeostasis of the cell (Osman and Cramer, 2020; Schier and Taatjes, 2020).

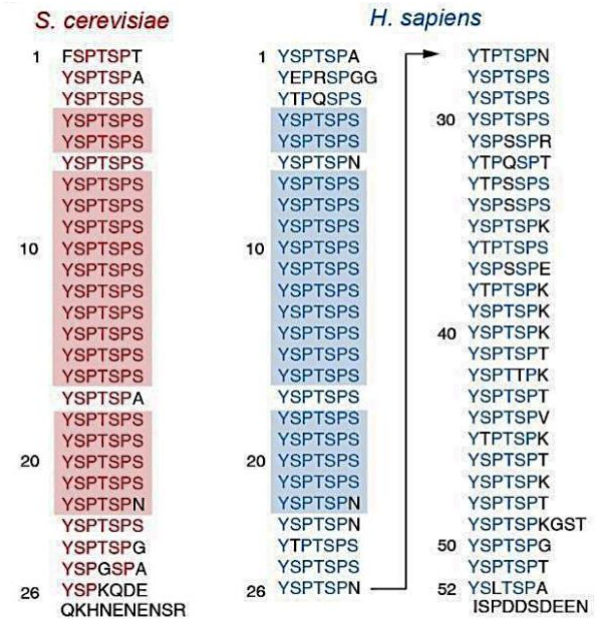


Figure 1.2 | The Pol II C-terminal domain. Comparison of yeast and human Pol II CTD sequences. Adapted from (Boehning *et al.*, 2018).

1.4 The RNA polymerase II (Pol II) transcription cycle

A transcription cycle is a collective set of biochemical reactions that control RNA polymerase activity, from promoter binding to polymerase recycling, at every active gene in any genome (Kang *et al.*, 2020). The transcription cycle integrates multiple sources of information serving as a command center to ensure that RNA synthesis across genomic loci is tailored precisely to the needs of the cell and organism. In eukaryotic organism, transcription by Pol II is orchestrates in three major parts, namely initiation, elongation, termination and re-initiation **(Figure 1.3)** (Hantsche and Cramer, 2016).

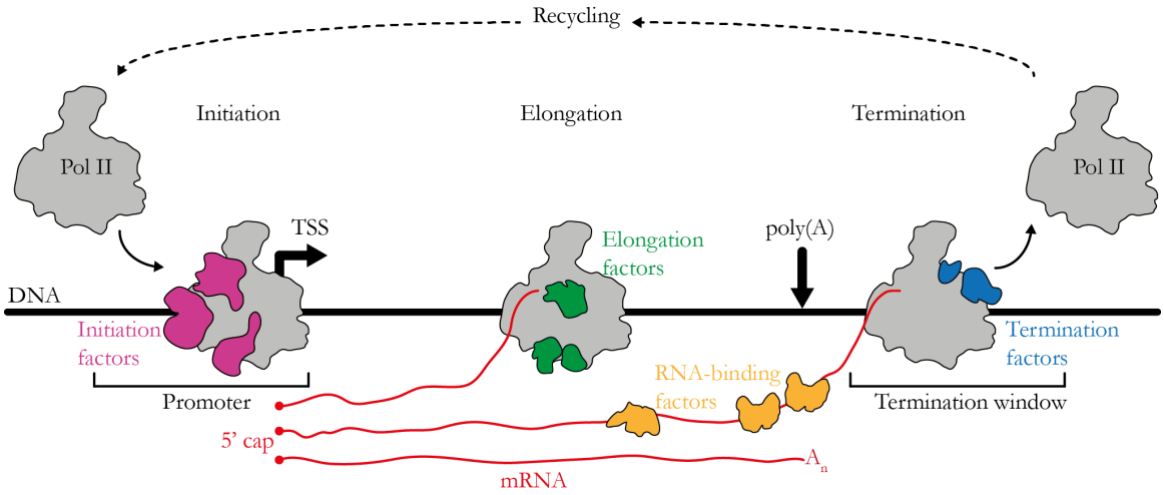


Figure 1.3 | Pol II transcription cycle. During transcription initiation, Pol II binds the promoter of a gene close to the transcription start site (TSS) with the help of general transcription factors. During elongation, the nascent mRNA chain (red) is extended. The polyadenylation (poly(A)) site marks the end of the gene where the mRNA is cleaved. Further downstream, Pol II is displaced from the DNA template and freed for a new round of transcription. The mRNA is co-transcriptionally processed through addition of a 5' cap (red dot) and a poly(A) tail (A_n) at the 3' end. The Pol II C-terminal domain is omitted in the structure for clarity. Adapted from (Hantsche and Cramer, 2016).

1.4.1 Initiation

Pol II transcribes through a chromatin environment and regulation of chromatin structure is linked to Pol II initiation (Kulaeva, Gaykalova and Studitsky, 2007). Transcription initiation requires access of Pol II to promoter DNA at the transcription start site of a gene (Nikolov and Burley, 1997). *In vivo*, DNA is compacted through binding to octameric histone complexes like "beads on a string" (Woodcock, Safer and Stanchfield, 1976). Each histone octamer is composed of two histone H2A/H2B dimers and one histone H3/H4 tetramer, which together with 147 bp of wrapped DNA form the nucleosome, the fundamental building block of chromatin (Luger *et al.*, 1997). The accessibility of the DNA template is strongly impaired by this packaging structure of DNA or chromatinization, and chromatin remodelers, chromatin-modifying enzymes, coactivators and mediator are recruited to

remodel the chromatin architecture to ensure accessibility (Zhang *et al.*, 2016). In active transcription units, promoters are generally located in nucleosome-depleted regions to provide access for the transcriptional machinery to the DNA (Kujirai and Kurumizaka, 2020).

Transcription initiation requires the assembly of the pre-initiation complex (PIC) comprising the general transcription factors (GTFs) TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH together with Pol II at the core promoter (Petrenko *et al.*, 2019). GTFs position Pol II on the promoter for transcription start site (TSS) selection to facilitate transcription initiation (Osman and Cramer, 2020). TFIIH plays a key role in this process - it opens the double-stranded promoter DNA at the TSS through its ATP-dependent translocase activity ('open PIC') which allows the translocation of the template strand into the Pol II active site, and promotes the polymerization of a complementary RNA strand via a conserved catalytic mechanism ('initially transcribing complex') (Steitz and Steitz, 1993; Brueckner, Ortiz and Cramer, 2009). Once the transcript exceeds a critical length, the growing RNA chain clashes with TFIIB, strongly destabilizing the PIC (Pal, Ponticelli and Luse, 2005). Simultaneously, the trimeric TFIIH kinase module containing the cyclin dependent kinase(CDK) 7 kinase (Kin28 in yeast) phosphorylates the Pol II CTD at the heptad positions Ser5 and Ser7, which further facilitates PIC disassembly and promoter escape (Hantsche and Cramer, 2016; Osman and Cramer, 2020; Schier and Taatjes, 2020).

1.4.2 Elongation and CTD phosphorylation

The transition from transcription initiation to early elongation happens when Pol II gets stable by the growing DNA-RNA hybrid and the initiation factors dissociate (Nechaev and Adelman, 2011). The transition is known as promoter clearance or promoter escape (Jonkers, Kwak and Lis, 2014). If not terminated by abortive initiation, Pol II enters into the productive elongation phase to form a processive transcription elongation complex (Nechaev and Adelman, 2011) . Upon promoter escape, Ser5-phosphorylated CTD recruits the capping enzyme to modify the 5'-end of the nascent mRNA with a stabilizing methylated guanosine nucleotide (Cho *et al.*, 1997). Thus, the nascent mRNA receives a 5'- cap when it reaches a length of 25-30 nucleotides (Ramanathan, Robb and Chan, 2016). In metazoans, the Pol II elongation complex temporarily pauses in the promoter-proximal region ~50 nt downstream of the TSS, representing a regulatory checkpoint for transcriptional control during elongation (Core and Adelman, 2019). Pol II pausing is stabilized by the negative elongation factor (NELF) and DRB sensitivity-inducing factor (DSIF) (Vos *et al.*, 2018). Pause release by the CDK9 kinase subunit of the positive elongation factor b (P-TEFb)

results in displacement of NELF by the elongation factor complex PAF, binding of the elongation factor SPT6, and Ser2-phosphorylation of the Pol II CTD (Vos *et al.*, 2018). While pause sites are generally located upstream of the first (+1) nucleosome, further transcription of the gene body necessitates Pol II passage through nucleosomes (Teves, Weber and Henikoff, 2014). Recruitment of positive elongation factors as the histone chaperone SPT6 to the Ser2/Ser5-hyperphosphorylated CTD can enable efficient nucleosome passage to facilitate elongation (Kasiliauskaite *et al.*, 2022). As transcription elongation proceeds through the gene body, the phosphorylated CTD coordinates co-transcriptional pre-mRNA maturation through the direct interaction with components of the splicing apparatus (Hantsche and Cramer, 2016; Cramer, 2019b; Osman and Cramer, 2020; Schier and Taatjes, 2020).

1.4.3 Termination and re-initiation

Termination of Pol II transcription, the final phase of the transcription cycle is reached when Pol II reaches the poly (A) site (pA site) of the transcribed gene (Tran *et al.*, 2001). The elongation factors are then replaced by termination factors (Kecman *et al.*, 2018). At the pA site the transcript is cleaved off, Pol II transcribes a little further downstream and is finally terminated (Proudfoot, 2016). During these steps, the mRNA receives a 3'-poly(A) tail (Wahle and Rüegsegger, 1999). After termination, transcription can be initiated again either by recruitment of the complete transcription machinery to the promoter region or by facilitated re-initiation due to the promoter bound scaffold complex (Kang *et al.*, 2020). This remainder of the initial transcription machinery, including TFIIA, TFIID, TFIIE, TFIIH and mediator, enables rapid PIC formation and thus efficient successive initiation (Shino and Takada, 2021). The underlying processes of termination are highly regulated and coupled to RNA processing events, which often occur far downstream of the poly (A) site (Cramer, 2019b; Osman and Cramer, 2020; Schier and Taatjes, 2020).

1.5 Regulation of the Pol II transcription cycle by CTD

The Pol II CTD is necessary for transcription cycle regulation (Hsin and Manley, 2012). Each of the seven amino acids in the highly conserved heptad peptide ($Y_1S_2P_3T_4S_5P_6S_7$) repeat of CTD may undergo different modifications and each repeat may have a different post-translational modification pattern (Zaborowska, Egloff and Murphy, 2016a). Therefore, the CTD potentially exhibits a large and complex pattern collectively called the CTD code **(Figure 1.4)** (Aristizabal and Kobor, 2016). The CTD code, in particular the levels of Ser2P, Ser5P, and Ser7P during three stages of transcription, is critical for transcriptional control (Kim *et al.*, 2010). In general, before transcription starts, Ser2, Ser5,

and Ser7 are all unphosphorylated, and the initiation of transcription requires Ser5P and Ser7P (Calvo and García, 2012). However, the Ser5P and Ser7P level declines when gene transcription enters the elongation stage (Z. L. Zheng, 2022). Ser2P levels increase during productive elongation, and a full-length RNA transcript is synthesized by a stable elongation complex (Zhou, Li and Price, 2012). All CTD phosphorylation decrease at the termination stage for Pol II to enter another transcriptional cycle (Z. L. Zheng, 2022). This can be referred to as the CTD phosphorylation cycle which is associated with the transcription cycle (Hsin and Manley, 2012). The dynamic CTD phosphorylation pattern during transcription is tightly regulated by various CDKs and CTD phosphatases (Zaborowska, Egloff and Murphy, 2016a). The CTD also functions as a general platform to recruit many regulatory factors involved in transcription, mRNA processing and histone modification (Heidemann *et al.*, 2013; Zaborowska, Egloff and Murphy, 2016c).

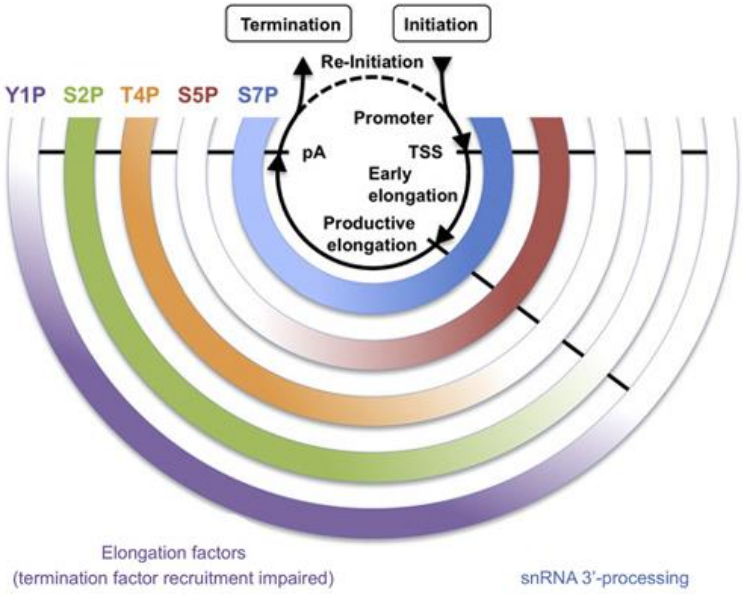


Figure 1.4 | Extended "CTD code" for transcription cycle coordination. During the cycle, levels of CTD phosphorylation at Tyr1, Ser2, Thr4, Ser5 and Ser7 residues change differently, as illustrated by gradients of violet (Tyr1P), green (Ser2P), orange (Thr4P), red (Ser5P) and blue (Ser7P) in five semicircles. Upon PIC assembly, Ser5 and Ser7 residues are phosphorylated next to the transcription start site (TSS). Ser5P facilitates promoter escape as well as recruitment of the capping enzyme. While Ser5P levels are diminished towards the 3' end, Ser2P, Thr4P and Tyr1P levels increase during elongation. Before the polyadenylation (pA) site, Tyr1P is erased, while Ser2P levels remain high and leads to mRNA 3' end processing and termination of transcription. After dephosphorylation of CTD, RNAPII is released and can reenter another round of transcription. Adapted from (Heidemann *et al.*, 2013).

1.6 Transcription-associated cyclin dependent kinases (tCDKs)

Cyclin dependent kinases (CDKs) are a family of approximately 20 serine/threonine kinases that are engaged in fundamental cellular processes such as cell cycle regulation, transcription,

epigenetic regulation, neuronal functions, metabolism, hematopoiesis, angiogenesis, DNA damage and repair, stem cell self-renewal, proteolysis and spermatogenesis (Malumbres, 2014a). They are broadly divided into two major subclasses: (i) cell cycle–associated CDKs (including CDK1, CDK2, CDK4, and CDK6) that directly regulate progression through the phases of the cell cycle and (ii) transcription-associated CDKs (including CDK7, CDK8, CDK9, CDK12, and CDK13) that regulate gene transcription by phosphorylating the CTD of Pol II as well as other targets (Espinosa, 2019a).

1.6.1 Evolution of transcription-associated CDKs (tCDKS)

CDKs in yeast and human have evolved in a structurally and functionally conserved manner. Kin28, Srb10, Bur1 and Ctk1 are the yeast orthologs of CDK7, CDK8, CDK9 and CDK12, respectively **(Figure 1.5)** (Malumbres, 2014a). The CDK20 and CDK10/CDK11 subfamilies are not represented in yeast. Orthologs of CDKs in yeast and human have almost identical Ser-specificity (Malumbres, 2014b; Z. Zheng, 2022).

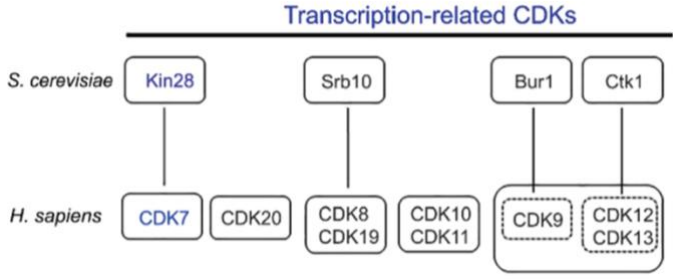


Figure 1.5 | Transcriptional CDKs from yeast (*S. cerevisiae*) and human (*H. sapiens*). The vertical line indicates an orthologous relationship. CDKs in color represent the two major regulators that are involved in cell cycle and transcriptional control. Adapted from (Zheng, 2022).

1.6.2 Activation of transcription-associated CDKs (tCDKS)

Each of the tCDKs binds to a specific activating cyclin partner to regulate gene transcription, which directs the activity of the tCDK **(Figure 1.6)**.

CDK7 binds to cyclin H and the accessory protein MAT1 to function as a CDK-activating kinase (CAK) (Peissert *et al.*, 2020). CDK8 and its paralog CDK19, associates with cyclin C, MED12, and MED13 to form a complex known as the CDK8/CDK19 module which can associate with the mediator complex, a multimeric transcriptional coactivator complex (Hoeppner, Baumli and Cramer, 2005). CDK9 binds to either cyclin T or cyclin K for its kinase activity. The CDK9 complex with cyclin T, referred to as the positive transcription elongation factor b (P-TEFb), is a general transcription factor (GTF) (Lin *et al.*, 2002).

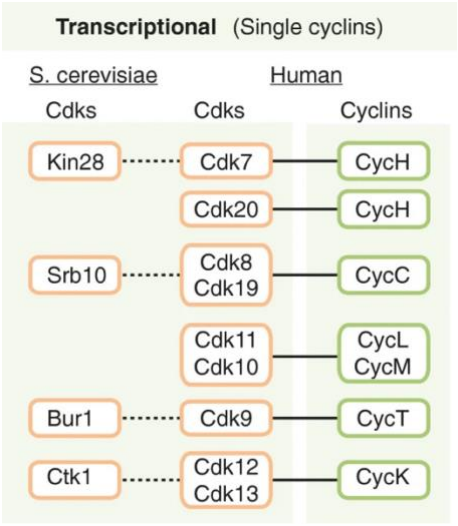


Figure 1.6 | Brief description of yeast and mammalian transcription-associated CDKs and corresponding cyclins for mammalian tCDKs. Adapted from (Malumbres, 2014b).

Amino-terminal lobe of the CDK12 and CDK13 kinase domain interfaces with the cyclin box of cyclin K to create heterodimers, which then further dimerize (Kohoutek and Blazek, 2012). CDK10 associates with Cyclin M as the cyclin partner and CDK11 associates with L-type cyclins with presumed roles in transcription. tCDKs share a core set of features and activation requires phosphorylation of their 'activating T-loops' by a CDK activating kinase (CAK) (Malumbres, 2014b; Z. Zheng, 2022).

1.6.3 Regulation of the Pol II transcription cycle by tCDK mediated CTD phosphorylation

Pol II transcription cycle are orchestrated by the enzymatic activity of transcription-associated CDKs, all of which cooperate to guide the Pol II through the nucleosome (Hsin and Manley, 2012). Functionally, transcription-associated CDKs and their cyclin partners control phosphorylation of the Pol II CTD (Z. L. Zheng, 2022). The pattern of phosphorylation of the Pol II CTD dictates the transition between the initiation, elongation, and termination phases of transcription (Buratowski, 2009). Yeast Kin28 and human CDK7 are subunits of transcription factor TFIIH, that preferentially phosphorylates Ser5 and Ser7 of the Pol II CTD (Rimel and Taatjes, 2018). These marks predominate on Pol II transcribing the promoter-proximal and upstream regions of genes, implicating functions early in the transcription cycle during initiation and promoter clearance. Following initiation, Pol II enters transcriptional pausing, during which NELF and DSIF are loaded onto RNA Pol II. The P-TEFb/CDK9 complex is then recruited to paused Pol II where CDK9 phosphorylates the CTD in Ser2, as well as DSIF and NELF, to release RNA Pol II for productive elongation. This cascade of events controls the switch from transcriptional

initiation to elongation of RNA Pol II. In addition to CDK7 and CDK9, CDK12 and CDK13 also preferentially phosphorylate the CTD at Ser2 and Ser5 when it is prephosphorylated at Ser7. CDK9 an ortholog of Bur1, also contributes to phosphorylation of the Ser2 mark at the 5' ends of genes **(Figure 1.7)**. Transcript termination results in dephosphorylation of Pol II, making it ready for another round of re-initiation (Zaborowska, Egloff and Murphy, 2016b; Chou *et al.*, 2020).

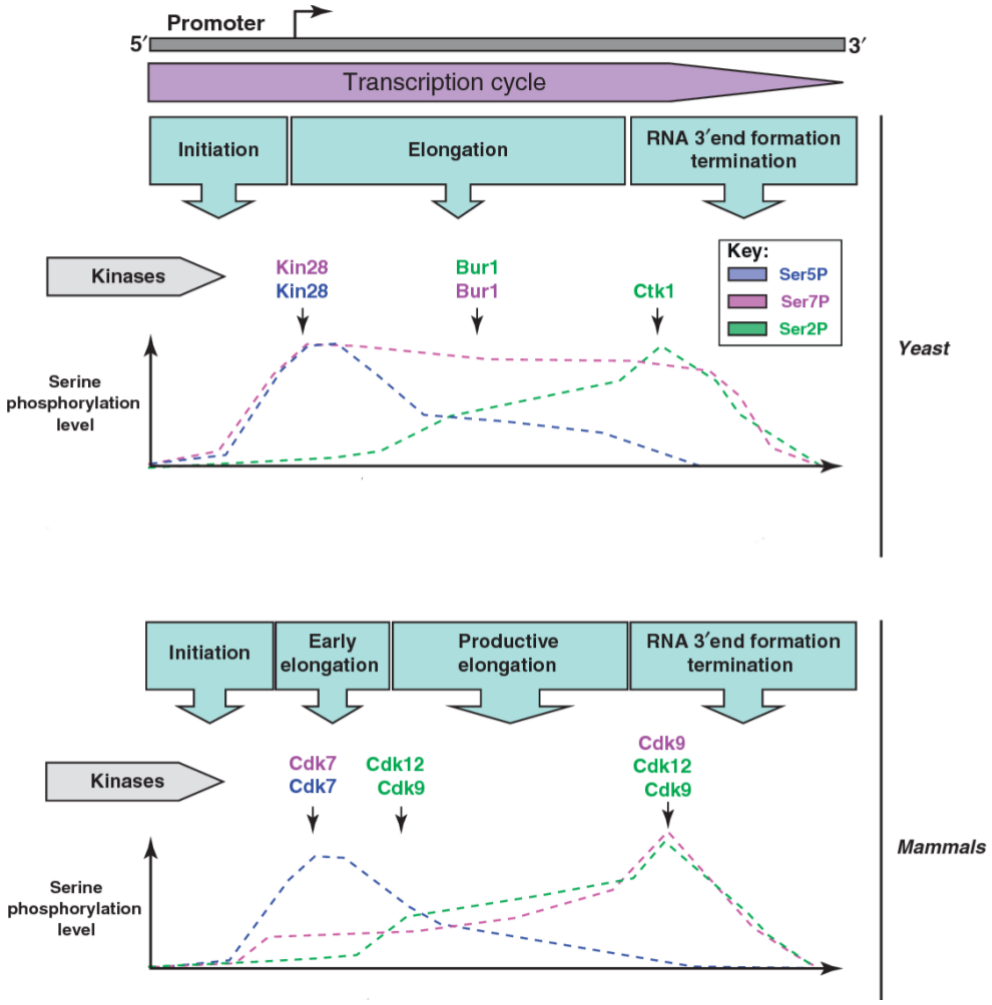


Figure 1.7 | Transcriptional CDKs contribute to the CTD phosphorylation pattern over the course of transcription. Adapted from (Egloff, Dienstbier and Murphy, 2012).

1.6.4 Regulation of the Pol II transcription cycle by tCDK beyond CTD phosphorylation

Transcription-associated CDKs has transcriptional targets apart from the Pol II CTD to regulate the transcription cycle, which are not completely understood. These targets have been shown to regulate splicing, intronic polyadenylation, genomic stability, epigenetic modifications, and translation (Bury *et al.*, 2021; Łukasik, Załuski and Gutowska, 2021).

Thus, transcription-associated CDKs function throughout the transcription cycle **(Figure 1.8**).

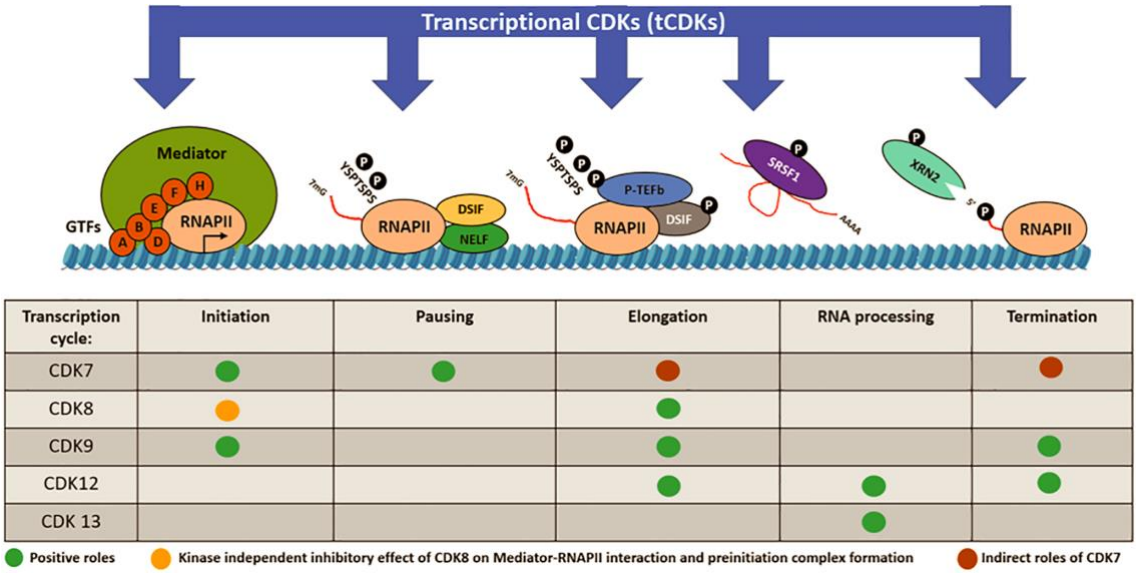


Figure 1.8 | Chief roles of the transcription-associated CDKs span throughout the Pol II transcription cycle. Adapted from (Galbraith, Bender and Espinosa, 2018; Sundar *et al.*, 2021).

CDK7 is able to phosphorylate and activate other CDKs, thus acting as a CDK-activating kinase (CAK) . Kin28 does not have this activity, which is mediated in yeast by a different kinase unrelated to CDKs, Cak1. As a CAK, CDK7 functions are executed throughout the Pol II transcription cycle, starting from promoter clearance, 5'-end capping of the nascent transcript and promoter-proximal pausing extending to mRNA 3´-end formation and termination. CDK7 and Kin28 also plays role in dissociating Pol II from the mediator complex that bridges the PIC and upstream regulatory elements (Fisher, 2019).

CDK9 along with cyclin T phosphorylates the DRB sensitivity–inducing factor (DSIF) and negative elongation factor (NELF), to relieve promoter pausing and promote transcription elongation. CDK9 can be described as a "signaling hub" for transcriptional control for the roles in gene and enhancer transcription, RNA processing, chromatin regulation (Anshabo *et al.*, 2021).

CDK8 and CDK19 are the mediator-associated kinases with roles in enhancer-promoter communication, transcriptional memory, metabolism, and, in the case of CDK19, kinase independent roles in transcriptional control. CDK8 associated with the mediator complex can transmits signals from transcription factors to RNA Pol II. The CDK8/Cyclin C complex also targets CDK7/cyclin H via phosphorylation, repressing the ability of TFIIH to activate transcription and its CTD kinase activity. CDK8 has also been described to

restrain activation of super enhancers, affecting global gene expression in a cell type–specific fashion. CDK19 can also associate with the mediator complex to regulate transcription in a gene-specific manner. CDK8 and CDK19 phosphorylate other proteins associated with chromatin modification, DNA repair, and transcription (Sundar *et al.*, 2021; Luyties and Taatjes, 2022).

Despite structural similarities between CDK12 and CDK13, each seems to regulate the expression of a distinct set of genes implying shared but nonoverlapping functions between CDK12 and CDK13. These tCDKs are perhaps the least understood members of this family having roles in mRNA processing and genome stability. In addition, although both CDK12 and CDK13 directly interact with splicing machinery, they affect the processing of different sets of genes and noncoding RNAs (Greenleaf, 2019).

Amongst other transcription-associated CDKs, CDK10 has been implicated as a tumor-suppressive kinase in estrogen-driven cancers. CDK11 associates with the splicing machinery as well as proteins involved in transcriptional initiation and elongation and is highly expressed in triple-negative breast cancer (TNBC), multiple myeloma, and liposarcoma (Pilarova, Herudek and Blazek, 2020).

1.6.5 Transcription-associated CDKs as targets and biomarkers for cancer therapy

Several recent studies implicate the role of transcription-associated CDKs (tCDKs) in driving and maintaining cancer cell growth, particularly in cancers primarily driven by dysregulated transcription factors (Galbraith, Bender and Espinosa, 2019). The tCDKs are less developed as therapeutic targets, and small-molecule inhibitors of tCDKs have not yet entered routine clinical use (Ettl, Schulz and Bauer, 2022). Mounting evidence suggests that inhibiting tCDKs may have important therapeutic relevance (Galbraith, Bender and Espinosa, 2019). Given the critical roles that tCDKs play in regulating gene expression, they have emerged as putative targets in cancer and other diseases (Galbraith, Bender and Espinosa, 2019; Bury *et al.*, 2021; Łukasik, Załuski and Gutowska, 2021).

1.7 Methods to study transcription-associated kinases

Protein kinases can be removed by genetic knockout or by RNA interference-mediated downregulation (Agrawal *et al.*, 2003). Alternatively, the activity of kinases can be inhibited by chemical inhibitors of varying specificity (Klaeger *et al.*, 2017). Such inhibitors are of high therapeutic interest, as many kinases are involved in human cancer (Bhullar *et al.*, 2018). Most of these inhibitors are structurally similar to ATP and compete for binding in the active site with this nucleotide (Klaeger *et al.*, 2017). Since the active site structure has been well

conserved through evolution, it can be difficult to find drugs that are specific for one particular protein kinase (Eswaran and Knapp, 2010). For example, a covalent CDK7 inhibitor called THZ1 provides some selectivity, but THZ1 also has weak activity against CDK9, CDK12, and CDK13 (Sava *et al.*, 2020). In addition, the biological functions of non-essential protein kinases can be studied with protein kinase-dead mutants (Rauch *et al.*, 2011). The protein kinases whose activities are indispensable for cell are considered as essential and analysis of biological functions of these kinases remains challenging (Wilson *et al.*, 2018). Furthermore, continuous perturbation or condition specific inactivation can mask the exact functions of the analyzed protein kinases making substrate identification highly unreliable (Wilson *et al.*, 2018). Loss of particular protein kinase activity forces cells to substitute the phosphorylation of affected proteins by other protein kinases, thus hampering the identification of protein kinase substrates (Jurcik *et al.*, 2020).

The development of analog sensitive kinase technology advances the understanding for the phosphorylation of a substrate by a specific kinase *in vivo*. Analog sensitive (AS) kinase technology, a chemical-genetic technique was conceptualized by (Shokat *et al.*, 2000) that enables systematic generation of highly specific inhibitors for individual kinases by taking advantage of structural information **(Figure 1.9)**. This biochemical engineering approach comprises of two parts: engineering functional analog sensitive kinases and identifying ATP analogs as small-molecule inhibitors to inhibit the kinase activity (Michael S. Lopez, Kliegman and Shokat, 2014). The structurally conserved position in the active site (ATP binding pocket) of most kinases have an amino acid residue with a large bulky group (methionine, leucine, phenylalanine, threonine, etc.) at the back of the pocket (Zhang *et al.*, 2013). This residue is known as the gatekeeper (Zhang *et al.*, 2013). To engineer an analog sensitive kinase, the gatekeeper residue is mutated from the natural amino acid to a residue with a smaller side chain (glycine or alanine) (Garske *et al.*, 2011). The mutation creates a space within the ATP binding pocket not found in wild type (WT) kinases without compromising kinase activity (Krishnamurty and Maly, 2010). A bulky ATP-analog that complements the shape of the enlarged mutant ATP pocket can potently and specifically inhibit the target kinase activity (Michael S. Lopez, Kliegman and Shokat, 2014). The bulky ATP-analog cannot inhibit any wild type kinases due to steric hindrance (Michael S. Lopez, Kliegman and Shokat, 2014). AS kinase technology is a simplistic and general approach for probing kinase-signaling pathways with small molecule inhibitors. The application of this method to study kinases has elucidated the physiological function of individual kinases and

the transcriptional events controlled by Pol II phosphorylation (Bishop *et al.*, 2000; Blethrow *et al.*, 2004; Michael S Lopez, Kliegman and Shokat, 2014).

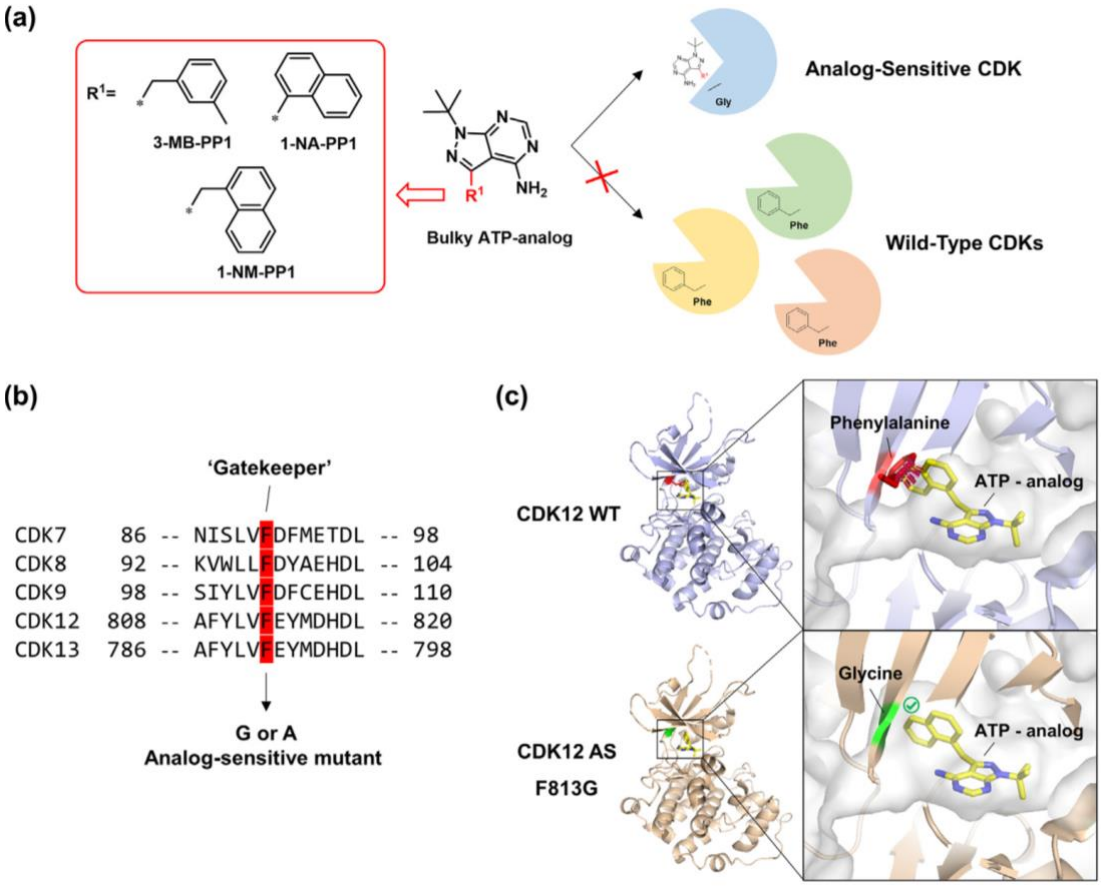


Figure 1.9 | Scheme of the analog sensitive kinase design method. (A) Chemical structure of bulky ATP analogs which selectively inhibit analog sensitive CDKs. **(B)** Sequence alignment shows highly conserved 'gate-keeper' phenylalanine residue among CDKs, mutated into glycine or alanine to create AS mutants. **(C)** Structural modeling of CDK12 WT (light blue ribbon) and AS mutant (wheat ribbon) illustrates how steric hindrance between the 'gate-keeper' residue and bulky ATP-analog (yellow stick) becomes a determinant for selective inhibition. Adapted from (Kim *et al.*, 2021).

1.8 Genome wide multi-omics approach to comprehend transcription-associated cyclin dependent kinase (tCDK) mediated transcriptional regulation

Transcriptional regulation by cyclin dependent kinase is a highly complex and dynamic mechanism comprising of a series of biochemical events governed by phosphorylation. The phosphorylation by CDKs synchronizes the transcription cycle which is interconnected with chromatin accessibility and other epigenetic mechanisms such as enhancer-promoter looping, which is necessary for a successful gene transcription. The additional complexity of transcriptional regulation is derived from different RNA maturation events, such as 5'-end capping, co-transcriptional splicing and polyadenylation as well as the involvement of non-coding RNAs (ncRNAs). The systematic operation of transcriptional regulation by cyclin dependent kinase can be explained by genome wide studies, rather than the necessary study

of individual genes and proteins, owing to the complexity (Casamassimi and Ciccodicola, 2019).

Most of the biological phenomena relating to transcriptional gene regulation can be measured genome wide using high-throughput experimental techniques. High-throughput methods aim to quantify or locate the genetic and epigenetic landscape that harbors the biological feature (expression, occupancy, etc.) of interest and is a prominent way to study biological function. Genome wide high-throughput techniques can answer complex biological questions and new variant of the existing techniques comes along in response to a new question. There are several genome wide high-throughput experiments coupled with massively parallel sequencing, known as next generation sequencing (NGS) in transcriptomics. A particular high throughput sequencing technology reveals a particular part of the complex transcriptional regulation. Thus, different high throughput sequencing techniques presents different ways to look at the underlying activities of the cells. For example, RNA-seq can measure global transcription expression profile, TT-seq can quantify newly synthesized RNA expression profile, transcription associated protein binding site can be detected by ChIP-seq whereas genome organization and contact can be revealed through Hi-C or Capture-C, and nucleosome occupancy profile can be determined from MNase-seq. NGS has completely revolutionized transcriptome analysis at an unprecedented level. These multi-dimensional transcriptomics NGS datasets and computational biology enables to integrate different layers of information from biophysical, biochemical, and molecular cell biology studies to untangle the complex transcriptional regulatory network (Churko *et al.*, 2013; Lightbody *et al.*, 2019; Mccord, Kaplan and Giorgetti, 2020; Prudêncio *et al.*, 2020; Furlan, De Pretis and Pelizzola, 2021). The complex regulatory networks can further be studied with kinetic model of transcription which gives us a quantitative output of transcriptional regulation (Gressel *et al.*, 2017; Gressel, Schwalb and Cramer, 2019; Caizzi *et al.*, 2021; Choi, Lee and Park, 2021; Mazzocca *et al.*, 2021; Žumer *et al.*, 2021; Shao *et al.*, 2022).

1.9 Rationale and aim of the thesis

Transcription-associated cyclin dependent kinases (tCDKs) play critical roles in regulating gene expression by phosphorylation at different phases during the transcription cycle. In principle, all these phases are rate limiting, adding an important layer to transcriptional regulation. Despite this critical importance, our understanding of the transcription cycle regulation by tCDKs is limited (Espinosa, 2019b). This lack of knowledge hampers our ability to manipulate transcriptional activity both in basic research and the applied sciences.

The main challenges remain to identify the relevant targets of individual tCDKs at each step of the transcription cycle which in turn provides us with a comprehensive understanding of the function. In addition, these tCDKs also have structural redundancies and they interact with each other to regulate transcription cycle. Discretely and as a family, the mechanism by which the tCDKs affect the transcriptional cycle remains to be fully elucidated.

This thesis aims to unravel the function of individual tCDKs in the transcription cycle. For this purpose, the concept of specific inhibition of tCDKs is exploited which would provoke the effects of the transcriptional deregulation. To eliminate the possibility of an off target effect by other kinases, analog sensitive (AS) kinase approach is used to generate analog sensitive kinases by CRISPR/Cas9 engineering which renders the specificity along with the possibility to study the primary effect of a particular kinase. AS kinase inhibition act on the upstream and downstream regulatory pathways of transcription which helps to uncover fundamental insights into the particular kinase mediated transcriptional regulation pathways. Combining the approach with genome wide high throughput sequencing techniques for multi-omics dataset for individual kinases helps to draw inferences which transcriptional process rely on which particular tCDKs in biological systems. This provides us with crucial and detailed understanding of the associated pathways and interactors and ultimately the mechanism of transcriptional regulation by tCDKs.

In this context, the main objective of this thesis is culminated as

- Identifying the primary functions and targets of the individual tCDKs by rapid, reversible, and highly specific inhibition of individual engineered kinases.
- Explore possible mechanisms of tCDKs in transcriptional cycle and thus better define the general landscape of kinase mediated transcriptional regulation.

The objective of the thesis is applied for three different tCDKs, CDK7 and CDK12 in human cells and CDK8 is yeast cell, to unravel the underlying mechanism of kinase mediated transcriptional regulation.

As transcription-associated CDKs are emerging as important targets and biomarkers in oncology, understanding the primary function of individual CDKs can be further exploited therapeutically.

CHAPTER 2

METHODS

The aim of this chapter is to familiarize the reader with the experimental setup and analysis performed to draw the inferences in this thesis. The chapter is divided into two sections.

- The first section contains general experimental setup, Next Generation Sequencing (NGS) analysis pipeline, kinetic parameters and mathematical model applied in the thesis.
- The second section contains details of project specific data analysis.

✓ Experiments were not performed by the author of this dissertation, but a general outline is included in this chapter for a comprehensive understanding of the data analysis pipeline.

✓ All the analysis were performed by the author of this dissertation unless otherwise stated. If applicable, a list of contributions can be found at the beginning of each subsection or paragraph.

SECTION ONE

2.1 Experimental design

2.1.1 Analog-sensitive kinase technology

2.1.1a CRISPR/Cas9 engineering of analog-sensitive kinase cell lines

The human analog-sensitive kinase (CDKas) cell lines were generated and validated by ***Dr. Shona Murphy (Murphy laboratory)*** using CRISPR/Cas9 engineering technology where the bulky amino acid Phenylalanine (F) was replaced with small amino acid Glycine (G) **(Figure 2.1).**

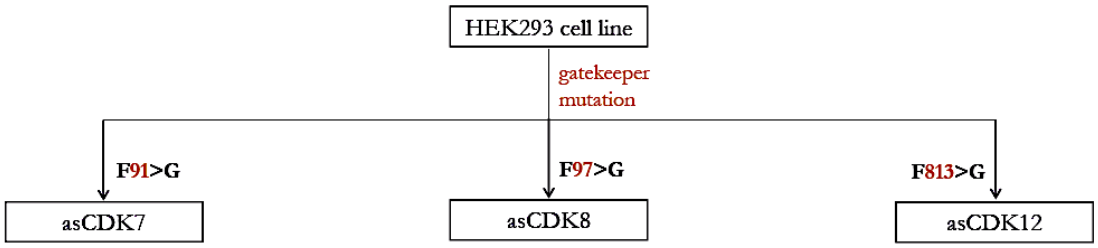


Figure 2.1 | CRISPR/Cas9 engineering to generate human analog-sensitive kinase (CDKas) cell lines.

The yeast CDK8 analog-sensitive kinase (CDKas) strain was provided by *Steven Hahn* (Liu et al., 2004) .

2.1.1b ATP analogs as small-molecule inhibitors to inhibit the kinase activity

The adenine analog 1-NM-PP1 was used for highly specific inhibition of individual engineered human analog-sensitive kinases and 1-NA-PP1 was used for yeast CDK8as (Srb10) inhibition.

2.1.2 Next-Generation Sequencing experiments

An overview of the NGS experiments included in this thesis is shown on **(Figure 2.2).**

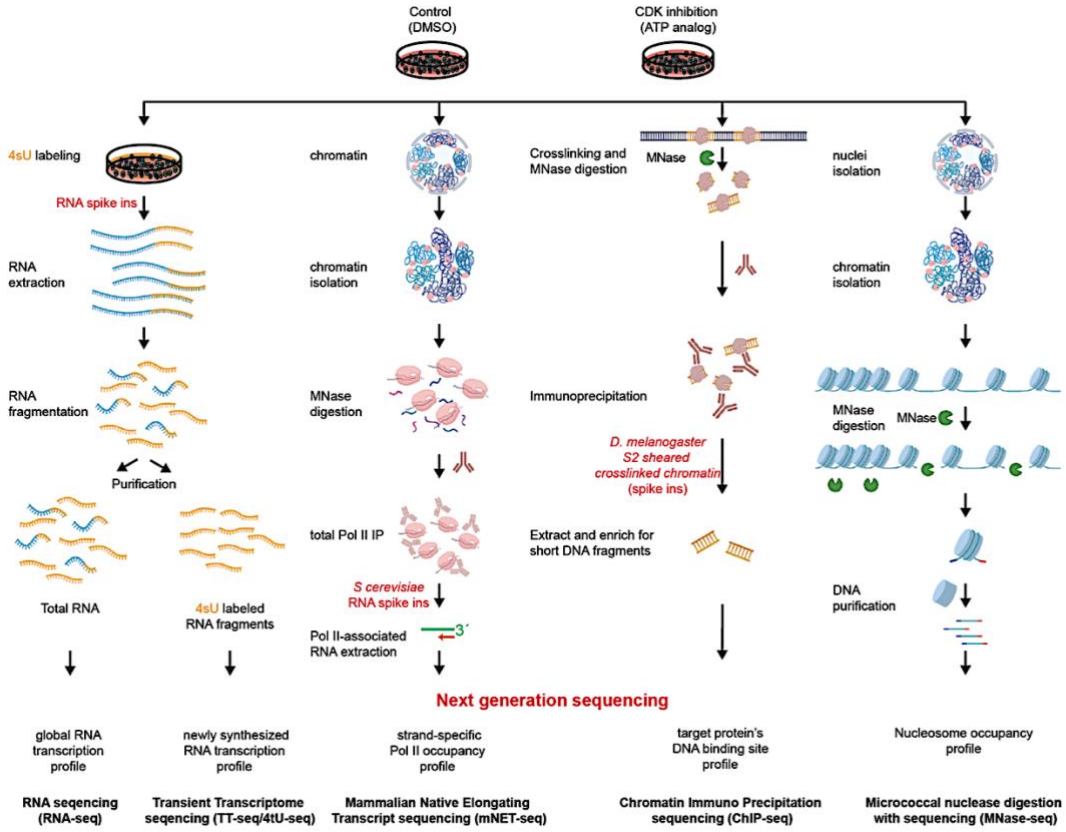


Figure 2.2 | An overview of the NGS experiments performed. RNA-seq, TT-seq, mNET-seq, ChIP-seq and MNase-seq were performed on CDKas cells subjected to control or inhibitor treatment.

2.2 Next-generation sequencing data analysis

Next Generation Sequencing produces enormous quantities of data that need careful bioinformatic analysis to extract the biological information on a genome-wide scale in a shorter time frame. A general analysis scheme for next-generation sequencing (NGS) dataset used in this thesis is shown on **(Figure 2.3).**

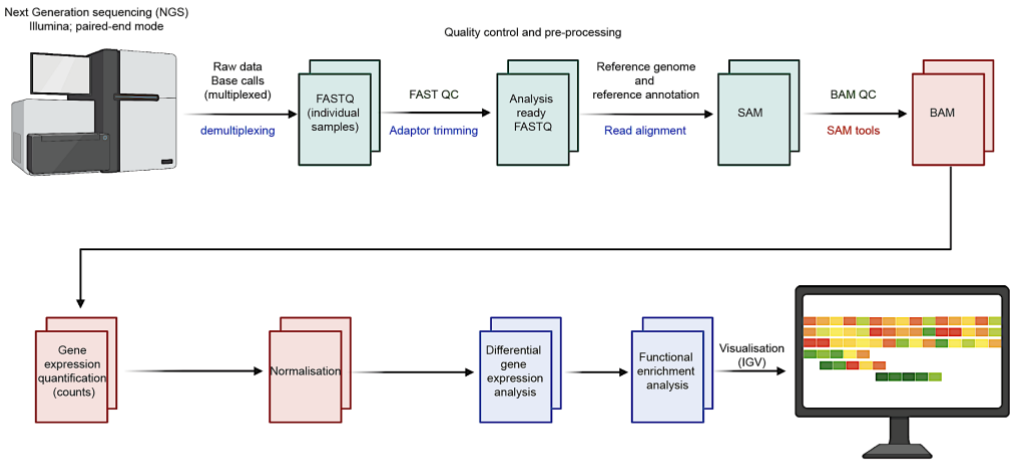


Figure 2.3 | A schematic overview of next-generation sequencing data analysis.

2.2.1 Sequencing platform: Illumina sequencing

All data presented in the thesis are sequenced in Illumina Next-seq 500 on paired-end mode unless otherwise stated.

Illumina sequencing was performed in-house by Dr. Kerstin Maier and Dr. Petra Rus (Max Planck Institute for Multidisciplinary Sciences, Department of Molecular Biology).

2.2.2 Sequencing data preprocessing

2.2.2a Demultiplexing

Sequencing raw data are demultiplexed with Illumina bcl2fastq demultiplexing software provided the barcodes for each of the samples.

2.2.2b Raw reads quality control and preprocessing

The paired-end reads generated in the sequencing process can be subjected to assess the quality to ensure good quality to begin with and increase reliability of downstream analysis. For this purpose, all raw sequencing data were quality checked with FastQC (Babraham-

Institute) which provides a general view on number and length of reads and quality of sequences (Andrews, 2010).

2.2.2c Trimming

To improve the initial quality of data to more accuracy, it is important to carry out pretreatment of data, such as trimming of low-quality bases, trimming of adapters, filter bases by quality scores, filter reads by lengths and filter duplicated reads. Cutadapt (Martin, 2011) is used to quality filter the sequencing data.

2.2.2d Read alignment

After checking the quality and necessary preprocessing of data and the next step is mapping, also called aligning, of reads against an annotated reference genome or reference transcriptome.

i. Reference genomes

Bioinformatics analysis pipelines rely on the use of a reference genome. For model organisms the reference genome can be downloaded. Several reference genome assembly and release version can be available for model organisms. For example, there are a couple of reference genomes for *Homo sapiens* e.g., *Homo sapiens* GRCh37 vs *Homo sapiens* GRCh38. The numbers correspond to versions (or "builds") of the reference genome. The higher the number, the more recent the version. To maintain consistency in analysis, the same genome version is used throughout an analysis.

ii. Reference genome annotation

Genome annotation is the process of identifying the locations of genes and all of the coding regions in a genome and assign the functions. The choice of a genome annotation has a big impact on the accuracy of downstream data analysis. On the basis of conception and hypothesis of an analysis, a reference annotation is chosen.

- **Reference genome annotation from databases**

Reference genome annotation can be downloaded from RefSeq (O'Leary et al., 2016) or GENCODE (Frankish et al., 2019). RefSeq is the oldest sequence database built by the National Center for Biotechnology Information (NCBI). GENCODE is the default gene annotation for the Ensembl project. The GENCODE Basic set is very similar to RefSeq but GENCODE Comprehensive set is richer in novel CDSs, novel exons, nonsense transcripts, such as long non-coding RNAs (lncRNAs), pseudogenes, and alternative splicing.

- **Transcript annotation from transcript expression dataset**

Transcript annotation can also be obtained from transcription expression dataset as described in the following subsections -

✓ **Major isoform annotation from RNA-seq**

The major isoform annotation can be done with Salmon (version 1.5.2) (Patro, Duggal, Love, Irizarry, & Kingsford, 2017) using RNA-seq data. Salmon calculates transcript abundances in transcripts per million (TPM) units by quantifying RNA-seq data for the human transcriptome. For curated RefSeq annotated transcript isoforms, the major isoform is determined as the one with maximum mean Transcripts Per Million (TPM) value across RNA-seq samples.

✓ **Transcript Unit annotation from TT-seq**

Annotation of different transcript classes, in particular enhancer RNAs (eRNAs), can be done with GenoSTAN (Zacher, Lidschreiber, Cramer, Gagneur, & Tresch, 2014) as in (Lidschreiber et al., 2021). In brief, genome-wide coverage is calculated from all TT-seq fragment midpoints in consecutive 200 bp bins throughout the genome. A two-state hidden Markov model with a Poisson-Log-Normal emission distribution is learned in order to segment the genome into 'transcribed' and 'untranscribed' states. Consecutive 'transcribed' states were joined, if its gaps were smaller than 200 bp, within an annotated mRNA or lincRNA, or showed uninterrupted coverage supported by all TT-seq samples. Resulting transcribed units (TUs) are further filtered using a minimal expression threshold that was defined based on overlap with genes annotated in GENCODE. The threshold is optimized using the Jaccard index criterion.

Transcript classification

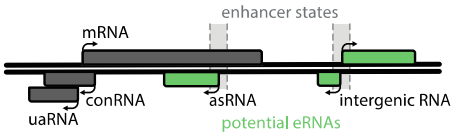


Figure 2.4 | Definition of transcript classes and putative enhancer RNA (eRNA) annotation. Adapted from (Lidschreiber et al., 2021).

TUs that overlapped at least 25% of an annotated protein-coding gene and overlapped with an annotated exon of the corresponding gene are classified as mRNAs. Remaining TUs are annotated as non-coding (nc)RNAs and further classified according to their genomic

location relative to protein-coding genes: upstream antisense RNA (uaRNA), convergent RNA (conRNA), antisense RNA (asRNA), and intergenic RNA. ncRNAs located on the opposite strand of an mRNA were classified as asRNA if the TSS was located > 1 kbp downstream of the sense TSS, as uaRNA if the TSS was located < 1 kbp upstream of the sense TSS, and as conRNA if the TSS was located < 1 kbp downstream of the sense TSS. All remaining ncRNAs are classified as intergenic **(Figure 2.4)**. Further classification of intergenic and asRNAs as eRNAs is done on an experiment specific manner.

iii. Mapping

Choosing the most appropriate mapping platform according to the requirements of the work is imperative and depends on the type of analysis to be carried out. Two types of mapping strategies - spliced mapping and unspliced mapping are applied for different dataset. STAR (Dobin et al., 2013) or Tophat2 (Kim et al., 2013) spliced mappers are used to map raw reads of transcriptomic expression (RNA-seq or TT-seq) data and Pol II occupancy (mNET-seq) data to a reference genome taking into account the presence of splice junctions. In comparison to Tophat2, STAR works fast with being very accurate and precise and compatible for both short and long reads. For ChIP occupancy data, to map only continuous reads to a reference genome, unspliced mapper Bowtie2 (Langmead, Trapnell, Pop, & Salzberg, 2009) is used. For nucleosome occupancy data, we exploited the function of STAR mapper (Dobin et al., 2013) as unspliced mapper by setting the parameter of maximum intron length to one.

2.2.2e Post alignment processing

After mapping of reads, mapping quality is checked with Samtools (Li et al., 2009), as some of the biases in the data only show up after the mapping step. Mapping quality can be improved with precision mapping and processing of the mapped reads. This greatly improves the accuracy and quality of further downstream analysis.

2.2.2f Quantification of features

HTseq (Anders, Pyl, & Huber, 2015) is used to compute feature counts from mapped reads. It takes as input a mapped reads file, and uses an annotated reference genome to produce a mapped reads counts file, indicating how many reads overlap each feature specified in the genome's annotation.

2.2.2g Calculation of the number of transcribed bases

Aligned duplicated fragments are discarded for each sample. Of the resulting unique fragment isoforms only, those are kept that exhibited a positive inner mate distance. The number of transcribed bases (*tbj*) for all samples is calculated as the sum of the coverage of evident (sequenced) fragment parts (read pairs only) for all fragments smaller than 500 bases in length and with an inner mate interval not entirely overlapping a Refseq annotated intron in addition to the sum of the coverage of non-evident fragment parts (entire fragment).

2.2.3 Downstream processing of NGS data

Following sequence alignment to a reference genome, the data needs to be analyzed in an experiment-specific fashion. Alignment serves as a basis for many types of downstream analysis such as calculating transcript /gene expression, differential gene expression, signal visualization in the genomic browsers, and so forth.

2.2.3a Visualizing mapped reads in Genome browser

The standard tool for visualization of mapped reads is the Genome Browser (Hahne & Ivanek, 2016). Genome Browser allows to browse a specific genomic position or a specific feature helping to better understand NGS data considering their nature and allows to focus on most important findings.

2.2.3b Normalization

Next generation sequencing of transcriptomics data involves the conversion of RNA to cDNA, fragmentation and amplification of double-stranded cDNA by polymerase chain reaction (PCR) as library preparation kits require specific volumes and concentrations of cDNA material as input for sequencing. These introduces technical noise, if not corrected, can conceal the biological variations between samples. Normalization is done to calibrate for technical noise between groups of samples.

i. Correction for global variations by sequencing depth

If spike-ins are included in an NGS experiment, spike-in normalization strategy is commonly used to account for global variations between samples from different experimental conditions. Spike-ins can be synthetic RNAs of known abundance or RNA from a related organism used as internal standard, such as *Saccharomyces cerevisiae* or *Drosophila melanogaster* RNA for *Homo Sapiens* expression or occupancy measurements. The spike in normalizations stated in this thesis are based on synthetic spike-ins for 4sUseq, TT-seq and RNA-seq data,

Saccharomyces cerevisiae spike-ins for mNET-seq data and *Drosophila melanogaster* spike-ins for ChIP-seq data.

Calculation of sequencing depth

Normalization strategy for 4tU/4sU-seq or TT-seq, along with total RNA (RNA-seq), uses spike-ins as described in (Schwalb et al., 2016). This spike-in normalization uses both unlabeled and labeled spike-ins where unlabeled spike-ins exemplifies total RNA fragments and labeled spike-ins imitate signifies newly synthesized RNA fragments. Sequencing depth (σ) as normalisation factors are calculated from these spike-in read counts.

- For TT-seq and RNA-seq experiments in human CDKas cell lines, sequencing depth (σ) from spike-in (RNAs) were calculated across libraries with rCube function SpikeinNormalization (Villamil, Wachutka, Cramer, & Gagneur, n.d.). The function fits negative binomial distribution and maximum likelihood (ML) using generalized linear model (GLM) to calculate normalization factor.
- For 4sU-Seq experiment in *Saccharomyces cerevisiae*, sequencing depth from spike-in (RNAs) were calculated for each sample *j* according to

$$\sigma_j = median_i\left(\frac{k_{ij}}{l_i}\right)$$

with read counts k_{ij} for the labelled spike-ins *i* in sample *j* and l_i for the length of labelled spike-ins *i*. Sequencing depth are further standardized by

$$\sigma_j = \frac{\sigma_j}{\sigma_1}$$

mNET-seq normalization factor was calculated from *Saccharomyces cerevisiae* spike-ins count data with DEseq2 package function estimateSizeFactors (Anders & Huber, 2010b; Love, Huber, & Anders, 2014) which fits negative binomial distribution and likelihood ratio test (LRT) using generalized linear model (GLM) to calculate normalization factor across libraries.

mNET-seq and ChIP-seq data was also normalized by scaling factor RPKM (reads per kilobase of per million reads mapped) and FPKM (fragments per kilobase of per million mapped fragments using Deeptools2 package function bamCoverage. The calculation of RPKM or FPKM feature *i* is performed as

$$RPKM_i or\ FPKM_i = \left(\frac{q_i}{l_i * \sum_j q_j}\right) * 10^9$$

where q_i are raw read or fragment counts, l_i is feature (i.e., gene or transcript) length, and $\sum_j q_j$ corresponds to the total number of mapped reads or fragments.

ii. Correction for antisense bias

Next generation sequencing of transcriptomics data is centered on cDNA-based library preparation where RNA to converted to cDNA by the reverse transcription reaction. This tends to produce false or spurious reads originating from the opposite strand. To correct for the experimental bias, antisense bias ratio c_j were calculated for each sample *j* according to

$$c_j = \underset{i}{\text{median}}\left(\frac{k_{ij}^{antisense}}{k_{ij}^{sense}}\right)$$

for all available spike-ins *i*. The antisense bias ratio c_j is used to correct read counts or coverage s_{ij} for any feature i in sample j as

$$s_{ij} = \frac{S_{ij} - c_j A_{ij}}{1 - c_j^2}$$

where S_{ij} and A_{ij} are respectively the observed read counts or coverage on the sense and antisense strand.

iii. Correction of TT-seq and RNA-seq data for labeling bias

NGS experiments with metabolically labeled RNA samples (4tU/4sU-seq/TT-seq) requires an additional normalization to account for cross contamination, proportion of unlabeled reads purified in the labeled RNA samples. Cross-contamination rate ϵ_j, was calculated for each sample *j* as

$$\epsilon_j = \underset{i}{\text{median}}\left(\frac{k_{ij}}{l_i}\right)/\sigma_j$$

using the unlabeled spike-ins *i* for TT-seq samples. ϵ_j is set to 1 for RNA-seq samples as RNA-seq samples do not undergo 4sU-labeled pull-down purification, and thus have maximal amounts of unlabeled RNA. The antisense bias corrected read counts or coverage s_{ij} for any feature i in sample j is corrected for cross-contamination as

$$t_{ij} = \frac{\frac{s_{ij}^{TT-seq}}{\sigma_j^{TT-seq}} - \epsilon_j\left(\frac{s_{ij}^{RNA-seq}}{\sigma_j^{RNA-seq}}\right)}{1 - \epsilon_j}$$

2.2.3c Expressed gene set for analysis

For downstream analyses, expressed feature sets are defined based on the antisense bias corrected **R**ead counts **P**er **K**ilobases (RPKs) as measure of expression. RPKs are calculated

using antisense bias corrected read counts (s_{ij}) falling into the region of feature divided by feature length in kilobases.

2.2.3d Reproducibility

NGS experiments are performed in two or three biological replicates to generate sample NGS libraries to ensure reproducibility. Consistency of replicates data were checked by spearman correlation coefficient by comparing counts of feature reads.

2.2.3e Statistical test

Statistical significance is performed using the non-parametric Wilcoxon rank-sum test to compare between samples.

2.2.4 Downstream analysis

2.2.4a Differential gene expression analysis

Mapped reads count, which quantifies the expression of each gene can be analyzed using software such as DESeq2 (Anders & Huber, 2010a), which runs a statistical analysis of gene expression under the different conditions studied, to identify differentially expressed genes. DESeq2 uses a negative binomial distribution to model the counts for each gene with a given set of parameters (i.e., normalization factor, dispersion) and fit the normalized count data to it. After model fitting, coefficients (log2 fold changes) are estimated for each sample group along with their standard error. The result is explained with log-fold change in base 2, the fold-change in average expression of a gene between two groups. DESeq2 also computes a p-value for each gene by the Wald test where a low p-value (typically less than 0.005) is viewed as evidence that the gene is differentially expressed**.** In addition, performing multiple testing allows by Benjamini and Hochberg method to assess significance as the false discovery rate. The False discovery rate (adj p-value) is a corrected version of the p-value, which accounts for multiple testing correction. Typically, an FDR less than 0.05 is good evidence that the gene is differentially expressed. For analysis purposes, differentially expressed genes can be filtered by maximum acceptable false discovery rate, up regulation or down regulation and minimum $\log_2$ fold change (Log_2FC). To observe global effects on RNA expression, spike-in–derived normalization factors were used for differential gene expression analysis.

2.2.4b Functional and pathway enrichment analysis

Functional annotation of target genes is based on Gene Ontology (GO)(Carbon et al., 2019).

2.2.5 Estimation of kinetic parameters

Kinetic modeling was designed by Dr. Björn Schwalb. The original model was published in (Gressel, Schwalb, & Cramer, 2019; Gressel et al., 2017). *The model is adapted for specific requirements of different experimental hypothesis.*

i. RNA amount per cell

A conversion factor RNA amount per cell [cell^{-1}] can be calculated given the known sequence and mixture of the RNA spike-ins as described. The sequence of spike-ins allows to calculate their molecular weight M (assuming perfect RNA extraction) as

$$M = A_n \times 329.2 + (1 - \tau) \times U_n \times 306.2 + C_n \times 305.2 + G_n \times 345.2 + \tau \times 4sU_n \times 322.26 + 159$$

where A_n, U_n, C_n, G_n and $4sU_n$ are the number of each respective nucleotide within each spike-in polynucleotide. For labeled spike-ins τ is set to 0.1 and otherwise 0. The addition of 159 corresponds to the molecular weight of a 5' triphosphate.

The number of spike-in molecules per cell N [cell^{-1}] was calculated as

$$\mathrm{N} = \frac{m}{M} N_A$$

where m is the number of spike-ins in a particular number of cells, M is the molecular weight of the spike-ins and N_A is the Avogadro constant.

The conversion factor to RNA amount per cell k [cell^{-1}] then can be calculated for all labeled spike-ins i with length L_i as

$$k = mean\left(median_i\left(\frac{tb_i}{L_i \cdot N}\right)\right)$$

ii. Productive initiation frequency

Productive initiation frequency is a quantification of Pol II initiation event that exits the promoter-proximal pause site to enter productive elongation measured by the transcribed bases. Productive initiation frequency I_i can be calculated for each feature i using sequencing depth and antisense bias corrected number of transcribed bases tb_i quantified from TT-seq coverage on exons spanning a defined window downstream of the TSS excluding the first exon to eliminate the possibility of splicing bias to downstream of the TSS, as

$$I_i = \frac{1}{k} \cdot \frac{tb_i}{t * L_i}$$

with labelling duration t and length L_i.

iii. Detection of Pol II pause or arrest site

Pol II pause or arrest site *m** is calculated by calling the maximum peak of mNET-seq signal ρ_i in a defined window *m* downstream of the TSS (exceeding three times the median signal p_{im}) for the sense strand

$$\rho_i = \max_m p_{im}$$

iv. Detection of nucleosome arrest site

Nucleosome arrest sites are calculated with the same principle described in section 2.4. For nucleosome arrest sites, the defined window *m* spans from 100 bp upstream of the annotated nucleosome dyad position to the nucleosome dyad position for the sense strand.

2.2.6 Multi-omics analysis

i. Pause duration calculation

The pause duration *d* was calculated for the expressed feature with annotated promoter proximal pause sites and nucleosome arrest sites. Pause duration is defined as the time a polymerase needs to pass through a defined 'pause window' around the pause site or arrest site. The pause duration *d* was calculated in the pause window as

$$d_i = s \cdot \frac{\sum P_i}{I_i}$$

where P_i is the mNET-seq coverage values in the pause window and I_i is the productive initiation frequency.

ii. Elongation velocity estimation

The elongation velocity was calculated as described (Caizzi et al., 2021; Žumer et al., 2021). Elongation velocity is defined as the RNA synthesis per pol II as a function of time. As TT-seq measures the nascent RNA synthesis profile of Pol II during the labeling time and mNET-seq measures the Pol II occupancy profile (Gressel et al., 2017), the TT-seq/mNET-seq signal ratio can be quantified as a proxy for elongation velocity. For each expressed feature *i,* the elongation velocity can be calculated for a window downstream of the TSS as

$$v_i = \frac{s}{t * k} * \frac{\sum tb_i}{\sum P_i}$$

where tb_i is the sequencing depth and antisense bias corrected TT-seq coverage and and P_i is the sequencing depth normalized mNET-seq coverage values in the defined window.

2.2.7 Kinetic modeling of transcription

Kinetic modeling was conceived by Dr. Björn Schwalb. The original model was published in (Gressel et al., 2017). *The model is extended and adapted by the author of the thesis for specific requirements of different experimental hypothesis.*

2.2.7a Kinetics of gene transcriptions

Transcription of a typical gene by RNA polymerase II (Pol II) initiates at transcription start site (TSS) with a certain elongation velocity. Elongation velocity denotes the speed with which a polymerase moves along a typical gene during transcription. Along the gene template polymerase encounters the first checkpoint in the promoter proximal pause region with a pause. The polymerase then moves along the gene body up to transcription termination site (TTS) to produce a functional mRNA. The elongation velocity along the gene template can be obtained by tracking of the movement of a single polymerase along the gene template over time. This tracking of single polymerase can be termed as elongation velocity profile (**Figure 2.5**).

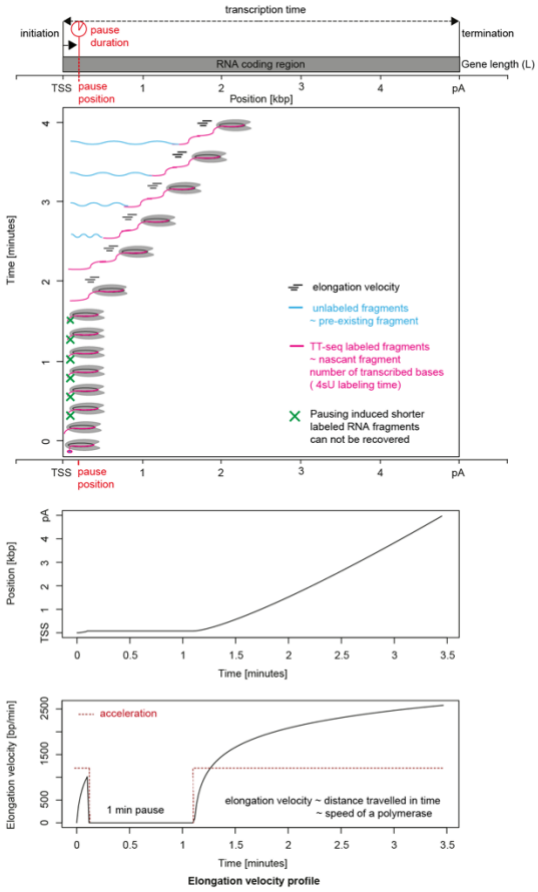


Figure 2.5 | A Schematic diagram illustrating transcription kinetics with single polymerase movement over time.

2.2.7b Simulating elongation velocity profile

The elongation velocity profiles $v(t)$ along a typical gene can be used to model the positioning of the polymerase on a DNA template. Elongation velocity denotes the speed with which a polymerase moves along a typical gene during transcription. This can be calculated as the number of transcribed bases/nucleotides a polymerase can synthesize/incorporate per time into RNA.

The positioning of the polymerases can further be used to calculate the RNA/transcription levels as coverage of a gene. The positioning of the polymerases can simulate the mNET-seq coverage whereas the RNA/transcription levels can simulate the TT-seq data **(Figure 2.6)**.

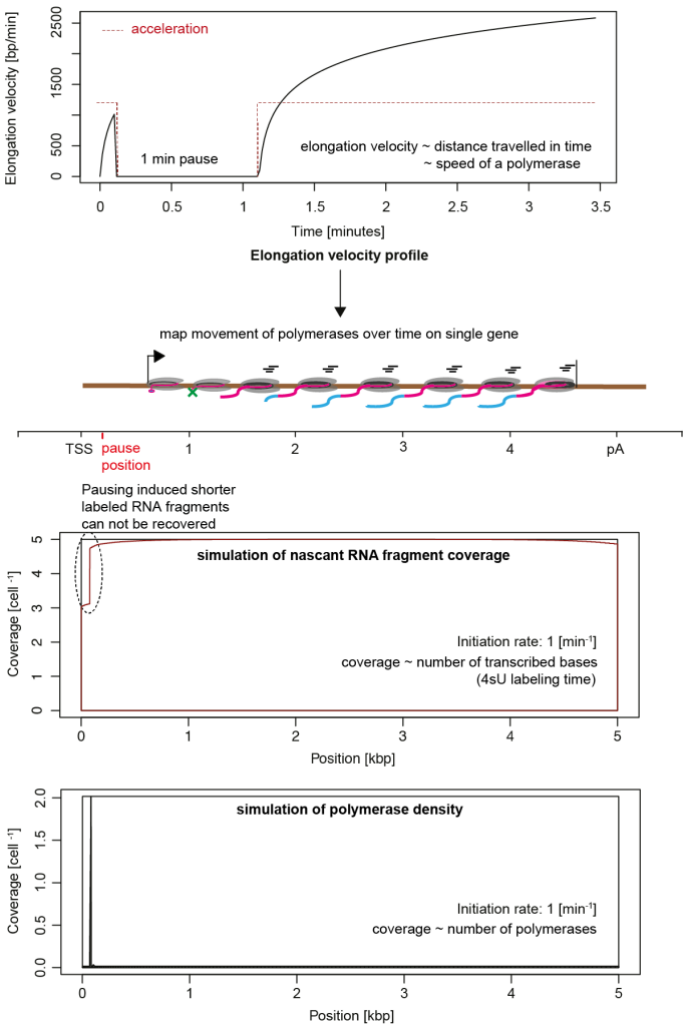


Figure 2.6 | Simulating elongation velocity profile to recapitulate transcription kinetics.

2.2.7c Implementation of the simulation model based on kinetic parameters

In order to mimic observed elongation velocity along a typical gene, we assume the acceleration (a) of the polymerase to be a logarithmic decreasing function

$$a(t) = \frac{\alpha}{(t+1)\log(10)} \quad (1)$$

with α to adjust the maximum elongation velocity in an asymptotic manner and log (10) to modulate the steepness of the acceleration curve. Thus, the velocity function can be derived as

$$v(t) = \int_0^t a(t)dt$$

which can be solved as

$$v(t) = \frac{\alpha \log(t+1)}{\log(10)} \quad (2)$$

The elongation velocity profile $v(t)$ can be used to calculate the number of elongated positions of the polymerase $\tau(t)$ at timepoint t as

$$\tau(t) = \int_0^t v(t)dt$$

which can be solved as

$$\tau(t) = \frac{\alpha\,[(t+1)\log(t+1) - t]}{\log(10)} \quad (3)$$

Given the transcription start site $\tau(0)$ the number of elongated positions $\tau(t)$ can be used to determine the end of an emerging nascent fragment f. Based on that we determined the start position of a fragment as $\tau\left(\max(t - t^{lab}, 0)\right)$ for each labeling duration t^{lab} as the position of the polymerase at the beginning of the labeling process. Given a pre-defined initiation rate, a numeric value indicating the number of polymerases initiating transcription per minute and the above defined integral, elongating polymerases are propagated at nucleotide resolution along the template in a time-dependent manner. Each position of elongating polymerases τ^* can be mapped to its respective time of transcription $t^*$with the elongation velocity profile $v(t)$ as

$$\tau^* = \int_0^{t^*} v(t)dt$$

which can be solved as

$$\tau^* = \frac{\alpha\,[(t^*+1)\log(t^*+1) - t^*]}{\log(10)} \quad (4)$$

Consequently, the progression of a polymerase at any given position τ^* with velocity $v(t^*)$ for time t_p can be calculated as

$$\tau_{t_p}(t) = \int_0^{t_p+t*} v(t)dt - \int_0^{t^*} v(t)dt = \int_{t*}^{t_p+t*} v(t)dt$$

which can be solved as

$$\tau_{t_p}(t) = \tau(t_p + t^*) - \tau(t^*) \quad (5)$$

while $v(t)$ is the current velocity function at absolute time.

The equation (5) can be solved as

$$t^* = e^{W\left(\frac{\tau^* \log(10) - \alpha}{\alpha e}\right)+1} - 1 \quad (6)$$

with $\alpha \neq 0$ and $\alpha \neq p^* \log(10)$ and where W is lambert W function, also called the product logarithm.

The equation in (7) can calculate the duration of transcription for any transcription elongation window.

To simulate polymerase pausing the elongation velocity profile can be adjusted for any given pause position and pause duration. Pause position is a numeric value indicating the pausing position of polymerase in promoter-proximal regions [set to 0 for no pausing] whereas pause duration is a numeric value indicating the duration of pausing of Polymerase in promoter-proximal regions [set to 0 for no pausing; set to a non-zero value as pausing duration in minutes/seconds]. The argument "breaks" is used to introduce pausing in the implementation.

The velocity function in (2) can be adjusted for pausing as

$$v(t) = \frac{\alpha \log(t - b + 1)}{\log(10)} \quad (7)$$

The feasibility of pause-initiation combination was calculated according to the ehrensberger inequality.

2.2.7d Simulation of TT-seq and mNET-seq data

By providing elongation velocity profiles $v(t)$, a labeling duration t_{lab} indicating the 4sU labeling period in minutes and a uracil content dependent labeling bias l_f on the model described the previous section, for a particular gene length in a given number of cells to be used in parallel where initial initiation events are forced to be equidistant among number of cells, we can simulate TT-seq and mNET-seq coverage values. Uracil content dependent labeling bias can be calculated as

$$l_f = 1 - (1 - p^{lab})^{\#u_f}$$

where p^{lab} denotes the labeling probability (set to 0.01) and $\#u_f$ the number of uracil residues of a given fragment f (set to 0.28 times fragment length).

2.2.7e Dynamic progression

i. Steady state transcription

The equation stated in (5) can track the movement of single polymerases over the course of labeling duration. Start and end positions of fragments can be derived given the positions of any given polymerase at $\tau(t, t - t^{lab})$ for a particular transcription time. The snapshot of the steady state transcription can thus be achieved by providing particular transcription duration.

ii. Perturbed/Altered transcription

If variable parameters such as the initiation rate, pause duration or elongation velocity is changed subsequently, the altered movement of single polymerases can also be derived from altered start and altered end positions of fragments for a given transcription time. These altered positions provide us with the snapshot of the altered transcription.

We also used the number of uracil residues present in the RNA fragment $\#u_f$ to weight the amount of coverage contributed by this fragment as l_f. Additionally, we applied a size selection similar to that in the original protocol for fragments below 80 bp in length with a sigmoidal curve that mimics a typical size selection spread.

SECTION TWO

Project-specific data analysis

The general bioinformatics workflow for next generation sequencing (NGS) data is adapted to cater to the outcome of each project.

This section list project specific data analysis for

3.0 CDK7 kinase activity promotes transcription factor exchange and the initiation-elongation transition

4.0 The CDK8 kinase module regulates Mediator-RNA polymerase II interaction

5.0 CDK12 globally stimulates RNA polymerase II transcription elongation and carboxyl-terminal domain phosphorylation

6.0 Kinetic modeling of transcription predicts dynamic RNA synthesis and polymerase occupancy profiles

All the analysis were performed by the author of this dissertation unless otherwise stated. If applicable, a list of contributions can be found at the beginning of each subsection or paragraph. Detailed author contributions can be found on publications.

3.0 CDK7 kinase activity promotes transcription factor exchange and the initiation-elongation transition

NGS libraries generated in this study

All NGS libraries were generated by Taras Velychko
(Max Planck Institute for Multidisciplinary Sciences, Department of Molecular Biology)

Experiment		Replicate	15 minutes	30 minutes	60 minutes
TT-seq		2	✓	✓	✓
RNA-seq		2	✓	✓	✓
mNET-seq		2	✓	✓	x
ChIP-seq	Pol II	3	x	✓	x
ChIP-seq	Med26	3	x	✓	x
ChIP-seq	TFIIB	3	x	✓	x
ChIP-seq	TFIIE	3	x	✓	x
ChIP-seq	NELF	3	x	✓	x
MNase-seq		2	✓	x	x

3.1 Reference genome and reference annotation

For all the analysis, the human genome, transcriptome and annotation were obtained for human genome assembly GRCh38.p13 (RefSeq assembly accession GCF_000001405.39) from NCBI.

i. Major isoform annotation

The major isoform annotation was done with Salmon (1.5.2) (Patro *et al.*, 2017) as described in **section 2.2.2d** subsection **ii** (Transcript annotation from transcript expression dataset - Major isoform annotation from RNA-seq). The final major isoform annotation excludes overlapping genes as well as isoforms located on chromosomes X, Y and M [n =12,585].

ii. GenoSTAN annotation of transcription units (TUs, CDK7as)

Transcription unit annotation from TT-seq samples for HEK293 CDKas was done with GenoSTAN as described in **section 2.2.2d** subsection **ii** (Transcript annotation from transcript expression dataset - Transcription unit annotation from TT-seq data) resulting in 32,789 TUs.

Transcript classification (HEK293 CDK7as)

Transcript classification was performed by Dr. Michael Lidschreiber
(Max Planck Institute for Multidisciplinary Sciences, Department of Molecular Biology)

Annotated TUs were classified as mRNAs, upstream antisense RNA (uaRNA), convergent RNA (conRNA), antisense RNA (asRNA), and intergenic RNA as defined in **Figure 2.4** and corresponding paragraph. Intergenic and asRNAs were further classified as eRNAs, if their TSS ± 500 bp overlapped with an enhancer state annotated by GenoSTAN (Zacher *et al.*, 2014). In addition, antisense eRNAs set was restricted to only those overlapping with publicly available DHS and H3K27ac peaks (ENCODE ENCFF680DCW and ENCFF451UZW), and intergenic eRNAs to only those originating from regions with TT-seq detected transcription on both strands.

3.2 Analysis of TT-seq data and RNA seq data

Raw FASTQ files of paired-end 75 base reads with additional 6 base reads of barcodes were obtained for each of the samples. Reads were demultiplexed and mapped using STAR (v2.6.1b) (Dobin *et al.*, 2013) to the GRCh38.p13 genome assembly merged with the synthetic RNA spike-in sequences (Schwalb et al. 2016) with maximum 2 percent mismatches and only unique alignments were retained. SAMtools (v1.3.1) (Li *et al.*, 2009) was used to quality filter SAM files, whereby alignments with MAPQ smaller than 7 (-q 7) were skipped and only proper pairs (-f2) were selected. Read counts for features and spike-ins were calculated using HTSeq (Anders, Pyl and Huber, 2015).

i. TT-seq and RNA-seq data processing with global normalization parameters

Global normalization parameters antisense bias ratio, sequencing depth and cross-contamination rate were calculated based on the spike-in (RNAs) normalization strategy described in (Schwalb, Michel, Zacher, Hauf, *et al.*, 2016). Calculations for each parameter are described in **section 2.2.3b** in detail. The number of transcribed bases or read counts for major isoform transcripts were corrected for antisense bias and normalized by sequencing depth as described in **section 2.2.3b**. The cross-contamination rate estimates for TT-seq in CDK7as HEK293 were very low and were thus not corrected for.

ii. Expressed gene set for analysis

For downstream analyses, the expressed gene set are defined as described in **section 2.2.3c**. Based on the antisense bias corrected RPKs, a group of expressed major transcript isoforms [n= 8,950] was defined to comprise all transcript isoforms with a median RPK of 20 or

higher in all of TT-seq DMSO samples. An RPK of 20 corresponds to approximately a coverage of 4 per sample due to an average fragment size of 200.

iii. Response ratio

For each condition *j* (control or CDK7as inhibited) the antisense bias corrected number of transcribed bases tb_{ij} was calculated for all expressed major transcript isoform *i* exceeding 10 kbp in length. 5´response ratio was calculated for a window from the TSS to 10 kbp downstream (excluding the first 500 bp) and 3' response ratio was calculated for a window pA-10kbp to pA for each expressed major transcript isoform *i* as

$$r_i = \frac{tb_i^{\text{CDK7as inhibited}}}{tb_i^{Control}}$$

iv. Response to inhibitor treatment

Response to inhibitor treatment was calculated for a window from the TSS to 10 kbp downstream (excluding the first 500 bp) for each expressed major transcript isoform *i* as

$$r_i = 1 - \frac{tb_i^{CDK7as\ inhibited}}{tb_i^{Control}}$$

where negative values were set to 0.

3.3 Analysis of mNET-seq data

For all samples, paired-end 45 base reads were trimmed for adapters (-a TGGAATTCTCGGGTGCCAAGG -A GATCGTCGGACT) and low-quality bases (< Q20) were removed with Cutadapt (v 1.9.1) (Martin, 2011). Reads were then mapped to the GRCh38.p13 genome assembly merged with the yeast genome version SacCer3 using STAR (v 2.5.2b) (Dobin *et al.*, 2013) with maximum 2 percent mismatches. Only unique alignments were retained. Samtools (Li *et al.*, 2009) was used to quality filter SAM files, where alignments with MAPQ smaller than 7 (-q 7) were skipped and only proper pairs (-f 2) were selected.

i. mNET-seq data normalization

Read counts (k_{ij}) for yeast genes for each condition *j* (control or CDK7as inhibited) were calculated using HTSeq (Anders, Pyl and Huber, 2015). The counts data was used to generate normalization factors with DESeq2 (Anders and Huber, 2010) to normalize the mNET-seq data as described in **section 2.2.3b**

3.4 Analysis of ChIP-seq data

All ChIP-seq data except PolII was analysed by Dr. Michael Lidschreiber.
(Max Planck Institute for Multidisciplinary Sciences, Department of Molecular Biology)

For all samples, paired-end 43 base reads were mapped using Bowtie2 (v 2.3.4.1) (Langmead *et al.*, 2009) to the GRCh38.p13 genome assembly merged with *Drosophila* (BDGP6.28) genome assembly. SAMtools(v 1.3.1) (Li *et al.*, 2009) was used to quality filter SAM files, whereby alignments with MAPQ smaller than 7 (-q 7) were skipped and only proper pairs (-f2) were selected. Further data processing was carried out using the R/Bioconductor environment. Duplicate fragments and fragments longer than 100 bp were excluded from further analysis. ChIP-seq coverages were obtained from piled-up fragment midpoint counts for every genomic position.

i. ChIP-seq data normalization

To adjust for differences in sequencing depth, Pol II coverages were normalized using a *Drosophila* chromatin spike-in approach. Normalization factors were obtained from total *Drosophila* fragment counts. For all other factors, which exhibited a peak-like binding behavior, normalization factors were obtained from total human fragment counts. To obtain normalized ChIP-seq coverages for subsequent analyses, coverages were divided by the respective normalization factors and then summed over replicates.

3.5 Analysis of MNase-seq data

For all samples, paired-end 45 base reads were trimmed for low quality bases, random matches and for a minimum length [trimming parameters: -q 20,20 -O 12 -m 25] with Cutadapt (v 1.9.1)(Martin, 2011). Processed reads were then mapped to the GRCh38.p13 genome assembly using STAR (v 2.6.1b) (Dobin *et al.*, 2013) with maximum 2 percent mismatches and only unique alignments were retained. Samtools (Li *et al.*, 2009) was used to quality filter SAM files, where alignments with MAPQ smaller than 7 (-q 7) were skipped and only proper pairs (-f 2) were selected. The resulting BAM files for each condition were merged and converted to BED files with bedtools (v 2.29.1)(Quinlan and Hall, 2010).

i. Detection of +1 and +2 nucleosome positions

The BED files were processed by DANPOS3 algorithm (dpos) to determine nucleosome positions for each condition. Further data processing was carried out using the R/Bioconductor environment. For annotation of the +1 nucleosome, the closest nucleosome that was found downstream of the TSS and within the window TSS+200 bp

for each gene was chosen. Similarly, the succeeding nucleosome downstream of the annotated +1 nucleosome and within the window TSS+400 bp for each gene was annotated as the +2 nucleosome. The midpoint of the annotated nucleosome positions was determined as the positions of the +1 and +2 nucleosome dyads. The final expressed gene set includes 7938 +1 and +2 nucleosome positions for further analysis.

ii. MNase-seq data normalization

Read counts for expressed major transcript isoforms were calculated using HTSeq (Anders, Pyl and Huber, 2015). Size factors for each sample were then estimated as described by Equation (5) in DESeq package (Anders and Huber, 2010) to correct for library size and variations as described in **section 2.2.3b**. These size factors were then used for normalizing the read coverage profiles.

3.6 Reproducibility

All treatments were performed in two biological replicates to generate sample NGS libraries, except ChIP-seq, for which three replicates were collected.

3.7 Downstream analysis

3.7a Differential expression analysis

Differential gene expression analysis was performed with the DESeq2 package (Anders and Huber, 2010) to observe changes in RNA synthesis (TT-seq) after inhibition. The sizeFactor parameter was set to the spike-in normalization strategy derived sequencing depth. The significant changes were denoted for the padj cutoff set to 0.05 and fold changes greater than 1.5-fold.

3.8 Kinetic parameter estimation

i. Productive initiation frequency

The productive initiation frequency was calculated as described in section **2.2.5 (i)** for all major transcript isoforms *i* exceeding 10 kbp in length. The productive initiation frequency was calculated for each condition *j* (control or CDK7as inhibited) using sequencing depth and antisense bias corrected TT-seq coverage on exons spanning a window from the TSS+500 to 10 kbp downstream, excluding the first exon, for 10-minute labelling duration.

ii. Detection of promoter proximal pause sites

The major transcript isoform set for this analysis includes expressed protein-coding isoforms with first exon greater than 30 bp (to exclude possibilities for 5' capping bias). Promoter proximal pause site *m** for Pol II was annotated as described in section **2.2.5 (iii)** of the control samples. For major transcript isoforms with +1 nucleosome position downstream of the first exon, *m* was defined from the TSS + 30 bp to the end of the first exon (excluding the last 5 bases) and for major transcript isoforms with +1 nucleosome within the first exon, the *m* was defined from the TSS + 30 bp to the +1 nucleosome dyad position.

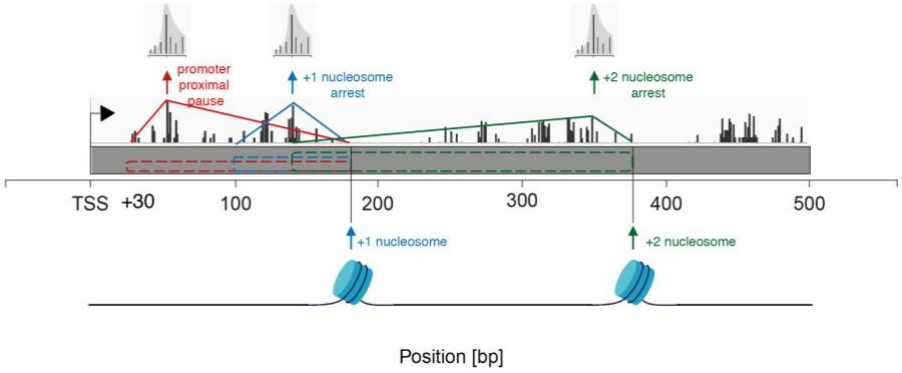


Figure 3.1 | Detection of promoter proximal pause and nucleosome arrest sites as described in (3.8-ii and -iii)

iii. Detection of nucleosome arrest sites

Annotation of nucleosome arrest sites was done as described in section **2.2.5 (iv)** for the expressed gene set with annotated +1 and +2 nucleosome positions. For +1 nucleosome arrest sites the defined window *m* was from 50 bp downstream of the annotated promoter proximal pause site to the +1 nucleosome dyad position and the +2 nucleosome arrest sites were calculated within a window *m* from 100 bp upstream of the annotated +2 nucleosome dyad position to the +2 nucleosome dyad position for the sense strand of the control samples.

3.9 Multi-omics analysis

i. Pause duration

The pause duration *d* was calculated as described in **section 2.2.6 (i)** for the expressed major isoform set with annotated promoter proximal pause sites and nucleosome arrest sites. Pause window is defined as a 40 bp window located ±20 bp around the promoter proximal pause site or nucleosome arrest site. For each condition *j* (control or CDK7as inhibited), the pause duration *d* was calculated in the pause window as

$$d_i = s * \frac{\sum_{\pm 20} P_i}{I_i}$$

where P_i is the mNET-seq coverage values in the pause window and I_i is the productive initiation frequency. The median pause duration (DMSO) was calibrated to resemble pause duration of previous experiments in K562 cells with a calibration factor s obtained in (Gressel *et al.*, 2017; Gressel, Schwalb and Cramer, 2019).

ii. Elongation velocity estimation

The elongation velocity was calculated as described in **section 2.2.6 (ii)**. For each expressed major transcript isoform *i* exceeding 10 kbp in length, the elongation velocity was calculated for a window from TSS to 10 kbp downstream (excluding the first 500 bp).

3.10 Visualization metagene profiles

For visualization, normalized read coverages were summed up over replicates and visualized using the LSD package.

4.0 The Cdk8 kinase module regulates Mediator-RNA polymerase II interaction

The study design and analysis presented in this section have been published (Osman *et al.*, 2021). The published text was adapted to match the style of this thesis and thus deviate from the published version.

This section is a modified excerpt from (Osman *et al.*, 2021)

NGS libraries generated in this study

All NGS libraries were generated by Dr. Sara Osman and Dr. Kerstin Meyer.
(Max Planck Institute for Multidisciplinary Sciences, Department of Molecular Biology)

Experiment	Replicate
4tU-seq (steady state)	2
4tU-seq (heat shock)	2

4.1 Reference genome

For all the analysis, *S. cerevisiae* genome (sacCer3, version 64.2.1) was used.

4.2 Reference annotation

TIF-seq derived TSS and pA site annotations for 5578 protein-coding genes for yeast genome was obtained from (Pelechano, Wei and Steinmetz, 2013)

4.3 Analysis of 4tU-seq data

Raw FASTQ files of Paired-end 75 base reads (for steady-state experiment) and 42 base reads (for heat shock experiment) with additional six base reads of barcodes were obtained for each of the samples. Reads were demultiplexed and low-quality bases were removed using Cutadapt (version 1.9.1) (Martin, 2011) with parameters -q 20,20 -o 12 -m 25. Trimmed reads were then mapped to the *S. cerevisiae* genome (sacCer3, version 64.2.1) merged with the synthetic spike-ins (Schwalb, Michel, Zacher, Hauf, *et al.*, 2016) using STAR (v 2.5.2b) (Dobin *et al.*, 2013). Samtools (Li *et al.*, 2009) was used to quality filter SAM files. Alignments with MAPQ smaller than 7 (−q 7) were skipped, and only proper pairs (−f 2) were selected. A spike-in (RNAs) normalization strategy as described in **section 2.2.3b** allows for observation of global changes in the 4tU-Seq signal. Read counts for protein-coding genes and spike-ins were calculated using HTSeq (Anders, Pyl and Huber, 2015). Further processing of the 4tU-Seq data was carried out using the R/Bioconductor environment.

4.4 Downstream analysis

i. Differential expression analysis

Differential gene expression analysis was performed with the DESeq2 package (Anders and Huber, 2010) using the spike-in–derived normalization factors. An adjusted p-value cutoff of 0.1 was used to call significant changes. Whether a gene's expression is significantly changed depends on both its $\log_2$ fold change and its dispersion estimate. If the dispersion estimates from replicates, quantified by within-group variability, are higher the significance of $\log_2$ fold changes decreases. This explains why some genes with similar $\log_2$ fold change can appear as nonsignificant and significant in **Figure 8.3, 8.6, 8.8**, respectively. To examine the effects of CDK8 inhibition during steady-state growth, the 1-NA-PP1 samples were compared with the DMSO samples. To verify induction of temperature response genes (Gene Ontology: 0009266) upon heat shock, the 12-min heat shock DMSO samples were compared with the steady-state DMSO samples. To examine the effect of CDK8 inhibition on gene expression during heat shock, the 12 min heat shock 1-NA-PP1 samples were compared with the 12-min heat shock DMSO samples. To examine the effect of CDK8 inhibition on induction of heat shock genes, the 12-min heat shock 1-NA-PP1 samples were compared with the steady-state 1-NA-PP1 samples.

5.0 CDK12 globally stimulates RNA polymerase II transcription elongation and carboxyl-terminal domain phosphorylation.

The study design and analysis presented in this section have been published (Tellier *et al.*, 2020). The published text was adapted to match the style of this thesis and thus deviate from the published version.

This section is a modified excerpt from (Tellier *et al.*, 2020)

NGS libraries generated in this study

The TT-seq and RNA-seq libraries were generated by Dr. Livia Caizzi and Taras Velychko. (Max Planck Institute for Multidisciplinary Sciences, Department of Molecular Biology)
All other libraries were generated by collaborators from Murphy laboratory.

Experiment		Replicate	15 minutes	30 minutes
TT-seq	Wild Type	2	✓	✓
TT-seq	CDK12as	2	✓	✓
RNA-seq		2	✓	✓
mNET-seq	Pol II	2	✓	x
mNET-seq	Ser2P	4	✓	x
mNET-seq	Ser5P	2	✓	x
ChIP-seq	WT Pol II	2	✓	x
ChIP-seq	CDK12as Pol II	2	✓	x
ChIP-seq	Ser2P	2	✓	x
ChIP-seq	Ser5P	2	✓	x
ChIP-seq	LEO1	2	✓	x
ChIP-seq	SPT6	2	✓	x
ChIP-seq	CPSF3	2	✓	x

5.1 Reference genome

For all the analysis, the human genome assembly GRCh38 was obtained from Human Genome Reference Consortium (https://www.ncbi.nlm.nih.gov/grc).

5.2 Genome annotation and definition of transcription units

The genome annotation was obtained for UCSC RefSeq genome assembly GRCh38. For each annotated gene, transcription units (RefSeq-TUs) were defined as the union of all existing inherent transcript isoforms.

5.3 Analysis of TT-seq data and RNA seq data

Raw FASTQ files of paired-end 75 base reads with additional 6 base reads of barcodes were obtained for each of the samples. Reads were demultiplexed and mapped using STAR (v2.5.2b)(Dobin *et al.*, 2013) to the merged human genome assembly with synthetic spike-ins (Schwalb, Michel, Zacher, Hauf, *et al.*, 2016) with maximum 2 percent mismatches and only unique alignments were retained. SAMtools was used to quality filter SAM files, whereby alignments with MAPQ smaller than 7 (-q 7) were skipped and only proper pairs (-f2) were selected. Read counts for RefSeq-TUs and spike-ins were calculated using HTSeq.

i. Expressed gene set for analysis

For downstream analyses, the expressed gene set was defined as described in **section 2.2.3c**. Based on the antisense bias corrected RPKs, a group of expressed TUS [n= 11,282] was defined to comprise all TUs with a median RPK of 20 or higher in two summarized replicates of TT-seq 15 minutes DMSO samples. An RPK of 20 corresponds to approximately a coverage of 2 per sample due to an average fragment size of 200.

ii. TT-seq and RNA-seq data processing with global normalization parameters

Global normalization parameters antisense bias ratio, sequencing depth and cross-contamination rate were calculated based on the spike-in (RNAs) normalization strategy described in (Schwalb, Michel, Zacher, Frühauf, *et al.*, 2016). Calculations for each parameter are described in **section 2.2.3b** in detail. The number of transcribed bases or read counts for RefSeq-TUs were corrected for antisense bias and normalized by sequencing depth as described in **section 2.2.3b**. The cross-contamination rate estimates for TT-seq in CDK12as HEK293 were very low and were thus not corrected for.

Additionally, all TT-seq samples were subjected to a more robust normalization procedure. First a subgroup of expressed RefSeq-TUs [n= 5631] were selected based on the antisense bias corrected RPKs of 100 or higher in two summarized replicates of TT-seq 15 minutes control samples. An RPK of 100 corresponds to approximately a coverage of 10 per sample due to an average fragment size of 200. Within this subgroup, a subset of TUs having a length greater than 50 kb [n=1867] was used to identify the non-Differentially Expressed (non-DE) TUs for both time points over the response of 1-NM-PP1(CDK12as inhibitor) using DESeq2 package (Anders and Huber, 2010).

The antisense bias corrected coverage s_{ij} of TU *i* in sample *j* was normalized for sequencing depth **section 2.2.3b** where σ_j was balanced between replicates via classical size factor normalization to gain statistical power in the differential expression analysis.

On the resulting 140 non-differentially expressed RefSeq-TUs, size factors for each sample were then estimated as described by estimateSizeFactors function in DESeq package (Anders and Huber, 2010) depth as described in **section 2.2.3b** to correct for library size and sequencing depth variations.

iii. Simulation of TT-seq data based on elongation velocity profiles

Based on the kinetic model described in **section 2.2.7** TT-seq coverage values were simulated for templates resembling genes of sizes ~0,1-2000 kbp. The resulting gene-wise RNA synthesis profiles were subsequently accumulated to yield meta-gene profiles.

5.4 Analysis of mNET-seq data

mNET-seq data were processed and analysed by collaborators from Murphy laboratory.
adapters were trimmed with Cutadapt in paired-end mode with the following parameters: -q 15, 10 –minimum-length 10 –A GATCGTCGGACTGTAGAACTCTGAAC –a AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC. Trimmed reads were mapped to the human hg38 reference sequence with Tophat2 and the parameters –g 1 –r 3000 –no-coverage-search. SAMtools was used to retain only properly paired and mapped reads (-f 3). A custom python script was used to obtain the 3' nucleotide of the second read and the strandedness of the first read. Strand-specific bam files were generated with SAMtools. FPKM normalized bigwig files were generated for each bam files with Deeptools2 bamCoverage tool (-bs 1 –p max –normalizeUsing RPKM).

i. mNET-seq normalization

The total pol II mNET-seq treated with DMSO and NM 15 min were re-normalized to a set of 140 genes found to be non-affected in the spiked-in TT-seq analysis after 15- and 30-min inhibition. The re-normalization factor was calculated from the average fold change on total pol II signal between DMSO and the NM conditions across the gene bodies of this set of non-affected genes found in TT-seq. For the total pol II mNET-seq metaprofiles and total pol II quantification, the re-normalization factors for NM 15 min R1 and R2 are 0.90 and 1.11, respectively. The Ser2P and Ser5P 15 min were normalized to the CTD phosphorylation signals in histone genes with the following normalization factors for the

merged biological replicates: Ser2P MABI0602: 0.626, Ser2P ab5095: 0.544 and Ser5P ab5131: 0.791.

5.5 Analysis of ChIP-seq data: processing and normalization

ChIP-seq data were processed and analysed by collaborators from Murphy laboratory.

Adapters were trimmed with Cutadapt in paired-end mode with the same parameters as mNET-seq. Obtained sequences were mapped to the human hg38 reference genome with Bowtie2. Properly paired and mapped reads were filtered with SAMtools. PCR duplicates were removed with Picard MarkDuplicates tool. FPKM normalized bigwig files were generated for each bam files with Deeptools2 bamCoverage tool (-bs 10 –p max –normalizeUsing RPKM –e). For the HEK293 ChIP-seq, the pol II signal was normalized to the KPNB1 gene body (TSS + 500 bp to poly(A) site, NM normalization factor: 0.494) as this gene was found to be non-affected by pol II ChIP-qPCR. CDK12as pol II ChIP-seq NM normalization factors are 1.345 and 0.840 for replicates 1 and 2, respectively. Ser2P NM normalization factors are 1 and 1.429 for replicates 1 and 2, respectively. Ser5P NM normalization factors are 1.275 and 1.269 for replicates 1 and 2, respectively.

5.6 mNET-seq and ChIP-seq metagene profiles

Data were analysed by collaborators from Murphy laboratory

Metagene profiles of genes scaled to the same length were then generated with Deeptools2 computeMatrix tool with a bin size of 10 bp and the plotting data obtained with plotProfile –outFileNameData tool. Graphs representing the (IP – Input) signal (ChIP-seq) or the mNET-seq signal were then created with GraphPad Prism 8.4.0. Metagene profiles are shown as the average of two biological replicates.

5.7 P-values and significance tests

P-values were computed with an unpaired two-tailed Student's t test. Statistical tests were performed in GraphPad Prism 8.4.0.

6.0 Kinetic modeling of transcription predicts dynamic RNA synthesis and polymerase occupancy profiles

6.1 Simulation of TT-seq and mNET-seq data as metagene profiles

The simulation according to the model described **section 2.2.7** for altered transcriptional parameters (initiation frequency, pausing duration and elongation velocity) was done for templates resembling genes of sizes ~0,1- 2000 kbp. The resulting gene-wise RNA synthesis profiles (TT-seq) and polymerase positioning profiles (mNET-seq) were subsequently accumulated to yield meta-gene profiles.

CHAPTER 3

RESULTS and DISCUSSIONS

The aim of this chapter is to present the reader with the inferences drawn from the experiments and analysis as results and discussions. The chapter is divided into four sections. Each section lists project specific results and discussions for

7.0 CDK7 kinase activity promotes transcription factor exchange and the initiation-elongation transition

8.0 The CDK8 kinase module regulates Mediator-RNA polymerase II interaction

9.0 CDK12 globally stimulates RNA polymerase II transcription elongation and carboxyl-terminal domain phosphorylation

10.0 Kinetic modeling of transcription predicts dynamic RNA synthesis and polymerase occupancy profiles

- ✓ Experiments were not performed by the author of this dissertation, but are included in this section for a coherent presentation of the obtained findings.
- ✓ All the analysis were performed by the author of this dissertation unless otherwise stated. If applicable, a list of contributions can be found at the beginning of each subsection or paragraph.

CDK7 kinase activity promotes transcription factor exchange and the initiation-elongation transition

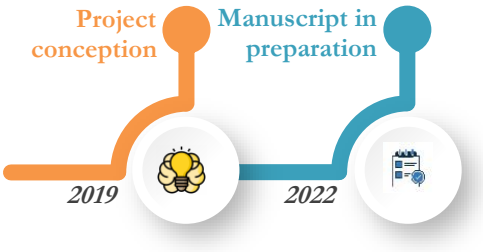
Taras Velychko, **Eusra Mohammad**, Michael Lidschreiber, Patrick Cramer

Distinctive contribution:

Contribution to the analysis and interpretation of

- TT-seq data
- mNET-seq data
- Pol II ChIP data
- Write-up of the manuscript

Manuscript publication status



The project was conceived as an ɪn'haʊs (internal) collaboration *(Department of Molecular Biology, Max Planck Institute for Multidisciplinary Sciences, Göttingen)*. The study is in the process of being culminated into the published manuscript.

Experiments and analysis that were not performed by the author of this dissertation, but are included in this section for a coherent presentation of the obtained findings, are stated at the beginning of each subsection (*in italic*).

PROJECT SUMMARY

Studies throughout the years have illuminated CDK7 functions during the Pol II transcription cycle, starting from promoter clearance, in both promoter-proximal pausing establishment and release, to mRNA 3´-end formation and termination (Fisher, 2019). However, the precise role CDK7 plays in transcription remains controversial. The project was conceived to elucidate the primary effect of CDK7 activity on transcription. To understand the function of the kinase activity of CDK7 in transcription, CRISPR/Cas9 gene engineering was used to mutate the endogenous CDK7 gene in HEK293 cells to produce an analog-sensitive CDK7 (CDK7as) that can be selectively inhibited by bulky ATP analogs. The use of CDK7as cell line was combined with functional genomics methods TT-seq, mNET-seq, ChIP-seq and MNase-seq to monitor the direct effects of rapid CDK7 inhibition on transcription activity, Pol II, transcription factor and nucleosome occupancy in human cells.

INTRODUCTION

CDK7 constitutes part of the human general transcription factor TFIIH in which it forms the cyclin activating kinase-subcomplex with cyclin H (CCNH) and CDK activating kinase assembly factor MAT1 (MAT1) (Roy et al. 1994; Serizawa et al. 1995) . TFIIH unwinds the double stranded DNA in the promoter region to facilitate passage of the Pol II on the template strand (Fishburn et al. 2015; Tsutakawa et al. 2020). The kinase module of TFIIH, CDK7, phosphorylates Pol II CTD in Ser5 and Ser7 in CTD (Christopher C Ebmeier et al. 2017; Glover-Cutter et al. 2009) which aids the recruitment of Pol II from phase-separated polymerase clusters to the promoter region of genes (Boehning et al. 2018). Current studies suggest CDK7 facilitates transcription initiation by phosphorylation of Ser5P on the Pol II CTD conferring an affinity reduction of the polymerase to the mediator complex ultimately resulting in the promoter escape (Wong, Jin, and Struhl 2014). Additional functions of CDK7 in transcription includes functioning as a TFIIH-free CDK7 complex form called CDK-activating kinase (CAK) that phosphorylates various other CDKs involved in

transcription, for example, CDK9 (Anshabo et al. 2021). Therefore, CDK7 can regulate promotor-proximal pausing (Fisher 2019a). CDK7 has also been shown to be important in mRNA 5' end capping (Nilson et al. 2015), termination of transcription and mRNA 3' end polyadenylation (Christopher C Ebmeier et al. 2017). CDK7 also has role in histone modifications (Christopher C. Ebmeier et al. 2017; Fisher 2019b). Outside of transcription, CAK complex is also needed to activate cell cycle-controlling CDKs, therefore indirect involvement of CDK7 in cell cycle is observed (Fisher 2019b).

In recent years CDK7 has further gained attention due to its therapeutic relevance in numerous oncological malignancies. The kinase is perceived as a regulator of various cancer-specific pathways and thus has been targeted by several small molecule inhibitors such as THZ1 and ICEC0942 (Liang et al. 2021). Importantly, ICEC0942 and THZ1-related SY-1365 are also being tested in phase I clinical trials (Sava et al. 2020). Despite the multitude of available CDK7 inhibitors, the use of many of these small molecules, including widely studied inhibitor THZ1, has been accompanied by off-target effects on other kinases such as CDK12 and CDK13 (Olson et al. 2019). Aiming to address the function of CDK7 in transcriptional regulation by highly specific inhibition, we utilize a chemical genetics approach for kinase inhibition (Lopez, Kliegman, and Shokat 2014) with multi-omics dataset. Our study suggests on CDK7 kinase function as a global regulator of transcription initiation. We also perform a proteome-wide regulation and interaction analysis accompanied by the determination of phosphorylation targets, which provides us insight with plausible underlying molecular mechanism by which CDK7 kinase activity controls transcriptional initiation.

RESULTS

7.1 Rapid and specific CDK7as inhibition in human cells

CDK7as HEK293 cell line, using CRISPR/Cas9, where the gatekeeper phenylalanine residue (codon TTT) is mutated to glycine (codon GGT) was generated and validated by Dr. Shona Murphy.

CDK7as HEK293 cell line allows rapid and selective inhibition of CDK7 kinase activity with a bulky ATP-analog 1-NM-PP1 (NM).

7.2 CDK7 inhibition results in global downregulation of new RNA synthesis

All experiments in this section were conducted by Taras Velychko
(Department of Molecular Biology, Max Planck Institute for Multidisciplinary Sciences, Göttingen)

To elucidate the primary role of CDK7 in transcription, we monitored changes in RNA synthesis by TT-seq (Schwalb *et al.*, 2016) before and after CDK7 inhibition. CDK7as cells

were treated with 7.5 µM of DMSO and 1-NM-PP1 for 15, 30 and 60 minutes with an RNA labeling time of 10 minute (**Figure 7.1A**). Across all time points, differential expression analysis upon CDK7 inhibition showed a global downregulation of newly synthesized RNA (**Figure 7.1B**). The percentage of significantly downregulated genes increased from 54% after 15 minutes to 97% after 60 minutes of inhibition.

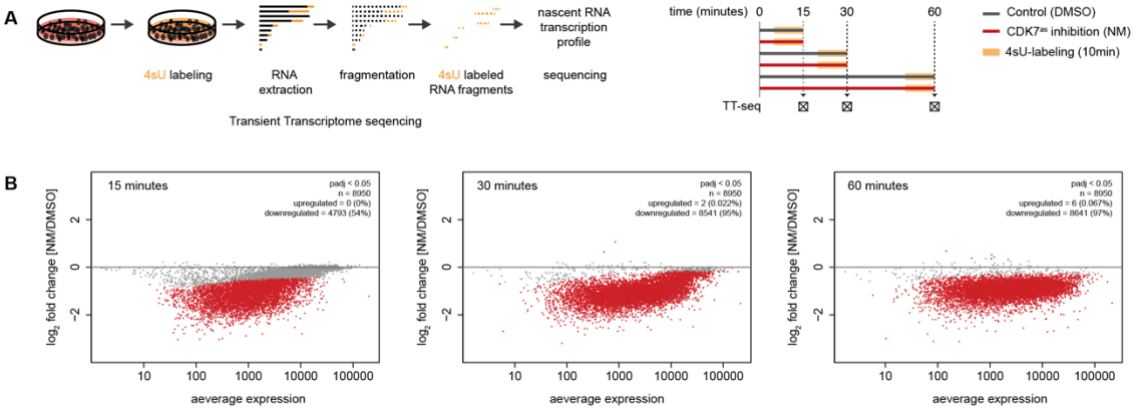


Figure 7.1 | CDK7as inhibition globally decreases RNA synthesis. **(A)** Experimental design of TT-seq **(B)** Log_2 fold change upon 1-NM-PP1(NM) treatment for 15 minute (left), 30 minute (middle) and 60 minutes (right) versus the average expression as normalized mean read count across replicates and conditions for CDK7as HEK293cells. Significantly up- or downregulated major transcript isoforms (adjusted p-value<0.05) are marked in red.

To analyze the pattern of global downregulation, the TT-seq signals averaged overexpressed genes in CDK7as cells were visualized as metagene profiles. Upon inhibition, metagene analysis of 15 minutes showed a stronger decrease in TT-seq signal at the 5' end compared to the 3' end of genes whereas a stably decreased TT-seq signal across gene bodies was observed for 30 and 60 minutes (**Figure 7.2A**). This observation indicates that less Pol II was initiating and/or released from the promoter-proximal pausing into gene bodies. To investigate this further, we calculated CDK7 inhibition response ratios comparing the change in TT-seq signal at the 5' and 3' end of short, medium and long genes (**Figure 7.2B**). The 5' end response ratios showed reduced RNA synthesis irrespective of gene length and CDK7 inhibition time, confirming defective RNA synthesis in the beginning of genes (**Figure 7.2B**). In contrast, the 3' ends of long genes were largely unaffected after 15 minutes of CDK7 inhibition, indicating continued RNA synthesis by Pol II that had been released before CDK7 inhibition. This is further exemplified at the *ATF7IP* gene, which showed downregulation of RNA synthesis at its 5' end already after 15 minutes of CDK7 inhibition, but it took until 60 minutes of inhibition to observe uniform downregulation up to the end of the gene (**Figure 7.2C**).

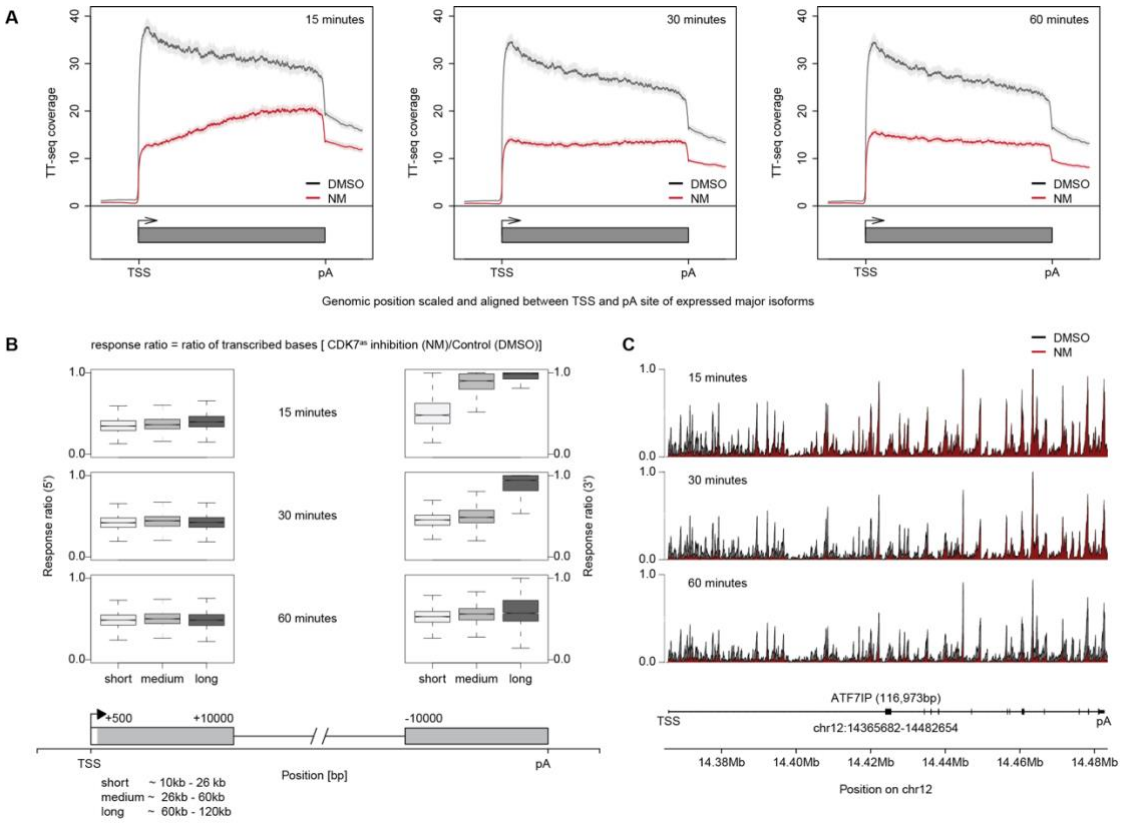


Figure 7.2 | CDK7as inhibition globally decreases RNA synthesis. **(A)** Metagene analysis of TT-seq signal for expressed genes after treatment of cells with DMSO (grey) versus 1-NM-PP1(NM) treatment (red) for 15 minute (left), 30 minute (middle) and 60 minute (right). The TT-seq coverage was merged and aligned at their transcription start sites (TSSs) and polyadenylation (pA)-sites. Shaded areas around the average signal (solid lines) indicate confidential intervals. **(B)** Box plot showing the response ratio for 5' end (right) and 3' end (left) of the major transcript isoform for 15 minute (upper panel), 30 minute (middle panel) and and 60 minute (lower panel) inhibition of CDK7as compared to control. **(C)** TT-seq signal before (grey) and after (red) CDK7 inhibition at the ATF7IP gene locus (116,937 [bp]) on chromosome 12 15 minute (upper panel), 30 minute (middle panel) and and 60 minutes (lower panel). Two biological replicates were merged.

Next, to determine the relative response of genes to CDK7 inhibition, we calculated response to inhibitor treatment within a scale of 0% to 100% where 100% indicates fully responding genes. The response to inhibitor treatment decreased over the time course indicating a recovery of RNA synthesis activity (**Figure 7.3A**). The recovery may be attributed to Pol II mediated RNA synthesis without CDK7 kinase activity, or it may stem from incomplete or reversible CDK7 inhibition. However, assuming that the inhibitor is evenly distributed across and within cells, the proportion of CDK7 that has not been fully inhibited must be very small.

TT-seq also allowed us to quantify 'productive initiation rate', that is the number of polymerases that initiated and successfully transitioned into productive elongation (Gressel *et al.*, 2017a; Gressel, Schwalb and Cramer, 2019). We observed a decrease in 'productive

initiation rate' for all time points after CDK7 inhibition (**Figure 7.3B**). Taken together, these results show that inhibition of CDK7 kinase activity causes transcriptional repression at the beginning of genes.

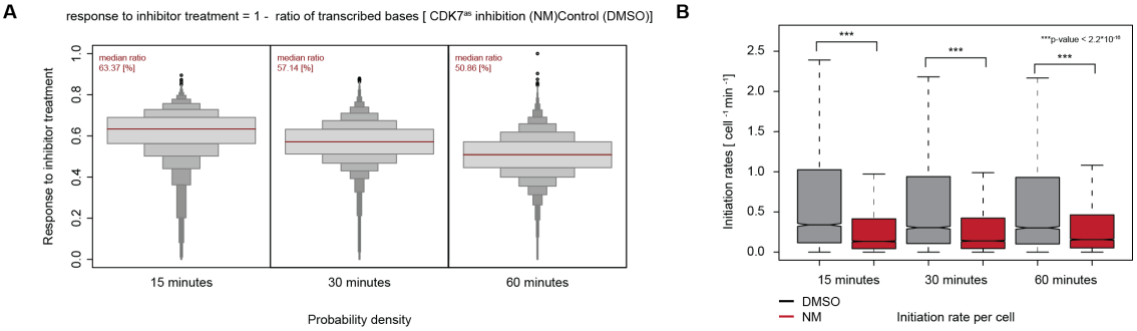


Figure 7.3 | CDK7as inhibition globally decreases RNA synthesis. (A) Violin plot showing the relative response to CDK7 inhibition for major transcript isoform for 15 minute (left), 30 minute (middle) and and 60 minutes (right) for a window from the TSS to 10 [kbp] downstream, excluding the first 500 [bp]. Red line indicates the median response. **(B)** Distributions of gene-wise initiation rates before (grey) and after (red) CDK9 inhibition for 15 minute (left), 30 minute (middle) and and 60 minute (right).

7.3 CDK7 inhibition results in decreased Pol II gene occupancy

All experiments in this section were conducted by Taras Velychko
(Department of Molecular Biology, Max Planck Institute for Multidisciplinary Sciences, Göttingen)

The observed reduction in TT-seq signal at the beginning of genes upon CDK7 inhibition may be the result of either a direct or indirect decrease in initiation events, where an indirect decrease would be due to an increase in promoter-proximal pause duration. To distinguish between the two possibilities, we performed mNET-seq to map actively engaged Pol II in CDK7as cells after 15 and 30 minute of inhibitor treatment (**Figure 7.4A**). Metagene profiles of mNET-seq signals indicated decreased mNET-seq signal at the beginning of genes (**Figure 7.4B**) as well as across the gene body (**Figure 7.4C**). In contrast, upon inhibiton of CDK9, which facilitates Pol II pause release, mNET-seq signal increased at the beginning of genes and decreased in the gene body, resulting in promoter proximal pausing mediated decrease in productive initiation events (Gressel *et al.*, 2017b; Gressel, Schwalb and Cramer, 2019). The striking dissimilarity between the Pol II occupancy profile changes prompted us to exclude the possibility of an indirect decrease in initiation events and emphasize on the fact that CDK7 regulates gene expression by directly altering the initiation frequency at human genes.

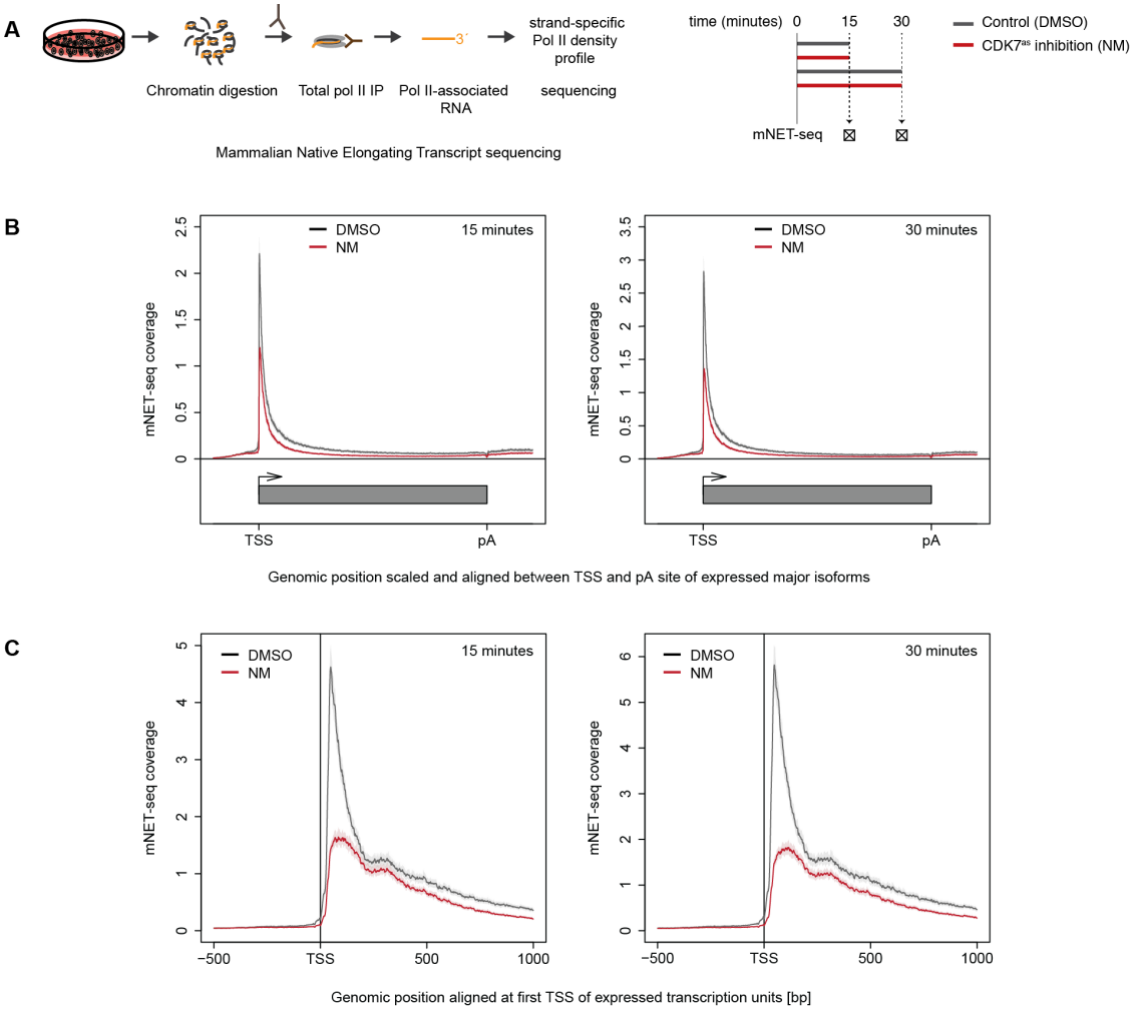


Figure 7.4 | Inhibition of CDK7as decreases Pol II gene occupancy. (A) Experimental design of mNET-seq. **(B)** Metagene analysis of mNET-seq signal for expressed genes after treatment of cells with DMSO (grey) versus 1-NM-PP1(NM) treatment (red) for 15 minute (left) and 30 minute (right). The mNET-seq coverage was merged and aligned at their transcription start sites (TSSs) and polyadenylation (pA)-sites**.(C)** Metagene analysis of mNET-seq signal for expressed genes after treatment of cells with DMSO (grey) versus 1-NM-PP1(NM) treatment (red) for 15 minute (left) and 30 minute (right). The mNET-seq coverage was merged and aligned at their transcription start sites (TSSs). Shaded areas around the average signal (solid lines) indicate confidential intervals for metagene profiles.

7.4 CDK7 inhibition leads to the accumulation of the preinitiation complex upstream of the TSS

All experiments in this section were conducted by Taras Velychko.
All ChIP data except Pol II was analysed by Dr. Michael Lidscreiber
(Department of Molecular Biology, Max Planck Institute for Multidisciplinary Sciences, Göttingen)

To further support this hypothesis, we performed MNase-ChIP-seq of Pol II after 30-minute inhibition of CDK7as cells (**Figure 7.5A**). The high resolution of MNase-ChIP-seq allowed us to distinguish Pol II complexes at pause sites from those upstream of the TSS. Metagene profiles of ChIP-seq signal centered around the TSSs of expressed genes showed the altered profile of Pol II that shifted from downstream pause sites toward the TSS

(**Figure 7.5B**). This transition of Pol II from pause sites to the TSS can be interpreted as the new "poised" preinitiation complexes that localize to the TSS but are unable to transition to the pause sites without CDK7 kinase activity. This is further exemplified at the *RPL10A* gene, which showed a similar overall Pol II occupancy change upon CDK7 inhibition (**Figure 7.5C**). Together, these data strongly support the role of CDK7 kinase activity in transcription initiation.

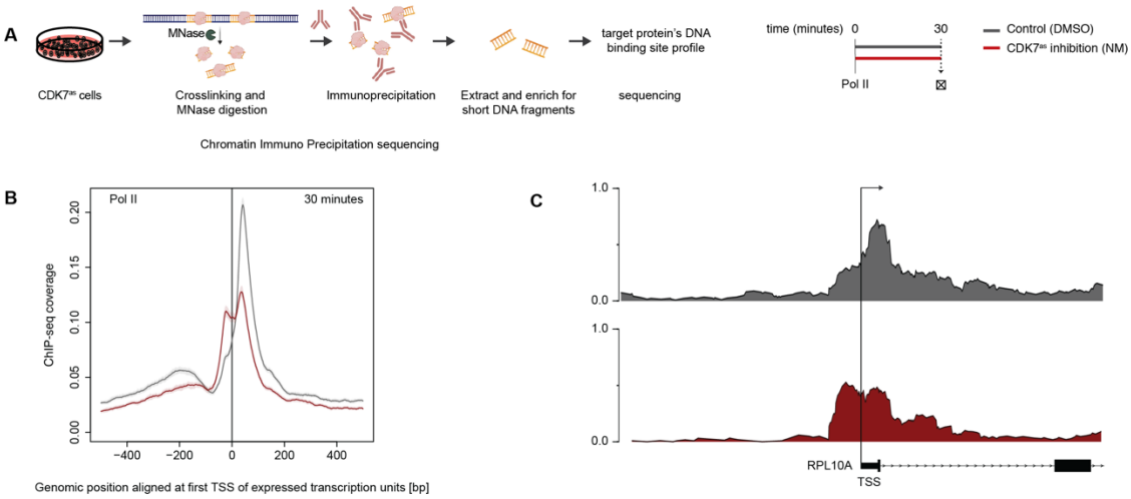


Figure 7.5 | Inhibition of CDK7 leads to the PIC accumulation upstream of the TSS. (A) Experimental design of MNase-ChIP-seq. **(B)** Metagene analysis of ChIP-seq signal for expressed genes after treatment of cells with DMSO (grey) versus 1-NM-PP1(NM) treatment (red) for 30 minute. The ChIP-seq coverage was merged and aligned at their transcription start sites (TSSs). Shaded areas around the average signal (solid lines) indicate confidential intervals for metagene profiles.**(C)** ChIP-seq signal around TSS at the RPL10A gene locus before (grey) and after (red) 30 minute CDK7 inhibition. Two biological replicates were merged.

7.5 Efficient release of preinitiation factors requires CDK7 activity

All experiments and analysis in this section were conducted by Dr. Iwan Parfentev. (Bioanalytical Mass Spectrometry Group, Max Planck Institute for Multidisciplinary Sciences ,Göttingen; Institute of Clinical Chemistry, University Medical Center Göttingen, 37075 Göttingen, Germany)

To gain further insights into CDK7 kinase function, we performed proteomics experiments with 30-minute inhibition of CDK7as cells (**Figure 7.6A**). Chromatin phosphoproteomics upon CDK7 inhibition allowed us to identify the changes in chromatin-associated protein composition and phosphorylation status of proteins in respect to inefficient transcription initiation events.

Chromatin proteomics analysis for each condition identified 4507 proteins amongst which 18 were most affected by inhibition of CDK7 (adj p-value < 0.05) (**Figure 7.6B**). We found a significant increase in all TFIIH kinase module components including CDK7, cyclin H, and MAT1. Moreover, upstream components of the transcription initiation machinery such

as general transcription factor IIE (TFIIE) and general transcription factor IIF (TFIIF) subunits were also significantly enriched. Collectively, the recruitment of these proteins highlights increased cellular demand for transcription initiation factors, potentially trying to overcome an impaired Pol II initiation. Curiously, the subset of significantly decreased proteins was strongly dominated by the NELF family proteins. Specifically, NELF A, NELF B, NELF C/D and NELF E were lost from the chromatin fraction hinting at putative compensations via Pol II promoter-proximal pausing release. Furthermore, we identify a reduction of anti-termination factors SCAF4 and SCAF8, which require Ser2P and Ser5P for Pol II association and suppress early polyA site selection and termination.

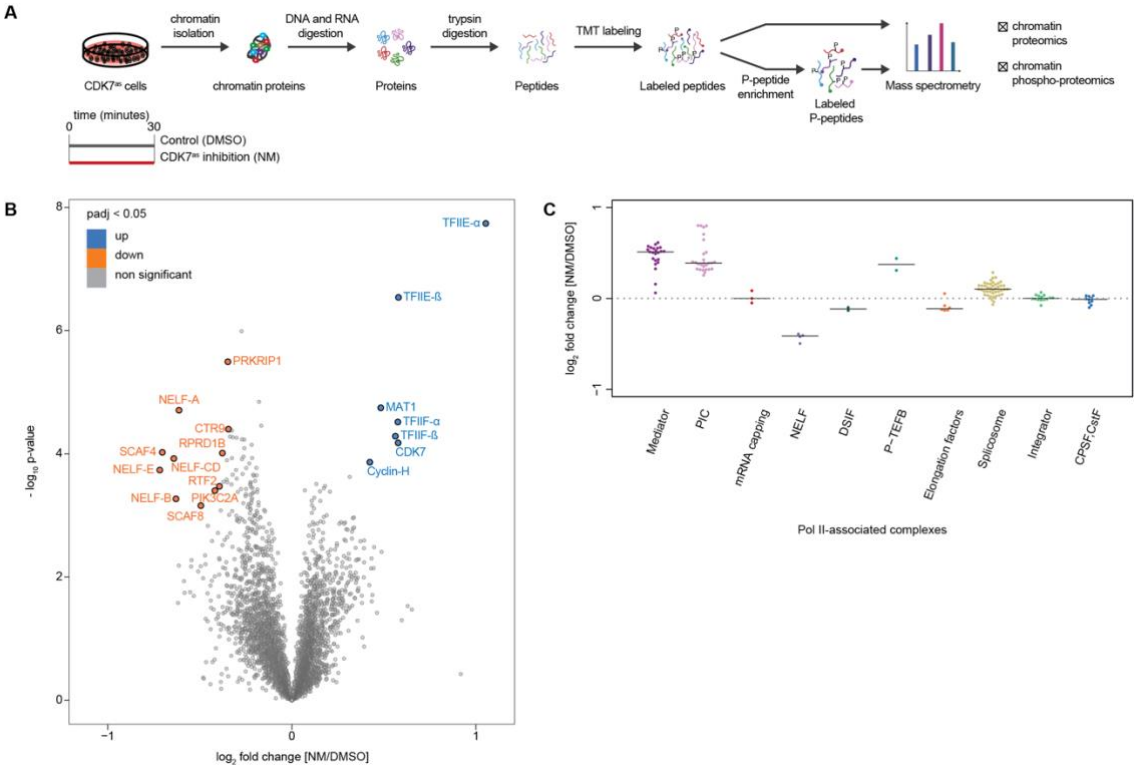


Figure 7.6 | Efficient release of preinitiation factors requires CDK7 activity. (A) Experimental design of quantitative proteomics. (B) Log_2 fold change versus average abundance ($-log_{10}$ p-value) plot showing chromatin proteomics analysis from 3 independent experiments. With adj p-value < 0.05, enriched proteins and lost proteins from chromatin are shown in blue and orange, respectively. Other proteins are shown in dark grey. (C) Log_2 fold change of Pol II-associated complexes. Each dot indicates a subunit of each complex that was detected in our proteomics data. Dashed lines indicate no change.

Next, we normalized protein levels to RNA Pol II abundance in each condition and examine relative changes of RNA Pol II-associated complexes (**Figure 7.6C**). Chromatin binding of mRNA capping, splicing, integrator and termination complexes were largely unchanged. In accordance with the previous results, levels of NELF, DSIF and elongation factors were decreased while mediator and PIC-associated GTFs were enriched in the chromatin.

Strikingly, we also observed enrichment of P-TEFB which might explain the significant loss of NELF from the chromatin as a potential compensation mechanism.

To understand how CDK7 kinase activity regulates transcription initiation, we analyzed phosphor-proteome of the chromatin-bound proteins upon its inhibition. We identified 13,030 phosphopeptides amongst which we found 281 phosphosites (100 proteins) to be significantly decreased (adj p-value < 0.05) upon inhibition of CDK7, whereas 92 phosphorylations (55 proteins) increased (**Figure 7.7A**). Analyzing all significantly changed phosphopeptides, we detected a significant reduction in RPB1 CTD phosphorylation, while the phosphorylation status of PIC-associated GTFs remained largely unchanged. The RPB1 phosphorylation changes are largely affecting serine residues aligning with the function of CDK7 as a Ser5 and Ser7 phosphorylating kinase. Specific mapping of the altered phosphosites to CTD positions remains impossible due to the inability of common MS setups to reliably identify CTD peptides. Focusing further on Pol II transcription-related proteins, we found that CDK7 phosphorylation is decreasing on Ser164 upon its inhibition. This residue constituting part of the CDK7 T-loop has been described as critical for CDK7-cyclin-H-MAT1-complex stability and CTD kinase activity *in vivo* (Larochelle *et al.*, 2001).

Contrastingly, we found increased phosphorylation levels of transcription elongation factor NELFE. NELFE phosphorylation on Ser115 was shown to facilitate its dissociation from the chromatin, aiding Pol II pause release and the transition to elongation (Borisova *et al.*, 2018). While no direct effect of CDK7 inhibition can be assumed, we predict a compensatory effect aiming to reduce pausing by NELF phosphorylation (Wu *et al.*, 2003; Lu *et al.*, 2016).

Aligned structures of PIC and paused transcription complexes showed that CDK7 as a part of CDK-activating kinase (CAK) binds the same Pol II interface as NELF, and pre-initiation factors TFIIE and TFIIF bind at the same Pol II location as DSIF (**Figure 7.7B**). Thus, for transitioning from initiation to productive elongation, release of pre-initiation factors is essential. In agreement with this model, *in silico* reconstruction of the structure of the PIC indicated a close location of the CAK to TFIIE (He *et al.*, 2013). After TFIIE release, DSIF would be loaded on Pol II together with NELF to induce Pol II pausing.

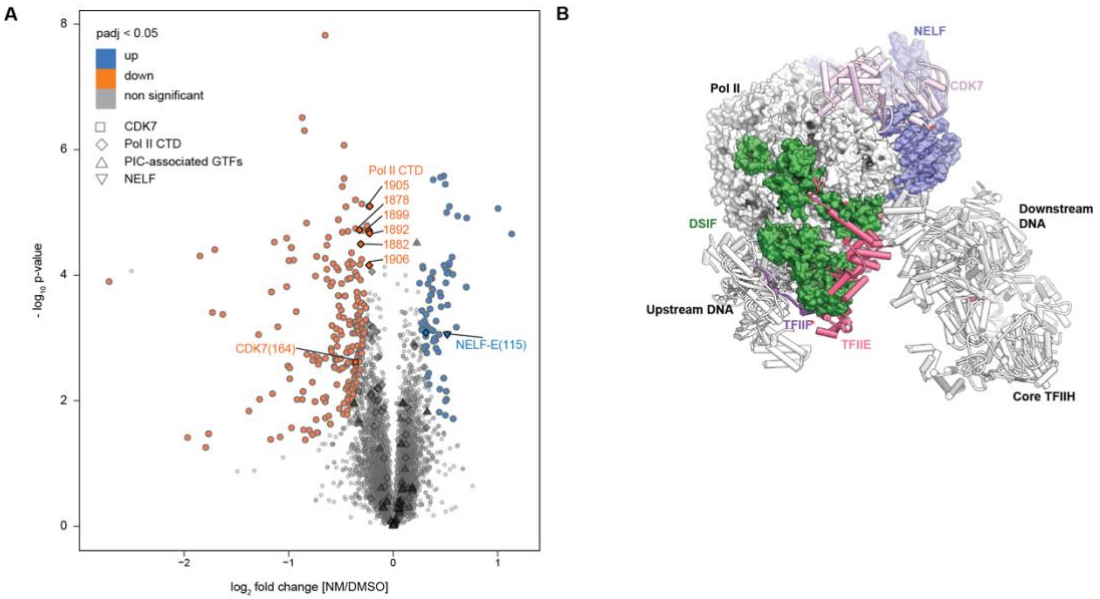


Figure 7.7 | Efficient release of preinitiation factors requires CDK7 activity. (A) Volcano plot showing the CDK7 phosphoproteome upon inhibition. Most significantly up- and downregulated (adj p-value < 0.05) phosphopeptides are annotated in blue and orange, respectively. (B) Aligned structures of PIC and paused transcription complexes.

Taken together, the proteomics analysis suggests a possible mechanism of CDK7 kinase activity for the efficient release of initiation factors with further recruitment of elongation factors. Inhibited CDK7 cannot phosphorylate CTD of Pol II and this hinders the dissociation of the pre-initiation factors. This lack of CDK7 activity also results in stalled Pol II around TSS and increased occupancies of TFIIE and Mediator downstream of TSS.

7.6 Defective elongating Pol II struggles to transcribe through nucleosomes

All experiments in this section were conducted by Taras Velychko.
(Department of Molecular Biology, Max Planck Institute for Multidisciplinary Sciences, Göttingen)

Next, we sought to gain insight into how the defective exchange of initiation factors for elongation factors affects the transcription kinetics of RNA Pol II during the early stages of transcription. Our proteomics data suggests that in the absence of CDK7 kinase activity DSIF and NELF loading on Pol II is compromised. Loading of DSIF together with NELF on Pol II induces Pol II promoter-proximal pausing and CDK7 has been implicated in two opposing functions with respect to pause establishment and release (Larochelle *et al.*, 2012; Kwiatkowski *et al.*, 2014; Ebmeier *et al.*, 2017). Downstream of the TSS, Pol II accumulation is an indicator of Pol II arrest or pausing, which can be observed with Pol II occupancy profiles obtained by mNET-seq.

To investigate the role of CDK7 in promoter-proximal pausing, we inspected our mNET-seq profiles downstream of the TSS (**Figure 7.8A**), where the highest Pol II accumulation

peak is shifted slightly downstream upon inhibition. Loss of NELF has previously been shown to regulate a distinct early elongation step associated with the +1 nucleosome acting as a barrier rather than with promoter proximal pausing (Aoi *et al.*, 2020). With respect to the loss of NELF subunits from our proteomics data, we performed MNase-seq after 15-minute inhibition of CDK7as cells. MNase-seq metagene profiles aligned at the TSS that the observed downstream shift may correspond to a nucleosome barrier during transcription (**Figure 7.8B**).

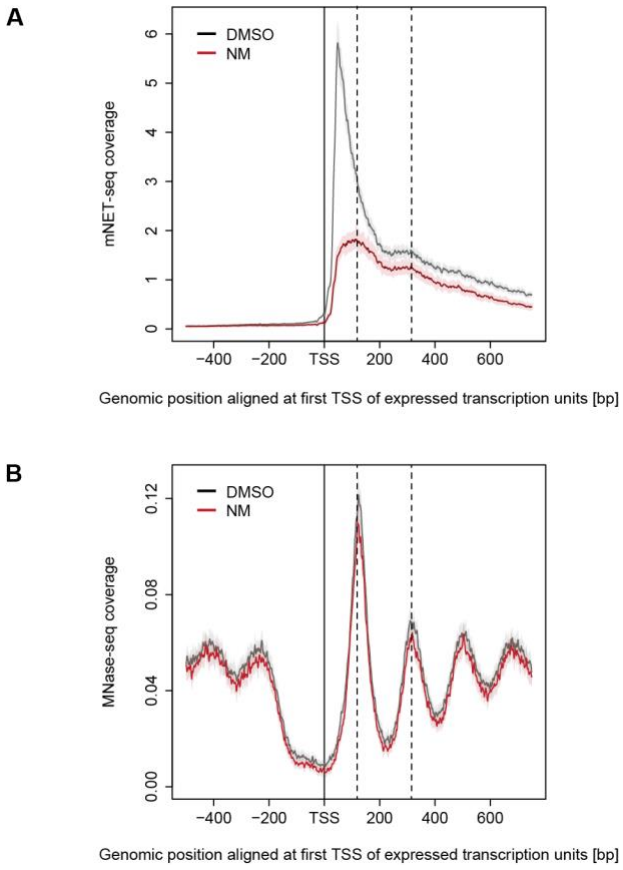


Figure 7.8 | Defective elongating Pol II struggles to transcribe through nucleosomes. (A) Metagene analysis of mNET-seq signal for expressed genes after treatment of cells with DMSO (grey) versus 1-NM-PP1(NM) treatment (red). The mNET-seq coverage was merged and aligned at their transcription start sites (TSSs). Shaded areas around the average signal (solid lines) indicate confidential intervals for metagene profiles. (B) Metagene analysis of MNase-seq signal for expressed genes after treatment of cells with DMSO (grey) versus 1-NM-PP1(NM) treatment (red). The mNET-seq coverage was merged and aligned at their transcription start sites (TSSs). Shaded areas around the average signal (solid lines) indicate confidential intervals for metagene profiles.

To further investigate this, we quantified the Pol II residence time in the promoter-proximal pause region and in the +1 nucleosome and +2 nucleosome arrest regions. Whereas Pol II residence time in the promoter proximal pause region decreased (**Figure 7.9A)**, the residence time upstream of +1 **(Figure 7.9B)** and +2 nucleosomes increased (**Figure 7.9C**).

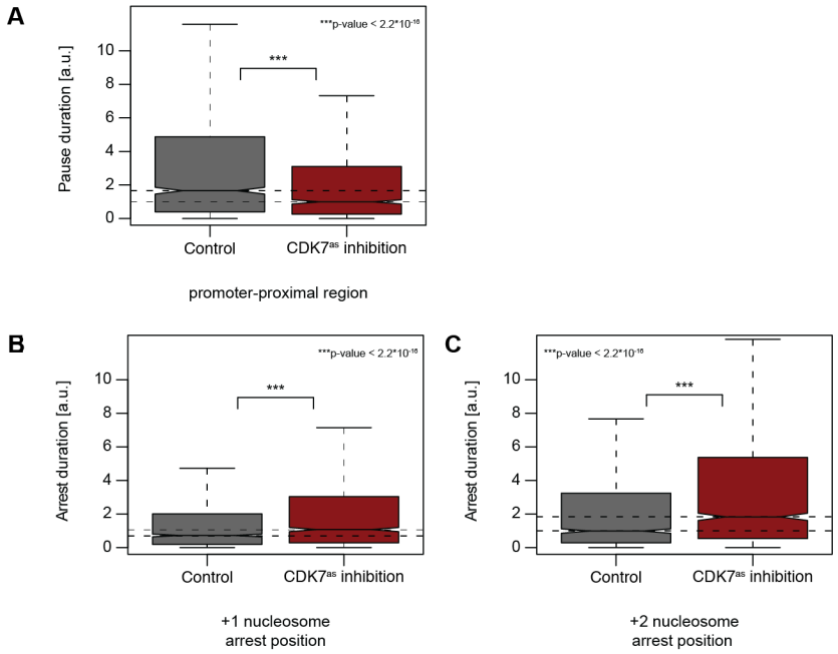


Figure 7.9 | Defective elongating Pol II struggles to transcribe through nucleosomes. (A) Distributions of gene-wise pause duration *d* [a.u.] before (control; grey) and after (red) CDK7 inhibition on promoter proximal pause position. (B) Distributions of gene-wise pause/arrest duration *d* [a.u.] before (control; grey) and after (red) CDK7 inhibition on+1 nucleosome arrest position. (C) Distributions of gene-wise pause/arrest duration *d* [a.u.] before (control; grey) and after (red) CDK7 inhibition on+2 nucleosome arrest position.

In accordance with our proteomics data, these results suggest that in absence of CDK7 kinase activity Pol II is unable to establish promoter proximal pausing due to defective exchange of initiation factors with elongation factors. These observations are further supported by ChIP coverage changes of Med26, initiation factors TFIIB and TFIIE and pausing factor NELF upon CDK7 inhibition (**Figure 7.10B-E**). Thus, Pol II travels with the associated mediator subunit, initiation factors TFIIE and TFIIB and consequently struggles to transcribe through nucleosomes for productive elongation.

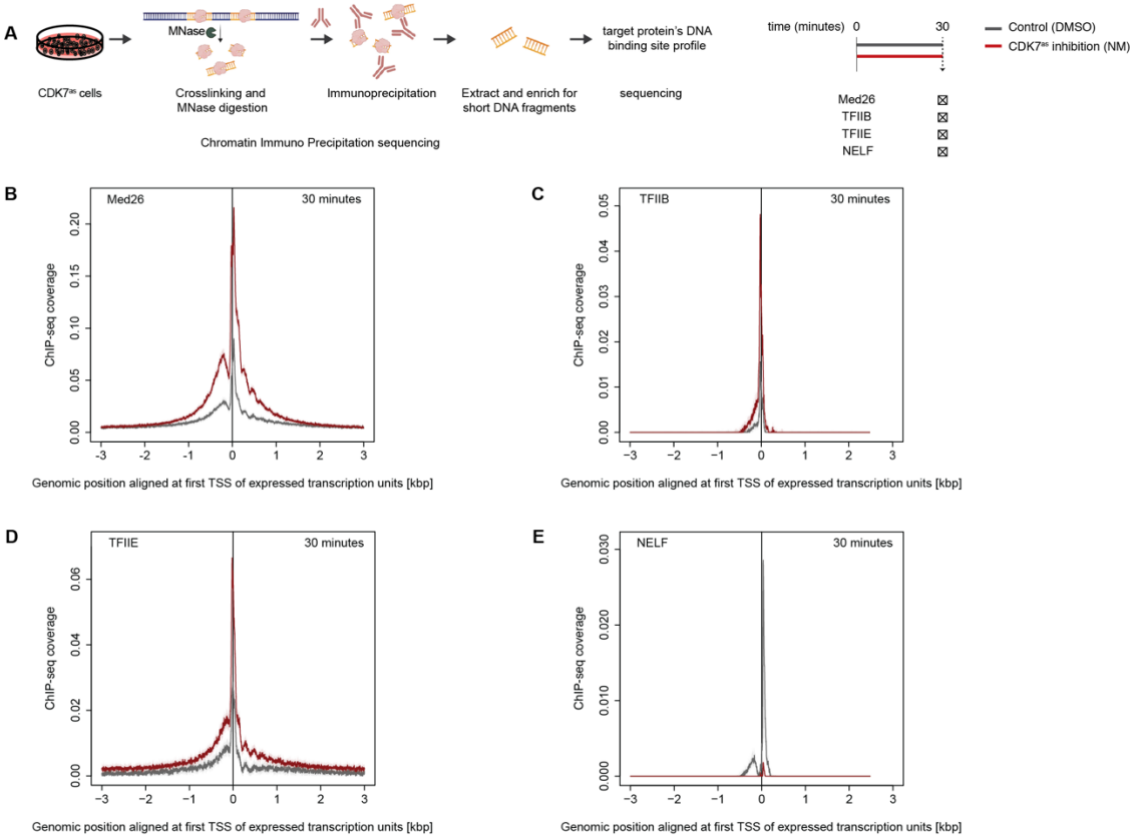


Figure 7.10 | Efficient release of preinitiation factors requires CDK7 activity. (A) Experimental design of MNase-ChIP-seq. (B-E) Metagene analysis of ChIP-seq signal for expressed genes after treatment of cells with DMSO (grey) versus 1-NM-PP1(NM) treatment (red) for 30 minute. The ChIP-seq coverage was merged and aligned at their transcription start sites (TSSs). Shaded areas around the average signal (solid lines) indicate confidential intervals for metagene profiles.

This loss of productive elongation is supported by a decreased elongation velocity (**Figure 7.11**).

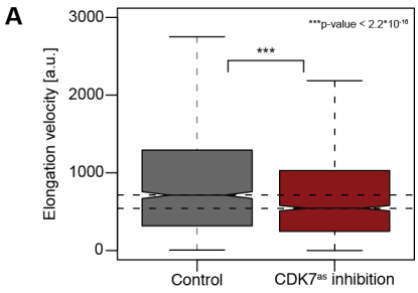


Figure 7.11 | Distributions of gene-wise elongation velocity [a.u.] before (control; grey) and after (red) CDK7 inhibition.

Together, our results show that CDK7 is a global regulator of efficient transcription initiation and suggest the kinase activity of CDK7 is required to maintain a productive transcription cycle by an efficient exchange of factors during the early stages of transcription.

DISCUSSION

In this study, we have used CDK analog sensitive cell line for specific inhibition of CDK7 in human cells. Multi omics analysis monitor genome-wide immediate changes in RNA synthesis from TT-seq, global changes from RNA-seq, actively transcribing Pol II from NET-seq and stalled Pol II from ChIP-seq. Our multi-omics data ranges from 15 minutes up to 60 minutes which can capture both primary and secondary effect on transcription dynamics upon CDK7 inhibition. In contrast to previous studies which have reported that CDK7 kinase activity is dispensable for global transcription (Ganuza et al. 2012; Kanin et al. 2007), our genome-wide experiment results suggests that CDK7 kinase activity is required for global transcription initiation. To further delineate the role of CDK7 kinase activity, we combined our multi-omics data with proteomics data. Quantitative proteomics results provide insight into a probable mechanism of CDK7 mediated transcription initiation regulation. Our proteomics results show an enrichment of initiation factors and depletion of pausing and elongation factors upon CDK7 inhibition. This result suggest that CDK7 can play role in exchange of initiation factors with elongation factors in agreement with previous predictions (Compe et al. 2019; He et al. 2013; Larochelle et al. 2012; Nilson et al. 2015). This result is also supported by genome wide transcription factor occupancy measurements by ChIP-seq which shows an increased occupancy of basal transcription factor Med26, TFIIB, TFIIE and decreased occupancy of NELF. This results is significant as mediator interacts with CDK7, TFIIE, and TFIIH and is involved in promoter-proximal pausing (Conaway and Conaway 2013; Rengachari et al. 2021). The role of CDK7 in promoter-proximal pausing is questionable (reviewed in (Coin and Egly 2015; Fisher 2019)). Our multi-omics analysis of Pol II occupancy profiles and nucleosome occupancy profiles suggests due to an inefficient exchange of initiation factors to elongation factors in the absence of CDK7 kinase activity, Poll II pause duration in promoter proximal region is decreased and Pol II arrest duration near +1 nucleosome increases. Based on all our observation and analysis, we suggest a model for CDK7 kinase activity (**Figure 7.12**). In the absence of CDK7 kinase activity, mediator, TFIIE and TFIIB cannot dissociate from PIC. This results in defective loading of NELF and DSIF on Pol II, which is unable to establish promoter proximal pause resulting in decreased pausing in this region. The PIC travels with basal transcription factors and gets arrested by +1 nucleosome due to the absence of elongation factors. Taken together, CDK7 is required in human cells to maintain efficient transcription regulation.

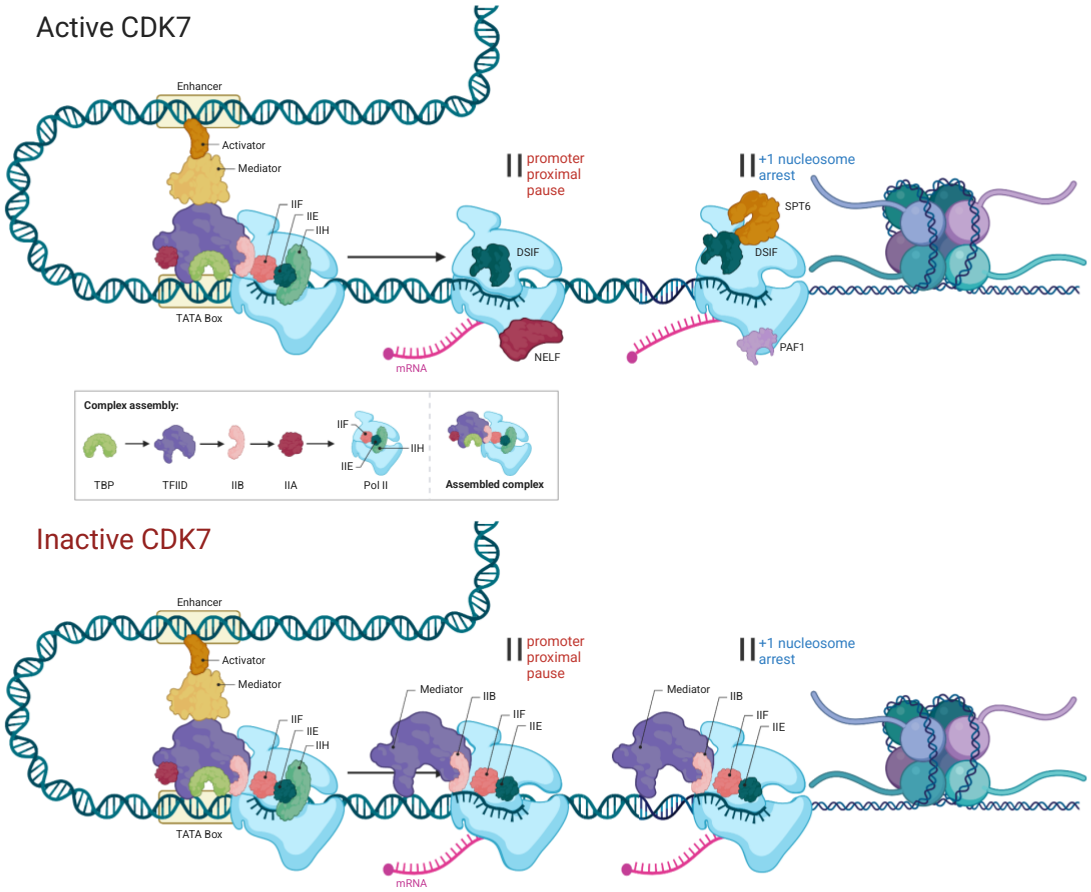


Figure 7.12 | Proposed model of CDK7 function in transcription (as described in discussion section).

CONCLUDING REMARKS

In conclusion, our data for the first time suggests that specific and rapid inhibition of CDK7 causes a global initiation defect demonstrated by multi-omics analysis of TT-seq and mNET-seq data which is further supported by Pol II ChIP-seq. Additional proteomics and ChIP studies imply a plausible mechanism for this defect. Taken together, our data suggests, phosphorylation by CDK7 prompts basal transcription factor exchange with elongation factors for the transition of transcription initiation to elongation. Future studies may gather additional justification for the initiation to elongation switch by CDK7. Finally, our results imply that CDK7 is a global regulator of Pol II mediated transcription initiation. Prospective structural and functional studies are necessary to validate our findings to unravel molecular mechanisms by which CDK7 contributes to global downregulation of initiation.

The CDK8 kinase module regulates Mediator-RNA polymerase II interaction

Sara Osman, **Eusra Mohammad**, Michael Lidschreiber, Alexandra Stuetzer, Fanni Bazsó, Kerstin Maier, Henning Urlaub, Patrick Cramer

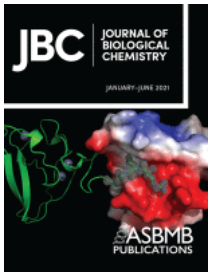
Distinctive contribution:

Co-authorship for contribution to

- the analysis of *in vivo* data (4tU-seq)
- the pertinent write-up of the analysis

Manuscript publication status

Manuscript availability



Journal of Biological Chemistry (JBC)
Volume 296
January-June 2021
Section: Gene Regulation

DOI: 10.1016/j.jbc.2021.100734

The project was initialized with focus on biochemical and structural *in vitro* study of CDK8 kinase module (CKM). In line with the observations of the aforementioned studies, the *in vivo* experiment and analysis were conceived as an ɪn'haʊs (internal) collaboration *(Department of Molecular Biology, Max Planck Institute for Multidisciplinary Sciences, Göttingen).* The collective *in vitro* and *in vivo* study was further culminated into the published manuscript.

The results presented in this section comprises of the *in vivo* experiment and analysis performed during the preparation of the manuscript. Experiments were not performed by the author of this dissertation, but are included in this section for a coherent presentation of the obtained findings. Contribution in conducting the experiments can be found at the beginning of each subsection (*in italic*). Detailed author contributions can be found on publications.

The published text was adapted to match the style of this thesis. Numbering and references to figures as well as references to the literature thus deviate from the published version.

PROJECT SUMMARY

This section is a modified excerpt from (Osman *et al.*, 2021)

All experiments in this section were conducted by Dr. Sara Osman (in vivo experiments were conducted together with Dr. Kerstin Meyer)
(Department of Molecular Biology, Max Planck Institute for Multidisciplinary Sciences, Göttingen)

The CDK8 kinase module (CKM) is a dissociable part of the coactivator complex mediator, which regulates gene transcription by RNA polymerase II. The CKM has both negative and positive functions in gene transcription that remain poorly understood at the mechanistic level. In order to reconstitute the role of the CKM in transcription initiation, we prepared recombinant CKM from the yeast *Saccharomyces cerevisiae.* We showed that CKM bound to the core mediator (cMed) complex, sterically inhibiting cMed from binding to the polymerase II preinitiation complex (PIC) *in vitro.* We further showed that the CDK8 kinase activity of the CKM weakened CKM-cMed interaction, thereby facilitating dissociation of the CKM and enabling mediator to bind the PIC in order to stimulate transcription initiation. We next advanced our biochemical and structural *in vitro* observations of CDK8 kinase module (CKM) with *in vivo* 4tU-seq experiments observations to explain a plausible model for the functions of the CDK8 kinase module (CKM) with mechanistic detail.

RESULTS

Text in this section is an unmodified excerpt from. (Osman *et al.*, 2021). Figures in this section are adapted from the manuscript version and presented in a simultaneous manner with results. Additional figures are included for comprehensive understanding of the results.

All experiments in this section were conducted by Dr. Kerstin Meyer and Dr. Sara Osman (Department of Molecular Biology, Max Planck Institute for Multidisciplinary Sciences, Göttingen)

8.1. CDK8 kinase activity is required for transcription activation during heat shock

To investigate the role of the CDK8 kinase activity *in vivo*, we used a CDK8 analog-sensitive yeast strain (Liu *et al.*, 2004a) allowing us to specifically inhibit CDK8 using the ATP analog 1-Naphthyl-PP1 (1-NA-PP1). We applied either 1-NA-PP1 at a final concentration of 6 µM or dimethyl sulfoxide (DMSO) for 12 min, followed by 5 min of RNA labeling with 4-thiouracil (4tU) (**Figure 8.1**).

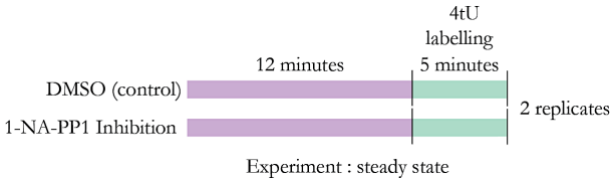


Figure 8.1 | Experimental setup of the steady state 4tU-seq experiment in *Saccharomyces cerevisiae*. 4tU-seq experiments were performed in two independent biological samples.

This allowed us to isolate newly transcribed RNA following the perturbation. We then sequenced the newly synthesized RNA and mapped it to the *Saccharomyces cerevisiae* genome (sacCer3). We performed each condition in two biological replicates and added spike-ins before RNA extraction to normalize RNA counts for quantification. The biological replicates correlated well (**Figure 8.2**).

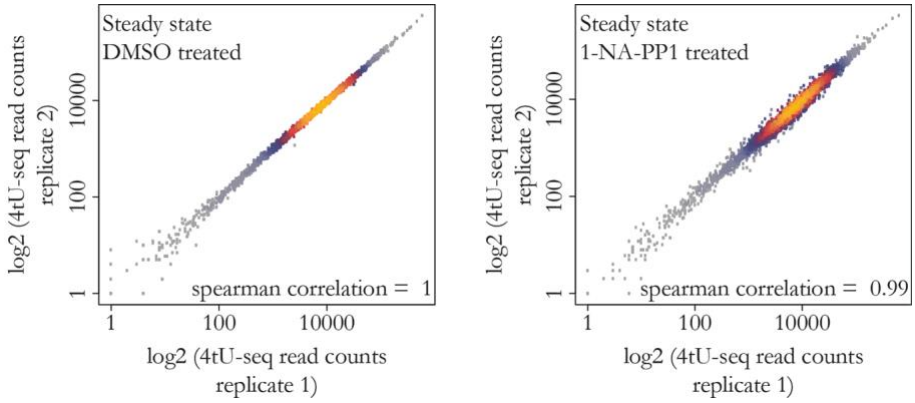


Figure 8.2 | Assessment of reproducibility of 4tU-Seq data (steady state). Comparison of replicate measurements for 4tU-seq of the DMSO samples (left panel) and 1-NA-PP1 samples (right panel) under steady state conditions. The scatterplot compares read counts of ORF-Ts. Spearman correlation coefficient is mentioned in the respective figures.

We found no statistically significant changes resulting from CDK8 inhibition under steady-state growth conditions, as demonstrated by genome-wide differential gene expression analysis comparing the inhibited sample to the DMSO control (**Figure 8.3**).

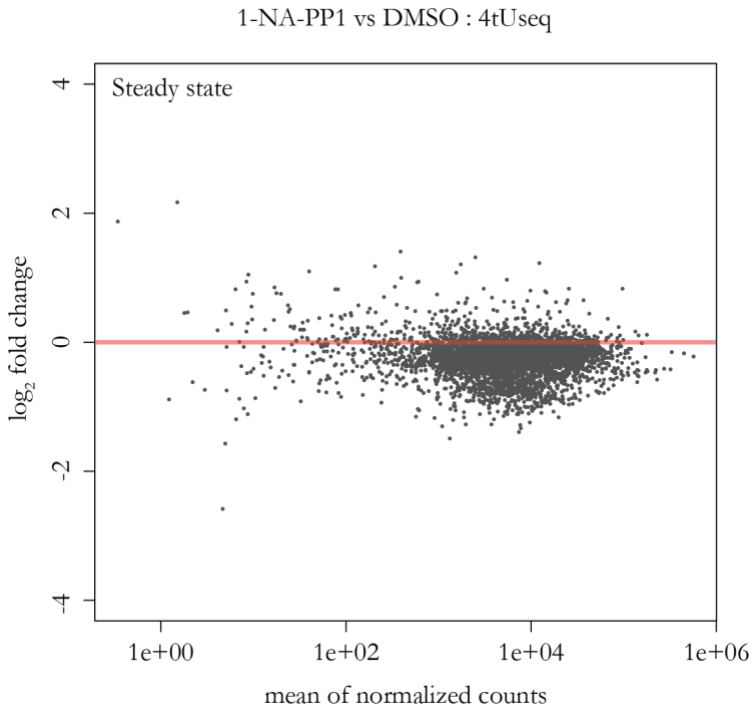


Figure 8.3 | MA plot shows differential gene expression between the CDK8 inhibited sample (1-NA-PP1) and DMSO control for all protein coding genes (n = 4928) under steady-state conditions. Log_2 fold changes are plotted against the mean of spike-in normalized counts over all samples. No significantly changed genes were detected (adjusted p value < 0.1).

We then performed a second experiment, where we once again applied 1-NA-PP1 or DMSO for 12 min and then raised the temperature of the yeast cultures to 37 °C to impose heat shock for 12 min (**Figure 8.4**). The biological replicates of the heat shock (HS) 4tU-seq experiment correlated well (**Figure 8.5**).

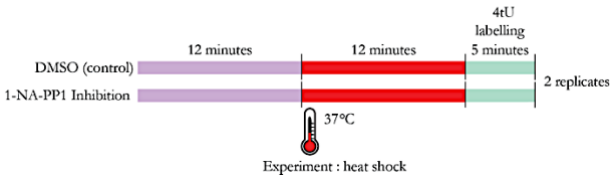


Figure 8.4 | Experimental setup of the heat shock (HS) 4tU-seq experiment (at 37 °C) in *Saccharomyces cerevisiae*. 4tU-seq experiments were performed in two independent biological samples.

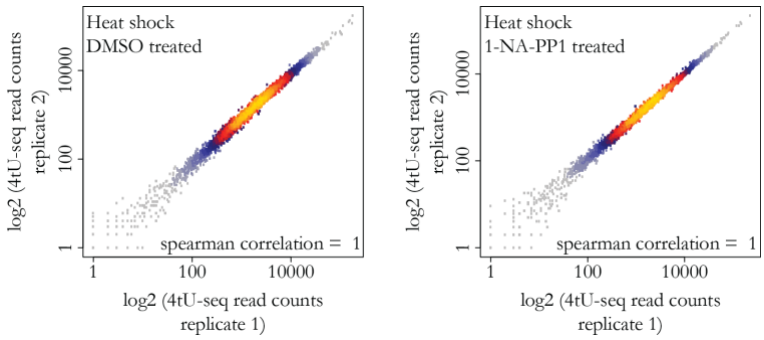


Figure 8.5 | Assessment of reproducibility of 4tU-Seq data upon heat shock (HS). Comparison of replicate measurements for 4tU-seq of the DMSO samples (left panel) and 1-NA-PP1 samples (right panel) under heat shock conditions. The scatterplot compares read counts of ORF-Ts. Spearman correlation coefficient is mentioned in the respective figures.

The effect of heat shock treatment was validated by genome-wide differential gene expression analysis comparing the heat shock DMSO sample to the steady-state DMSO sample (**Figure 8.6**). Differential gene expression analysis identified significant upregulation of 7.7% and downregulation of 37% of genes, matching the expected pattern for the heat shock response in yeast.

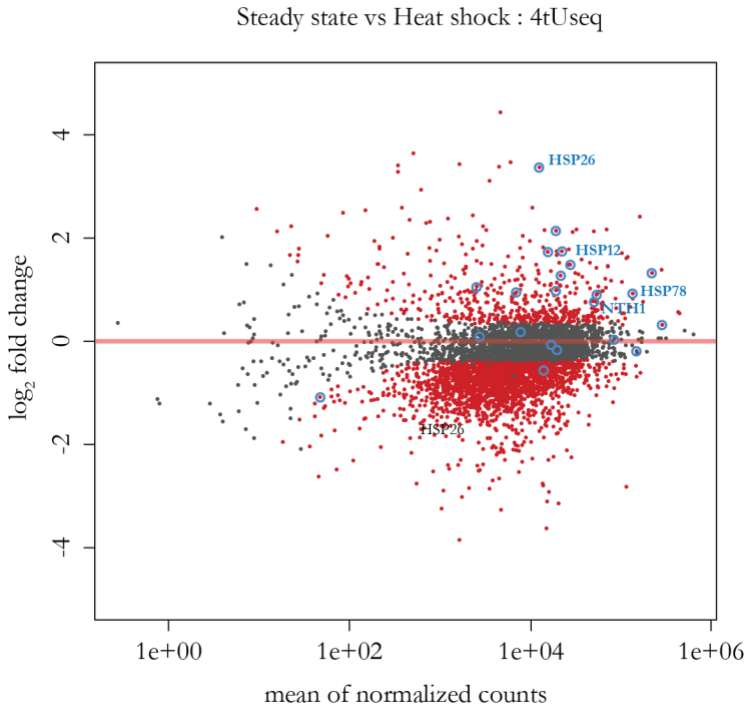


Figure 8.6 | MA plot shows differential gene expression between 12-min heat shock DMSO samples and the steady-state DMSO samples. Significantly changed genes are shown in red (adjusted p value < 0.1). Temperature response genes (Gene Ontology: 0009266) are encircled in blue.

Gene ontology analysis showed that heat shock genes (encircled in blue) are enriched within the induced group (**Figure 8.7**).

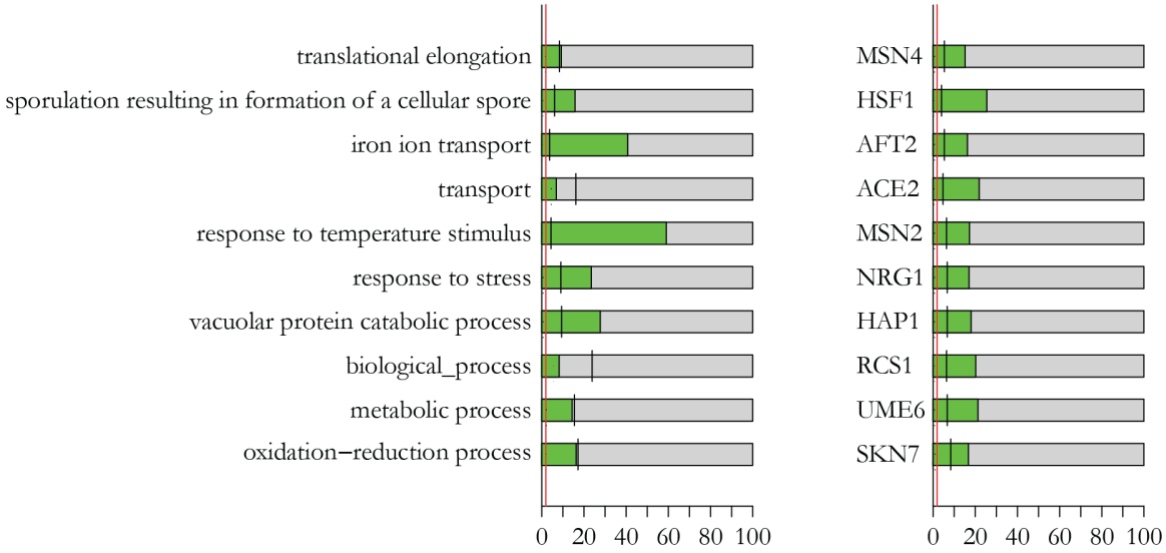


Figure 8.7 | Gene ontology analysis of genes induced upon application of heat shock shows that temperature stimulus response genes (GO:0009266) are enriched in this group (left). Specific transcription factor target gene groups that were found to be enriched are shown on the right.

Under this condition, we found that CDK8 inhibition resulted in a global downregulation of gene transcription by ~1.5-fold (**Figure 8.8**), as indicated by differential expression analysis of the inhibited sample compared with the DMSO control. More specifically, the group of genes induced during the heat shock response in yeast fails to be induced upon CDK8 inhibition (**Figure 8.9**).

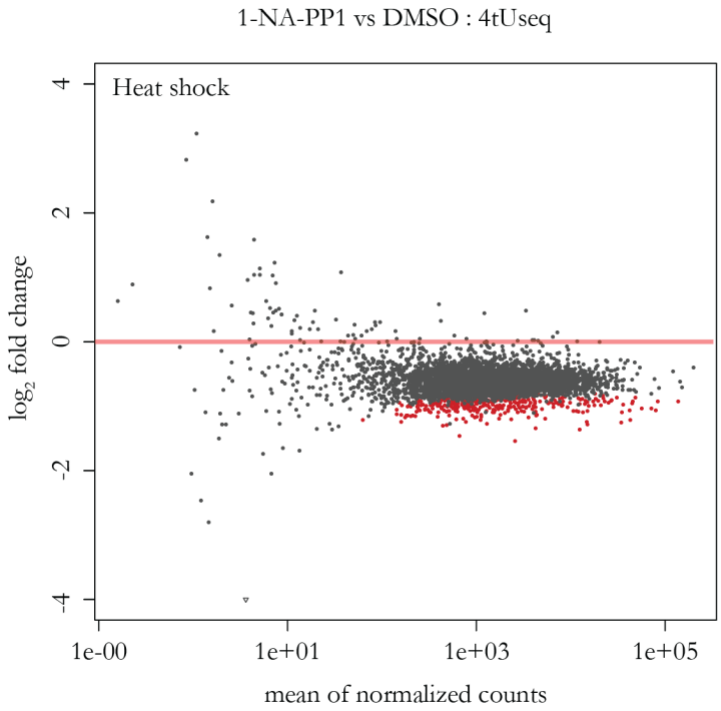


Figure 8.8 | MA plot shows differential gene expression between the CDK8 inhibited sample and DMSO control under heat shock conditions. Significantly changed genes are shown in red (adjusted p value < 0.1).

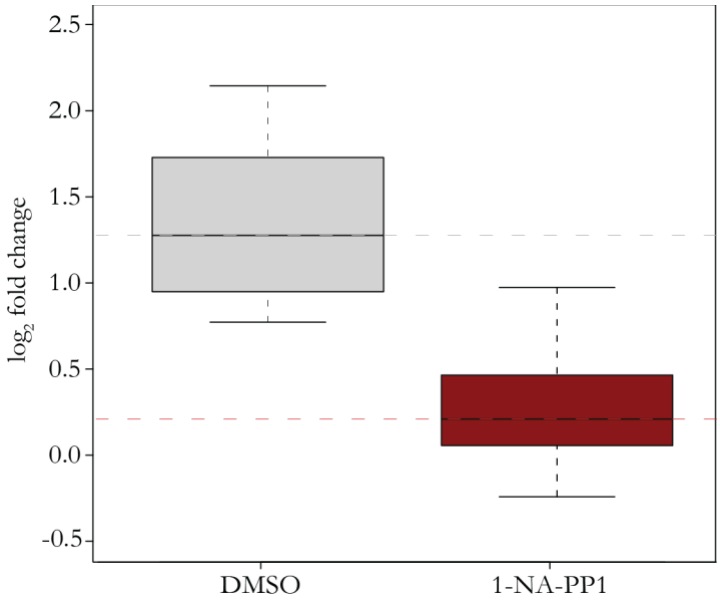


Figure 8.9 | CDK8 activity is required for transcription activation during heat shock. Box plots showing log_2 fold changes in expression of induced heat shock response genes (**Figure 8.6**) comparing 12-min heat shock DMSO samples to the steady-state DMSO samples (left) and comparing the 12-min heat shock CDK8 inhibited samples to the steady state CDK8 inhibited samples (right) (Wilcoxon rank sum test, p value = 0.0002).

These observations show that CDK8 kinase activity is required for heat shock gene activation during the heat shock response.

DISCUSSION

Text and figure in this section are a modified excerpt from. (Osman *et al.*, 2021).

The catalytic subunit of the CKM, the CDK8 kinase, has previously been shown to phosphorylate a host of targets. CDK8 phosphorylates the pol II CTD in yeast (Hengartner *et al.*, 1998; Borggrefe *et al.*, 2002; Liu *et al.*, 2004b) and human (Knuesel *et al.*, 2009). In addition, CDK8 phosphorylates various activators in both organisms. CDK8 phosphorylation of the activator Gal4 is necessary for galactose-inducible transcription (Sadowski *et al.*, 1991; Liao *et al.*, 1995; I, C and R, 1996; Hirst *et al.*, 1999; Rohde, Trinh and Sadowski, 2000; Ansari *et al.*, 2002). CDK8 phosphorylation of activators involved in gluconeogenesis and stress response have also been shown to result in gene activation (Vincent *et al.*, 2001; Lenssen *et al.*, 2007). Conversely, CDK8 phosphorylation of other activators involved in pathways of starvation and stress response has been found to increase their turnover or nuclear exclusion (Chi *et al.*, 2001; Nelson *et al.*, 2003; Raithatha *et al.*, 2012). It has remained unclear how the presence and phosphorylation activity of the CKM regulates transcription. The CKM is generally construed as a repressive molecule, but a large body of evidence has affirmed a positive role of its CDK8 kinase on gene transcription in both yeast (Sadowski *et al.*, 1991; I, C and R, 1996; Hirst *et al.*, 1999; Rohde, Trinh and Sadowski, 2000; Vincent *et al.*, 2001; Liu *et al.*, 2004b; Lenssen *et al.*, 2007) and human (Chen *et al.*, 2017; Galbraith *et al.*, 2017; Garibaldi, Carranza and Hertel, 2017; Steinparzer *et al.*, 2019). Indeed, we report that the kinase activity of CDK8 is required for gene activation during the stressful condition of heat shock *in vivo* but not under steady-state growth conditions. We found that CDK8 inhibition impaired induction of heat shock genes in yeast, corroborating a positive role in gene activation. This apparent paradox is mitigated by our biochemical observation that the CDK8 kinase activity facilitates the release of the CKM from mediator, which would free up its pol II-interacting surface needed for gene activation. As such, active CDK8 ensures that mediator is free to establish its interaction with the PIC to stimulate transcription initiation.

In summary, the CKM plays a dual function—a repressive structural function and an activating CDK8-dependent function that counteracts repression during stress response. Our results converge with previous literature on a simple model for the role of the CKM module in gene activation (**Figure 8.10**) (Allen and Taatjes, 2015). A stable complex of activators, mediator and the CKM is present at UASs (or at enhancers in metazoan cells).

This complex may represent an inactive form of mediator that is unable to activate genes because of masking of its pol II-interacting surface by the CKM. This may explain the presence of the mediator complex on UASs or enhancers, irrespective of the activity status of their target genes, in both yeast (Andrau *et al.*, 2006) and human (Grünberg *et al.*, 2016).

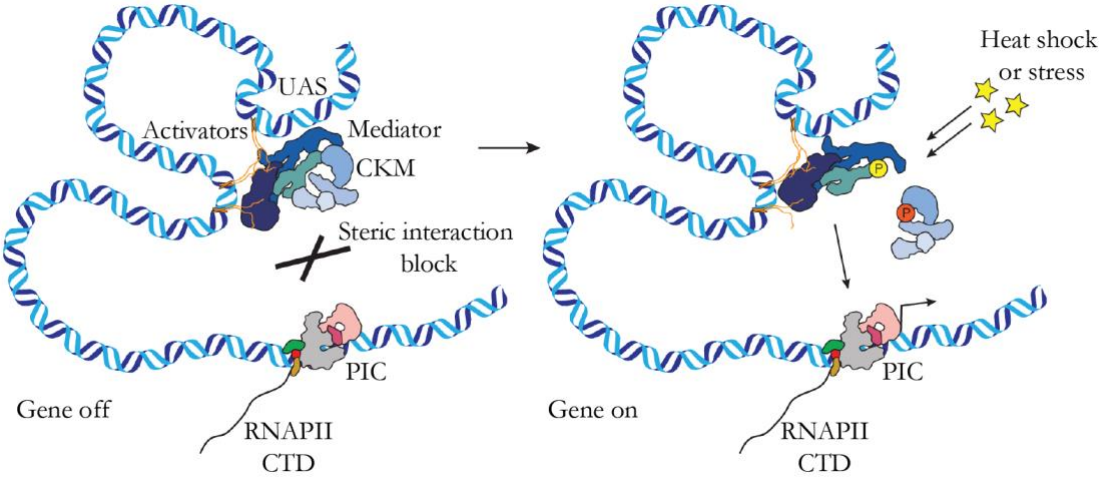


Figure 8.10 | Model of CKM steric repression and Cdk8-mediated release during stress response. CKM occludes mediator's PIC interacting interface resulting in a "gene off" status (left). Upon heat shock, Cdk8-dependent phosphorylation frees up the mediator-PIC interaction block, resulting in a "gene on" status (right).

In our model, the CKM serves to uncouple mediator presence from mediator activity at target genes by steric obstruction. This may be important to retain mediator near target genes and enable their rapid activation upon an external signal. Finally, we suggest that gene activation may involve activation of the CDK8 kinase, which weakens CKM-mediator interaction, thereby liberating mediator and enabling its binding to the PIC. Our observation that CDK8 inhibition produces more pronounced transcriptomic effects under conditions of heat shock is consistent with similar observations during nutrient stress (Holstege *et al.*, 1998) , oxidative stress (Stieg, Cooper and Strich, 2020), and in human cells (Galbraith *et al.*, 2013, 2017; Chen *et al.*, 2017) and argues that the CDK8 kinase is not constitutively active. Indeed, stress and developmental signaling cascades culminate in phosphorylating various mediator or CKM complex subunits (Chang, Howard and Herman, 2004; Van De Peppel *et al.*, 2005; Kim *et al.*, 2006). Signaling-dependent exchange of activator-bound mediator from its CKM bound to its free form has previously been reported (Van De Peppel *et al.*, 2005). The mediator complex and the rest of the CKM may thus act as a conduit of signals to the CDK8. CDK8 activity is sensitive to the location of a Med12 activation helix (Klatt *et al.*, 2020), an observation suggestive of allosteric signal transduction translated into an output of CDK8 activity. This

may allow CDK8 to support rapid activation of genes within specific transcriptional programs, imparting improved adaptability to stress, and accounting for its early occurrence and high degree of evolutionary conservation (Bourbon, 2008).

Our model describes an important facet of the complex events taking place at the UAS-promoter axis. We have distilled a general functional principle of the CKM from the yeast system, in which no additional homologs of this complex are present. However, more of the CDK8 functions in metazoans still remain unaccounted for, in congruence with the added regulatory complexity underpinning multicellularity and development (Allen and Taatjes, 2015). The evolution of homologs may have allowed a lateral diversification of regulatory programs, as evidenced by the observation that CDK8 and its homolog CDK19 control different gene sets (Steinparzer *et al.*, 2019). It also remains unclear what role the other phosphorylation targets of CDK8 in metazoans (Poss *et al.*, 2016) including histone proteins and combinatorial effects on the TFIIH kinase CDK7 (Knuesel *et al.*, 2009) and the PIC (this study) may play. In addition, CDK8/19 inhibition was found to further activate gene expression at superenhancers in pluripotent and embryonic stem cells in mouse and human (Pelish *et al.*, 2015; Lynch *et al.*, 2020), implying a different function in the superenhancer context.

CONCLUDING REMARKS

This section is a modified excerpt from (Osman *et al.*, 2021)

In summary, our observation of CDK8-dependent release of the CKM-mediator occlusion provides a crucial missing link that explains how the CKM's steric repression and the observed activating effects of its CDK8 kinase on target gene transcription can be reconciled. Based on these results, we propose a model in which the CKM negatively regulates mediator function at upstream-activating sequences by preventing mediator binding to the PIC at the gene promoter. However, during gene activation in response to stress, the CDK8 kinase activity of the CKM may release mediator and allow its binding to the PIC, thereby accounting for the positive function of CKM. This may impart improved adaptability to stress by allowing a rapid transcriptional response to environmental changes, and we speculate that a similar mechanism in metazoans may allow the precise timing of developmental transcription programs.

CDK12 globally stimulates RNA polymerase II transcription elongation and carboxyl-terminal domain phosphorylation.

Michael Tellier, Justyna Zaborowska, Livia Caizzi, **Eusra Mohammad**, Taras Velychko, Björn Schwalb, Ivan Ferrer-Vicens, Daniel Blears, Takayuki Nojima, Patrick Cramer and Shona Murphy

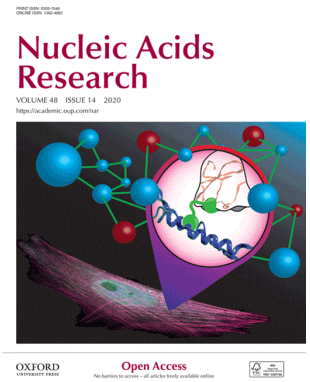
Distinctive contribution:

Co-authorship for contribution to the analysis and interpretation of

- TT-seq data
- mNET-seq data (not included in the manuscript version)
- Kinetic model of transcription
- the pertinent write-up of the analysis

Manuscript publication status

Manuscript availability



Nucleic Acids Research
Volume 48, Issue 14, 20 August 2020, Pages 7712–7727
Section: GENE REGULATION, CHROMATIN AND EPIGENETICS
DOI: https://doi.org/10.1093/nar/gkaa514

The project was conceived in collaboration with Murphy laboratory of Dr. Shona Murphy, Sir William Dunn School of Pathology, University of Oxford, United Kingdom. Experiments and analysis that were not performed by the author of this dissertation, but are included in this section for a coherent presentation of the obtained findings, are stated at the beginning of each subsection (*in italic*). Detailed author contributions can be found on publications.

The published text was adapted to match the style of this thesis. Numbering and references to figures as well as references to the literature thus deviate from the published version.

PROJECT SUMMARY

This section is a modified excerpt from (Tellier *et al.*, 2020)

Cyclin-dependent kinase 12 (CDK12) phosphorylates the carboxyl-terminal domain (CTD) of RNA polymerase II (pol II) but its roles in transcription beyond the expression of DNA damage response genes remain unclear. The project was conceived to elucidate the primary effect of CDK12 activity on transcription. To understand the function of the kinase activity of CDK12 in transcription, CRISPR/Cas9 gene engineering was used to mutate the endogenous CDK12 gene in HEK293 cells to produce an analog-sensitive CDK12 (CDK12as) that can be selectively inhibited by bulky ATP analogs (Blethrow *et al.*, 2004). The use of CDK12as cell line was combined with two complementary functional genomics methods TT-seq (Schwalb *et al.*, 2016) and mNET-seq (Nojima, Gomes, Grosso, Kimura, Michael J Dye, *et al.*, 2015; Nojima *et al.*, 2016) to monitor the direct effects of rapid CDK12 inhibition on transcription activity and CTD phosphorylation in human cells.

RESULTS

The text of this section is an unmodified excerpt from (Tellier *et al.*, 2020). Figures in this section are adapted from the manuscript version and presented in a simultaneous manner with results. Additional figures are included for comprehensive understanding of the results.

All experiments are performed by the collaborators from Murphy laboratory except TT-seq.
TT-seq was conducted by Dr. Livia Caizzi and Taras Velychko.
(Department of Molecular Biology, Max Planck Institute for Multidisciplinary Sciences, Göttingen)

9.1 Rapid and specific CDK12as inhibition in human cells

Several studies using RNAi-mediated knockdown of CDK12 have highlighted roles for this kinase in CTD phosphorylation and expression of DDR genes (Bartkowiak *et al.*, 2010; Blazek *et al.*, 2011; Davidson, Muniz and West, 2014). To assess whether these effects are

mediated by CDK12 kinase activity, the loss of CDK12 itself, or indirect effects due to the long time-frame of knockdown, we produced a CDK12as HEK293 cell line, using CRISPR/Cas9, where the gatekeeper phenylalanine residue (codon TTT) is mutated to glycine (codon GGT) (**Figure 9.1A**). This allows rapid and selective inhibition of CDK12 kinase activity with a bulky ATP-analog 1-NM-PP1 (NM) (Lopez, Kliegman and Shokat, 2014; Bartkowiak and Greenleaf, 2015; Bartkowiak, Yan and Greenleaf, 2015), without affecting levels of the protein itself, which has been shown to interact with other components of the transcription/RNA processing machinery (Liang *et al.*, 2015; Yu *et al.*, 2015; Tien *et al.*, 2017). Importantly, the growth rate of the CDK12as cells is not lower than that of the parental HEK293 cell line (**Figure 9.1B**).

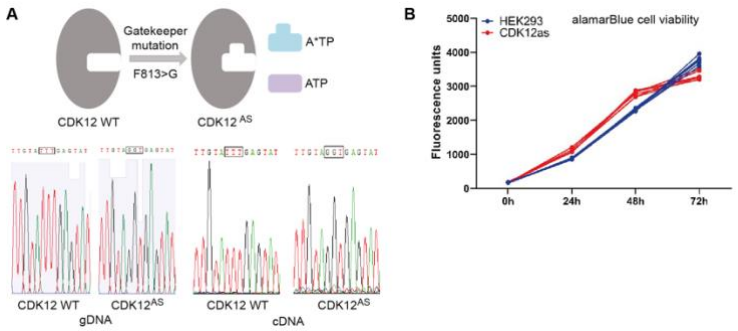


Figure 9.1 | Characterization of the CDK12as HEK293 cell line (A) Cartoon depicting the creation of CDK12as by altering the 'gate-keeper' residue in the kinase active site (left, top). CRISPR/Cas9 genome engineering was used to mutate F (TTT) to G (GGT) (left, bottom). Sanger sequencing tracks of gDNA and cDNA from CDK12 WT HEK293 control and CDK12as are shown. **(B)** Time-course of the growth of viable HEK293 and CDK12as cells after seeding, as measured using alamarBlue HS. n = 6 biological replicates.

Also, CDK12 and Cyclin K protein levels are unchanged in CDK12as cells compared to the parental cell line, as measured by western blot analyses (**Figures 9.2A and 9.2B**).

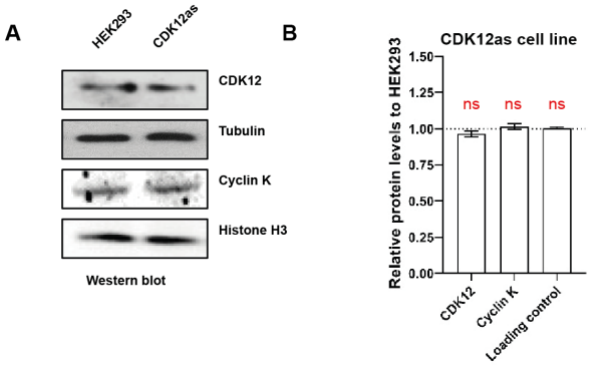


Figure 9.2 | Characterization of the CDK12as HEK293 cell line (A) Western blots of parental HEK293, and CDK12as cell whole cell extracts with the antibodies noted at the right. Tubulin and histone H3 were used as loading controls. **(B)** Relative protein levels of CDK12, Cyclin K, and loading controls in CDK12as cells compared to HEK293 cells. Error bars = s.e.m. (n = 2 biological replicates). Statistical test: two-tailed unpaired t test, ns = not significant.

To determine which concentration of NM to use, we followed the growth of CDK12as and the parental HEK293 cells after the addition of 5, 7.5, and 10 μM of NM to the medium using either an alamarBlue HS cell viability assay or continuous label-free monitoring by xCelligence (**Figures 9.3A and 9.3B**). In both cases, 7.5 μM NM affects the viability and growth of CDK12as cells with less effect on HEK293, whereas 10 μM NM affects both cell lines. This is in agreement with previously published data on CDK12as inhibition in HeLa cells (Bartkowiak, Yan and Greenleaf, 2015).

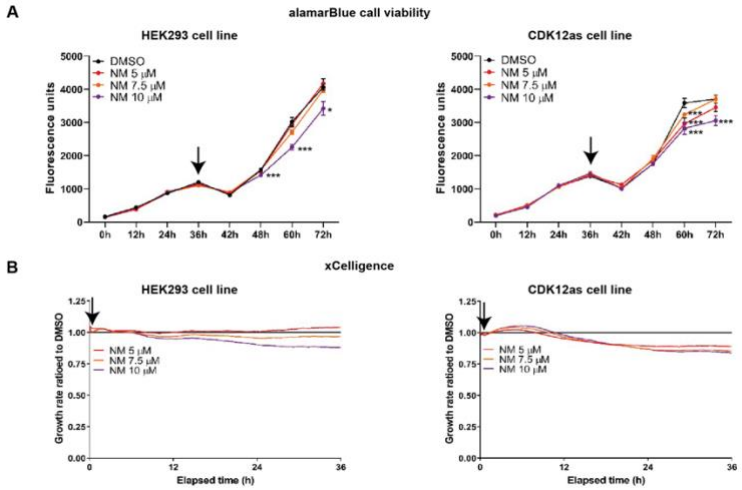


Figure 9.3 | Characterization of the CDK12as HEK293 cell line (A) Time- course of the growth of viable HEK293 and CDK12as cells after seeding using alamarBlue HS. Error bars= s.e.m. (n = 6 biological replicates). Statistical test: two- tailed unpaired t test, *** $p < 0.001$. The noted concentrations of NM were added at the 36 hours' time point as indicated by the arrows. **(B)** Time-course of the growth of the HEK293 and CDK12as cells measured by xCelligence (n = 6 biological replicates). NM was added 36 hours after seeding, indicated as the 0 hour time point on the figure and indicated by the arrows.

Importantly, short-term treatment of HEK293 and CDK12as cells with 7.5 μM NM does not affect the protein levels of CDK12 and Cyclin K (**Figures 9.4A and 9.4B**) and takes many hours to affect cell viability (**Figure 9.3A**).

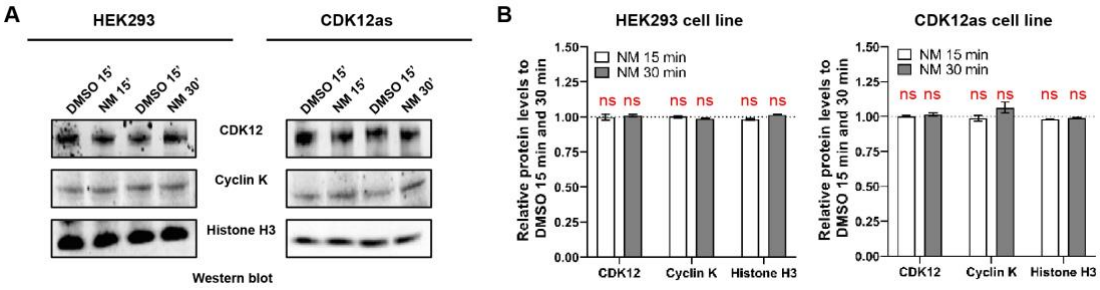


Figure 9.4 | Characterization of the CDK12as HEK293 cell line (A) Western blots of parental HEK293, and CDK12as cell whole cell extracts treated with DMSO or 7.5 μM NM for 15 or 30

minutes with the antibodies noted in the middle. Histone H3 was used as a loading control. **(B)** Quantitation of protein levels of CDK12, Cyclin K, and histone H3 in HEK293 and CDK12as cells after NM treatment relative to the control. Error bars= s.e.m. (n = 3 biological replicates). Statistical test: two-tailed unpaired t test, ns = not significant.

We have therefore treated cells with this inhibitor concentration for 15 or 30 minutes, in order to determine the immediate effect of CDK12 inhibition. Importantly, pol II distribution on expressed transcription units in the parental cells is not affected by short-term NM treatment as measured by pol II chromatin immunoprecipitation followed by quantitative PCR (ChIP-qPCR) of *KPNB1* and ChIP-sequencing (ChIP-seq) (**Figures 9.5A and 9.5B**).

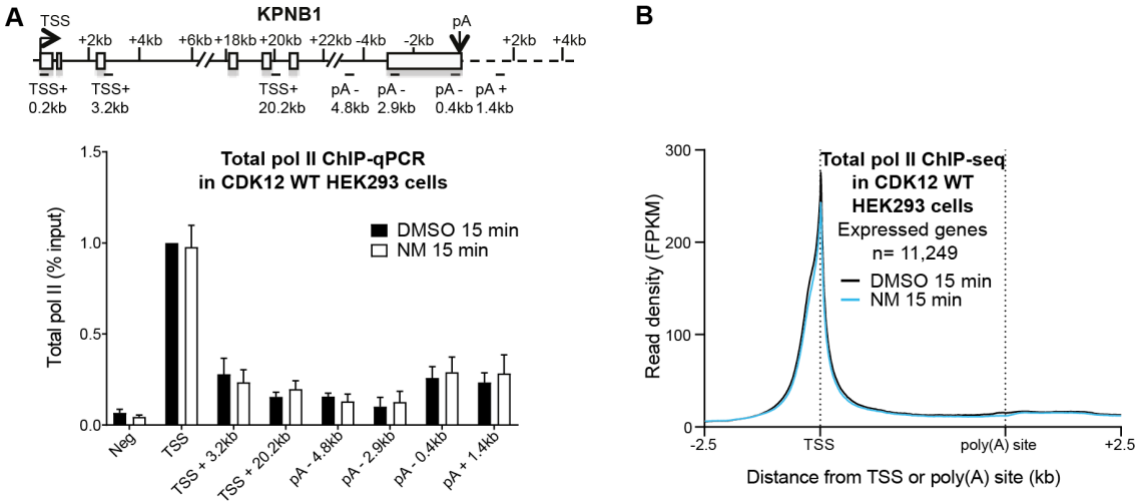


Figure 9.5 | Characterization of the CDK12as HEK293 cell line (A) ChIP-qPCR of pol II on *KPNB1* after treatment of HEK293 cells with DMSO or 7.5 μM NM for 15 minutes. A schematic of *KPNB1* is shown above. **(B)** Meta-analysis of Pol II ChIP-seq of scaled expressed genes performed on the parental HEK293 cells treated with DMSO (black) or with 7.5 μM NM (blue) for 15 minutes.

9.2 CDK12as inhibition globally decreases RNA synthesis

To monitor changes in RNA synthesis upon CDK12as inhibition, we performed TT-seq with RNA spike-ins after 15 and 30 minutes of NM treatment (**Figure 9.7A**). Also using this analysis, treatment with NM has no effect on transcription in the parental cells (**Figures 9.6B and 9.6C**).

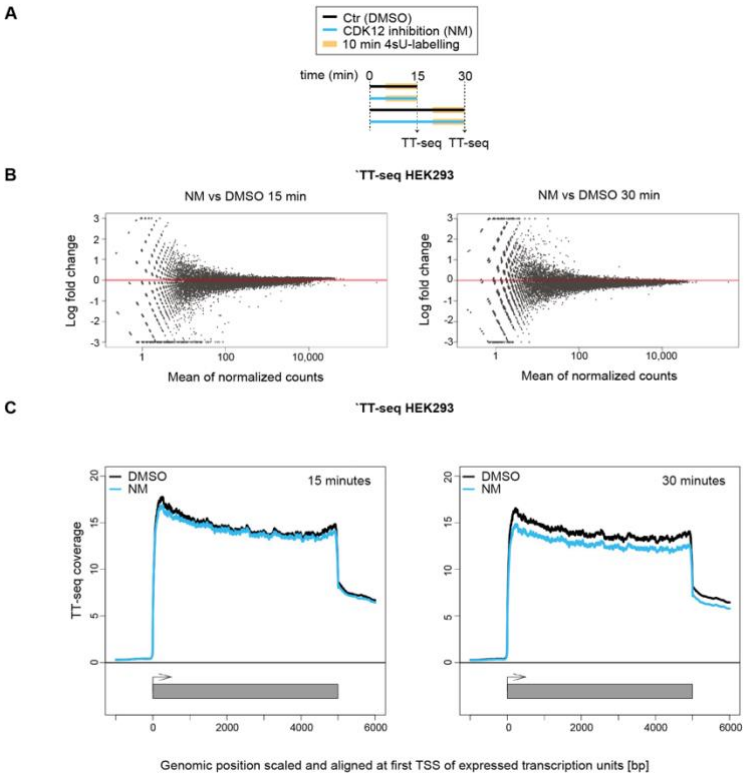


Figure 9.6 | Characterization of the HEK293 WT cell line (A) Experimental design. **(B)** Log fold change upon 7.5 μM NM treatment for 15 min (left) 30 min (right) versus mean of normalized counts across replicates in parental HEK293 cells. **(C)** Meta-analysis of TT-seq data of expressed genes after treatment of HEK293 cells with DMSO (black) or 7.5 μM NM (blue) for 15 or 30 minutes.

However, metagene profiles of TT-seq signals averaged over expressed genes in CDK12as cells show a decreased TT-seq signal across gene bodies after 15 minutes of inhibition, indicating reduced RNA synthesis, either as the result of reduced elongation or loss of polymerase, in 11,182 transcription units, including 10,393 protein-coding genes and 789 unclassified transcription units (**Figures 9.7B-D**). CDK12as inhibition for 30 minutes instead leads to a recovery of RNA synthesis activity at the beginning of genes (**Figures 9.7C, right, 9.7D lower panel**).

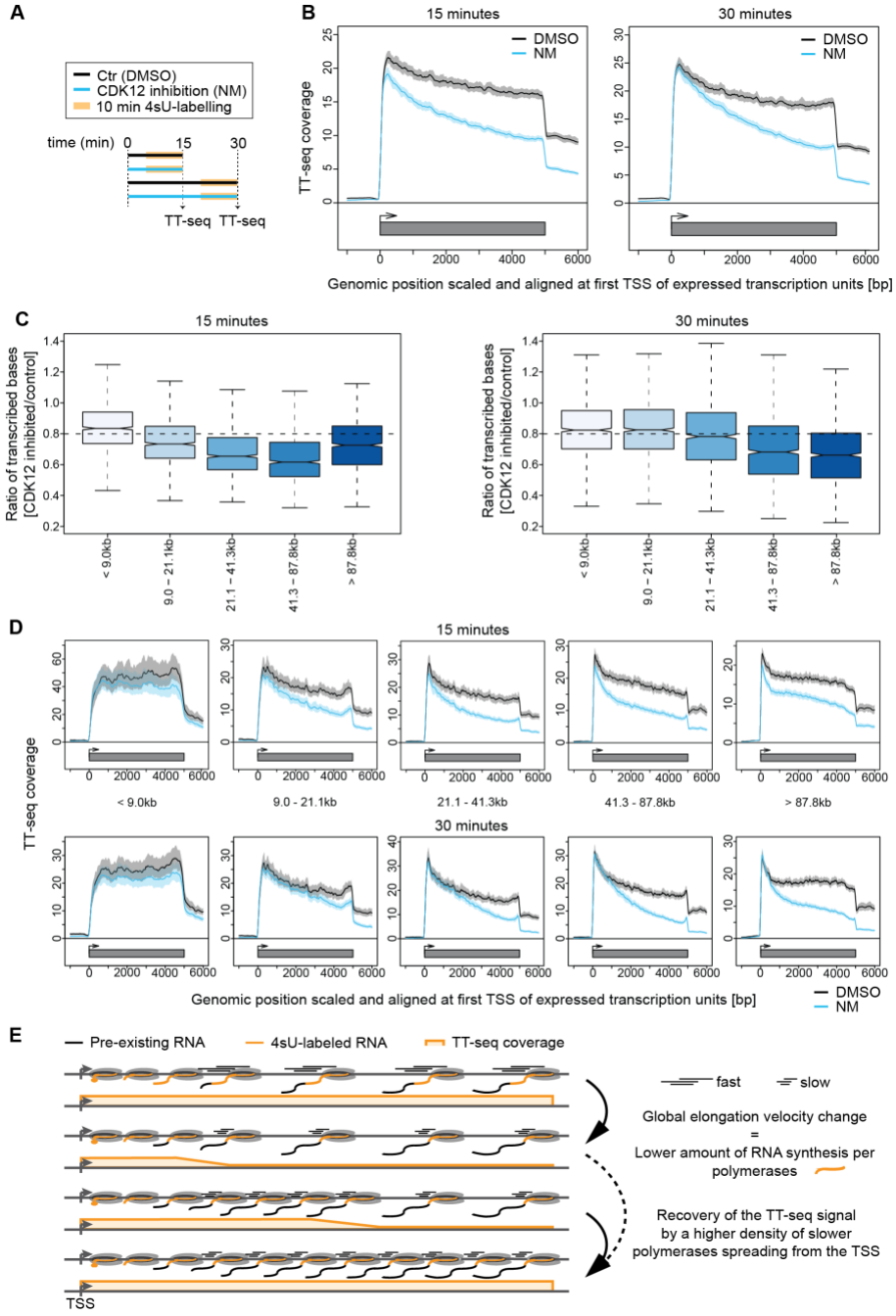


Figure 9.7 | CDK12as inhibition globally decreases RNA synthesis. (A) Experimental design. **(B)** Metagene analysis of TT-seq signal for expressed genes after treatment of cells with DMSO (black) versus 7.5 μM NM treatment (blue) for 15 min (left) or 30 min (right). The TT-seq coverage was averaged and aligned at their transcription start sites (TSSs) and polyadenylation (pA)-sites. Shaded areas around the average signal (solid lines) indicate confidential intervals. **(C)** Box plots of different length classes show the ratio of transcribed bases after 15 min (left) and 30 min (right) inhibition of CDK12as compared to control. **(D)** Metagene analysis for different length classes comparing the average TT-seq signal before (DMSO treatment, black) and after CDK12as inhibition (7.5 μM NM treatment, blue) for 15 min (upper panel) and 30 min (lower panel). The TT-seq coverage was averaged and aligned at their transcription start sites (TSSs) and polyadenylation (pA)-sites. Shaded areas around the average signal (solid lines) indicate confidential intervals. **(E)** Schematic representation of TT-seq signal changes along the gene body upon elongation velocity change. Upper panel: steady state transcription. Lower panels show TT-seq signal recovery spreading from the TSS.

Recovery of the TT-seq signal after 30 minutes could be explained by a higher density of slower polymerases spreading from the transcription start site and gradually populating the gene body (**Figure 9.7E**). This effect would not require any changes in initiation frequencies, but could stem only from an increase in polymerase density over genes due to slower elongation. After CDK12as inhibition, polymerases dramatically slow down, resulting in a lower amount of RNA synthesis, which may be successively restored by a higher number of polymerases occupying the gene body provided the same number of initiation events. This model predicts that recovery of RNA synthesis takes longer for long genes and shorter for short genes (**Figures 9.7C-E**). As human genes take on average longer than 30 minutes to be transcribed, recovery of RNA synthesis and TT-seq signal would be limited to the 5'-region of genes and not observed for many 3'-regions, leading to the observed slope in the TT-seq metagene profile (**Figures 9.7C-E**).

To test this model further, we simulated TT-seq metagene profiles based on a kinetic model that computes RNA synthesis at each gene position based on initiation rates and elongation velocities. The model readily recapitulates the observed TT-seq profiles if we assume that CDK12 inhibition induces rapid and global downregulation of RNA elongation without affecting transcription initiation frequency (**Figures 9.8A and 9.8B**).

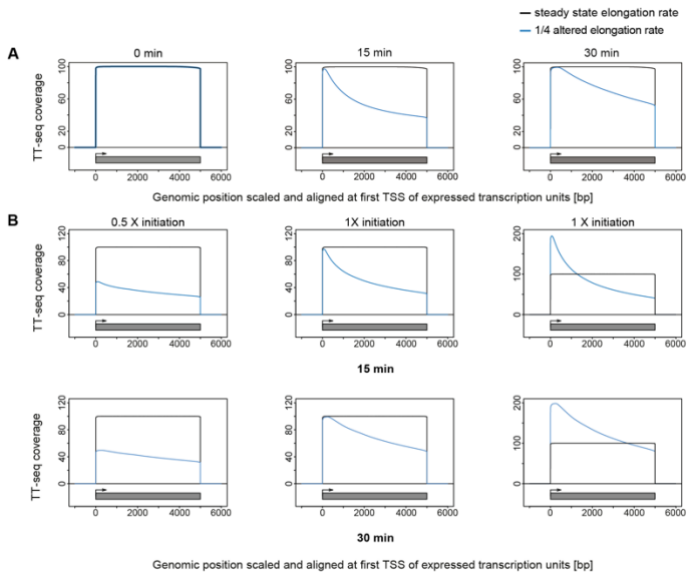


Figure 9.8 | CDK12as inhibition globally decreases RNA synthesis. (A) Simulated metagene profiles of expected RNA synthesis (TT-seq signal) in steady-state (left) and upon changes in elongation rate for 15 minutes (middle) and 30 minutes (right). **(B)** Simulated metagene profiles of expected RNA synthesis (TT-seq signal) upon changes in initiation frequencies, 0.5X (left), 1X (middle), and 2X (right), and elongation rate for 15 minutes (top) and 30 minutes (bottom). In comparison to the steady state elongation rate profile (black), altered 1⁄4 elongation rate profile (blue) shows lower coverage of different magnitude for above stated initiation frequencies. With an initiation frequency 2X (right) the altered elongation rate profile (blue) shows a higher coverage close to the transcription start site which decreases along the gene body for the given time window.

Thus, TT-seq analysis and kinetic modelling supports the notion that CDK12 is required for normal transcription elongation, and the elongation defect resulting from its inactivation is readily observed after rapid inhibition followed by immediate monitoring of RNA synthesis.

9.3 Inhibition of CDK12as affects transcription elongation

To confirm that the observed decrease in RNA synthesis activity results from reduced elongation, we performed mNET-seq of total pol II in CDK12as cells after 15 minutes of NM treatment (**Figure 9.9A**). mNET-seq identifies the last nucleotide transcribed by sequencing the RNA present in the active site of immunoprecipitated pol II (Nojima, Gomes, Grosso, Kimura, Michael J. Dye, *et al.*, 2015). The mNET-seq data were normalized to a set of genes found to be unaffected in TT-seq.

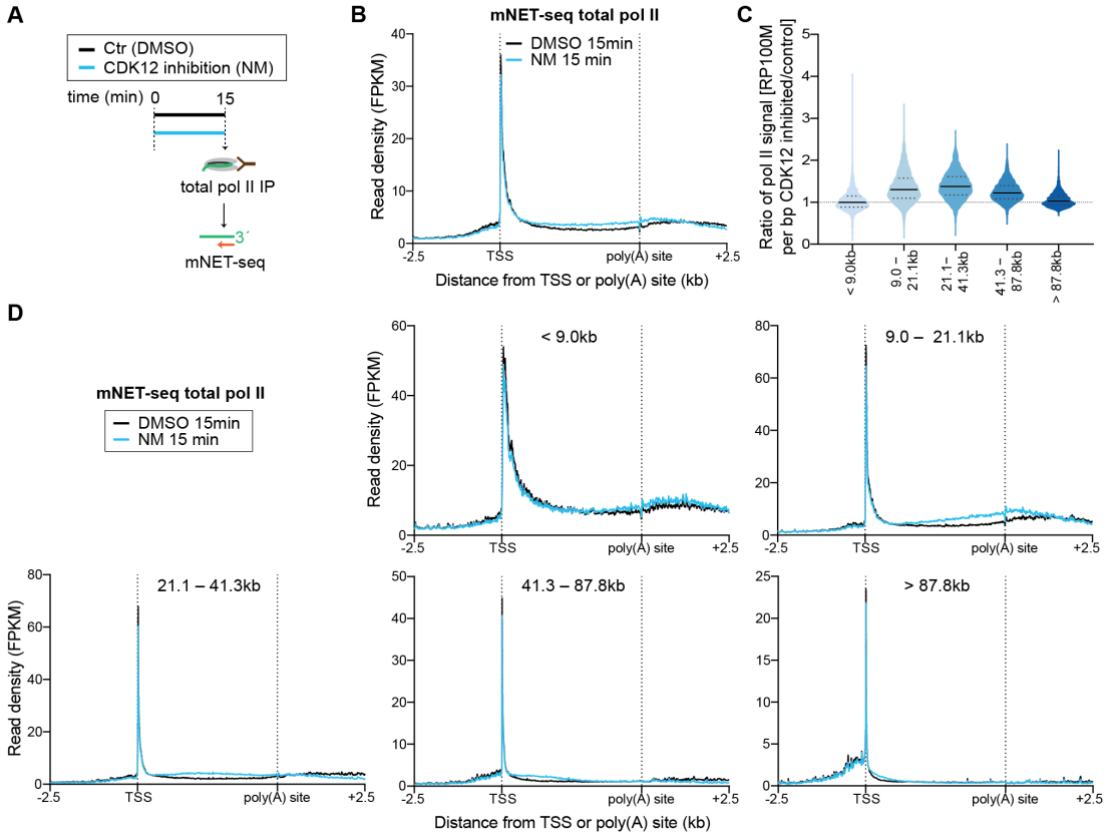


Figure 9.9 | Inhibition of CDK12as affects transcription elongation. (**A**) Experimental design. mNET-seq using an antibody against total pol II was performed in CDK12as HEK293 cells after treatment with 15 min of DMSO solvent control or 7.5 μM NM. (**B**) Metagene analysis comparing the average mNET-seq signal before and after CDK12as inhibition for 15 min of expressed genes. (**C**) Violin plots of the different length classes noted show the ratio of mNET-seq signal after 15 min inhibition of CDK12as compared to control. (**D**) Metagene analysis for the different length classes noted comparing the average mNET-seq signal before and after CDK12as inhibition for 15 min. The mNET-seq reads were averaged and aligned at their TSSs and (pA)-sites.

Inhibition of CDK12as results in an increase of pol II signal in the gene body of the vast majority of expressed pol II-transcribed genes within 15 minutes of inhibition (**Figures 9.9B and 9.9C**), consistent with higher densities of more slowly-elongating pol II. ChIP-seq of pol II gives a similar picture after 15 minutes inhibition of CDK12as (**Figure 9.10A**). We have confirmed this increase of pol II occupancy in the gene body of a model gene, *KPNB1*, by performing pol II ChIP-qPCR after 15- and 30-minutes treatment of CDK12as cells with NM (**Figure 9.10B**). The higher pol II signal results from an increased pol II residence time in gene bodies due to slower elongation, and not from an increased amount of pol II entering elongation, as shown by the decrease in RNA synthesis in TT-seq (**Figure 9.7**).

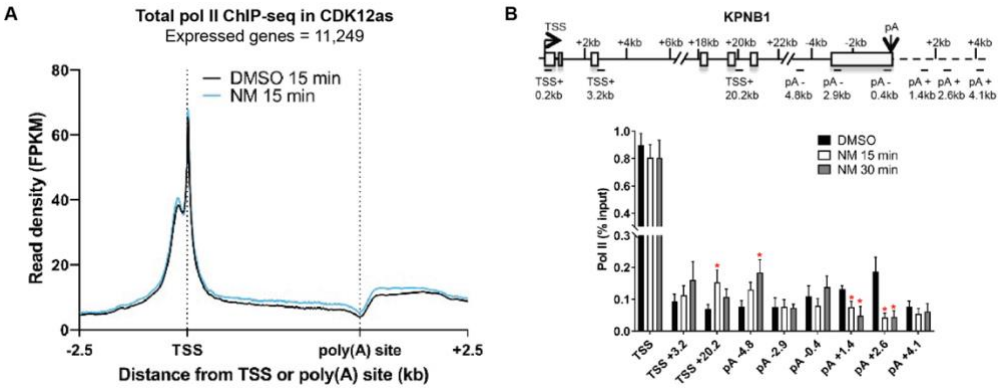


Figure 9.10 | CDK12 inhibition affects elongation of transcription beyond the EEC. (A) Meta-analysis of Pol II ChIP-seq of scaled expressed genes performed on the CDK12as cells treated with DMSO (black) or with 7.5 μM NM (blue) for 15 minutes. **(B)** ChIP-qPCR of pol II on *KPNB1* after treatment of CDK12as cells with DMSO or 7.5 μM NM for 15 or 30 minutes. Error bars= s.e.m. (n = 3 biological replicates). Asterisks indicate statistical significance (* $p < 0.05$), based on unpaired, two-tailed Student's t test. A schematic of *KPNB1* is shown on the top.

Further analysis shows that transcription of pol II-transcribed genes of all lengths is affected by CDK12as inhibition, with a wave of elongation-compromised pol II reaching the ends of the shortest genes first (**Figure 9.9D**). Genes <9 kb have a less pronounced elongation defect after CDK12as inhibition than genes >9 kb as measured by both TT-seq and mNET-seq (**Figures 9.7D and 9.9D**). In addition, the elongation defect increases with increasing distance from the TSS for all gene length classes (**Figure 9.7D**). This is more apparent after analysis of TT-seq and mNET-seq data for the first few kilobases of genes >6.5 kb (**Figures 9.11A and 9.11B**). Thus, the effect of CDK12as inhibition starts early in the transcription cycle and increases with increasing distance from the TSS (**Figures 9.11A and 9.11B**). Many of the genes <9 kb will therefore be too small to exhibit a major effect of CDK12as inhibition on elongation as pol II will reach the end of genes soon after the elongation defect starts to occur.

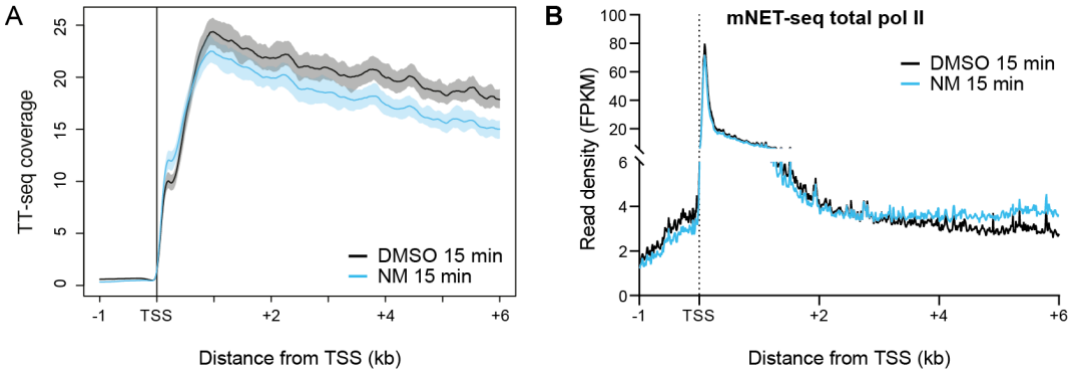


Figure 9.11 | CDK12 inhibition affects elongation of transcription beyond the EEC. (A) Meta-analysis of TT- seq data for the first 6 kb of expressed genes > 6.5 kb in length after treatment of CDK12as cells with DMSO (black) or 7.5 μM NM (blue) for 15 minutes. **(B)** Meta-analysis of mNET-seq data for the first 6kb of expressed genes > 6.5 kb in length after treatment of CDK12as cells with DMSO (black) or 7.5 μM NM (blue) for 15 minutes.

Interestingly, on genes longer than 21.1 kb, pol II increases in the gene body but is specifically reduced downstream of the poly(A) site (**Figure 9.9D**), before the wave of slower polymerases reaches this region. Single gene examples of this phenomenon are shown in (**Figure 9.12)** and this can also be observed by pol II ChIP-qPCR of the *KPNB1* gene (**Figure 9.10B**).

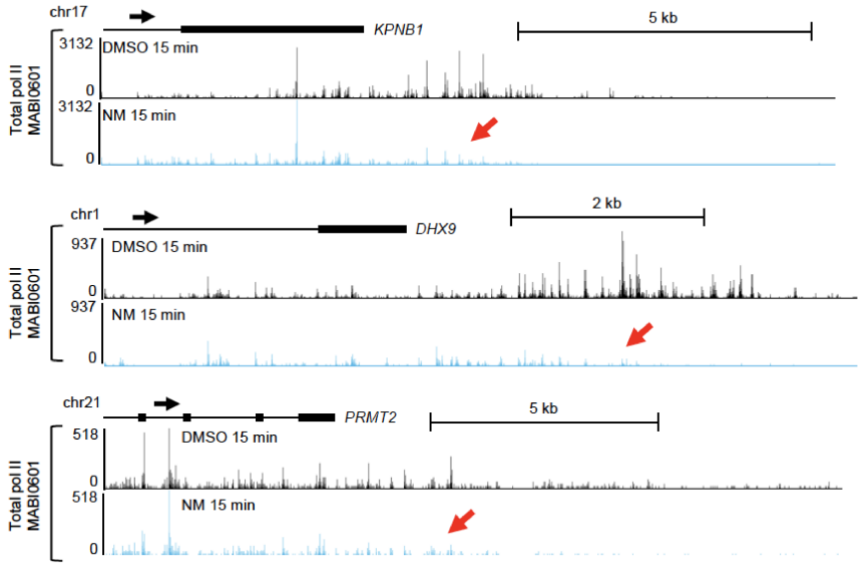


Figure 9.12 | Total pol II mNET-seq profiles across the 3'end of *KPNB1*, *DHX9*, and *PRMT2*.

Since transcription of intron-containing protein-coding genes is affected by short-term inhibition of CDK12as, we investigated whether transcription of intronless (n = 5,396), histone (n = 118), and snRNAs (n = 36) genes is also altered (**Figures 9.13A and 9.13B**).

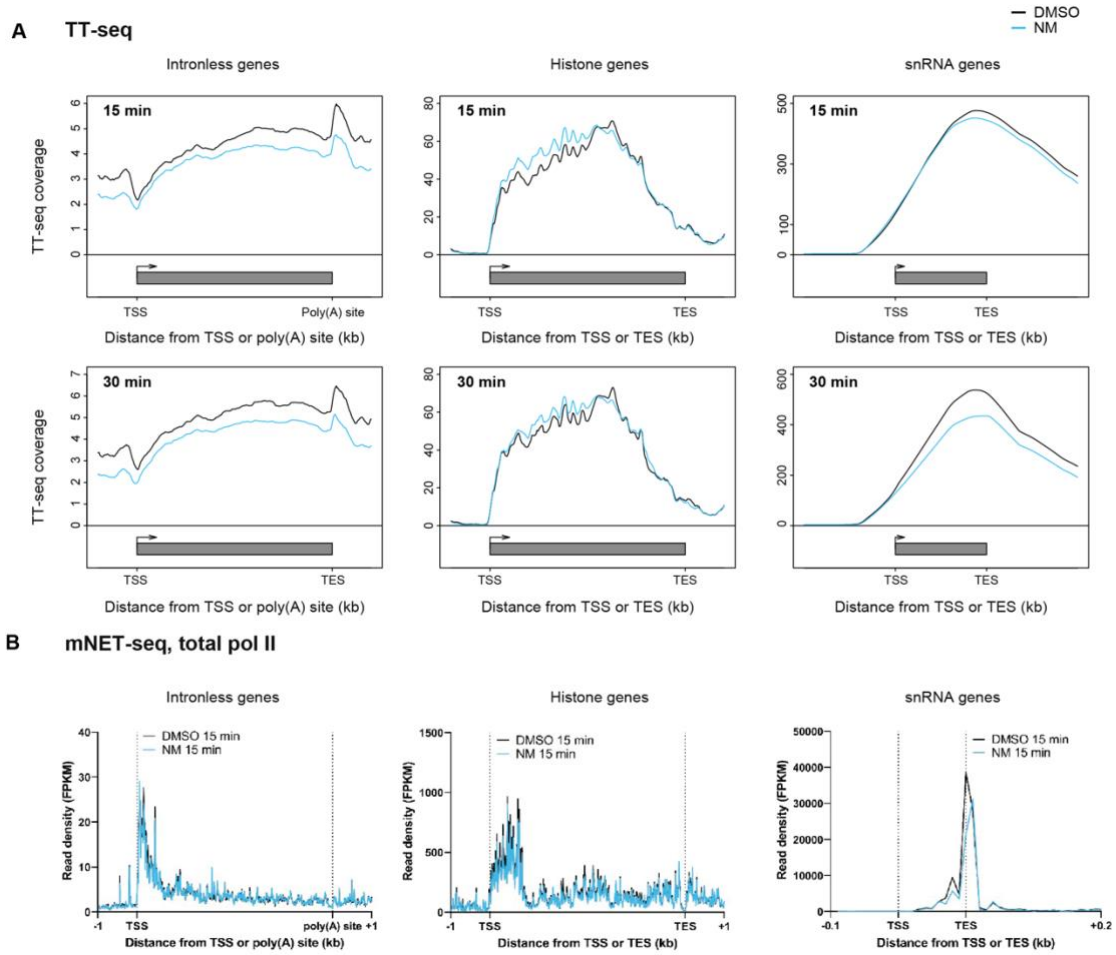


Figure 9.13 | CDK12 inhibition does not affect transcription of histone genes. (**A)** Meta-analysis of TT-seq data of intronless, histone, and snRNA genes after treatment of CDK12as cells with DMSO (black) or 7.5 μM NM (blue) for 15 or 30 minutes. (**B)** Meta-analysis of total pol II mNET-seq data of intronless, histone, and snRNA genes after treatment of CDK12as cells with DMSO (black) or 7.5 μM NM (blue) for 15 minutes.

TT-seq signals for intronless and pol II-transcribed snRNA genes decrease after 15- or 30-minutes inhibition, whereas mNET-seq signals on intronless genes are not markedly affected and slightly reduced on snRNA genes. We interpret this to mean that pol II is slower on these genes but pol II is not building up as they are relatively short (the intronless and snRNAs have an average size of 948 bp and 150 bp, respectively). TT-seq signals for histone genes instead are either unaffected or slightly increased and mNET-seq signals unaffected by CDK12as inhibition. We therefore conclude that transcription of histone genes does not require CDK12 activity. We have confirmed this by pol II ChIP-qPCR for five different histone genes (**Figure 9.14**).

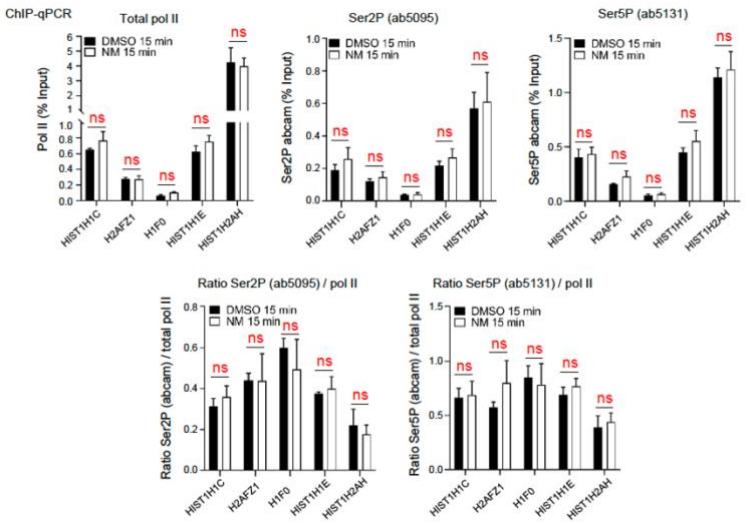


Figure 9.14 | CDK12 inhibition does not affect transcription of histone genes. ChIP-qPCR of pol II, Ser2P (ab5095), and Ser5P (ab5131) on the gene body of five different histone genes after treatment of CDK12as cells with DMSO or 7.5 μM NM for 15 minutes. Error bars= s.e.m. (n = 3 biological replicates). Asterisks indicate statistical significance (ns: not significant), based on unpaired, two- tailed Student's t test.

The primary effects of CDK12as inhibition on transcription of intron-containing protein-coding genes are therefore an elongation defect in gene bodies and loss of pol II downstream of the poly(A) site. Taken together, the results indicate that CDK12 plays a role in maintaining efficient transcription elongation velocity and processivity on these human genes. Efficient transcription of intronless genes and snRNA genes also appears to require CDK12, although the defects are less apparent. Histone genes, on the other hand, are transcribed efficiently in the absence of CDK12.

9.4 CDK12 phosphorylates transcribing pol II

mNET-seq can also be carried out with CTD phospho-site-specific antibodies to detect nascent RNA associated with particular pol II phospho-isoforms (Nojima *et al.*, 2016). Accordingly, we have used mNET-seq to investigate the primary effect of CDK12as inhibition on CTD phosphorylation. As short-term inhibition of CDK12as is sufficient to cause global changes in transcription (**Figures 9.7 and 9.9**), we treated CDK12as with NM for 15 minutes and obtained mNET-seq profiles for pol II CTD phosphorylation at Ser2P and Ser5P using antibodies against each of these phospho-forms (**Figure 9.15**).

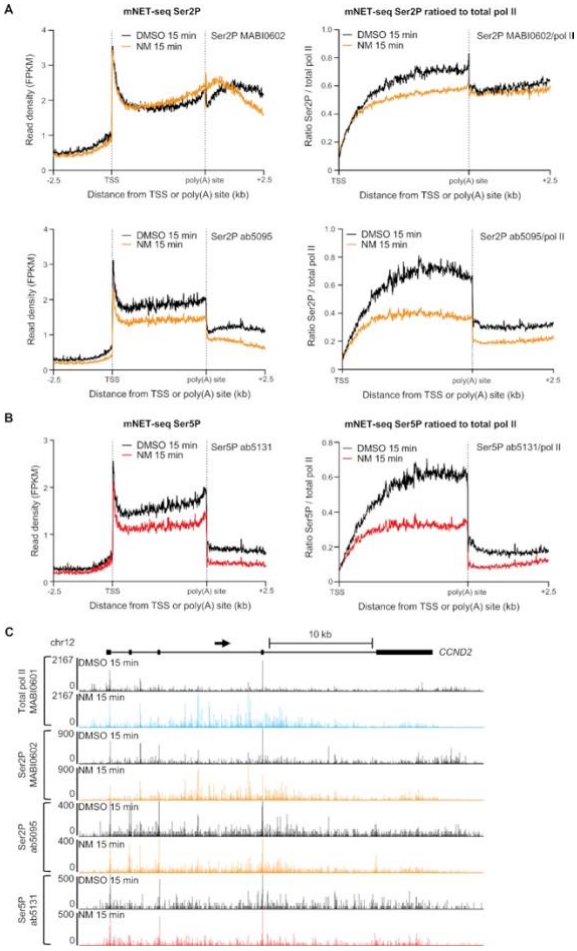


Figure 9.15 | CDK12 phosphorylates transcribing pol II. (A) Meta-analyses of scaled expressed genes of mNET-seq for Ser2P with and without normalization to pol II after treatment of CDK12as cells with DMSO (black) or 7.5μM NM (orange) for 15 min. **(B)** Meta-analyses of scaled expressed genes of mNET-seq for Ser5P with and without normalization to pol II after treatment of CDK12as cells with DMSO (black) or 7.5 μM NM (red) for 15 min. **(C)** mNET-seq profiles across *CCND2* using a total pol II antibody, two Ser2P antibodies and one Ser5P antibody.

As transcription of histone genes is relatively unaffected by CDK12as inhibition, we have investigated whether pol II CTD Ser2P and Ser5P levels are affected by CDK12as inhibition. Ser2P and Ser5P ChIP-qPCR for five different histone genes indicates that Ser2P and Ser5P, ratioed and unratioed to pol II levels, is unaffected by NM treatment (**Figure 9.14**), indicating no requirement for CDK12 for CTD phosphorylation during transcription of these genes. Interestingly, the level of Ser2P relative to pol II on histone is generally much lower than on other protein-coding genes (a ratio of up to 0.6 on histone genes and a ratio of >40 at the end of *KPNB1* with ab5095) (**Figures 9.13 , 9.14 and 9.16**), which is consistent with the relative lack of requirement for a Ser2 kinase.

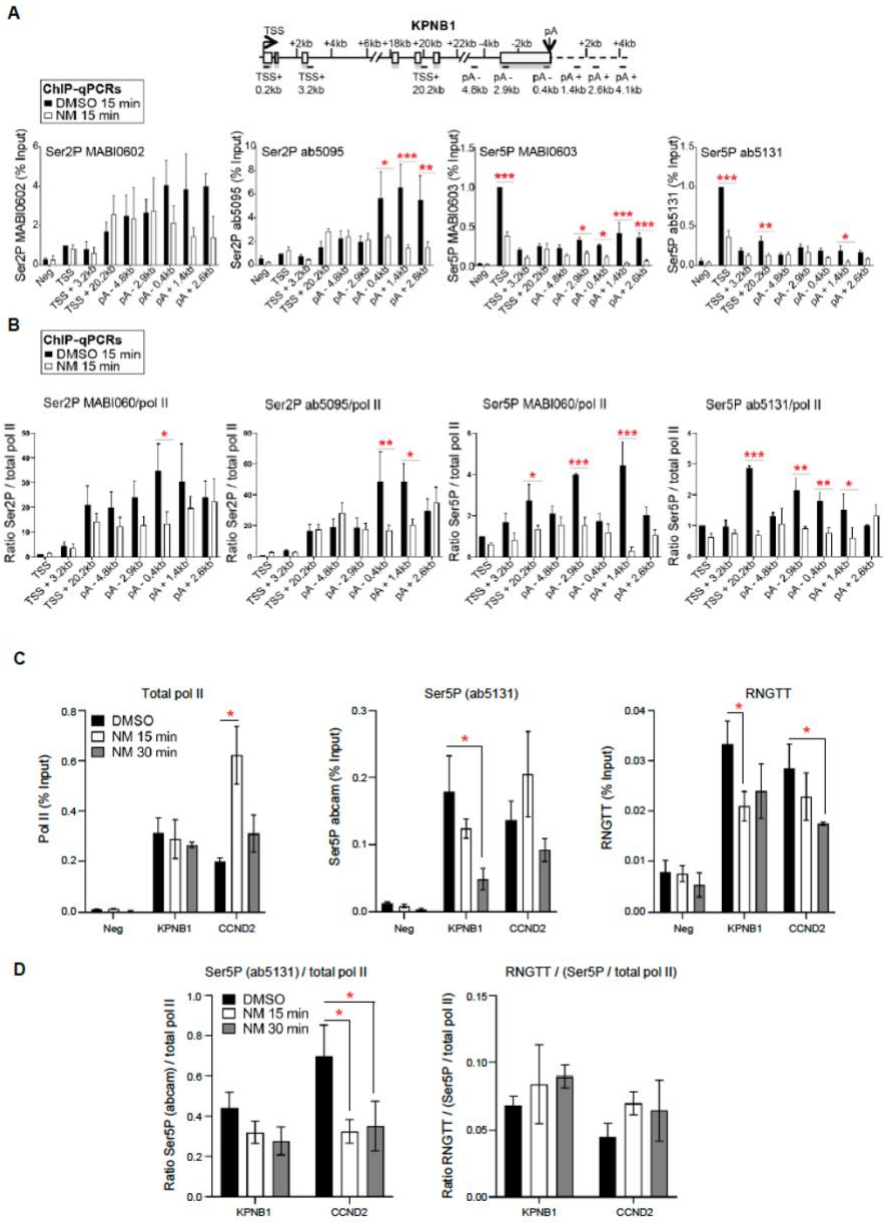


Figure 9.16 | CDK12 regulates phosphorylation of the CTD of engaged pol II. (A) ChIP-qPCR of Ser2P and Ser5P after treatment of CDK12as cells with DMSO or 7.5 μM NM for 15 minutes. Error bars= s.e.m. (n = 3 biological replicates). Asterisks indicate statistical significance (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$), based on unpaired, two-tailed Student's t test. A schematic of *KPNB1* is shown above. **(B)** Ratio of the ChIP-qPCR for Ser2P and Ser5P to pol II after treatment of CDK12as cells with DMSO or 7.5 μM NM for 15 minutes. Error bars= s.e.m. (n = 3 biological replicates). Asterisks indicate statistical significance (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$), based on unpaired, two-tailed Student's t test. **(C)** ChIP-qPCR of total pol II, Ser5P, and RNGTT on the TSS of *KPNB1* and *CCND2* after treatment of CDK12as cells with DMSO or 7.5 μM NM for 15 or 30 minutes. Error bars= s.e.m. (n = 3 biological replicates). Asterisks indicate statistical significance (* $p < 0.05$), based on unpaired, two-tailed Student's t test. **(D)** Ratio of the ChIP-qPCR of Ser5P to pol II or of RNGTT ratioed to Ser5P/pol II after treatment of CDK12as cells with DMSO or 7.5 μM NM for 15 or 30 minutes. Error bars = s.e.m. (n = 3 biological replicates). Asterisks indicate statistical significance (* $p < 0.05$), based on unpaired, two-tailed Student's t test.

We have therefore used the levels of Ser2P and Ser5P on histone genes to normalize our mNET-seq Ser2P and Ser5P data. We have displayed the results with and without normalisation of the mNET-seq signals to total pol II levels to take changes of pol II occupancy into account (**Figures 9.15A and 9.15B**). (**Figure 9.15C)** shows the genome browser track of the unratioed mNET-seq results for the *CCND2* gene, a ~32 kb long gene displaying increased pol II signal in the gene body and a loss of pol II downstream of the poly(A) site. The mNET-seq signals for two different Ser2P antibodies and for the Ser5P antibody show a decrease after CDK12as inhibition when ratioed to the total pol II level (**Figure 9.15A and 9.15B**). The same effect of CDK12as inhibition on Ser2P and Ser5P phosphorylation is seen using ChIP-seq of Ser2P and Ser5P and ChIP-qPCR on *KPNB1* (**Figures 9.17A, 9.17B, 9.16A and 9.16B**).

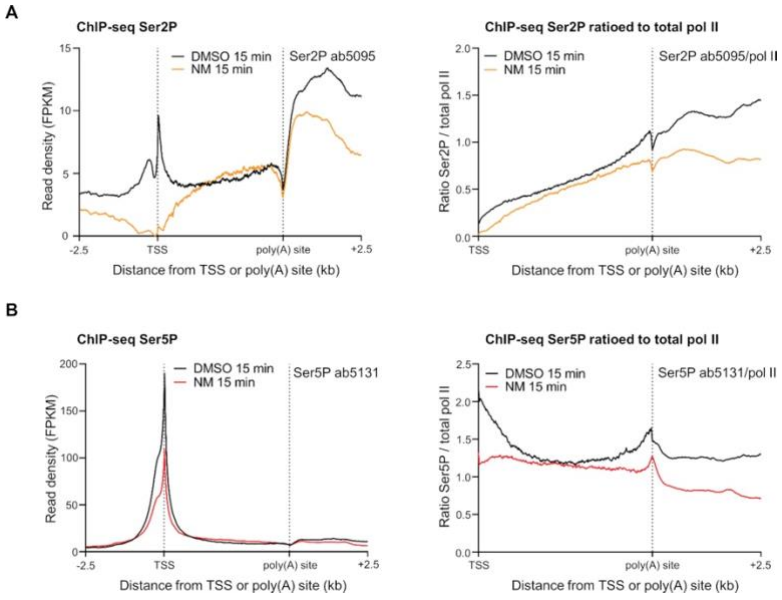
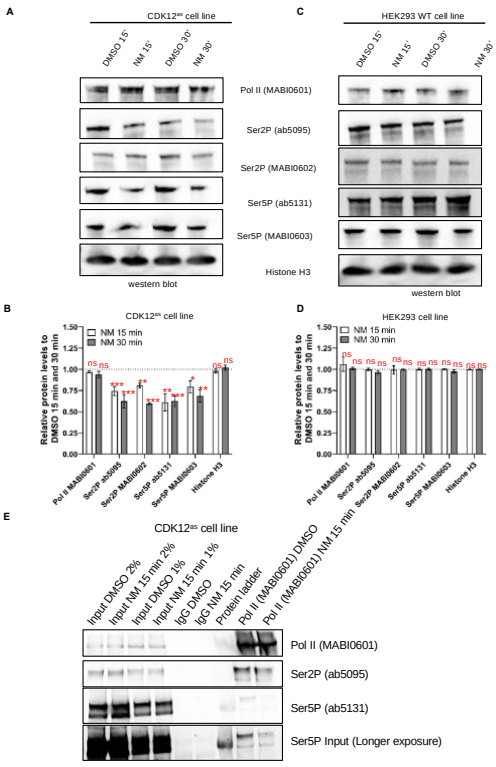


Figure 9.17 | ChIP-seq analysis of CDK12 phosphorylation of transcribing pol II. (A) Meta-analyses of scaled expressed genes of ChIP-seq for Ser2P (ab5095) with and without normalization to pol II after treatment of CDK12^{as} cells with DMSO (black) or 7.5μM NM (orange) for 15 min. **(B)** Meta-analyses of scaled expressed genes of ChIP-seq for Ser5P (ab5131) with and without normalization to pol II after treatment of CDK12^{as} cells with DMSO (black) or 7.5 μM NM (red) for 15 min.

Interestingly, we observed a loss of Ser5P phosphorylation at the TSS of genes after CDK12as inhibition, indicating that CDK12 acts as a Ser5 kinase at this step and could affect mRNA capping, which relies on Ser5P (Ho and Shuman, 1999) (**Figure 9.16B**). We first confirmed the loss of Ser5P by ChIP-qPCR on the TSS of *KPNB1* and *CCND2* after 15- and 30-minutes inhibition of CDK12as (**Figures 9.16C and 9.16D**). To investigate a possible effect on mRNA capping, we performed ChIP-qPCR of RNGTT, the enzyme responsible for the first two catalytic steps of the cap formation, and found a decrease in

signal at the TSS following CDK12as inhibition consistent with the reduction in Ser5P relative to total pol II (**Figures 9.16C and 9.16D**).

We also carried out western blot analyses on chromatin-associated pol II with antibodies against different pol II CTD phospho-isoforms after treating cells with NM for 15 and 30 minutes (**Figures 9.18A and 9.18B**). In line with the results of mNET-seq analysis (**Figures 9.15A and 9.15B**), inhibition of CDK12as markedly affects Ser2 and Ser5 phosphorylation within 15 minutes, as measured with two different antibodies against these modifications (**Figures 9.18A and 9.18B**). Both Ser2P and Ser5P continue to be affected after 30 minutes, with a further decrease in phosphorylation observed for Ser2P MABI0602 and Ser5P MABI0603 (**Figures 9.18A and 9.18B**). In contrast, no loss of Ser2P or Ser5P occurs when HEK293 cells are treated with NM for 15 or 30 minutes (**Figures 9.18C and 9.18D**). Loss of Ser2P and Ser5P after 15 minutes of inhibition of CDK12as was also confirmed by performing immunoprecipitation of total pol II followed by western blotting with a Ser2P (ab5095) and Ser5P (ab5131) antibody (**Figure 9.18E**). These results are consistent with previous *in vitro* kinase assays and analysis of CDK12as HeLa cells (Bösken *et al.*, 2014; Bartkowiak, Yan and Greenleaf, 2015). Taken together, the western and mNET-seq analyses indicate that CDK12 contributes to phosphorylation of the CTD of engaged pol II with activity towards both Ser2 and Ser5.



(continued to next page)

Figure 9.18 | CDK12as inhibition causes changes to CTD phosphorylation. (A) Western blots of CDK12as cell chromatin extracts. Cells are either treated with 7.5 μM 1-NM-PP1 or DMSO for 15 and 30 minutes as noted. The antibodies used are indicated on the left. Histone H3 was used as a loading control. **(B)** Quantitation of protein levels of total pol II, Ser2P, Ser5P and histone H3 in CDK12as cells relative to DMSO controls. Error bars = s.e.m. (n = 3 biological replicates). Statistical test: two- tailed unpaired t test, ns = not significant, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. **(C)** Western blots of chromatin extracts of DMSO-treated and NM-treated parental HEK293 cells with the antibodies noted on the left. Histone H3 was used as a loading control. **(D)** Quantitation of protein levels of total pol II, Ser2P, Ser5P and histone H3 in HEK293 cells relative to DMSO controls. Error bars = s.e.m. (n = 3 biological replicates). Statistical test: two-tailed unpaired t test, ns = not significant. **(E)** Co- immunoprepication of total pol II (MABI0601) followed by western blot with total pol II (MABI0601), Ser2P (ab5095), and Ser5P (ab5131) antibodies.

CDK12 activity is required for stable association of elongation and termination factors

Our results indicate that CDK12 is required for normal transcription elongation of most pol II-transcribed genes (**Figures 9.7 and 9.9**). However, the mechanism remains unclear. CDK12 activity may be required for the recruitment of elongation factors to transcribing pol II or for the stabilization of their interactions with pol II. We have therefore tested by ChIP-seq and ChIP-qPCR whether the levels of the LEO1 and CDC73 subunits of the elongation factor PAF1 complex (PAF1C) and the elongation factor SPT6 detected in chromatin are affected by CDK12as inhibition (**Figure 9.19A-C and 9.20**). We have displayed the LEO1 and SPT6 ChIP-seq results with or without normalisation to pol II, to take the changes in pol II levels into account (**Figure 9.19A and 9.19B**). The level of LEO1 and SPT6 associated with expressed genes appears to be affected genome-wide. In addition, LEO1 appears to be selectively lost from the newly-elongating pol II as long genes (e.g. >41.3 kb) have a reduction in the ratio of LEO1 to total pol II at the 5' end, but not at the 3' end, where the pol II that initiated before treatment is still elongating (**Figure 9.19C**). In contrast, SPT6 is reduced more globally as a decrease in signal is also observed at the 3'end of long genes, albeit less than at the 5' end of genes (**Figure 9.19D**). In addition, the association of CDC73 and SPT6 by ChIP-qPCR with *KPNB1*, with or without normalisation to pol II, is reduced after 15 min inhibition of CDK12as (**Figure 9.20C**).

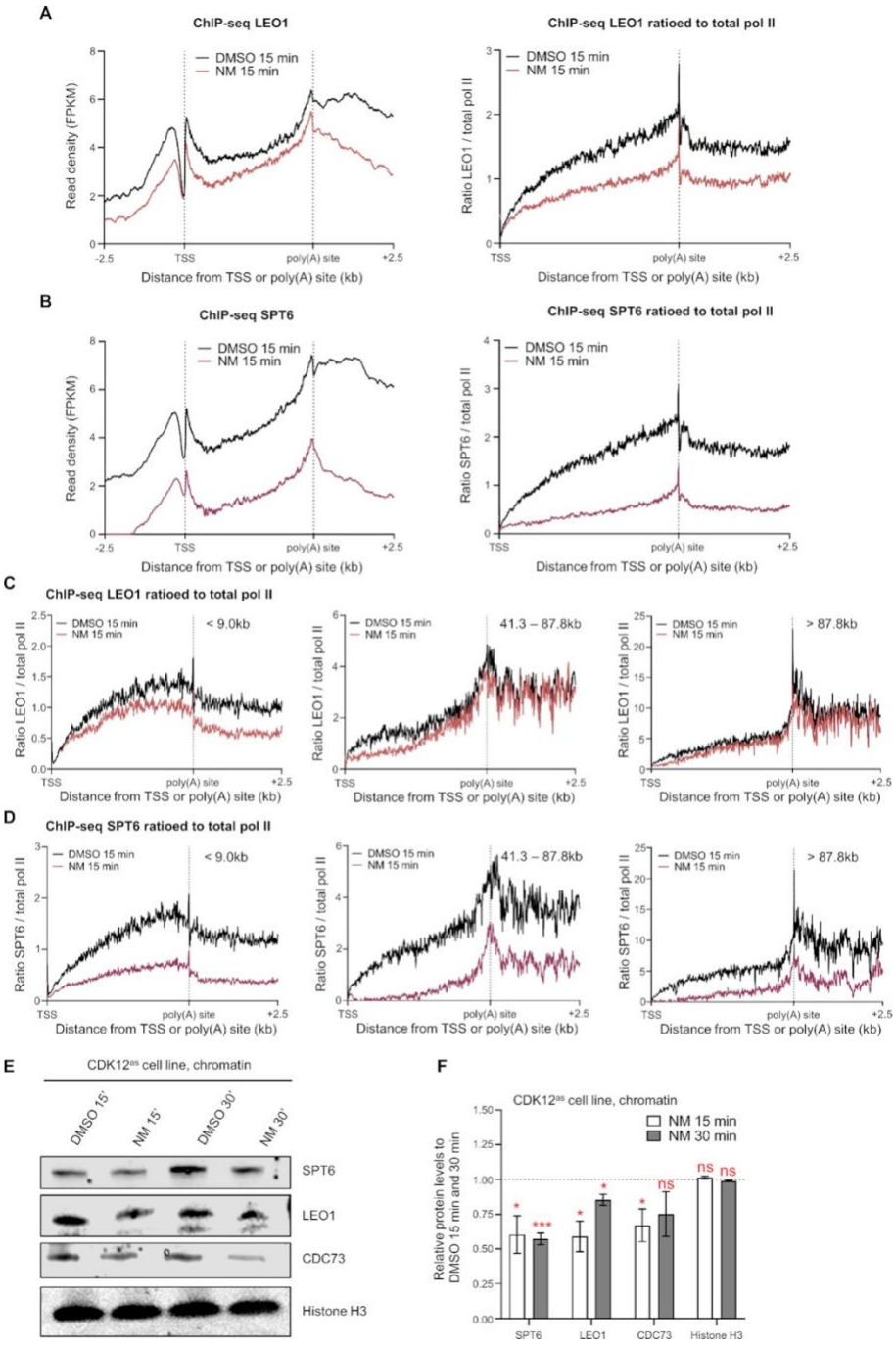


Figure 9.19 | CDK12 activity is required for stable association of elongation factors with chromatin. (A) Meta-analysis of LEO1 subunit of the PAF1 complex (PAF1C) ChIP-seq across scaled expressed genes. **(B)** Meta-analysis of SPT6 ChIP-seq across scaled expressed genes. **(C)** Meta-analyses of LEO1 ChIP-seq ratioed to the total pol II signal before and after CDK12^{as} inhibition for 15 min across scaled expressed genes for the different length classes noted. **(D)** Meta-analyses of SPT6 ChIP-seq ratioed to the total pol II signal before and after CDK12^{as} inhibition for 15 min across scaled expressed genes for the different length classes noted. **(E)** Western blots of CDK12^{as} cell chromatin extracts. Cells are either treated with 7.5 μM 1-NM-PP1 or DMSO for 15 and 30 min as noted. The antibodies used are indicated on the right. Histone H3 was used as a loading control. **(F)** Quantitation of chromatin protein levels of SPT6, LEO1, CDC73, and histone H3 in CDK12^{as} cells relative to DMSO controls. Error bars = s.e.m. (n = 3 biological replicates). Statistical test: two-tailed unpaired t test, ns = not significant, * $P < 0.05$, *** $P < 0.001$.

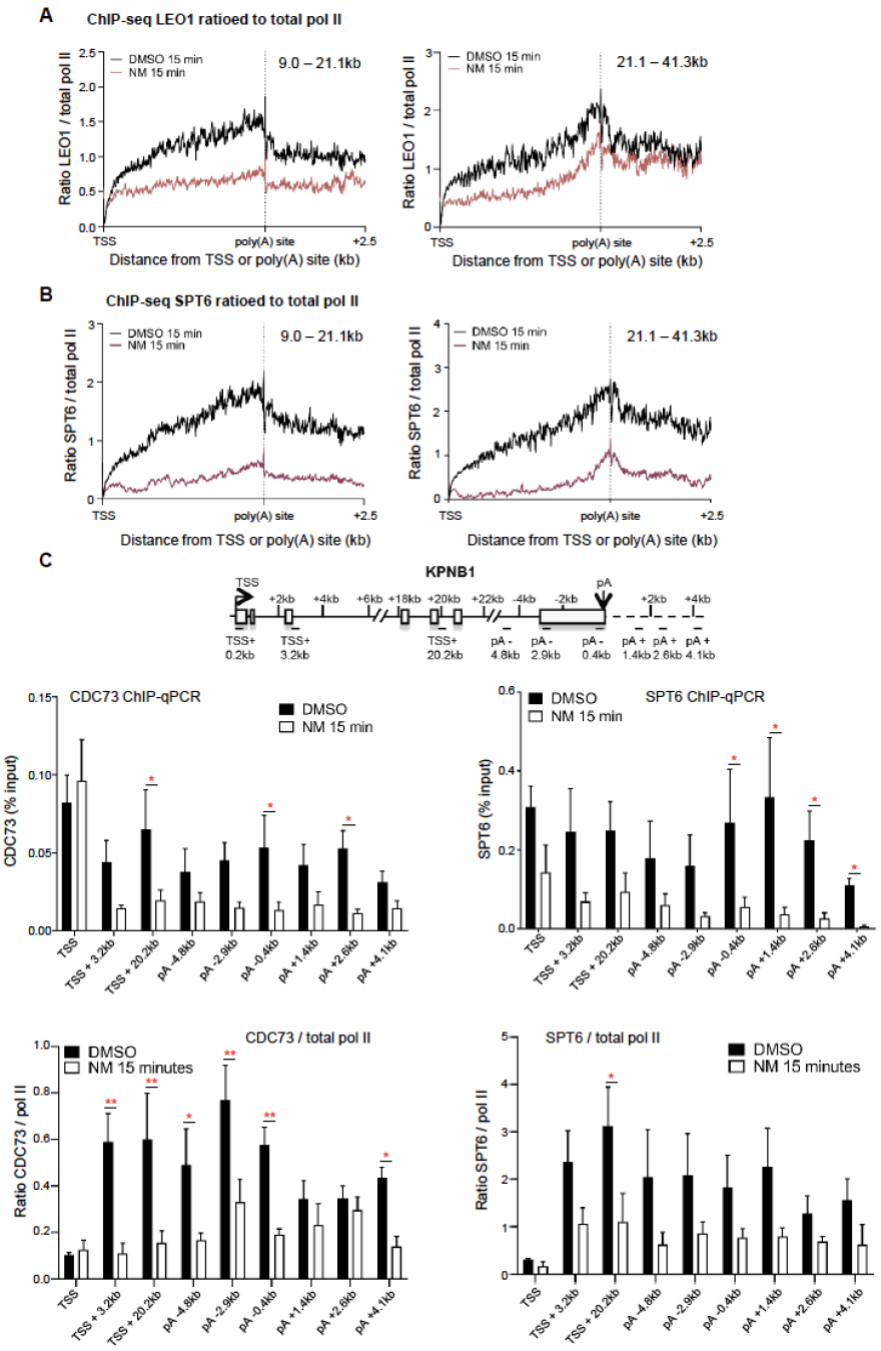


Figure 9.20 | CDK12 activity is required for stable association of elongation factors with chromatin. (A) Meta-analyses of LEO1 ChIP-seq ratioed to the total pol II signal across scaled expressed genes for the different length classes noted. **(B)** Meta-analyses of SPT6 ChIP-seq ratioed to the total pol II signal across scaled expressed genes for the different length classes noted. **(C)** CDC73 and SPT6 ChIP-qPCR unratioed (top) or ratioed to pol II (bottom) across *KPNB1* gene after treatment of CDK12as cells with DMSO or 7.5 μM NM for 15 minutes. Error bars = s.e.m. (n = 4 biological replicates). Asterisks indicate statistical significance (* $p < 0.05$, ** $p < 0.01$), based on unpaired, two-tailed Student's t test.

Similarly to LEO1, the level of CDC73 on *KPNB1*, which is ~34 kb long, is decreased in the gene body but not at the 3' end by CDK12as inhibition. Western analyses also indicate

that inhibition of CDK12as decreases the association of SPT6, LEO1, and CDC73 with chromatin (**Figure 9.19E and 9.19F**).

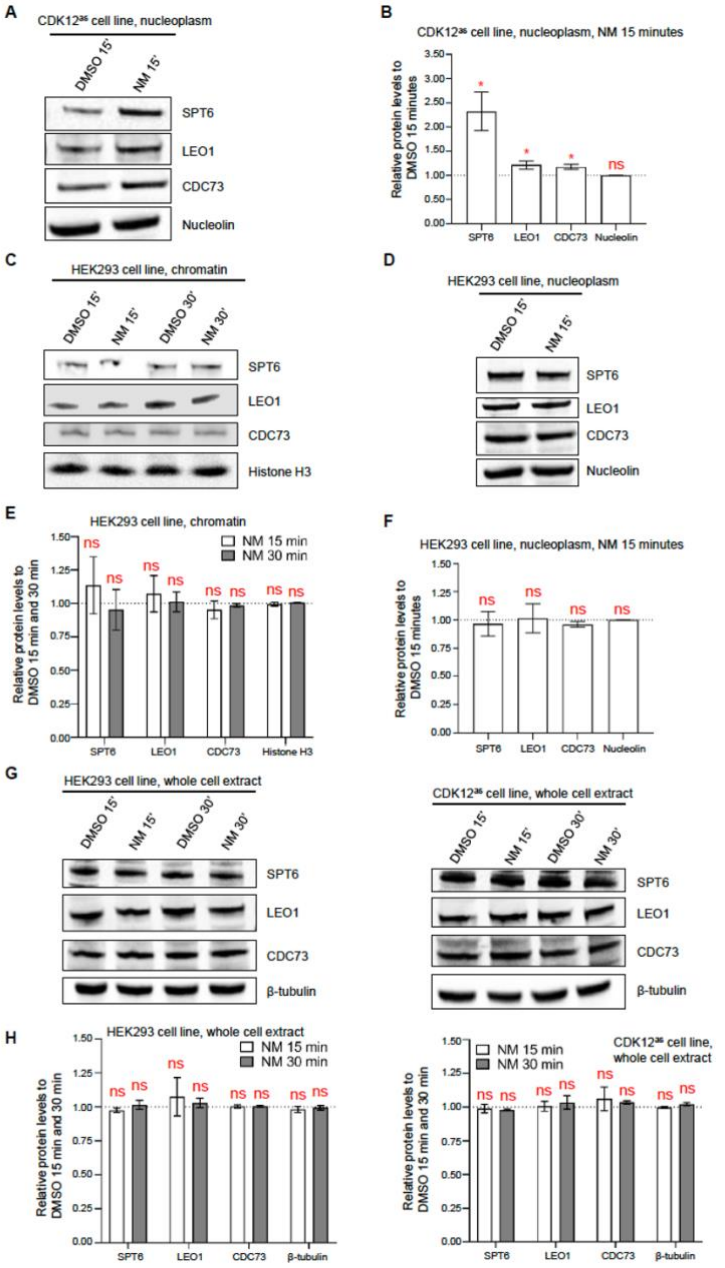


Figure 9.21 | CDK12 activity is required for stable association of elongation factors with chromatin. (A) Western blots of CDK12as cell nucleoplasmic extracts. Cells are either treated with 7.5 μM 1- NM-PP1 or DMSO for 15 minutes as noted. The antibodies used are indicated on the right. Nucleolin was used as a loading control. **(B)** Quantitation of nucleoplasmic protein levels of SPT6, LEO1, CDC73, and nucleolin in CDK12as cells relative to DMSO controls. Error bars = s.e.m. (n = 3 biological replicates). Statistical test: two-tailed unpaired t test, ns = not significant, * $p < 0.05$, **(C)** Western blots of HEK293 cell chromatin extracts. Cells are either treated with 7.5 μM 1-NM-PP1 or DMSO for 15 and 30 minutes as noted. The antibodies used are indicated on the right. Histone H3 was used as a loading control. **(D)** Western blots of HEK293 cell nucleoplasmic extracts. Cells are either treated with 7.5 μM 1-NM-PP1 or DMSO for 15 minutes as noted. The antibodies used are indicated on the right. Nucleolin was used as a loading control. **(E)** Quantitation of chromatin protein levels of SPT6, LEO1, CDC73, and histone H3 in HEK293

cells relative to DMSO controls. Error bars = s.e.m. (n = 3 biological replicates). Statistical test: two-tailed unpaired t test, ns = not significant. **(F)** Quantitation of nucleoplamsic protein levels of SPT6, LEO1, CDC73, and nucleolin in HEK293 cells relative to DMSO controls. Error bars = s.e.m. (n = 3 biological replicates). Statistical test: two-tailed unpaired t test, ns = not significant. **(G)** Western blots of HEK293 (left) or CDK12as (right) cell whole cell extracts. Cells are either treated with 7.5 μM 1-NM-PP1 or DMSO for 15 and 30 minutes as noted. The antibodies used are indicated in the middle. β-tubulin was used as a loading control. **(H)** Quantitation of whole cell extract protein levels of SPT6, LEO1, CDC73, and β-tubulin in HEK293 (left) or CDK12as (right) cells relative to DMSO controls. Error bars = s.e.m. (n = 3 biological replicates). Statistical test: two-tailed unpaired t test, ns = not significant.

The loss of SPT6, LEO1, and CDC73 from chromatin is mirrored by an increase in the level of these proteins in nucleoplasm fraction after CDK12as inhibition, with a particularly marked increase in nucleoplasmic SPT6 (**Figure 9.21A and 9.21B**). These results are in line with a loss of association of these factors with chromatin, although we cannot rule out that changes in post-translation modifications are affecting antibody reactivity. Importantly, no loss of these factors from chromatin is observed when HEK293 cells are treated with NM (**Figures 9.21C–F**). Also, the decrease on chromatin of SPT6, LEO1, and CDC73 is not due to protein degradation, as the global protein level of these elongation factors remains stable after NM treatment, both in HEK293 and CDK12as cells (**Figure 9.21G and 9.21H**).

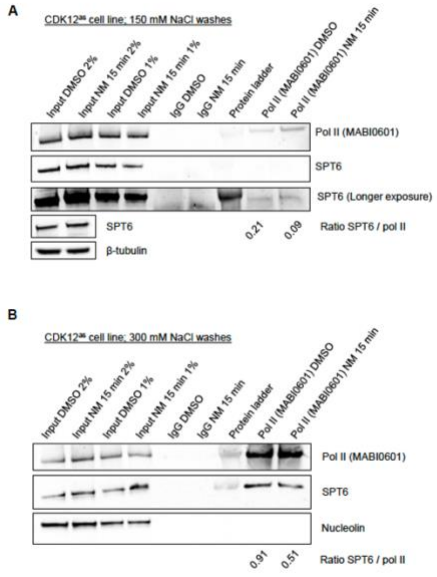


Figure 9.22 | CDK12 inhibition causes loss of detectable SPT6 associated with pol II. Co-immunoprecipitation of total pol II (MABI0601) from CDK12as cells treated with DMSO or 7.5 μM NM for 15 minutes and washed with 150 mM (**A/**) or 300 mM (**B/**) NaCl followed by western blot with total pol II (MABI0601) and SPT6 antibodies. β- tubulin and nucleolin are used as loading controls. Ratios of SPT6 to total pol II are indicated below each western blot.

The results of immunoprecipitation of total pol II from cell extracts followed by western analyses of total pol II and SPT6 after 15 min inhibition of CDK12as also suggest that SPT6 association with pol II is reduced (**Figure 9.22**).

Taken together, these results support the notion that CDK12 activity plays a key role in ensuring that the critical elongation factors PAF1C and SPT6 are part of the pol II elongation complex (Yu *et al.*, 2015; Yang *et al.*, 2016; Vos *et al.*, 2018). Loss of these factors could readily explain the elongation defect that we observe after CDK12 inhibition.

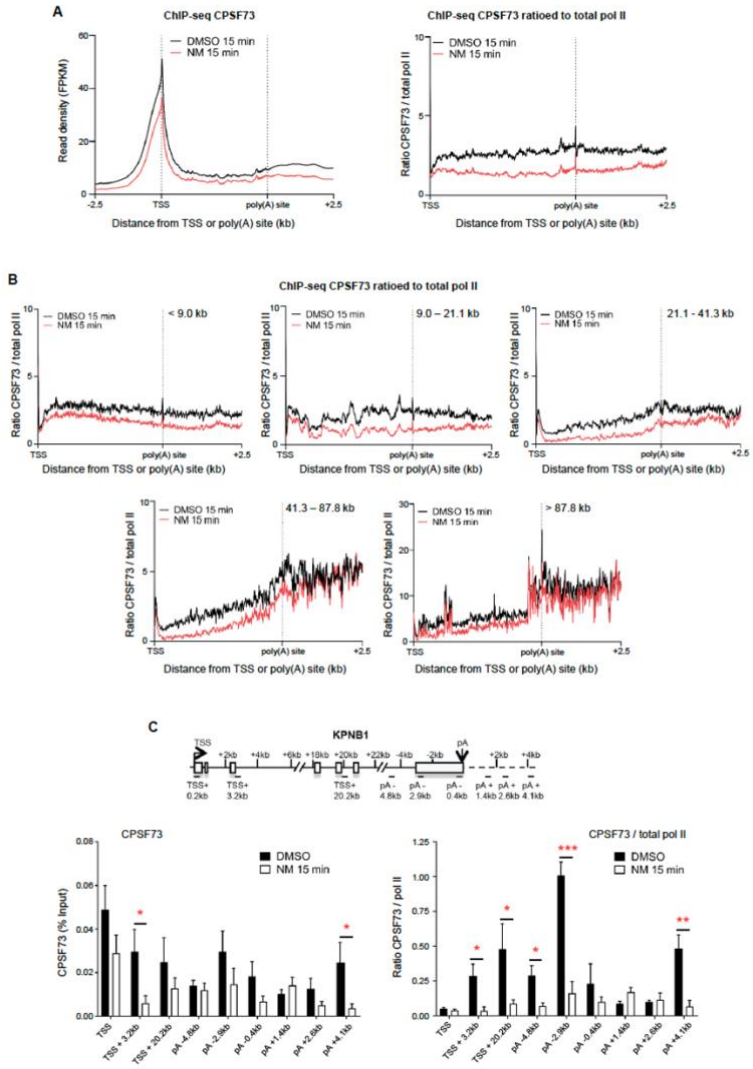


Figure 9.23 | CDK12 activity is required for stable association of the polyadenylation factor CPSF73. (A) Meta-analyses of scaled expressed genes of ChIP-seq for CPSF73 with and without normalization to pol II after treatment of CDK12as cells with DMSO (black) or 7.5μM NM (red) for 15 minutes **(B)** Metagene analyses for the different length classes noted comparing the CPSF73 signal ratioed to the total pol II mNET-seq signal before and after CDK12as inhibition for 15 minutes. **(C)** CPSF73 ChIP-qPCR unratioed (left) or ratioed to pol II (right) across *KPNB1* gene after treatment of CDK12as cells with DMSO or 7.5 μM NM for 15 minutes. Error bars = s.e.m. (n = 3 biological replicates). Asterisks indicate statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$), based on unpaired, two-tailed Student's t test.

It was previously shown that CDK12 is required for the recruitment of polyadenylation/transcription termination factors (Davidson, Muniz and West, 2014; Eifler *et al.*, 2015). To investigate further whether the loss of Ser2P and PAF1C following CDK12as inhibition affects the recruitment of termination factors, we performed ChIP-seq of CPSF73, the enzyme responsible for the cleavage of the pre-mRNA at the poly(A) site (**Figure 9.23**). We observed a general loss of CPSF73 levels after 15 min of CDK12as inhibition (**Figure 9.23A**). Importantly, the level of CPSF73 is unaffected at the 3'end of long genes (> 41.3 kb), suggesting that, similarly to LEO1 and CDC73, CPSF73 is lost from the newly-elongating pol II (**Figure 9.23B**). We confirmed the drop in CPSF73 levels detected at the same point as newly-initiated pol II after CDK12as inhibition by ChIP-qPCR on the *KPNB1* gene (**Figure 9.23C**).

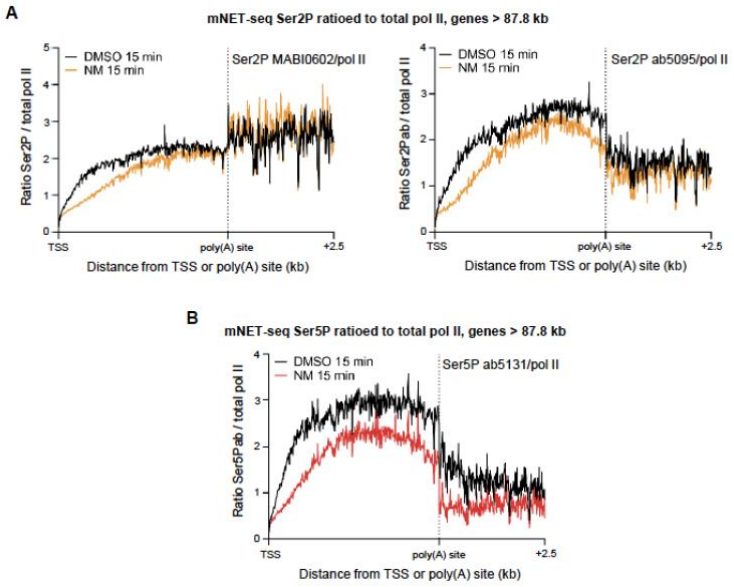


Figure 9.24 | Ser2P is unaffected at the 3'end of long genes. (A) Meta-analyses of mNET-seq for Ser2P (MABI0602 and ab5095) ratioed to pol II across genes > 87.8 kb after treatment of CDK12as cells with DMSO (black) or 7.5μM NM (orange) for 15 minutes. **(B)** Meta-analyses of mNET-seq for Ser5P (ab5131) ratioed to pol II across genes > 87.8 kb after treatment of CDK12as cells with DMSO (black) or 7.5μM NM (red) for 15 minutes.

These results are in agreement with the finding that Ser2P is retained relative to pol II at the 3' end of long genes (**Figure 9.24**). In contrast Ser5P is reduced relative to pol II at the end of long genes (**Figure 9.24**).

DISCUSSION

This section is a modified excerpt from (Tellier *et al.*, 2020)

Here, we have established a protocol to rapidly and specifically inhibit CDK12 in human cells and immediately monitor changes in RNA synthesis and the occupancy of engaged pol II genome-wide. We have restricted our analysis to the first 30 minutes after inhibition to capture the primary effect of loss of CDK12 kinase activity. An additional benefit of our strategy is the potential to lose the activity of the kinase independently of losing the protein itself. In contrast to previous studies, our results indicate that CDK12 activity is required for efficient elongation of transcription of the vast majority of expressed pol II-dependent genes, indicating that it is a general transcription kinase, rather than playing a specific role in transcription of genes longer than 45 kb (Krajewska *et al.*, 2019). Our results further suggest that CDK12 plays an increasing role in elongation after pol II has passed the CDK9-regulated early-elongation checkpoint (EEC) and entered productive elongation (**Figures 9.11A and 9.11B**) (Jonkers and Lis, 2015). It has been shown that the pol II elongation rate increases during transcription (Jonkers, Kwak and Lis, 2014) and CDK12 may therefore play a key role in pol II speeding up as it travels away from the EEC, explaining the increase in the elongation defect as pol II moves further from the TSS. Expression of long genes will therefore be disproportionately affected by an elongation defect and this will be amplified over time. This would therefore explain why CDK12 knockdown experiments show a loss of expression of long genes, in particular DDR genes. In addition, as CDK12 makes important contacts with RNA processing and elongation factors (Eifler *et al.*, 2015; Liang *et al.*, 2015; Yu *et al.*, 2015), the long-term loss of the protein itself may contribute to the gene-type-specific defects in the level of mRNA from long DDR genes. A recent study with CDK12as HCT116 cells also indicates that CDK12 kinase activity is specifically required for transcription of long DDR genes (Chirackal Manavalan *et al.*, 2019). However, in this case, pol II ChIP-seq was performed after 4.5 hours inhibition, which will result in a bias in the analysis towards detection of a defect on long genes.

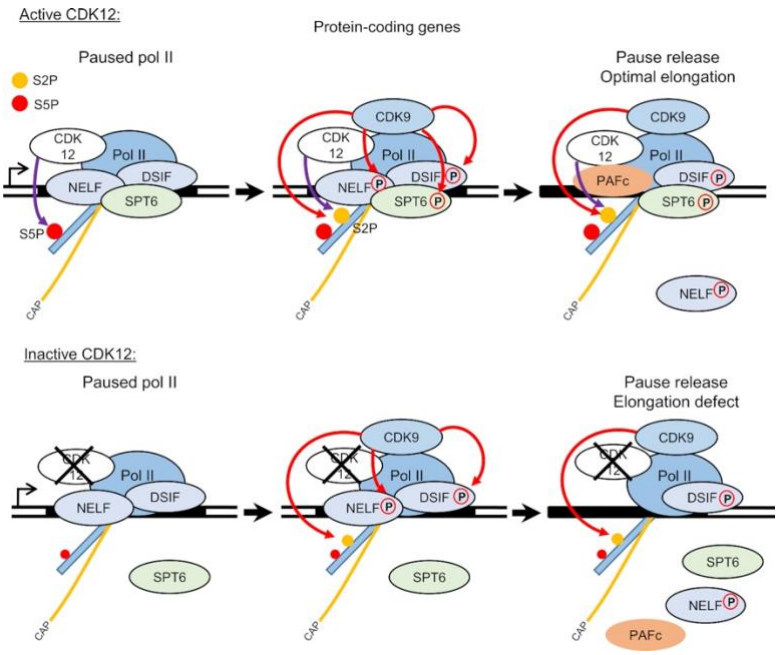


Figure 9.25 | Model of CDK12 function in transcription. During the transcription cycle, CDK12 phosphorylates Ser2 and Ser5 of the pol II CTD and possibly other factors involved in transcription such as SPT6. Inhibition of CDK12 affects the pol II elongation rate by affecting the recruitment and/or stability on the chromatin of LEO1 and CDC73 from the PAF1 complex (PAF1C) and SPT6. CDK12 inhibition does not affect pol II pause release, which is mediated by CDK9. Orange line with 'cap': capped mRNA; orange and red dot: Ser2P and Ser5P, respectively. Red and violet arrows: phosphorylation mediated by CDK9 and CDK12, respectively.

The effect of CDK12as inhibition on elongation of transcription is readily explained by the loss of association of LEO1 and CDC73, which are key components of the PAF1C elongation complex, and the elongation factor SPT6. We cannot rule out that CDK12as inhibition causes rapid nuclear export of these factors or that their association with chromatin appears reduced due to epitope masking caused by changes in post-translational modifications. However, CDK12as inhibition not only causes a decrease in the level of elongation factors on chromatin but an increased level in the nucleoplasm as would be expected if CDK12 activity is required for their effective recruitment to and/or stable association with pol II. CDK12 has already been shown to interact with the PAF1 complex (Yu *et al.*, 2015). It has also recently been shown that the CDK12/CDK13 inhibitor, THZ531, affects phosphorylation of SPT6 (Krajewska *et al.*, 2019). Phosphorylation of SPT6 by CDK12 could therefore play a role in contacts between SPT6 and other components of the elongation complex, including pol II and PAF1C (Jonkers and Lis, 2015). In turn, loss of SPT6 association may further destabilize the elongation complex (Nojima *et al.*, 2018). We therefore favour the notion that CDK12 inhibition causes reduced association of elongation factors with pol II through the loss of phosphorylation of one or more of its targets. Both loss of CDK12 and inhibition of

CDK12 and CDK13 by THZ531 lead to the activation of gene-internal poly(A) sites and premature termination of transcription (Dubbury, Boutz and Sharp, 2018; Krajewska *et al.*, 2019), which may be the direct result of slower elongation as suggested (Krajewska *et al.*, 2019) (**Figure 9.25**).

We also observed a loss of the polyadenylation factor CPSF73 from newly-elongating pol II (**Figure 9.22**). It was previously shown that CDC73 can interact with CPSF73 (Rozenblatt-Rosen *et al.*, 2009) and our data support this observation as both PAF1C and CPSF73 are lost from the newly-elongating pol II. CTD Ser2P also plays a role in CPSF73 recruitment (Davidson, Muniz and West, 2014). At the 3' end of long genes, the retention of CDC73 and Ser2P would therefore be sufficient to recruit CPSF73 to/stabilize its interaction with the pol II elongation complex, in the absence of CDK12 activity. As Ser5P is somewhat lost from the 3' end of long genes, this mark is likely to be less important for PAF1C/CPSF recruitment at this point.

Notably, the effect of inhibiting CDK12 differs markedly from the effect of CDK9 inhibition, which causes a drastic loss of elongating pol II downstream of the early-elongation checkpoint (EEC) (Laitem *et al.*, 2015; Gressel *et al.*, 2017). CDK9 and CDK12 therefore play non-redundant roles in ensuring efficient transcription of pol II-dependent genes. However, these two kinases are intimately connected as CDK9 activity is required for PAF1C recruitment and PAF1C in turn recruits CDK12 (Yu *et al.*, 2015; Zaborowska, Egloff and Murphy, 2016). In addition, inhibition of both CDK12 and CDK9 causes premature termination close to poly(A) sites where Ser2P is peaking (Laitem *et al.*, 2015; Tellier, Ferrer-Vicens and Murphy, 2016), indicating that the activity of both kinases is required for the correct transition between elongation and termination, through their phosphorylation of Ser2 or other targets.

CDK12 knockdown and THZ531 inhibition studies indicate that CDK12 phosphorylates Ser2 of the human pol II CTD (Bartkowiak *et al.*, 2010; Blazek *et al.*, 2011; Zhang *et al.*, 2016), whereas *in vitro* CDK12 has robust Ser5 kinase activity (Bösken *et al.*, 2014; Bartkowiak and Greenleaf, 2015). We find that detectable losses of Ser2P and Ser5P occur rapidly after CDK12as inhibition, with the biggest effect on Ser5P. Relevant to this, specific inhibition of CDK12as in HeLa cells results in a decrease in detection of pol II by both the H14 and H5 antibodies, which recognize Ser5P and Ser2P, respectively (Chapman *et al.*, 2007; Heidemann *et al.*, 2013; Bartkowiak, Yan and Greenleaf, 2015). A recent study with a CDK12as HCT116 cell line observed a decrease in Ser5P but not of

Ser2P following CDK12 inhibition (Chirackal Manavalan *et al.*, 2019). We have shown here that association of a Ser5P-binding capping factor with genes is affected by CDK12as inhibition. Our results therefore support the notion that CDK12 is a major Ser5 kinase. However, based on previous findings, we would have expected inhibition of CDK12as to cause a greater loss of Ser2P. Thus, in at least some human cells, the Ser2 kinase activity of CDK12 may be somewhat redundant with other Ser2 kinases, such as CDK9.

CONCLUDING REMARKS

This section is a modified excerpt from (Tellier *et al.*, 2020)

In conclusion, we have shown that short-term inhibition of CDK12 causes a genome-wide defect in pol II mediated transcription elongation by decrease of pol II elongation velocity and processivity as measured by TT-seq and loss of PAF1C components (elongation factors LEO1 and CDC73) and SPT6 from the newly-elongating pol II. Our findings also indicate CDK12 as a global regulator of Ser2 and Ser5 phosphorylation of pol II CTD *in vivo* as measured by mNET-seq and western blot. Taken together, our results imply that CDK12 is a general activator of pol II transcription elongation. It remains to be determined whether CDK12 function in transcription elongation is mediated through pol II CTD phosphorylation and/or through phosphorylation of other targets such as SPT6.

Kinetic modeling of transcription predicts dynamic RNA synthesis and polymerase occupancy profiles

Distinctive contribution:

Implementation of kinetic model described in **section 2.2.7** for experimental dataset

- the pertinent write-up of the analysis

INTRODUCTION

Transcription of the genome is a dynamic process that is regulated mainly during the initiation and elongation stages. Insights into the temporal dynamics of transcription are readily obtained by sequencing newly synthesized RNA and by localizing the transcribing enzyme, RNA polymerase II (Pol II), over the genome at different time points after cellular perturbation. Understanding such highly complex genome-wide data requires kinetic modeling of transcription. Available transcription models however generally focus on particular stages of the process and mostly built on stochastic parameters that usually do not take dynamics into account (Golding, Paulsson, Zawilski, & Cox, 2005; Honkela et al., 2015; Xu, Skinner, Sokac, & Golding, 2016; Gorin, Wang, Golding, & Xu, 2020). This necessitates the implementation of a more complete kinetic model of transcription that can explain and predict changes in genome-wide profiles as a function of time.

Here we present a kinetic model and simulation platform that can capture the dynamics of gene transcription. We introduce a mathematical model for calculating elongation velocity profiles $v(t)$ along a typical gene of length *L*. The mathematical model is adopted on the basis of classical mechanics which describes the behavior of a biological system in terms of movement as a function of time. If the dynamics of a biological system is known, the equations are the solutions for the differential equations describing the dynamics. Implementing the mathematical model with an instance of initial constant acceleration, elongation velocity profiles $v(t)$ can be derived for the polymerases for a defined time frame. As a function of time, the elongation velocity profiles subsequently provide us with a snapshot of the polymerase positions. These parameters modeled as a function of time captures the dynamics of transcription kinetics. The simulation with an initial set of constant parameters can capture the instance of steady state transcription in biological system.

The model also includes variable parameters such as the initiation frequency, the pause duration and the RNA labeling time. The mathematical model can recapitulate the nascent RNA expression profiles or RNA pol II occupancy profiles obtained from the experimental dataset for similar conditions. The nascent RNA expression profiles are obtained from transient transcriptome sequencing (TT-seq) (Schwalb et al., 2016) where the nascent RNA corresponds to the fragmented RNA obtained during 4sU labeling duration. The RNA occupancy profiles are obtained from mNET-seq which detects the

actively transcribing RNA through the capture of 3' end of the RNA (Churchman & Weissman, 2011; Nojima et al., 2015).

We have previously used kinetic modeling of transcription to investigate a number of biological phenomena. In particular, we have investigated three major parameters of transcription regulation, the frequency of initiation (*i.e.* how many polymerases initiate transcription at a given gene promoter per minute), the duration of polymerase pausing in the promoter-proximal region (i.e. how long polymerases spend on average within a pause window downstream of the transcription start site) and the velocity of transcription elongation (i.e. the speed of the polymerase within gene bodies). The model was first conceived in investigating promoter-proximal pausing (Gressel, Schwalb, & Cramer, 2019; Gressel et al., 2017). Further, we demonstrate the applicability of the model with the use of datasets obtained after rapid inactivation of three different cyclin dependent kinases (CDK7, CDK9, and CDK12) that have been implicated in regulating transcription initiation, pausing, and elongation, respectively. This study compared the nascent RNA expression profiles from TT-seq data for CDK's with the profiles from the mathematical model which are highly relatable. Thus, the reliability of the mathematical model can be emphasized. Using the simulation, a user can input simulation parameters and examine time dependent dynamic behavior of transcription. This approach presents a new mathematical model of time dependent dynamic behaviors of gene expression as profiles for multiple variable parameters. With the validation with time-series data, the simulation can be implemented to interpret complex dynamic transcription behaviors without making assumptions.

RESULTS

10.1 Kinetic modeling provides insight into the effect of altered transcriptional parameters (initiation frequency, pausing duration and elongation velocity) on transcription

The model presumes that within a cell, steady-state transcript levels are maintained with a constant initiation frequency, pausing duration and elongation velocity over time. The kinetic model described in the methods section **2.2.7** can therefore calculate the steady state transcription level with a set of variable transcriptional parameters to initialize with and subsequently any alteration in transcriptional levels in response to the change in transcriptional parameters. The change in the parameters can be assigned by either increasing or decreasing the value. The altered transcriptional parameters can be visualized

as TT-seq coverage which simulates the RNA/transcription levels and mNET-seq coverage which simulates the positioning of the polymerases.

10.2 Simulating initiation defect: perturbing transcription by altering initiation rate

Transcription can be perturbed on the level of initiation by altering the initiation frequency (**Figure 10.1**). An altered lower initiation frequency implies a lower number of polymerases initiating transcription (**Figure 10.1A**) whereas higher initiation frequency indicates initiation event by a higher number of polymerases (**Figure 10.1B**). As less polymerase initiates transcription when initiation frequency is lower, initially it results in a decrease of the transcription level around the transcription start site and less number of polymerases in the gene body. The steady-state level of transcription in the gene body implies the transcription of residing polymerases before the initiation frequency is altered. Eventually the residing polymerases are terminated and the gene body is populated with polymerases with the new initiation frequency which results in an overall decrease of transcription level along the gene body (**Figure 10.1, left panel**). On the contrary, with an altered higher initiation frequency an increase of the transcription level and polymerase occupancy around the transcription start site can be observed initially and the signal will gradually spread throughout the gene body as the population of new polymerases accumulate (**Figure 10.1, right panel**).

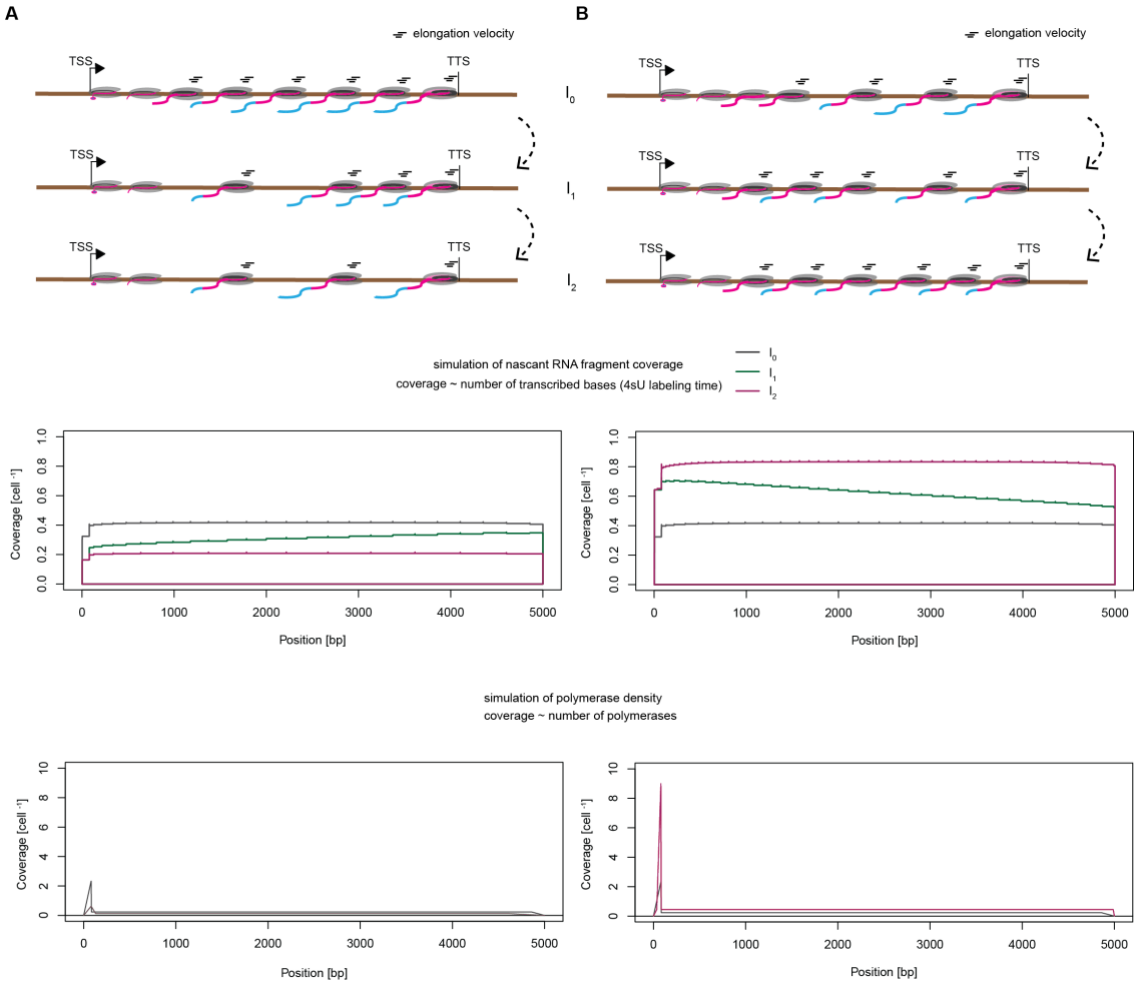


Figure 10.1 | Influence of altered initiation rate (*I*) on RNA synthesis profiles and polymerase positioning profiles on a single gene. I_0, I_1 and I_2 denotes sequential increasing time duration. (A) Profiles for altered decreased initiation rate (B) Profiles for altered increased initiation rate

10.3 Simulating pausing defect: perturbing transcription by altering pause duration

Change in the pause duration can also alter transcriptional gene expression (**Figure 10.2**). A decreased pause duration indicates shorter residing time of paused polymerases in the promoter-proximal region (**Figure 10.2A**). This results in an elevated level of transcription from the transcription start site up to the pause position due to the possibility of more polymerases initiating transcription as a result of less populated promoter-proximal pause region. On contrary, longer residing time of paused polymerases in the promoter-proximal region is implied by increased pause duration. This results in a densely populated promoter-proximal pause region which inhibits new transcription resulting in a decrease of the transcription level from the transcription start site to the pause position (**Figure 10.2B**).

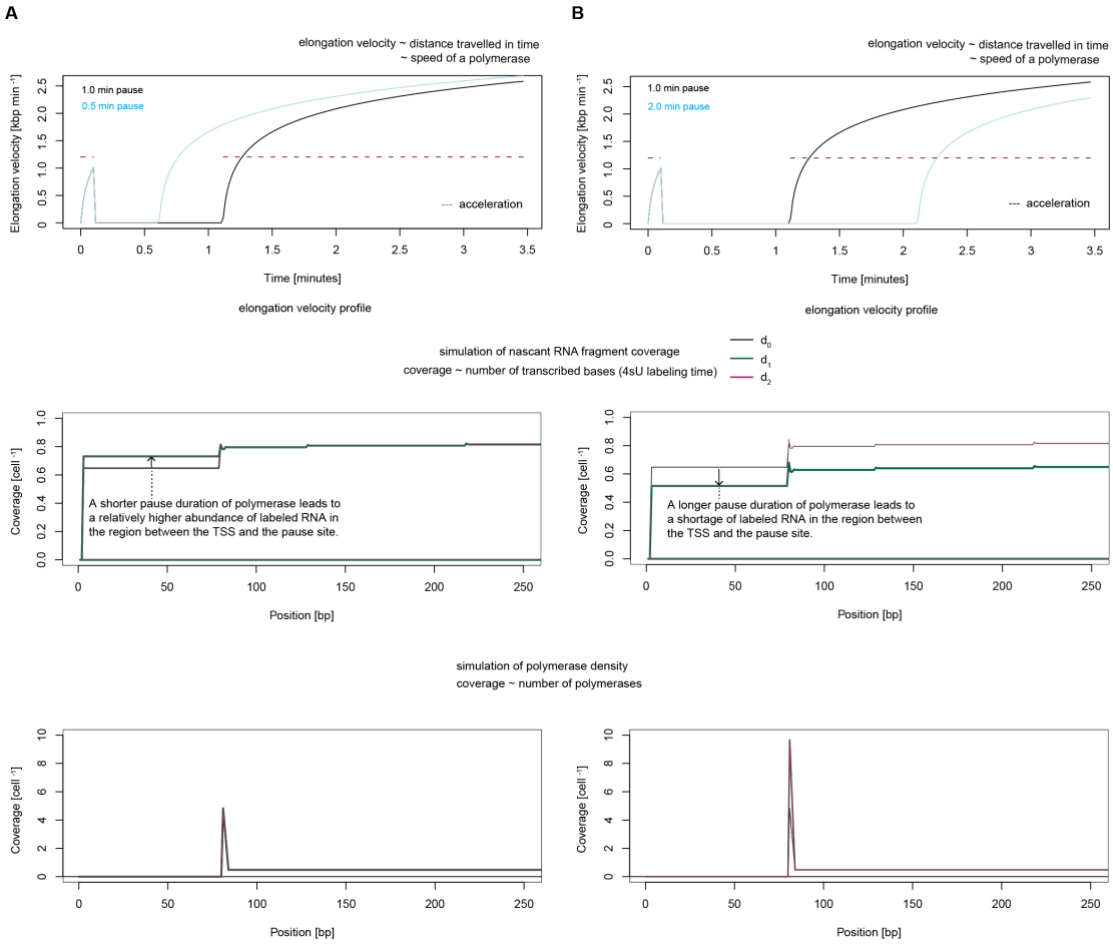


Figure 10.2 | Influence of altered pause duration (*d*) on RNA synthesis profiles and polymerase positioning profiles on a single gene. d_0, d_1 and d_2 denotes sequential increasing time duration. (A) Profiles for altered decreased pause duration (B) Profiles for altered increased pause duration

10.4 Simulating elongation defect: perturbing transcription by altering elongation velocity

By altering the elongation velocity, transcription can also be perturbed on the level of elongation (**Figure 10.3**). An altered lower elongation velocity initially decreases the overall transcription level as the elongating polymerases residing in the gene body adapt the altered elongation velocity profile. As new polymerases with the slower elongation velocity profile initiates transcription, a higher number of slow elongating polymerases accumulate along the gene body resulting in elevation of the transcript levels to the initial steady-state (**Figure 10.3A**). Conversely, the overall transcription level along the gene body is increased with an alteration to higher elongation velocity as residing polymerases elongate faster. A flattening of the transcript levels to the initial steady-state will be in effect as the gene body is gradually populated with faster polymerases (**Figure 10.3B**).

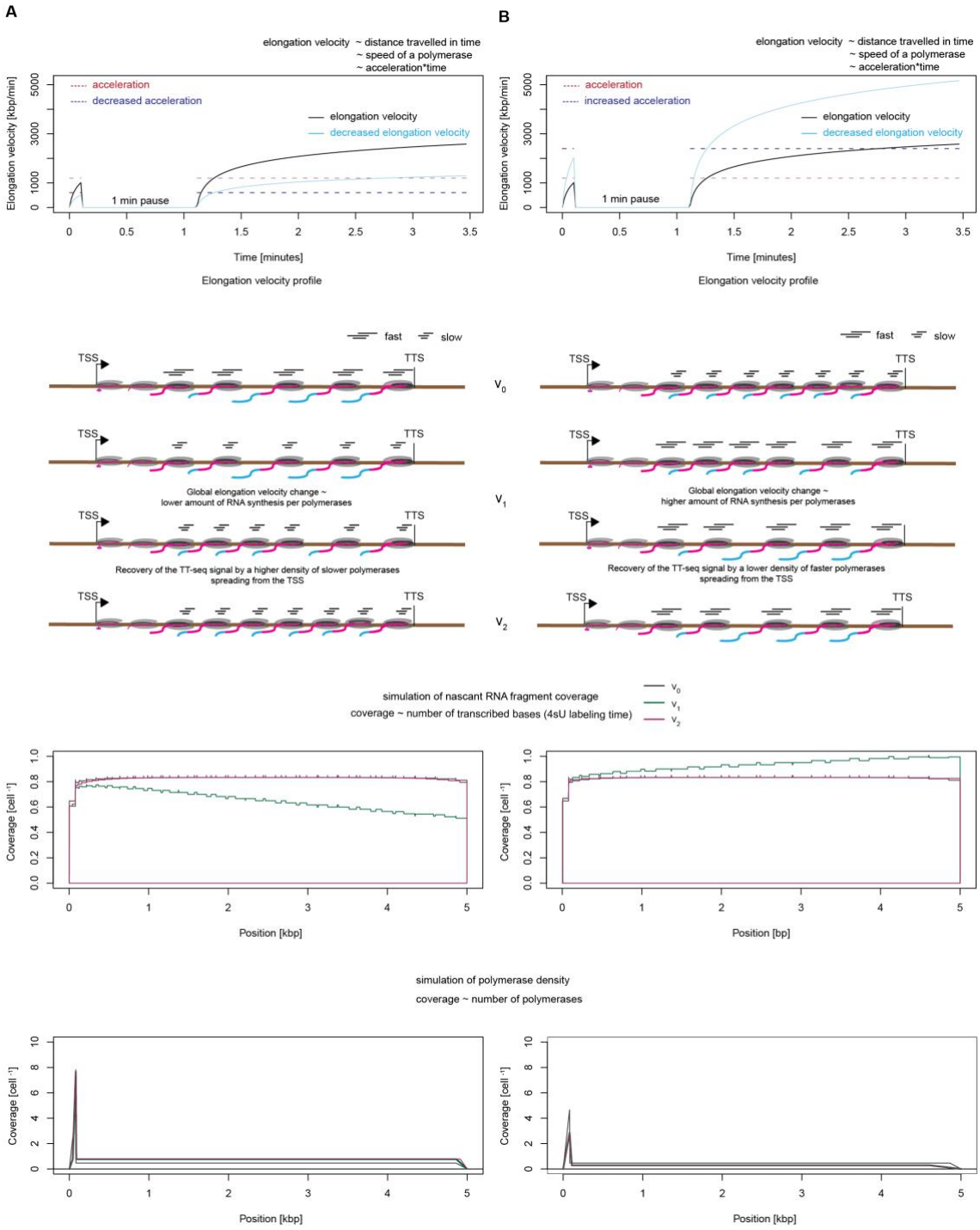


Figure 10.3 | Influence of altered elongation velocity (*v*) on RNA synthesis profiles and polymerase positioning profiles on a single gene. v_0, v_1 and v_2 denotes sequential increasing time duration. (A) Profiles for altered decreased elongation velocity (B) Profiles for altered increased elongation velocity

10.5 Simulation of TT-seq and mNET-seq data as metagene profiles

The simulation model described in the methods section **2.2.7** was implemented for steady state and altered transcriptional parameters (initiation frequency, pausing duration and elongation velocity) for gene length templates resembling genes of sizes ~0,1- 2000 kbp.

The resulting gene-wise RNA synthesis profiles (TT-seq) were subsequently accumulated to yield meta-gene profiles (**Figure 10.4**).

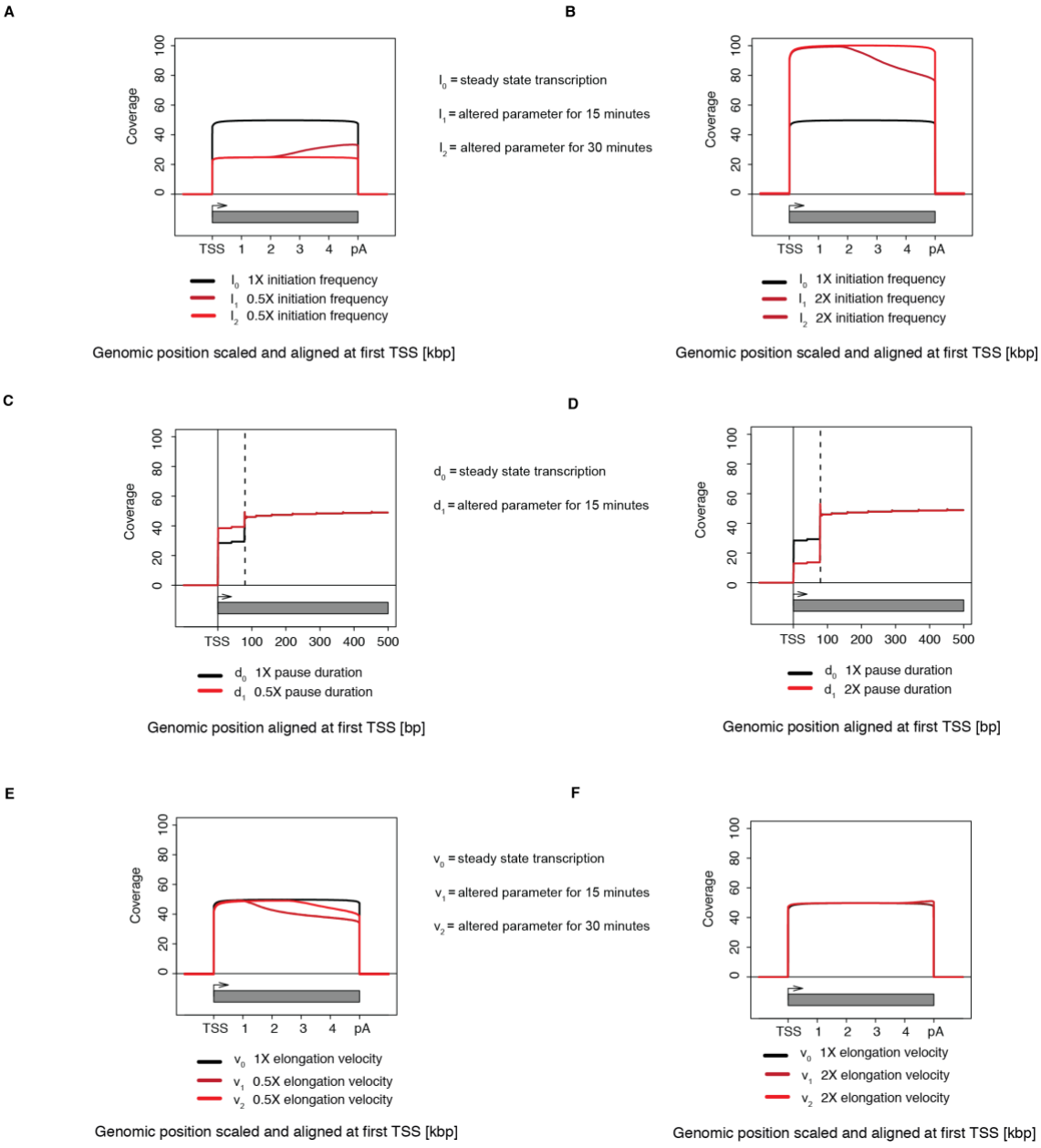


Figure 10.4 | TT-seq metagene profiles from the kinetic model for altered kinetic parameters. (A) TT-seq metagene profiles comparing steady state transcription (black) and altered 0.5X initiation frequency (red). **(B)** TT-seq metagene profiles comparing steady state transcription (black) and altered 2X initiation frequency (red). Time progression for **(A)** and **(B)** is denoted as I_o, I_1 and I_2. **(C)** TT-seq metagene profiles comparing steady state transcription (black) and altered 0.5X pause duration (red). **(D)**TT-seq metagene profiles comparing steady state transcription (black), and altered 2X pause duration (red). Time progression for **(C) and (D)** is denoted as d_o and d_1 and the pause position is indicated in dotted line. **(E)** TT-seq metagene profiles comparing steady state transcription (black), and altered 0.5X elongation velocity (red). **(F)** TT-seq metagene profiles comparing steady state transcription (black), and altered 2X elongation velocity (red). Time progression for **(E)** and **(F)** is denoted as v_o, v_1 and v_2.

10.6 Comparing simulation data with experimental data

To evaluate if the kinetic model described in **section 10.5** can recapitulate and explain profiles from experimental dataset, we compared the RNA synthesis profiles (TT-seq) from simulation (**Figures 10.1-10.4**) with TT-seq metagene profiles from time course experimental dataset generated for analog sensitive kinases inhibited with small molecule inhibitors (**Figure 10.5-10.7**). Here we compare the effect of inhibition on three different kinases implicated in transcription initiation CDK7, promoter-proximal pause release CDK9 and elongation CDK12.

10.7 Metagene profiles of CDK7as inhibition mimics the profiles of lower transcription initiation obtained from kinetic modeling

TT-seq was performed with RNA spike-ins after 15 and 30 minutes of NM-PP1 treatment on CDK7as HEK293 cells to monitor the immediate changes in RNA synthesis (**Figure 10.5A**). Metagene profiles of TT-seq signals averaged over expressed genes in CDK7as cells show a decreased TT-seq signal at the beginning of genes after 15 minutes of inhibition (**Figure 10.5B, left panel**). CDK7as inhibition for 30 minutes in turn shows a decreased TT-seq signal across gene bodies (**Figure 10.5B, right panel**). The observed profile of decreased RNA synthesis can be explained by the TT-seq profiles obtained from the kinetic model with decreased initiation frequencies (**Figures 10.1A and 10.4A**) which readily recapitulates the observed TT-seq profiles. Thus, comparison of TT-seq analysis and kinetic modelling supports the notion that CDK7 is involved in transcription initiation.

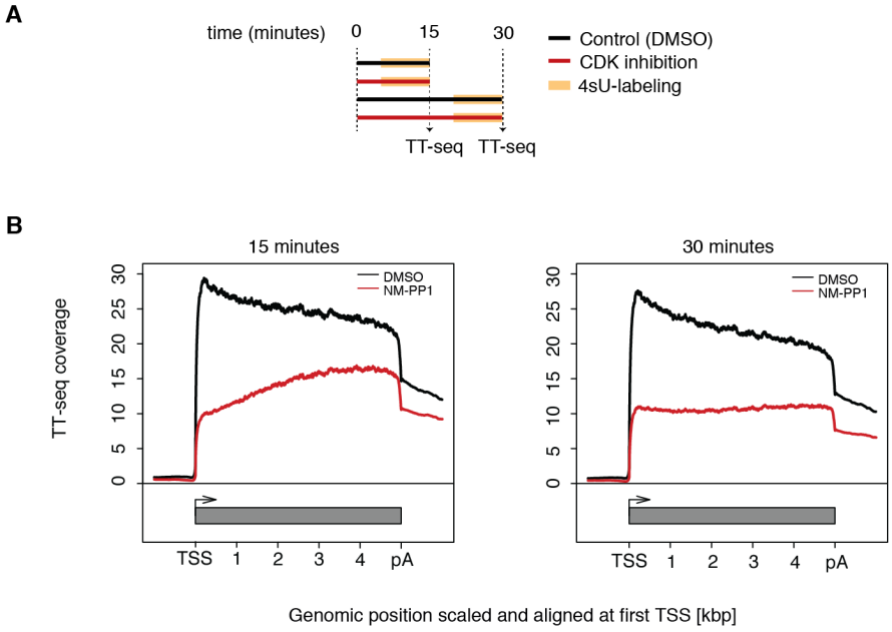


Figure 10.5 | TT-seq metagene profiles from time course experiment for analog sensitive CDK7 kinase inhibited with small molecule inhibitors. (A) Experimental setup for TT-seq for inhibition of CDK7 kinase with small molecule inhibitors in analog sensitive cell line. **(B)** Metagene profiles of TT-seq signal for expressed genes in CDK7as HEK293 cell line after DMSO treatment (black) and NM-PP1 treatment (red) for 15 minutes (left) or 30 minutes (right). The TT-seq coverage is averaged and aligned at their transcription start sites (TSSs) and polyadenylation (pA)-sites.

10.8 Metagene profiles of CDK9as inhibition replicates the profiles of increased paused duration generated with kinetic modeling

To capture the changes in actively transcribed RNA, TT-seq was performed with RNA spike-ins after 15 minutes of NA-PP1 treatment on CDK9as Raji cells (**Figure 10.6A**). A decreased TT-seq signal averaged over expressed genes in CDK9as cells is observed from the beginning of genes after 15 minutes of inhibition (**Figure 10.6B**). The decreased RNA synthesis profile is comparable with the TT-seq profiles obtained from the kinetic model with increase pause duration (**Figures 10.2B and 10.4D**). From this comparison, it can be inferred that CDK9 plays a role in promoter proximal pausing.

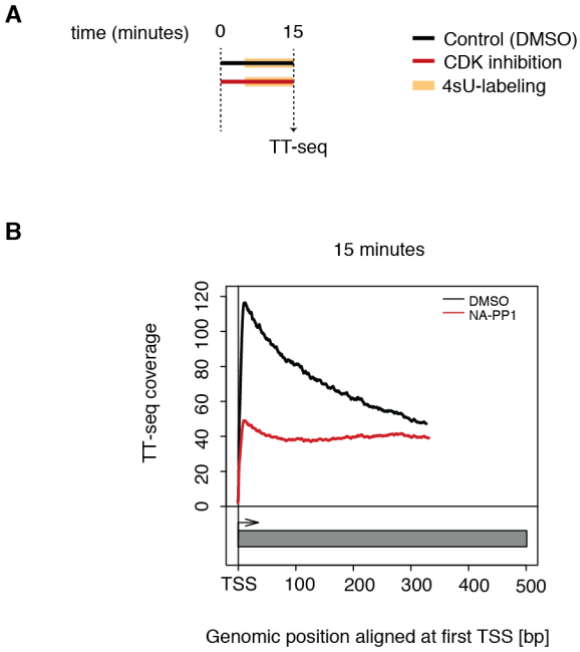


Figure 10.6 | TT-seq metagene profiles from time course experiment for analog sensitive CDK9 kinase inhibited with small molecule inhibitors. (A) Experimental setup for TT-seq for inhibition of CDK9 kinase with small molecule inhibitors in analog sensitive cell line. **(B)** Metagene profiles of TT-seq signal for expressed genes in CDK9as Raji cell line after DMSO treatment (black) and NM-PP1 treatment (red) for 15 minutes. The TT-seq coverage is aligned at their transcription start sites (TSSs).

10.9 The effect of CDK12as inhibition can be explained by the profiles generated for altered lower elongation velocity

To monitor changes in nascent RNA synthesis upon CDK12as inhibition on HEK293 cells, TT-seq was performed with RNA spike-ins after 15 and 30 minutes of NM-PP1 treatment (**Figure 10.7A**). Metagene profiles of TT-seq signals averaged over expressed genes in CDK12as cells show a decreased TT-seq signal across gene bodies after 15 minutes of inhibition, indicating reduced RNA synthesis, either as the result of reduced elongation or loss of polymerase (**Figure 10.7B, left panel**). CDK12as inhibition for 30 minutes instead leads to a recovery of RNA synthesis activity at the beginning of genes (**Figure 10.7B, right panel**). This effect can be recapitulated with the kinetic model with a decreased elongation velocity and a constant initiation frequency (Figure 5F/8C). Taken together, CDK12 is required for normal transcription elongation is supported by both TT-seq analysis and kinetic modelling.

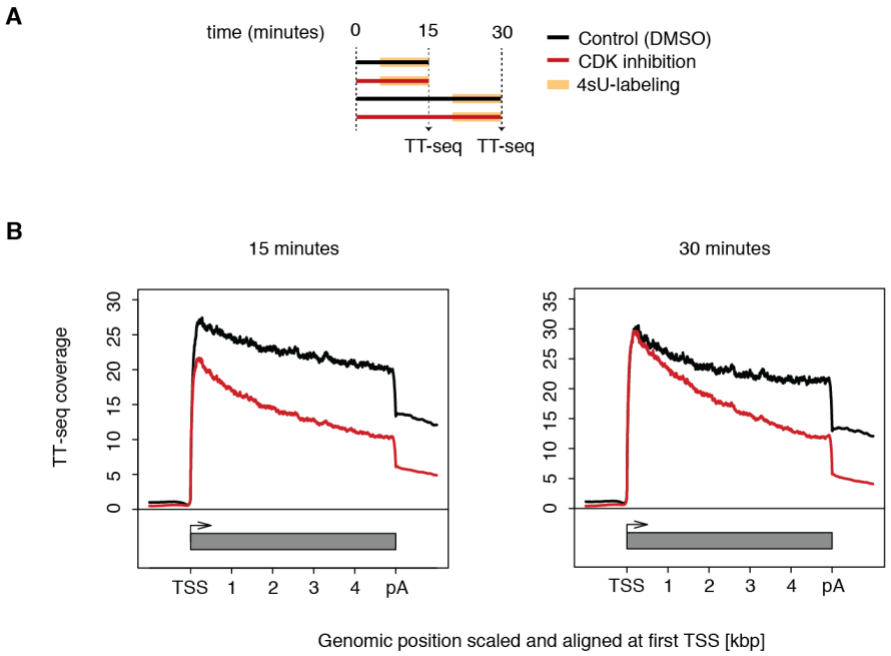


Figure 10.7 | TT-seq metagene profiles from time course experiment for analog sensitive CDK12 kinase inhibited with small molecule inhibitors. (A) Experimental setup for TT-seq for inhibition of CDK9 kinase with small molecule inhibitors in analog sensitive cell line. **(B)** Metagene profiles of TT-seq signal for expressed genes in CDK12as HEK293 cell line after DMSO treatment (black) and NM-PP1 treatment (red) for 15 minutes (left) or 30 minutes (right). The TT-seq coverage is averaged and aligned at their transcription start sites (TSSs) and polyadenylation (pA)-sites.

CHAPTER 4

FUTURE PERSPECTIVES

The chapter concludes the thesis with the authors view on regulation of transcription by cyclin dependent kinases.

Transcriptional cyclin dependent kinases have developed complex regulatory mechanisms by phosphorylation of their targets to maintain an active transcription cycle. Since the discovery of tCDKs, the structure and function of kinases is of great interest in both basic and applied research. The underlying mechanism of individual tCDKs activity still remains a fundamental question in biology. In addition, the interaction network between these CDKs to regulate biological processes also remains to be more fully elucidated. To understand the precise biological role and deconvolute kinase mediated transcriptional network it is necessary to identify specific targets and correlate expression pattern in biological system. Work presented in this thesis aimed at investigating these mechanisms behind the regulation of transcription cycle by the CDKs. This thesis work has provided new insight into the molecular mechanism of how tCDKs can stimulate the expression of Pol II transcribed genes. Some of the challenges arising in investigating individual transcriptional kinase activity are discussed in following paragraphs.

11.1 Evolution of cyclin dependent kinases: All for one and one for all

tCDKs share considerable structural similarities in their catalytic domain which results in overlapping function in regulating different transcriptional cycle (Zheng, 2022). However, sometimes despite being highly conserved, tCDKs use very different mechanisms of transcriptional regulation. The striking similarity with opposing functions introduces complexity in dissecting the function of individual kinases. Additionally, tCDKs are normally activated by their binding partner cyclins. Despite the structural similarity of tCDKs, the cyclins share little similarity (Malumbres, 2014). Taken together, disentanglement of tCDK network is a challenging issue. For decades, scientists have been trying to unravel the enduring mystery of individual kinase function that can provide us with knowledge to piece together the transcriptional kinase network. But the structural similarity in the catalytic domain makes it very difficult to achieve desired level of kinase selectivity for the kinase inhibition.

11.2 Solving the transcriptional kinase puzzle: one kinase at a time

The most common approach to study the function of individual transcriptional cyclin dependent kinases is the use of chemical inhibitors (Ettl, Schulz and Bauer, 2022). But chemical inhibition has low specificity for its substrate in other words broad kinase inhibition profile which deludes us from finding the primary function of a kinase. Additionally, longer inhibition duration can introduce secondary effect to circumvent the primary function of individual kinases. The discovery of analog sensitive kinase inhibition opened a new door to

systematically analyze the primary function of individual kinases (Lopez, Kliegman and Shokat, 2014). The success of these method largely depends on synthesizing inhibitors to specifically inhibit kinase function in addition to functional mutation of the kinase. Most of the inhibitors in this chemical-genetic approach are designed to be highly specific, but toxicity of an inhibitor can affect the function of a kinase. Also, most of the inhibitors are rapid in their effect of inhibiting a kinase but the reversibility is not well studied. Thus, even with this powerful technology, long term inhibition may provide us with information of secondary effect of a kinase. To differentiate between the primary effects from the secondary one, an additional RNA interference-mediated time-course inhibition can be combined with the analog sensitive approach. The efficacy of the inhibitor is also a crucial point while applying this technique to study the function of kinases. Despite the limitations, analog sensitive kinase technology is the best available technology to study the function of individual kinases in recent times. A short inhibition duration combined with appropriate methodology can provide us with valuable insight into individual kinase function and eventually, piecing together the pieces of puzzle, in a larger frame the transcriptional kinase network.

11.3 Combining -omics with structural studies: two are better than one if two act as one

The information encoded in multicellular organism is multilayered. Transcription is the first step to read and transfer the encoded information in a reliable and deterministic manner. To systematically characterize the transcriptional regulation, the combination of different -omics approach namely multi-omics is absolutely crucial. Combining transcriptomics with proteomics and structural genomics gives us novel insight into the exact functionality of individual CDKs. Transcriptomics, also known as functional genomics is the study of RNA functions and interactions. Different functional genomics methods are available to reveal different layer of information. Careful integration of this layered information is necessary for accurate interpretation of underlying molecular mechanism mediated by CDKs. Proteomics study helps to identify particular targets of an individual kinase which helps us to unravel the associated mechanism combining with transcriptomics. Additionally, structural studies are an important link to decipher the interaction-based activity of a CDK. Combining structural studies can help us to determine how structure influences the function of a CDK and vice versa. Knowing exact molecular mechanism will helps us to understand its impact on transcriptional regulation. Taken together, -omics studies can give us novel insight of the human kinome.

11.4 Transferring basic research to applied science: knowledge shared is power multiplied

The knowledge generated by leveraging the multi-omics data is necessary to understand the underlying mechanism of a kinase to exploit its therapeutic potential. This knowledge leads to identify and test selective inhibitors for therapeutic intervention and gain insight into drug resistance. The deregulation of tCDK activity is implicated in cancer extensively in recent times (Bury *et al.*, 2021; Jhaveri *et al.*, 2021; Thoma, Neurath and Waldner, 2021; Zhang *et al.*, 2021; Panagiotou *et al.*, 2022). Thus, tCDK family has been an intense area of research to develop more specific chemical inhibitors as drugs to target the tCDK. One prominent advancement in recent time in this respect is identifying a highly selective inhibitor Q901 which only inhibits CDK7 in the human kinome *in vivo*. This selective inhibition kills cancer cells with high specificity by targeting aberrant cell cycle and transcription regulation. This novel cancer therapy is now on clinical trial (https://www.max-planck-innovation.com/max-planck-innovation/news/press-releases/press-release/novel-cancer-therapy-enters-clinical-trials.html). Convincingly, in the coming years, tCDKs biology would continue receiving considerable attention promise in oncology.

CHAPTER 5

REFERENCES

The chapter includes the references cited in this thesis.

Agrawal, N. *et al.* (2003) 'RNA Interference: Biology, Mechanism, and Applications', *Microbiology and Molecular Biology Reviews*. American Society for Microbiology (ASM), 67(4), p. 657. doi: 10.1128/MMBR.67.4.657-685.2003.

Allen, B. L. and Taatjes, D. J. (2015) 'The Mediator complex: a central integrator of transcription', *Nature reviews. Molecular cell biology*. Nat Rev Mol Cell Biol, 16(3), pp. 155–166. doi: 10.1038/NRM3951.

Anders, S. and Huber, W. (2010a) 'Differential expression analysis for sequence count data.', *Genome biology*, 11(10), p. R106. doi: 10.1186/gb-2010-11-10-r106.

Anders, S. and Huber, W. (2010b) 'Differential expression analysis for sequence count data', *Genome Biology*, 11(10), p. R106. doi: 10.1186/gb-2010-11-10-r106.

Anders, S., Pyl, P. T. and Huber, W. (2015a) 'HTSeq--a Python framework to work with high-throughput sequencing data', *Bioinformatics (Oxford, England)*. Bioinformatics, 31(2), pp. 166–169. doi: 10.1093/BIOINFORMATICS/BTU638.

Anders, S., Pyl, P. T. and Huber, W. (2015b) 'HTSeq - a Python framework to work with high-throughput sequencing data', *Bioinformatics*. Oxford University Press, 31(2), pp. 166–169. doi: 10.1093/bioinformatics/btu638.

Andrau, J. C. *et al.* (2006) 'Genome-wide location of the coactivator mediator: Binding without activation and transient Cdk8 interaction on DNA', *Molecular cell*. Mol Cell, 22(2), pp. 179–192. doi: 10.1016/J.MOLCEL.2006.03.023.

Andrews, S. (2010) 'FastQC: a quality control tool for high throughput sequence data'.

Ansari, A. Z. *et al.* (2002) 'Transcriptional activating regions target a cyclin-dependent kinase', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 99(23), pp. 14706–14709. doi: 10.1073/PNAS.232573899.

Anshabo, A. T. *et al.* (2021) 'CDK9: A Comprehensive Review of Its Biology, and Its Role as a Potential Target for Anti-Cancer Agents', *Frontiers in Oncology*. Frontiers Media S.A., 11, p. 1573. doi: 10.3389/FONC.2021.678559/BIBTEX.

Aoi, Y. *et al.* (2020) 'NELF Regulates a Promoter-Proximal Step Distinct from RNA Pol II Pause-Release', *Molecular cell*. Mol Cell, 78(2), pp. 261-274.e5. doi: 10.1016/J.MOLCEL.2020.02.014.

Aristizabal, M. J. and Kobor, M. S. (2016) 'A single flexible RNAPII-CTD integrates many different transcriptional programs', *Transcription*. Transcription, 7(2), pp. 50–56. doi: 10.1080/21541264.2016.1163451.

Barba-Aliaga, M., Alepuz, P. and Pérez-Ortín, J. E. (2021) 'Eukaryotic RNA Polymerases: The Many Ways to Transcribe a Gene', *Frontiers in Molecular Biosciences*. Frontiers Media S.A.,

8, p. 207. doi: 10.3389/FMOLB.2021.663209/BIBTEX.

Bartkowiak, B. *et al.* (2010) 'CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1', *Genes & development*. Genes Dev, 24(20), pp. 2303–2316. doi: 10.1101/GAD.1968210.

Bartkowiak, B. and Greenleaf, A. L. (2015) 'Expression, purification, and identification of associated proteins of the full-length hCDK12/CyclinK complex', *The Journal of biological chemistry*. J Biol Chem, 290(3), pp. 1786–1795. doi: 10.1074/JBC.M114.612226.

Bartkowiak, B., Yan, C. and Greenleaf, A. L. (2015) 'Engineering an analog-sensitive CDK12 cell line using CRISPR/Cas', *Biochimica et biophysica acta*. Biochim Biophys Acta, 1849(9), pp. 1179–1187. doi: 10.1016/J.BBAGRM.2015.07.010.

Bhatti, G. K. *et al.* (2021) 'Emerging role of non-coding RNA in health and disease', *Metabolic Brain Disease 2021 36:6*. Springer, 36(6), pp. 1119–1134. doi: 10.1007/S11011-021-00739-Y.

Bhullar, K. S. *et al.* (2018) 'Kinase-targeted cancer therapies: progress, challenges and future directions', *Molecular Cancer 2018 17:1*. BioMed Central, 17(1), pp. 1–20. doi: 10.1186/S12943-018-0804-2.

Bishop, A. C. *et al.* (2000) 'A chemical switch for inhibitor-sensitive alleles of any protein kinase', *Nature*, 407(6802), pp. 395–401. doi: 10.1038/35030148.

Blazek, D. *et al.* (2011) 'The Cyclin K / Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes', 1, pp. 2158–2172. doi: 10.1101/gad.16962311.Phosphorylation.

Blethrow, J. *et al.* (2004) 'Design and Use of Analog-Sensitive Protein Kinases', *Current Protocols in Molecular Biology*. doi: 10.1002/0471142727.mb1811s66.

Boehning, M. *et al.* (2018) 'RNA polymerase II clustering through carboxy-terminal domain phase separation', *Nature Structural & Molecular Biology 2018 25:9*. Nature Publishing Group, 25(9), pp. 833–840. doi: 10.1038/s41594-018-0112-y.

Borggrefe, T. *et al.* (2002) 'A complex of the Srb8, -9, -10, and -11 transcriptional regulatory proteins from yeast', *The Journal of biological chemistry*. J Biol Chem, 277(46), pp. 44202–44207. doi: 10.1074/JBC.M207195200.

Borisova, M. E. *et al.* (2018) 'p38-MK2 signaling axis regulates RNA metabolism after UV-light-induced DNA damage', *Nature Communications 2018 9:1*. Nature Publishing Group, 9(1), pp. 1–16. doi: 10.1038/s41467-018-03417-3.

Bösken, C. A. *et al.* (2014) 'The structure and substrate specificity of human Cdk12/Cyclin K', *Nature communications*. Nat Commun, 5. doi: 10.1038/NCOMMS4505.

Bourbon, H. M. (2008) 'Comparative genomics supports a deep evolutionary origin for the

large, four-module transcriptional mediator complex', *Nucleic acids research*. Nucleic Acids Res, 36(12), pp. 3993–4008. doi: 10.1093/NAR/GKN349.

Brueckner, F., Ortiz, J. and Cramer, P. (2009) 'A movie of the RNA polymerase nucleotide addition cycle', *Current opinion in structural biology*. Curr Opin Struct Biol, 19(3), pp. 294–299. doi: 10.1016/J.SBI.2009.04.005.

Buratowski, S. (2009) 'Progression through the RNA polymerase II CTD cycle', *Molecular cell*. NIH Public Access, 36(4), p. 541. doi: 10.1016/J.MOLCEL.2009.10.019.

Bury, M. *et al.* (2021) 'New Insights into CDK Regulators: Novel Opportunities for Cancer Therapy', *Trends in Cell Biology*. Elsevier, 31(5), pp. 331–344. doi: 10.1016/J.TCB.2021.01.010.

Caizzi, L. *et al.* (2021) 'Efficient RNA polymerase II pause release requires U2 snRNP function', *Molecular cell*. Mol Cell, 81(9), pp. 1920-1934.e9. doi: 10.1016/J.MOLCEL.2021.02.016.

Calvo, O. and García, A. (2012) 'RNA Polymerase II Phosphorylation and Gene Expression Regulation', *Protein Phosphorylation in Human Health*. IntechOpen. doi: 10.5772/48490.

Carbon, S. *et al.* (2019) 'The Gene Ontology Resource: 20 years and still GOing strong', *Nucleic Acids Research*. Oxford Academic, 47(D1), pp. D330–D338. doi: 10.1093/NAR/GKY1055.

Carter, R. and Drouin, G. (2009) 'Structural differentiation of the three eukaryotic RNA polymerases', *Genomics*. Academic Press, 94(6), pp. 388–396. doi: 10.1016/J.YGENO.2009.08.011.

Casamassimi, A. and Ciccodicola, A. (2019) 'Transcriptional Regulation: Molecules, Involved Mechanisms, and Misregulation', *International Journal of Molecular Sciences*. Multidisciplinary Digital Publishing Institute (MDPI), 20(6). doi: 10.3390/IJMS20061281.

Chang, Y. W., Howard, S. C. and Herman, P. K. (2004) 'The Ras/PKA signaling pathway directly targets the Srb9 protein, a component of the general RNA polymerase II transcription apparatus', *Molecular cell*. Mol Cell, 15(1), pp. 107–116. doi: 10.1016/J.MOLCEL.2004.05.021.

Chapman, R. D. *et al.* (2007) 'Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7', *Science (New York, N.Y.)*. Science, 318(5857), pp. 1780–1782. doi: 10.1126/SCIENCE.1145977.

Chen, M. *et al.* (2017) 'CDK8/19 Mediator kinases potentiate induction of transcription by NFκB', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 114(38), pp. 10208–10213. doi: 10.1073/PNAS.1710467114.

Chi, Y. *et al.* (2001) 'Negative regulation of Gcn4 and Msn2 transcription factors by Srb10

cyclin-dependent kinase', *Genes & development*. Genes Dev, 15(9), pp. 1078–1092. doi: 10.1101/GAD.867501.

Chirackal Manavalan, A. P. *et al.* (2019) 'CDK12 controls G1/S progression by regulating RNAPII processivity at core DNA replication genes', *EMBO reports*. EMBO Rep, 20(9). doi: 10.15252/EMBR.201847592.

Cho, E. J. *et al.* (1997) 'mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain', *Genes & Development*. Cold Spring Harbor Laboratory Press, 11(24), p. 3319. doi: 10.1101/GAD.11.24.3319.

Choi, H., Lee, B. H. and Park, H. Y. (2021) 'Time-resolved analysis of transcription kinetics in single live mammalian cells', *bioRxiv*. Cold Spring Harbor Laboratory, p. 2021.12.23.474066. doi: 10.1101/2021.12.23.474066.

Chou, J. *et al.* (2020) 'Transcription-associated cyclin-dependent kinases as targets and biomarkers for cancer therapy', *Cancer Discovery*. American Association for Cancer Research Inc., 10(3), pp. 351–370. doi: 10.1158/2159-8290.CD-19-0528/43963/P/TRANSCRIPTION-ASSOCIATED-CYCLIN-DEPENDENT-KINASES.

Churchman, L. S. and Weissman, J. S. (2011) 'Nascent transcript sequencing visualizes transcription at nucleotide resolution', *Nature*. Nature Publishing Group, 469(7330), pp. 368–373. doi: 10.1038/nature09652.

Churko, J. M. *et al.* (2013) 'Overview of High Throughput Sequencing Technologies to Elucidate Molecular Pathways in Cardiovascular Diseases', *Circulation research*. NIH Public Access, 112(12), pp. 1613–1623. doi: 10.1161/CIRCRESAHA.113.300939.

Coin, F. and Egly, J. M. (2015) 'Revisiting the Function of CDK7 in Transcription by Virtue of a Recently Described TFIIH Kinase Inhibitor', *Molecular Cell*. Cell Press, 59(4), pp. 513–514. doi: 10.1016/J.MOLCEL.2015.08.006.

Compe, E. *et al.* (2019) 'TFIIE orchestrates the recruitment of the TFIIH kinase module at promoter before release during transcription', *Nature Communications*. Nature Publishing Group, 10(1). doi: 10.1038/S41467-019-10131-1.

Conaway, R. C. and Conaway, J. W. (2013) 'The Mediator complex and transcription elongation', *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. Elsevier, 1829(1), pp. 69–75. doi: 10.1016/J.BBAGRM.2012.08.017.

Corden, J. L. *et al.* (1985) 'A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 82(23), p. 7934. doi:

10.1073/PNAS.82.23.7934.

Core, L. and Adelman, K. (2019) 'Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation', *Genes & Development*. Cold Spring Harbor Laboratory Press, 33(15–16), p. 960. doi: 10.1101/GAD.325142.119.

Cramer, P. (2019a) 'Eukaryotic Transcription Turns 50', *Cell*. Cell Press, 179(4), pp. 808–812. doi: 10.1016/J.CELL.2019.09.018.

Cramer, P. (2019b) 'Organization and regulation of gene transcription', *Nature*. Nature Publishing Group, pp. 45–54. doi: 10.1038/s41586-019-1517-4.

Cramer, P., Bushnell, D. A. and Kornberg, R. D. (2001) 'Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution', *Science (New York, N.Y.)*. Science, 292(5523), pp. 1863–1876. doi: 10.1126/SCIENCE.1059493.

Crick, F. (1970) 'Central Dogma of Molecular Biology', *Nature 1970 227:5258*. Nature Publishing Group, 227(5258), pp. 561–563. doi: 10.1038/227561a0.

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 & development*. Genes Dev, 28(4), pp. 342–356. doi: 10.1101/GAD.231274.113.

Dobin, A. *et al.* (2013) 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*. Oxford University Press, 29(1), pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Dubbury, S. J., Boutz, P. L. and Sharp, P. A. (2018) 'CDK12 regulates DNA repair genes by suppressing intronic polyadenylation', *Nature*. Nature, 564(7734), pp. 141–145. doi: 10.1038/S41586-018-0758-Y.

Ebmeier, Christopher C *et al.* (2017) 'Human TFIIH Kinase CDK7 Regulates Transcription-Associated Chromatin Modifications', *CellReports*, 20, pp. 1173–1186. doi: 10.1016/j.celrep.2017.07.021.

Ebmeier, Christopher C. *et al.* (2017) 'Human TFIIH Kinase CDK7 Regulates Transcription-Associated Chromatin Modifications', *Cell Reports*. ElsevierCompany., 20(5), pp. 1173–1186. doi: 10.1016/j.celrep.2017.07.021.

Egloff, S., Dienstbier, M. and Murphy, S. (2012) 'Updating the RNA polymerase CTD code: Adding gene-specific layers', *Trends in Genetics*, 28(7), pp. 333–341. doi: 10.1016/J.TIG.2012.03.007.

Eifler, T. T. *et al.* (2015) 'Cyclin-dependent kinase 12 increases 3' end processing of growth factor-induced c-FOS transcripts', *Molecular and cellular biology*. Mol Cell Biol, 35(2), pp. 468–478. doi: 10.1128/MCB.01157-14.

Espinosa, J. M. (2019a) 'Transcriptional CDKs in the spotlight',

https://doi.org/10.1080/21541264.2019.1597479. Taylor & Francis, 10(2), pp. 45–46. doi: 10.1080/21541264.2019.1597479.

Espinosa, J. M. (2019b) 'Transcriptional CDKs in the spotlight', *Transcription*. Taylor and Francis Inc., 10(2), pp. 45–46. doi: 10.1080/21541264.2019.1597479.

Eswaran, J. and Knapp, S. (2010) 'Insights into protein kinase regulation and inhibition by large scale structural comparison', *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*. Elsevier, 1804(3), pp. 429–432. doi: 10.1016/J.BBAPAP.2009.10.013.

Ettl, T., Schulz, D. and Bauer, R. J. (2022) 'The Renaissance of Cyclin Dependent Kinase Inhibitors', *Cancers*. MDPI, 14(2). doi: 10.3390/CANCERS14020293.

Fishburn, J. *et al.* (2015) 'Double-stranded DNA translocase activity of transcription factor TFIIH and the mechanism of RNA polymerase II open complex formation', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 112(13), pp. 3961–3966. doi: 10.1073/PNAS.1417709112/-/DCSUPPLEMENTAL.

Fisher, R. P. (2019a) 'Cdk7: a kinase at the core of transcription and in the crosshairs of cancer drug discovery', *Transcription*. Taylor & Francis, 10(2), pp. 47–56. doi: 10.1080/21541264.2018.1553483.

Fisher, R. P. (2019b) 'Cdk7: a kinase at the core of transcription and in the crosshairs of cancer drug discovery', *Transcription*. Taylor & Francis, 10(2), pp. 47–56. doi: 10.1080/21541264.2018.1553483.

Frankish, A. *et al.* (2019) 'GENCODE reference annotation for the human and mouse genomes', *Nucleic Acids Research*. Oxford Academic, 47(D1), pp. D766–D773. doi: 10.1093/NAR/GKY955.

Furlan, M., De Pretis, S. and Pelizzola, M. (2021) 'Dynamics of transcriptional and post-transcriptional regulation', *Briefings in Bioinformatics*. Oxford Academic, 22(4), pp. 1–13. doi: 10.1093/BIB/BBAA389.

Galbraith, M. D. *et al.* (2013) 'HIF1A employs CDK8-mediator to stimulate RNAPII elongation in response to hypoxia', *Cell*. Cell, 153(6), p. 1327. doi: 10.1016/J.CELL.2013.04.048.

Galbraith, M. D. *et al.* (2017) 'CDK8 Kinase Activity Promotes Glycolysis', *Cell reports*. Cell Rep, 21(6), pp. 1495–1506. doi: 10.1016/J.CELREP.2017.10.058.

Galbraith, M. D., Bender, H. and Espinosa, J. M. (2018) 'Therapeutic targeting of transcriptional cyclin-dependent kinases'. doi: 10.1080/21541264.2018.1539615.

Galbraith, M. D., Bender, H. and Espinosa, J. M. (2019) 'Therapeutic targeting of transcriptional cyclin-dependent kinases', *Transcription*. Taylor & Francis, 10(2), p. 118. doi:

10.1080/21541264.2018.1539615.

Ganuza, M. *et al.* (2012) 'Genetic inactivation of Cdk7 leads to cell cycle arrest and induces premature aging due to adult stem cell exhaustion', *The EMBO journal*. EMBO J, 31(11), pp. 2498–2510. doi: 10.1038/EMBOJ.2012.94.

Garibaldi, A., Carranza, F. and Hertel, K. J. (2017) 'Isolation of Newly Transcribed RNA Using the Metabolic Label 4-Thiouridine', *Methods in molecular biology (Clifton, N.J.)*. Methods Mol Biol, 1648, pp. 169–176. doi: 10.1007/978-1-4939-7204-3_13.

Garske, A. L. *et al.* (2011) 'Chemical genetic strategy for targeting protein kinases based on covalent complementarity', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 108(37), pp. 15046–15052. doi: 10.1073/PNAS.1111239108/-/DCSUPPLEMENTAL.

Glover-Cutter, K. *et al.* (2009) 'TFIIH-Associated Cdk7 Kinase Functions in Phosphorylation of C-Terminal Domain Ser7 Residues, Promoter-Proximal Pausing, and Termination by RNA Polymerase II', *Molecular and Cellular Biology*. American Society for Microbiology, 29(20), pp. 5455–5464. doi: 10.1128/MCB.00637-09/SUPPL_FILE/GLOVER_CUTTER_REV_SUPP_LEGENDS.PDF.

Golding, I. *et al.* (2005) 'Real-time kinetics of gene activity in individual bacteria', *Cell*, 123(6), pp. 1025–1036. doi: 10.1016/j.cell.2005.09.031.

Gorin, G. *et al.* (2020) 'Stochastic simulation and statistical inference platform for visualization and estimation of transcriptional kinetics', *PLOS ONE*. Edited by J. Garcia-Ojalvo. Public Library of Science, 15(3), p. e0230736. doi: 10.1371/journal.pone.0230736.

Greenleaf, A. L. (2019) 'Human CDK12 and CDK13, multi-tasking CTD kinases for the new millenium', *Transcription*. Transcription, 10(2), pp. 91–110. doi: 10.1080/21541264.2018.1535211.

Gressel, S. *et al.* (2017a) 'CDK9-dependent RNA polymerase II pausing controls transcription initiation', *eLife*. eLife Sciences Publications Ltd, 6. doi: 10.7554/eLife.29736.

Gressel, S. *et al.* (2017b) 'CDK9-dependent RNA polymerase II pausing controls transcription initiation', *eLife*. eLife Sciences Publications Limited, 6, p. e29736. doi: 10.7554/eLife.29736.

Gressel, S., Schwalb, B. and Cramer, P. (2019) 'The pause-initiation limit restricts transcription activation in human cells', *Nature Communications*. Nature Publishing Group, 10(1), pp. 1–12. doi: 10.1038/s41467-019-11536-8.

Grünberg, S. *et al.* (2016) 'Mediator binding to UASs is broadly uncoupled from transcription and cooperative with TFIID recruitment to promoters', *The EMBO journal*. EMBO J, 35(22),

pp. 2435–2446. doi: 10.15252/EMBJ.201695020.

Hahne, F. and Ivanek, R. (2016) 'Visualizing genomic data using Gviz and bioconductor', in *Methods in Molecular Biology*. Humana Press Inc., pp. 335–351. doi: 10.1007/978-1-4939-3578-9_16.

Hantsche, M. and Cramer, P. (2016) 'The Structural Basis of Transcription: 10 Years After the Nobel Prize in Chemistry', *Angewandte Chemie International Edition*. John Wiley & Sons, Ltd, 55(52), pp. 15972–15981. doi: 10.1002/ANIE.201608066.

He, Y. *et al.* (2013) 'Structural visualization of key steps in human transcription initiation', *Nature*. Nature, 495(7442), pp. 481–486. doi: 10.1038/NATURE11991.

Heidemann, M. *et al.* (2013) 'Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription', *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. Elsevier, 1829(1), pp. 55–62. doi: 10.1016/J.BBAGRM.2012.08.013.

Hengartner, C. J. *et al.* (1998) 'Temporal regulation of RNA polymerase II by Srb10 and Kin28 cyclin-dependent kinases', *Molecular cell*. Mol Cell, 2(1), pp. 43–53. doi: 10.1016/S1097-2765(00)80112-4.

Hirst, M. *et al.* (1999) 'GAL4 is regulated by the RNA polymerase II holoenzyme-associated cyclin-dependent protein kinase SRB10/CDK8', *Molecular cell*. Mol Cell, 3(5), pp. 673–678. doi: 10.1016/S1097-2765(00)80360-3.

Ho, C. K. and Shuman, S. (1999) 'Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme', *Molecular cell*. Mol Cell, 3(3), pp. 405–411. doi: 10.1016/S1097-2765(00)80468-2.

Hoeppner, S., Baumli, S. and Cramer, P. (2005) 'Structure of the mediator subunit cyclin C and its implications for CDK8 function', *Journal of molecular biology*. J Mol Biol, 350(5), pp. 833–842. doi: 10.1016/J.JMB.2005.05.041.

Holstege, F. C. P. *et al.* (1998) 'Dissecting the regulatory circuitry of a eukaryotic genome', *Cell*. Cell, 95(5), pp. 717–728. doi: 10.1016/S0092-8674(00)81641-4.

Honkela, A. *et al.* (2015) 'Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 112(42), pp. 13115–13120. doi: 10.1073/pnas.1420404112.

Hsin, J. P. and Manley, J. L. (2012) 'The RNA polymerase II CTD coordinates transcription and RNA processing', *Genes & Development*. Cold Spring Harbor Laboratory Press, 26(19), p. 2119. doi: 10.1101/GAD.200303.112.

Hurwitz, J. (2005) 'The Discovery of RNA Polymerase', *Journal of Biological Chemistry*. Elsevier,

280(52), pp. 42477–42485. doi: 10.1074/JBC.X500006200.

I, S., C, C. and R, D. (1996) 'Phosphorylation of Ga14p at a single C-terminal residue is necessary for galactose-inducible transcription', *Molecular and cellular biology*. Mol Cell Biol, 16(9), pp. 4879–4887. doi: 10.1128/MCB.16.9.4879.

Jhaveri, K. *et al.* (2021) 'The evolution of cyclin dependent kinase inhibitors in the treatment of cancer', *Expert Review of Anticancer Therapy*. Taylor and Francis Ltd., 21(10), pp. 1105–1124. doi: 10.1080/14737140.2021.1944109/SUPPL_FILE/IERY_A_1944109_SM4502.DOCX.

Jonkers, I., Kwak, H. and Lis, J. T. (2014) 'Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons', *eLife*. Elife, 3(3). doi: 10.7554/ELIFE.02407.

Jonkers, I. and Lis, J. T. (2015) 'Getting up to speed with transcription elongation by RNA polymerase II', *Nature reviews. Molecular cell biology*. Nat Rev Mol Cell Biol, 16(3), pp. 167–177. doi: 10.1038/NRM3953.

Jurcik, J. *et al.* (2020) 'Phosphoproteomics Meets Chemical Genetics: Approaches for Global Mapping and Deciphering the Phosphoproteome', *International Journal of Molecular Sciences 2020, Vol. 21, Page 7637*. Multidisciplinary Digital Publishing Institute, 21(20), p. 7637. doi: 10.3390/IJMS21207637.

Kang, W. *et al.* (2020) 'Transcription reinitiation by recycling RNA polymerase that diffuses on DNA after releasing terminated RNA', *Nature Communications 2020 11:1*. Nature Publishing Group, 11(1), pp. 1–9. doi: 10.1038/s41467-019-14200-3.

Kanin, E. I. *et al.* (2007) 'Chemical inhibition of the TFIIH-associated kinase Cdk7/Kin28 does not impair global mRNA synthesis', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 104(14), pp. 5812–5817. doi: 10.1073/PNAS.0611505104.

Kasiliauskaite, A. *et al.* (2022) 'Cooperation between intrinsically disordered and ordered regions of Spt6 regulates nucleosome and Pol II CTD binding, and nucleosome assembly', *Nucleic Acids Research*. Oxford Academic, 50(10), pp. 5961–5973. doi: 10.1093/NAR/GKAC451.

Kecman, T. *et al.* (2018) 'Elongation/Termination Factor Exchange Mediated by PP1 Phosphatase Orchestrates Transcription Termination', *Cell reports*. Cell Rep, 25(1), pp. 259-269.e5. doi: 10.1016/J.CELREP.2018.09.007.

Kim, D. *et al.* (2013) 'TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions', *Genome Biology*. BioMed Central, 14(4), pp. 1–13. doi:

10.1186/GB-2013-14-4-R36/FIGURES/6.

Kim, H. *et al.* (2010) 'Gene-specific RNA pol II phosphorylation and the "CTD code"', *Nature structural & molecular biology*. NIH Public Access, 17(10), p. 1279. doi: 10.1038/NSMB.1913.

Kim, S. *et al.* (2006) 'Mediator is a transducer of Wnt/beta-catenin signaling', *The Journal of biological chemistry*. J Biol Chem, 281(20), pp. 14066–14075. doi: 10.1074/JBC.M602696200.

Kim, W. *et al.* (2021) 'Advancements in chemical biology targeting the kinases and phosphatases of RNA polymerase II-mediated transcription', *Current Opinion in Chemical Biology*. Elsevier Ltd, 63, pp. 68–77. doi: 10.1016/j.cbpa.2021.02.002.

Klaeger, S. *et al.* (2017) 'The target landscape of clinical kinase drugs', *Science (New York, N.Y.)*. Europe PMC Funders, 358(6367). doi: 10.1126/SCIENCE.AAN4368.

Klatt, F. *et al.* (2020) 'A precisely positioned MED12 activation helix stimulates CDK8 kinase activity', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 117(6), pp. 2894–2905. doi: 10.1073/PNAS.1917635117/-/DCSUPPLEMENTAL.

Knuesel, M. T. *et al.* (2009) 'The human CDK8 subcomplex is a histone kinase that requires Med12 for activity and can function independently of mediator', *Molecular and cellular biology*. Mol Cell Biol, 29(3), pp. 650–661. doi: 10.1128/MCB.00993-08.

Kohoutek, J. and Blazek, D. (2012) 'Cyclin K goes with Cdk12 and Cdk13', *Cell Division 2012 7:1*. BioMed Central, 7(1), pp. 1–10. doi: 10.1186/1747-1028-7-12.

Kornberg, R. D. (1999) 'Eukaryotic transcriptional control', *Trends in Cell Biology*. Elsevier Current Trends, 9(12), pp. M46–M49. doi: 10.1016/S0962-8924(99)01679-7.

Krajewska, M. *et al.* (2019) 'CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation'. doi: 10.1038/s41467-019-09703-y.

Krishnamurty, R. and Maly, D. J. (2010) 'Biochemical Mechanisms of Resistance to Small-Molecule Protein Kinase Inhibitors', *ACS chemical biology*. NIH Public Access, 5(1), p. 121. doi: 10.1021/CB9002656.

Kujirai, T. and Kurumizaka, H. (2020) 'Transcription through the nucleosome', *Current Opinion in Structural Biology*. Elsevier Current Trends, 61, pp. 42–49. doi: 10.1016/J.SBI.2019.10.007.

Kulaeva, O. I., Gaykalova, D. A. and Studitsky, V. M. (2007) 'Transcription Through Chromatin by RNA polymerase II: Histone Displacement and Exchange', *Mutation research*. NIH Public Access, 618(1–2), p. 116. doi: 10.1016/J.MRFMMM.2006.05.040.

Kwiatkowski, N. *et al.* (2014) 'Targeting transcription regulation in cancer with a covalent

CDK7 inhibitor', *Nature*. NIH Public Access, 511(7511), p. 616. doi: 10.1038/NATURE13393.

Laitem, C. *et al.* (2015) 'CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II-transcribed genes', *Nature structural & molecular biology*. Nat Struct Mol Biol, 22(5), pp. 396–403. doi: 10.1038/NSMB.3000.

Langmead, B. *et al.* (2009) 'Ultrafast and memory-efficient alignment of short DNA sequences to the human genome', *Genome Biology*. BioMed Central, 10(3), p. R25. doi: 10.1186/gb-2009-10-3-r25.

Larochelle, S. *et al.* (2001) 'T-loop phosphorylation stabilizes the CDK7-cyclin H-MAT1 complex in vivo and regulates its CTD kinase activity', *The EMBO journal*. EMBO J, 20(14), pp. 3749–3759. doi: 10.1093/EMBOJ/20.14.3749.

Larochelle, S. *et al.* (2012) 'Cyclin-dependent kinase control of the initiation-to-elongation switch of RNA polymerase II', *Nature Structural & Molecular Biology 2012 19:11*. Nature Publishing Group, 19(11), pp. 1108–1115. doi: 10.1038/nsmb.2399.

Lenssen, E. *et al.* (2007) 'The Ccr4-not complex regulates Skn7 through Srb10 kinase', *Eukaryotic cell*. Eukaryot Cell, 6(12), pp. 2251–2259. doi: 10.1128/EC.00327-06.

Li, Heng *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools.', *Bioinformatics (Oxford, England)*, 25(16), pp. 2078–9. doi: 10.1093/bioinformatics/btp352.

Li, H. *et al.* (2009) 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*. Oxford University Press, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

Li, J. and Liu, C. (2019) 'Coding or noncoding, the converging concepts of RNAs', *Frontiers in Genetics*. Frontiers Media S.A., 10(MAY), p. 496. doi: 10.3389/FGENE.2019.00496/BIBTEX.

Liang, H. *et al.* (2021) 'Recent progress in development of cyclin-dependent kinase 7 inhibitors for cancer therapy', *https://doi.org/10.1080/13543784.2021.1850693*. Taylor & Francis, 30(1), pp. 61–76. doi: 10.1080/13543784.2021.1850693.

Liang, K. *et al.* (2015) 'Characterization of human cyclin-dependent kinase 12 (CDK12) and CDK13 complexes in C-terminal domain phosphorylation, gene transcription, and RNA processing', *Molecular and cellular biology*. Mol Cell Biol, 35(6), pp. 928–938. doi: 10.1128/MCB.01426-14.

Liao, S. M. *et al.* (1995) 'A kinase-cyclin pair in the RNA polymerase II holoenzyme', *Nature*. Nature, 374(6518), pp. 193–196. doi: 10.1038/374193A0.

Lidschreiber, K. *et al.* (2021) 'Transcriptionally active enhancers in human cancer cells', *Molecular Systems Biology*. John Wiley & Sons, Ltd, 17(1), p. e9873. doi:

10.15252/MSB.20209873.

Lightbody, G. *et al.* (2019) 'Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application', *Briefings in Bioinformatics*. Oxford Academic, 20(5), pp. 1795–1811. doi: 10.1093/BIB/BBY051.

Lin, X. *et al.* (2002) 'P-TEFb containing cyclin K and Cdk9 can activate transcription via RNA', *The Journal of biological chemistry*. J Biol Chem, 277(19), pp. 16873–16878. doi: 10.1074/JBC.M200117200.

Liu, Y. *et al.* (2004a) 'Two cyclin-dependent kinases promote RNA polymerase II transcription and formation of the scaffold complex', *Molecular and cellular biology*. Mol Cell Biol, 24(4), pp. 1721–1735. doi: 10.1128/MCB.24.4.1721-1735.2004.

Liu, Y. *et al.* (2004b) 'Two cyclin-dependent kinases promote RNA polymerase II transcription and formation of the scaffold complex', *Molecular and cellular biology*. Mol Cell Biol, 24(4), pp. 1721–1735. doi: 10.1128/MCB.24.4.1721-1735.2004.

Lopez, Michael S., Kliegman, J. I. and Shokat, K. M. (2014) *The logic and design of analog-sensitive kinases and their small molecule inhibitors*. 1st edn, *Methods in Enzymology*. 1st edn. Elsevier Inc. doi: 10.1016/B978-0-12-397918-6.00008-2.

Lopez, Michael S, Kliegman, J. I. and Shokat, K. M. (2014) 'The Logic and Design of Analog-Sensitive Kinases and Their Small Molecule Inhibitors', *Protein Kinase Inhibitors in Research and Medicine*, 548, pp. 189–213. doi: 10.1016/B978-0-12-397918-6.00008-2.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Lu, X. *et al.* (2016) 'Multiple P-TEFbs cooperatively regulate the release of promoter-proximally paused RNA polymerase II', *Nucleic acids research*. Nucleic Acids Res, 44(14), pp. 6853–6867. doi: 10.1093/NAR/GKW571.

Luger, K. *et al.* (1997) 'Crystal structure of the nucleosome core particle at 2.8 Å resolution', *Nature 1997 389:6648*. Nature Publishing Group, 389(6648), pp. 251–260. doi: 10.1038/38444.

Łukasik, P., Załuski, M. and Gutowska, I. (2021) 'Cyclin-Dependent Kinases (CDK) and Their Role in Diseases Development–Review', *International Journal of Molecular Sciences*. Multidisciplinary Digital Publishing Institute (MDPI), 22(6), pp. 1–33. doi: 10.3390/IJMS22062935.

Luyties, O. and Taatjes, D. J. (2022) 'The Mediator kinase module: an interface between cell

signaling and transcription', *Trends in Biochemical Sciences*. Elsevier Ltd, 47(4), pp. 314–327. doi: 10.1016/J.TIBS.2022.01.002/ATTACHMENT/6CCA52E3-8E5F-4AFA-85D2-7A71FC85FEA2/MMC1.DOCX.

Lynch, C. J. *et al.* (2020) 'Global hyperactivation of enhancers stabilizes human and mouse naive pluripotency through inhibition of CDK8/19 Mediator kinases', *Nature cell biology*. Nat Cell Biol, 22(10), pp. 1223–1238. doi: 10.1038/S41556-020-0573-1.

Malumbres, M. (2014a) 'Cyclin-dependent kinases', *Genome Biology*. BioMed Central Ltd., 15(6), pp. 1–10. doi: 10.1186/GB4184/FIGURES/4.

Malumbres, M. (2014b) 'Cyclin-dependent kinases', *Genome Biology*. BioMed Central Ltd., 15(6), pp. 1–10. doi: 10.1186/GB4184/FIGURES/4.

Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), p. 10. doi: 10.14806/ej.17.1.200.

Mazzocca, M. *et al.* (2021) 'Transcription factor binding kinetics and transcriptional bursting: What do we really know?', *Current opinion in structural biology*. Curr Opin Struct Biol, 71, pp. 239–248. doi: 10.1016/J.SBI.2021.08.002.

Mccord, R. P., Kaplan, N. and Giorgetti, L. (2020) 'Molecular Cell Review Chromosome Conformation Capture and Beyond: Toward an Integrative View of Chromosome Structure and Function'. doi: 10.1016/j.molcel.2019.12.021.

Nechaev, S. and Adelman, K. (2011) 'Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation', *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*. Elsevier, 1809(1), pp. 34–45. doi: 10.1016/J.BBAGRM.2010.11.001.

Nelson, C. *et al.* (2003) 'Srb10/Cdk8 regulates yeast filamentous growth by phosphorylating the transcription factor Ste12', *Nature*. Nature, 421(6919), pp. 187–190. doi: 10.1038/NATURE01243.

Nikolov, D. B. and Burley, S. K. (1997) 'RNA polymerase II transcription initiation: A structural view', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 94(1), p. 15. doi: 10.1073/PNAS.94.1.15.

Nilson, Kyle A. *et al.* (2015) 'THZ1 reveals roles for Cdk7 in co-transcriptional capping and pausing', *Molecular cell*. NIH Public Access, 59(4), p. 576. doi: 10.1016/J.MOLCEL.2015.06.032.

Nilson, Kyle A *et al.* (2015) 'THZ1 reveals roles for Cdk7 in co-transcriptional capping and pausing HHS Public Access', *Mol Cell*, 59(4), pp. 576–587. doi: 10.1016/j.molcel.2015.06.032.

Nojima, T., Gomes, T., Grosso, A. R. F., Kimura, H., Dye, Michael J, *et al.* (2015) 'Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing.', *Cell*. Elsevier, 161(3), pp. 526–540. doi: 10.1016/j.cell.2015.03.027.

Nojima, T., Gomes, T., Grosso, A. R. F., Kimura, H., Dye, Michael J., *et al.* (2015) 'Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing', *Cell*. Cell Press, 161(3), pp. 526–540. doi: 10.1016/j.cell.2015.03.027.

Nojima, T. *et al.* (2016) 'Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide', *Nature Protocols*. Nature Publishing Group, 11(3), pp. 413–428. doi: 10.1038/nprot.2016.012.

Nojima, T. *et al.* (2018) 'Deregulated Expression of Mammalian lncRNA through Loss of SPT6 Induces R-Loop Formation, Replication Stress, and Cellular Senescence', *Molecular cell*. Mol Cell, 72(6), pp. 970-984.e7. doi: 10.1016/J.MOLCEL.2018.10.011.

O'Leary, N. A. *et al.* (2016) 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic acids research*. Nucleic Acids Res, 44(D1), pp. D733–D745. doi: 10.1093/NAR/GKV1189.

Olson, C. M. *et al.* (2019) 'Development of a Selective CDK7 Covalent Inhibitor Reveals Predominant Cell-Cycle Phenotype', *Cell chemical biology*. Cell Chem Biol, 26(6), pp. 792-803.e10. doi: 10.1016/J.CHEMBIOL.2019.02.012.

Osman, S. *et al.* (2021) 'The Cdk8 kinase module regulates interaction of the mediator complex with RNA polymerase II', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., 296. doi: 10.1016/J.JBC.2021.100734/ATTACHMENT/549E5C1A-F3DF-463B-B427-80A758024EA8/MMC4.XLSX.

Osman, S. and Cramer, P. (2020) 'Structural Biology of RNA Polymerase II Transcription: 20 Years On', *https://doi.org/10.1146/annurev-cellbio-042020-021954*. Annual Reviews , 36, pp. 1–34. doi: 10.1146/ANNUREV-CELLBIO-042020-021954.

Pal, M., Ponticelli, A. S. and Luse, D. S. (2005) 'The role of the transcription bubble and TFIIB in promoter clearance by RNA polymerase II', *Molecular cell*. Mol Cell, 19(1), pp. 101–110. doi: 10.1016/J.MOLCEL.2005.05.024.

Panagiotou, E. *et al.* (2022) 'Cyclin-dependent kinase (CDK) inhibitors in solid tumors: a review of clinical trials', *Clinical and Translational Oncology*. Springer Science and Business Media Deutschland GmbH, 24(2), pp. 161–192. doi: 10.1007/S12094-021-02688-5.

Patro, R. *et al.* (2017) 'Salmon provides fast and bias-aware quantification of transcript expression', *Nature Methods*. Nature Publishing Group, 14(4), pp. 417–419. doi:

10.1038/nmeth.4197.

Peissert, S. *et al.* (2020) 'Structural basis for CDK7 activation by MAT1 and Cyclin H', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 117(43), pp. 26739–26748. doi: 10.1073/PNAS.2010885117.

Pelechano, V., Wei, W. and Steinmetz, L. M. (2013) 'Extensive transcriptional heterogeneity revealed by isoform profiling', *Nature*. Nature, 497(7447), pp. 127–131. doi: 10.1038/nature12121.

Pelish, H. E. *et al.* (2015) 'Mediator kinase inhibition further activates super-enhancer-associated genes in AML', *Nature*. Nature, 526(7572), pp. 273–276. doi: 10.1038/NATURE14904.

Van De Peppel, J. *et al.* (2005) 'Mediator expression profiling epistasis reveals a signal transduction pathway with antagonistic submodules and highly specific downstream targets', *Molecular cell*. Mol Cell, 19(4), pp. 511–522. doi: 10.1016/J.MOLCEL.2005.06.033.

Petrenko, N. *et al.* (2019) 'Requirements for rna polymerase ii preinitiation complex formation in vivo', *eLife*. eLife Sciences Publications Ltd, 8. doi: 10.7554/ELIFE.43654.

Pilarova, K., Herudek, J. and Blazek, D. (2020) 'CDK12: cellular functions and therapeutic potential of versatile player in cancer', *NAR Cancer*. Oxford Academic, 2(1). doi: 10.1093/NARCAN/ZCAA003.

Poss, Z. C. *et al.* (2016) 'Identification of Mediator Kinase Substrates in Human Cells using Cortistatin A and Quantitative Phosphoproteomics', *Cell reports*. Cell Rep, 15(2), pp. 436–450. doi: 10.1016/J.CELREP.2016.03.030.

Proudfoot, N. J. (2016) 'Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut', *Science (New York, N.Y.)*. Europe PMC Funders, 352(6291), p. aad9926. doi: 10.1126/SCIENCE.AAD9926.

Prudêncio, P. *et al.* (2020) 'Analysis of Mammalian Native Elongating Transcript sequencing (mNET-seq) high-throughput data', *Methods*. Academic Press, 178, pp. 89–95. doi: 10.1016/J.YMETH.2019.09.003.

Quinlan, A. R. and Hall, I. M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features.', *Bioinformatics (Oxford, England)*, 26(6), pp. 841–2. doi: 10.1093/bioinformatics/btq033.

Raithatha, S. *et al.* (2012) 'Cdk8 regulates stability of the transcription factor Phd1 to control pseudohyphal differentiation of Saccharomyces cerevisiae', *Molecular and cellular biology*. Mol Cell Biol, 32(3), pp. 664–674. doi: 10.1128/MCB.05420-11.

Ramanathan, A., Robb, G. B. and Chan, S. H. (2016) 'mRNA capping: biological functions

and applications', *Nucleic Acids Research*. Oxford University Press, 44(16), p. 7511. doi: 10.1093/NAR/GKW551.

Rauch, J. *et al.* (2011) 'The secret life of kinases: Functions beyond catalysis', *Cell Communication and Signaling*. BioMed Central, 9(1), pp. 1–28. doi: 10.1186/1478-811X-9-23/FIGURES/7.

Rengachari, S. *et al.* (2021) 'Structure of the human Mediator–RNA polymerase II pre-initiation complex', *Nature 2021 594:7861*. Nature Publishing Group, 594(7861), pp. 129–133. doi: 10.1038/s41586-021-03555-7.

Rimel, J. K. and Taatjes, D. J. (2018) 'The essential and multifunctional TFIIH complex', *Protein Science*. John Wiley & Sons, Ltd, 27(6), pp. 1018–1037. doi: 10.1002/PRO.3424.

Rohde, J. R., Trinh, J. and Sadowski, I. (2000) 'Multiple signals regulate GAL transcription in yeast', *Molecular and cellular biology*. Mol Cell Biol, 20(11), pp. 3880–3886. doi: 10.1128/MCB.20.11.3880-3886.2000.

Roy, R. *et al.* (1994) 'The MO15 cell cycle kinase is associated with the TFIIH transcription-DNA repair factor', *Cell*. Cell, 79(6), pp. 1093–1101. doi: 10.1016/0092-8674(94)90039-6.

Rozenblatt-Rosen, O. *et al.* (2009) 'The tumor suppressor Cdc73 functionally associates with CPSF and CstF 3' mRNA processing factors', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 106(3), pp. 755–760. doi: 10.1073/PNAS.0812023106.

Sadowski, I. *et al.* (1991) 'GAL4 is phosphorylated as a consequence of transcriptional activation', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 88(23), pp. 10510–10514. doi: 10.1073/PNAS.88.23.10510.

Sava, G. P. *et al.* (2020) 'CDK7 inhibitors as anticancer drugs', *Cancer Metastasis Reviews*. Springer, 39(3), p. 805. doi: 10.1007/S10555-020-09885-8.

Schier, A. C. and Taatjes, D. J. (2020) 'Structure and mechanism of the RNA polymerase II transcription machinery', *Genes & Development*. Cold Spring Harbor Laboratory Press, 34(7–8), pp. 465–488. doi: 10.1101/GAD.335679.119.

Schwalb, B., Michel, M., Zacher, B., Frühauf, K., *et al.* (2016) 'TT-seq maps the human transient transcriptome.', *Science (New York, N.Y.)*, 352(6290), pp. 1225–8. doi: 10.1126/science.aad9841.

Schwalb, B., Michel, M., Zacher, B., Hauf, K. F., *et al.* (2016) 'TT-seq maps the human transient transcriptome', *Science (New York, N.Y.)*. Science, 352(6290), pp. 1225–1228. doi: 10.1126/SCIENCE.AAD9841.

Serizawa, H. *et al.* (1995) 'Association of Cdk-activating kinase subunits with transcription

factor TFIIH', *Nature 1995 374:6519*. Nature Publishing Group, 374(6519), pp. 280–282. doi: 10.1038/374280a0.

Shao, R. *et al.* (2022) 'Distinct transcription kinetics of pluripotent cell states', *Molecular Systems Biology*. John Wiley & Sons, Ltd, 18(1), p. e10407. doi: 10.15252/MSB.202110407.

Shino, G. and Takada, S. (2021) 'Modeling DNA Opening in the Eukaryotic Transcription Initiation Complexes via Coarse-Grained Models', *Frontiers in Molecular Biosciences*. Frontiers Media S.A., 8, p. 1021. doi: 10.3389/FMOLB.2021.772486/BIBTEX.

Statello, L. *et al.* (2020) 'Gene regulation by long non-coding RNAs and its biological functions', *Nature Reviews Molecular Cell Biology 2020 22:2*. Nature Publishing Group, 22(2), pp. 96–118. doi: 10.1038/s41580-020-00315-9.

Steinparzer, I. *et al.* (2019) 'Transcriptional Responses to IFN-γ Require Mediator Kinase-Dependent Pause Release and Mechanistically Distinct CDK8 and CDK19 Functions', *Molecular cell*. Mol Cell, 76(3), pp. 485-499.e8. doi: 10.1016/J.MOLCEL.2019.07.034.

Steitz, T. A. and Steitz, J. A. (1993) 'A general two-metal-ion mechanism for catalytic RNA.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 90(14), p. 6498. doi: 10.1073/PNAS.90.14.6498.

Stieg, D. C., Cooper, K. F. and Strich, R. (2020) 'The extent of cyclin C promoter occupancy directs changes in stress-dependent transcription', *The Journal of biological chemistry*. J Biol Chem, 295(48), pp. 16280–16291. doi: 10.1074/JBC.RA120.015215.

Stiller, J. W. and Hall, B. D. (2002) 'Evolution of the RNA polymerase II C-terminal domain', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 99(9), p. 6091. doi: 10.1073/PNAS.082646199.

Sundar, V. *et al.* (2021) 'Transcriptional cyclin-dependent kinases as the mediators of inflammation-a review', *Gene*. Elsevier B.V., 769(October), p. 145200. doi: 10.1016/j.gene.2020.145200.

Tellier, M. *et al.* (2020) 'CDK12 globally stimulates RNA polymerase II transcription elongation and carboxyl-terminal domain phosphorylation', *Nucleic Acids Research*. Oxford University Press, 48(14), pp. 7712–7727. doi: 10.1093/nar/gkaa514.

Tellier, M., Ferrer-Vicens, I. and Murphy, S. (2016) 'The point of no return: The poly(A)-associated elongation checkpoint', *RNA biology*. RNA Biol, 13(3), pp. 265–271. doi: 10.1080/15476286.2016.1142037.

Teves, S. S., Weber, C. M. and Henikoff, S. (2014) 'Transcribing through the nucleosome', *Trends in Biochemical Sciences*, 39(12), pp. 577–586. doi: 10.1016/j.tibs.2014.10.004.

Thoma, O. M., Neurath, M. F. and Waldner, M. J. (2021) 'Cyclin-Dependent Kinase

Inhibitors and Their Therapeutic Potential in Colorectal Cancer Treatment', *Frontiers in Pharmacology*. Frontiers Media S.A., 12, p. 3673. doi: 10.3389/FPHAR.2021.757120/BIBTEX.

Tien, J. F. *et al.* (2017) 'CDK12 regulates alternative last exon mRNA splicing and promotes breast cancer cell invasion', *Nucleic acids research*. Nucleic Acids Res, 45(11), pp. 6698–6716. doi: 10.1093/NAR/GKX187.

Tran, D. P. *et al.* (2001) 'Mechanism of Poly(A) Signal Transduction to RNA Polymerase II In Vitro', *MOLECULAR AND CELLULAR BIOLOGY*, 21(21), pp. 7495–7508. doi: 10.1128/MCB.21.21.7495-7508.2001.

Tsutakawa, S. E. *et al.* (2020) 'Envisioning how the prototypic molecular machine TFIIH functions in transcription initiation and DNA repair', *DNA Repair*. Elsevier, 96, p. 102972. doi: 10.1016/J.DNAREP.2020.102972.

Villamil, G. *et al.* (no date) 'Transient transcriptome sequencing: computational pipeline to quantify genome-wide RNA kinetic parameters and transcriptional enhancer activity'. doi: 10.1101/659912.

Vincent, O. *et al.* (2001) 'Interaction of the Srb10 kinase with Sip4, a transcriptional activator of gluconeogenic genes in Saccharomyces cerevisiae', *Molecular and cellular biology*. Mol Cell Biol, 21(17), pp. 5790–5796. doi: 10.1128/MCB.21.17.5790-5796.2001.

Vos, S. M. *et al.* (2018) 'Structure of activated transcription complex Pol II-DSIF-PAF-SPT6', *Nature*. Nature, 560(7720), pp. 607–612. doi: 10.1038/S41586-018-0440-4.

Wahle, E. and Rüegsegger, U. (1999) '3'-End processing of pre-mRNA in eukaryotes', *FEMS Microbiology Reviews*. Oxford Academic, 23(3), pp. 277–295. doi: 10.1111/J.1574-6976.1999.TB00400.X.

Watson, J. D. and Crick, F. H. C. (1953) 'Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid', *Nature 1953 171:4356*. Nature Publishing Group, 171(4356), pp. 737–738. doi: 10.1038/171737a0.

Wilson, L. J. *et al.* (2018) 'New Perspectives, opportunities, and challenges in exploring the human protein kinome', *Cancer Research*. American Association for Cancer Research Inc., 78(1), pp. 15–29. doi: 10.1158/0008-5472.CAN-17-2291/661230/P/NEW-PERSPECTIVES-OPPORTUNITIES-AND-CHALLENGES-IN.

Wong, K. H., Jin, Y. and Struhl, K. (2014) 'TFIIH phosphorylation of the Pol II CTD stimulates Mediator dissociation from the preinitiation complex and promoter escape', *Molecular cell*. NIH Public Access, 54(4), p. 601. doi: 10.1016/J.MOLCEL.2014.03.024.

Woodcock, C. L. F., Safer, J. P. and Stanchfield, J. E. (1976) 'Structural repeating units in

chromatin: I. Evidence for their general occurrence', *Experimental Cell Research*. Academic Press, 97(1), pp. 101–110. doi: 10.1016/0014-4827(76)90659-5.

Wu, C. H. *et al.* (2003) 'NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila', *Genes & Development*. Cold Spring Harbor Laboratory Press, 17(11), p. 1402. doi: 10.1101/GAD.1091403.

Xu, H. *et al.* (2016) 'Stochastic Kinetics of Nascent RNA', *Physical Review Letters*. American Physical Society, 117(12). doi: 10.1103/PhysRevLett.117.128101.

Yang, Y. *et al.* (2016) 'PAF Complex Plays Novel Subunit-Specific Roles in Alternative Cleavage and Polyadenylation', *PLoS genetics*. PLoS Genet, 12(1). doi: 10.1371/JOURNAL.PGEN.1005794.

Yu, M. *et al.* (2015) 'RNA polymerase II-associated factor 1 regulates the release and phosphorylation of paused RNA polymerase II', *Science (New York, N.Y.)*. Science, 350(6266), pp. 1383–1386. doi: 10.1126/SCIENCE.AAD2338.

Zaborowska, J., Egloff, S. and Murphy, S. (2016a) 'The pol II CTD: new twists in the tail', *Nature Structural & Molecular Biology 2016 23:9*. Nature Publishing Group, 23(9), pp. 771–777. doi: 10.1038/nsmb.3285.

Zaborowska, J., Egloff, S. and Murphy, S. (2016b) 'The pol II CTD: new twists in the tail', *Nature Structural & Molecular Biology*. Nature Publishing Group, 23(9), pp. 771–777. doi: 10.1038/nsmb.3285.

Zaborowska, J., Egloff, S. and Murphy, S. (2016c) 'The pol II CTD: New twists in the tail', *Nature Structural and Molecular Biology*. doi: 10.1038/nsmb.3285.

Zacher, Benedikt *et al.* (2014) 'Annotation of genomics data using bidirectional hidden Markov models unveils variations in Pol II transcription cycle', *Molecular Systems Biology*. John Wiley & Sons, Ltd, 10(12), p. 768. doi: 10.15252/msb.20145654.

Zacher, B *et al.* (2014) 'The genomic STate ANnotation package'.

Zhang, C. *et al.* (2013) 'Structure-guided Inhibitor Design Expands the Scope of Analog-Sensitive Kinase Technology', *ACS chemical biology*. NIH Public Access, 8(9), p. 1931. doi: 10.1021/CB400376P.

Zhang, M. *et al.* (2021) 'CDK inhibitors in cancer therapy, an overview of recent development', *American Journal of Cancer Research*. e-Century Publishing Corporation, 11(5), p. 1913. Available at: /pmc/articles/PMC8167670/ (Accessed: 31 March 2022).

Zhang, P. *et al.* (2016) 'An Overview of Chromatin-Regulating Proteins in Cells', *Current protein & peptide science*. NIH Public Access, 17(5), p. 401. doi: 10.2174/1389203717666160122120310.

Zhang, T. *et al.* (2016) 'Covalent targeting of remote cysteine residues to develop CDK12 and CDK13 inhibitors', *Nature chemical biology*. Nat Chem Biol, 12(10), pp. 876–884. doi: 10.1038/NCHEMBIO.2166.

Zheng, Z. (2022) 'Cycle Transcriptional Control : Conservation across Eukaryotic'.

Zheng, Z. L. (2022) 'Cyclin-Dependent Kinases and CTD Phosphatases in Cell Cycle Transcriptional Control: Conservation across Eukaryotic Kingdoms and Uniqueness to Plants', *Cells*. Multidisciplinary Digital Publishing Institute (MDPI), 11(2). doi: 10.3390/CELLS11020279.

Zhou, M. and Law, J. A. (2015) 'RNA Pol IV and V in Gene Silencing: Rebel Polymerases Evolving Away From Pol II's Rules', *Current opinion in plant biology*. NIH Public Access, 27, p. 154. doi: 10.1016/J.PBI.2015.07.005.

Zhou, Q., Li, T. and Price, D. H. (2012) 'RNA Polymerase II Elongation Control', *Annual review of biochemistry*. NIH Public Access, 81, p. 119. doi: 10.1146/ANNUREV-BIOCHEM-052610-095910.

Žumer, K. *et al.* (2021) 'Two distinct mechanisms of RNA polymerase II elongation stimulation in vivo', *Molecular cell*. Mol Cell, 81(15), pp. 3096-3109.e8. doi: 10.1016/J.MOLCEL.2021.05.028.

Żylicz, J. J. *et al.* (2019) 'The Implication of Early Chromatin Changes in X Chromosome Inactivation', *Cell*. Cell Press, 176(1–2), pp. 182-197.e23. doi: 10.1016/J.CELL.2018.11.041.