

*Multi-omics analysis of transcription  
kinetics in human cells*

---

DISSERTATION

FOR THE AWARD OF THE DEGREE  
“DOCTOR RERUM NATURALIUM”  
OF THE GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN

WITHIN THE DOCTORAL PROGRAM  
MOLECULAR BIOLOGY OF CELLS  
OF THE GEORG-AUGUST UNIVERSITY SCHOOL OF SCIENCE (GAUSS)

SUBMITTED BY

SASKIA-BERENICE GRESSEL

FROM BERLIN, GERMANY



GÖTTINGEN 2019



## **Thesis advisory committee (TAC)**

Prof. Dr. Patrick Cramer

Max Planck Institute for Biophysical Chemistry (MPI-bpc), *Department of Molecular Biology*, Göttingen, Germany

Prof. Dr. Halyna Shcherbata

MPI-bpc, *Department of Gene Expression and Signaling*, Göttingen;  
Hannover Medical School, Institute of Cell Biochemistry, Hannover, Germany

Prof. Dr. Julien Gagneur

Technical University of Munich, *Department of Informatics*, Germany

## **Members of the examination board**

First reviewer: Prof. Dr. Patrick Cramer

MPI-bpc, *Department of Molecular Biology*, Göttingen, Germany

Second reviewer: Prof. Dr. Markus Bohnsack

University Medical Center Göttingen, *Department of Molecular Biology*, Germany

## **Other members of the examination board**

Prof. Dr. Herbert Jäckle

MPI-bpc, *Department of Molecular Developmental Biology*, Göttingen, Germany

Prof. Dr. med. Matthias Dobbelsstein

University Medical Center Göttingen, *Department of Molecular Oncology*, Germany

Prof. Dr. Tim Beißbarth

University Medical Center Göttingen, *Department of Medical Biostatistics*, Germany

Dr. Ufuk Günesdogan

Göttingen Center for Molecular Biology, *Department of Developmental Biology*, Göttingen, Germany

Date of thesis submission: March 28, 2019

Date of thesis defense: May 6, 2019

# Promovierenden-Erklärung der Georg-August-Universität Göttingen

Name: Gressel, Saskia-Berenice  
(Name, Vorname)

Ich beabsichtige, eine Dissertation zum Thema „Multi-omics analysis of transcription kinetics in human cells“ an der Georg-August-Universität Göttingen anzufertigen. Dabei werde ich von Herrn Prof. Dr. Patrick Cramer betreut.

Ich gebe folgende Erklärung ab:

1. Die Gelegenheit zum vorliegenden Promotionsvorhaben ist mir nicht kommerziell vermittelt worden. Insbesondere habe ich keine Organisation eingeschaltet, die gegen Entgelt Betreuerinnen und Betreuer für die Anfertigung von Dissertationen sucht oder die mir obliegenden Pflichten hinsichtlich der Prüfungsleistungen für mich ganz oder teilweise erledigt.
2. Hilfe Dritter wurde bis jetzt und wird auch künftig nur in wissenschaftlich vertretbarem und prüfungsrechtlich zulässigem Ausmaß in Anspruch genommen. Insbesondere werden alle Teile der Dissertation selbst angefertigt; unzulässige fremde Hilfe habe ich dazu weder unentgeltlich noch entgeltlich entgegengenommen und werde dies auch zukünftig so halten.
3. Die Richtlinien zur Sicherung der guten wissenschaftlichen Praxis an der Universität Göttingen werden von mir beachtet.
4. Eine entsprechende Promotion wurde an keiner anderen Hochschule im In- oder Ausland beantragt; die eingereichte Dissertation oder Teile von ihr wurden nicht für ein anderes Promotionsvorhaben verwendet.

Mir ist bekannt, dass unrichtige Angaben die Zulassung zur Promotion ausschließen bzw. später zum Verfahrensabbruch oder zur Rücknahme des erlangten Grades führen.

Göttingen, den 28. März 2019

---

(Unterschrift)

# Acknowledgements

During this graduate work, I have had the pleasure to work with many remarkable people and I wish to express my sincere gratitude to them. As I am finishing my thesis, I realize how much these last four years influenced me both scientifically and on a personal level and I am deeply grateful for the lessons I learned and the perspectives I gained. I would like to specifically thank the following people:

- My thesis advisor Prof. Dr. Patrick Cramer for the offer to pursue this challenging project in a fantastic scientific environment. He supported my freedom to explore and develop my own project and ideas while keeping them focused towards the big picture. His mentorship and trust inspired me since we first met during my Master's program in Munich. Special thanks for his encouragement and support of my independence as a scientist by letting me travel to conferences, for training me in writing & co-reviewing and for giving me the opportunity to organize our annual symposium in Berlin.
- This thesis was a collaboration with a brilliant mathematician in the lab, Dr. Björn Schwalb. I have to sincerely thank him for cleverly reformulating my hypotheses into algorithms that produced answers. All the computations presented in this dissertation were performed by him, and without his contributions, I would still be buried in unanswered questions. We interpreted results together almost daily, and formulated new hypotheses that had never been asked before. Together, we have learned a great deal about how transcription of the genome might be regulated.
- My committee members, Prof. Dr. Halyna Shcherbata and Prof. Dr. Julien Gagneur, for keeping track of my progress, and offering useful advice along the way.
- Prof. Dr. Dirk Eick, Dr. Tim M. Decker, Prof. Dr. Heinrich Leonhardt, and Dr. Weihua Qin for our collaboration on the P-TEFb kinase. It was a delightful learning experience and I am thankful for their efforts.
- Dr. Katja Lidschreiber for invaluable TT-seq advice throughout my graduate work, for her support and friendship.
- My colleagues in the Cramer group for their direct or indirect input to my work and providing an extraordinary working atmosphere. It was always a great pleasure working with you! Special thanks to Dr. Margaux Michel who introduced me to TT-seq and got me started in the lab, to Dr. Seychelle Vos for getting me started with background reading on pausing and thought-provoking discussions, to Dr. Livia Caizzi for being a caring bench neighbor and my favorite co-organizer of our weekly get-togethers. Many thanks to Goran Kokić for fantastic discussions and for sharing an office while tackling this document.

- My rotation and Master student Kseniia Lysakovskaia for helping me to grow as a teacher. It was a pleasure working with you, and I wish you the best of luck for the challenges ahead.
- Kerstin Maier, Petra Rus and Till Hülsmann for their great support in our sequencing ‘facility’. Dr. Stefan Krebs and Dr. Helmut Blum for initial help with sequencing.
- Kirsten Backs, Janine Blümel, Kerstin Maier, Petra Rus, Thomas Schulz, Sigurd Hille, Angelika Kruse and Manuela Wenzel for keeping our lab running.
- Dr. Anna Sawicka, Dr. Kristina Žumer, Salma Sohrabi-Jahromi, Dr. Merle Hantsche, Marco Dombrowski, Dr. Hauke Hillen, Sara Osman, Felix Wagner, Marc Böhning, Taras Velychko, Dr. Christian Dienemann, Dr. Svetlana Dodonova, Dr. Carlo Bäjén, Dr. Carina Demel, Dr. Michael Lidschreiber, Dr. Sofia Battaglia, Katharina Hofmann, Dr. Lisa Anna Jung, Dr. Jinmi Choi, Noah Wulff Mottelson, Dr. Johannes Söding, Dr. Juliane Liepe and Prof. Dr. Henning Urlaub for support and feedback during this work.
- The entire GGNB team for putting up with many organizational questions, their advice and constant support during the last 4 years, especially during my time as a student representative for the CELLS program. I was very lucky to profit from multiple interesting workshops and to be awarded travel fellowships of our program. The scientific as well as personal support of the CELLS program and our spokesperson Prof. Dr. Markus Bohnsack was truly outstanding.
- The Max Planck Society for providing an outstanding working atmosphere. Special thanks to our institute’s administration (especially Helena Miletic), our PhD/PostDoc Community, and the Meioses department (especially Dr. Tommaso Cavazza) for supporting our student initiative of weekly get-togethers.
- My examination committee members for their interest in my work: Prof. Dr. Patrick Cramer, Prof. Dr. Markus Bohnsack, Prof. Dr. Tim Beißbarth, Prof. Dr. Matthias Dobbstein, Prof. Dr. Herbert Jäckle and Dr. Ufuk Günesdogan.
- Salma Sohrabi-Jahromi, Dr. Björn Schwalb, Dr. Christopher Thomas and Dr. Katja Lidschreiber for reviewing parts of this document.
- Prof. Dr. Jennifer Doudna, Prof. Dr. Thomas Ott, Prof. Dr. Erich Wanker, Prof. Dr. Patrick Cramer, and Prof. Dr. Klaus Förstemann for their scientific mentorship during my undergraduate time in Munich, Berlin and Berkeley from which I found my excitement for science.

Above all, I am very thankful to my boyfriend, my sister, my parents and all my friends who constantly supported me with their love and encouragement. **Thank you**, for sharing ups and downs, and having a mutual understanding of my fields working conditions! It is fantastic, that you take an interest in my work and help me to see my research in new and creative ways.

## Summary

Transcription of the eukaryotic genome is a highly regulated process which is accomplished in a number of steps, known as the transcription cycle. One of these steps is promoter-proximal pausing, a regulatory halt of the RNA polymerase II (Pol II) shortly after transcription initiation that is released by the P-TEFb kinase CDK9. A kinetic model of transcription predicted that pause duration delimits the initiation frequency and suggested that paused Pol II sterically interferes with initiation. The relationship between promoter-proximal pausing and initiation could thus far not be tested, as no method was available to measure initiation frequencies *in vivo*. Other kinetic parameters such as pause duration and elongation velocity could likewise not be described genome-wide.

Here we show that in human hematopoietic cells the ‘pause-initiation limit’ restricts transcription in steady state, upon perturbation of the CDK9 kinase, and upon heat shock.

To elucidate this relationship, we developed a novel multi-omics approach that quantifies transcription kinetic parameters by combining two state-of-the-art sequencing methods with novel kinetic modeling. Specifically, mammalian native elongating transcript sequencing (mNET-seq) and transient transcriptome sequencing (TT-seq). mNET-seq maps the position and amount of Pol II in high-resolution across each strand of the entire human genome. TT-seq allows to distinguish newly synthesized from pre-existing RNA, and thus, measures immediate transcription activity as transcribed nucleotides per time. Combining measurements of TT-seq and mNET-seq enables us to derive productive initiation frequency, pause duration and elongation velocity genome-wide.

For highly specific and fast inhibition of the pause release kinase CDK9, we engineered an analog-sensitive human cell line using CRISPR/Cas9. Upon CDK9 kinase inhibition, pause duration increased and productive initiation frequency decreased genome-wide. This shows that CDK9 activity stimulates the release of paused polymerase and activates transcription by increasing the number of transcribing polymerases and thus increases the amount of RNA synthesized per time. We found that highly CDK9 responsive genes are associated with long-range chromatin interactions. We showed that human pause sites are located ~50 bp downstream of the transcription start site and show an enrichment for G/C-C/G dinucleotides. Furthermore, transcribed RNA of genes with longer pause durations shows higher secondary structure propensity upstream of the pause site.

We next explored whether the pause-initiation limit applies to transcription activation upon heat shock. To this end, we annotated protein-coding RNAs, and six major long noncoding transcript classes in human hematopoietic cells. Using our multi-omics approach, we observed a reciprocal behavior of productive initiation frequency and pause duration in steady state for all gene classes except enhancers. We showed that upon heat shock the pause-initiation limit indeed restricts transcription activation at most genes. Surprisingly, enhancer elements are not limited by pausing and depend less on CDK9 activity than protein-coding genes.

Together these results suggest a model in which transcription might be activated by an increase of the productive initiation frequency at enhancer elements, accompanied by a decrease in pause duration at the target gene. This allows for an increase of productive initiation events at the target gene. Our multi-omics approach can now be used to further dissect the effect of other known Pol II pause and release factors in a quantitative and genome-wide manner, ultimately revealing the nature of gene regulation in human cells. This will pave the way for novel treatments of diseases with transcriptional malignancies such as cancers, HIV infection or cardiac hypertrophy.

## Publications

Part of this work has been published or is in the process of publication:

**2019 The pause-initiation limit restricts transcription activation in human cells**

**S. Gressel** \*, B. Schwalb \* and P. Cramer

(\* joint first authorship)

*Nature Communications, in revision*

Author contributions: SG conceived and carried out all experiments and contributed to bioinformatics analysis. BS designed and carried out all bioinformatics analysis. BS and PC designed and supervised research. Figures were prepared by SG. SG, BS and PC prepared the manuscript.

**2019 Transient transcriptome sequencing: experimental protocol to monitor genome-wide RNA synthesis including enhancer transcription**

**S. Gressel** \*, K. Lidschreiber \*, and P. Cramer

(\* joint first authorship)

Forthcoming publication in Springer, series *Methods in Molecular Biology*

Book title: Nascent RNA: Methods and Protocols

Preprint at *protocols.io* 2019 Jun 13. doi: [dx.doi.org/10.17504/protocols.io.3vzgn76](https://doi.org/10.17504/protocols.io.3vzgn76)

Author contributions: Figures were prepared by SG and KL. SG, KL and PC prepared the manuscript.

**2017 CDK9-dependent RNA polymerase II pausing controls transcription initiation**

**S. Gressel** \*, B. Schwalb \*, T. M. Decker, W. Qin, H. Leonhardt, D. Eick and P. Cramer

(\* joint first authorship)

*eLife* 2017 Oct 10. doi: [10.7554/eLife.29736](https://doi.org/10.7554/eLife.29736)

Author contributions: SG optimized and carried out TT-seq and mNET-seq experiments and contributed to the design of bioinformatics analysis. BS designed and carried out bioinformatics analysis. TMD carried out cellular and biochemical characterization of the CDK9<sup>as</sup> strain and contributed to TT-seq experiments of CDK9<sup>as</sup> strain. WQ generated and validated the CDK9<sup>as</sup> cell line. DE and HL designed and supervised CDK9<sup>as</sup> strain generation, validation and characterization. BS and PC designed and supervised research. Figures were prepared by SG and BS. SG, BS and PC prepared the manuscript, with input from all authors.

A detailed summary of chapters and items excerpted from the publications listed above can be found in the Appendix (see page 163). All co-authors listed above approved the use of our manuscript(s) in this document. Furthermore, a list of contributions can be found at the beginning of each subsection in the Methods chapter.

In addition, I contributed to the following publications:

**2019 Native molecule sequencing by nano-ID reveals synthesis and stability of RNA isoforms**

K. Maier, **S. Gressel**, P. Cramer and B. Schwalb

Preprint available at *bioRxiv 2019 May 11*. doi: 10.1101/601856

Author contributions: KM, BS and SG carried out experiments. BS designed and carried out all bioinformatics analysis. BS conceptualized, designed and supervised research. BS and PC prepared the manuscript, with input from all authors.

**2019 Transcriptome maps of general eukaryotic RNA degradation factors**

S. Sohrabi-Jahromi\*, K. B. Hofmann\*, A. Boltendahl, C. Roth, **S. Gressel**, C. Baejen, J. Soeding, and P. Cramer

(\* joint first authorship)

*eLife 2019 May 28*. doi: 10.7554/eLife.47040

Author contributions: SSJ designed and performed bioinformatics analysis. KH optimized and carried out PAR-CLIP experiments. AB, CB, SG carried out PAR-CLIP experiments. CR contributed to bioinformatics analysis. JS and PC designed and supervised research. SSJ, KH, JS and PC prepared the manuscript, with input from other authors.

---

## Table of Contents

Promovierenden-Erklärung der Georg-August-Universität Göttingen.....	II
Acknowledgements.....	III
Summary.....	IV
Publications.....	VII

### I. Introduction

1. Chromatin transcription and regulation in human cells.....	1
1.1 Key concepts of Pol II transcription regulation.....	2
1.2 The definition of a gene.....	5
1.3 Case study for transcription regulation: heat shock.....	7
2. Pol II promoter-proximal pausing.....	9
2.1 Positive transcription elongation factor b (P-TEFb).....	10
2.2 Integration of pausing models.....	10
3. Studying transcription kinetics genome-wide.....	12
3.1 Background reading on transcription kinetics.....	12
3.2 A novel multi-omics approach to quantify transcription kinetics.....	15
4. Motivation and aims of this thesis.....	17

### II. Materials and Methods

1. Materials.....	19
1.1 Chemicals, enzymes, kits and other consumables.....	19
1.2 Plasmids and oligonucleotides.....	23
1.3 Primers.....	23
1.4 Antibodies for Western blot and IP.....	25
1.5 Cell culture: cell lines, medium and consumables.....	26
1.6 Buffers.....	27
1.7 Software and databases.....	30
2. Methods.....	31
2.1 General methods.....	31
2.2 Multi-omics.....	35
2.3 CDK9-dependent Pol II pausing controls transcription initiation.....	46
2.4 The pause-initiation limit restricts transcription activation.....	54

### III. Results

1. Establishing mNET-seq in human suspension cells.....	62
2. CDK9-dependent Pol II pausing controls transcription initiation .....	65
2.1 CRISPR/Cas9 engineered mutation allows for specific CDK9 inhibition.....	66
2.2 TT-seq monitors immediate response to CDK9 inhibition .....	66
2.3 Pol II elongation velocity is gene-specific .....	68
2.4 Characteristics of the promoter-proximal pause site .....	69
2.5 Multi-omics analysis provides pause duration $d$ and initiation frequency $I$ .....	70
2.6 Human genes have a ‘pause-initiation limit’.....	70
2.7 Determinants of promoter-proximal pausing .....	73
3. The pause-initiation limit restricts transcription activation.....	75
3.1 Multi-omics analysis and annotation of transcription units.....	77
3.2 Transcription kinetics differ between gene classes.....	79
3.3 Monitoring transcription kinetics during a natural transcription response.....	79
3.4 The pause-initiation limit restricts transcription activation .....	80
3.5 CDK9 activity lowers the pause-initiation limit for gene activation.....	82
3.6 Enhancer transcription is generally not pause limited.....	84
3.7 Enhancer transcription is less dependent of CDK9.....	84

### IV. Discussion and Outlook

1. Our contributions to the field.....	86
1.1 A novel multi-omics approach .....	86
1.2 The pause-initiation limit .....	88
1.3 DNA sequence determines pause site but not pause duration .....	89
1.4 Transcription of enhancer elements is generally not pause-limited.....	91
1.5 Towards a model for gene activation .....	91
2. Technical improvements.....	93
2.1 TT-seq .....	93
2.2 mNET-seq.....	94
3. Future explorations.....	96
3.1 Gene-specific identification of (un)known pausing factors.....	97
3.2 Perturbation tools with high kinetic resolution.....	97
3.3 The ratio of non-productive to productive initiation frequency .....	98
3.4 Elongation velocity at high resolution across the gene.....	99

---

V. Supplementary Information

1. Introduction ..... 102

2. Materials and Methods ..... 108

    2.1 Sequencing data pre-processing ..... 108

    2.2 Additional tables and figures..... 110

3. Results ..... 113

    3.1 mNET-seq optimization in suspension cells ..... 113

    3.2 CDK9-dependent Pol II pausing controls transcription initiation ..... 116

    3.3 The pause-initiation limit restricts transcription activation..... 119

4. Discussion ..... 129

References ..... 131

Appendix

Listed Contents..... 160

    List of figures ..... 160

    List of tables ..... 161

    List of items from publications..... 163

    Abbreviations ..... 166

Sequences and plasmid maps ..... 169

Curriculum Vitae ..... 176



# I. Introduction

Three billion base pairs (bp) of deoxyribonucleic acid (DNA)<sup>1</sup> contain the molecular information that is fundamental to growth, development and survival of the human organism<sup>2,3</sup>. This information is encoded in its genome, spanning 23 pairs of chromosomes, each one a linear and complementary DNA molecule organized in a superior layer of three-dimensional (3D) nuclear architecture contained in each of its cells<sup>4,5</sup>. Its linear nucleic acid complementarity<sup>6</sup> and structure is essential for the process of transcription which can manage the genome's wealth of information and deliver it to the organism. At the heart of this process is an enzyme called DNA-dependent RNA polymerase<sup>7-9</sup> which catalyzes the transcription of double-stranded DNA to single-stranded ribonucleic acid (RNA) molecules by employing ribonucleoside triphosphates as substrates<sup>10,11</sup>. The copying process from so-called coding regions of the DNA, in which RNA serves as the messenger of genetic information for the production of a polypeptide sequence which then folds into a functional protein is also referred to as 'central dogma of molecular biology'<sup>12-14</sup>. In the last couple of years, it has come to light that noncoding RNAs contribute equally to human growth, development and survival<sup>15,16</sup>. Transcription of coding and noncoding RNAs is accomplished in a number of steps, known as the transcription cycle, during which the cell can regulate and control RNA outcome<sup>17</sup>. One of these steps is promoter-proximal pausing, a regulatory halt of the RNA polymerase during the active transcription process<sup>18,19</sup>. In this thesis, I will discuss new insights we have gained into how promoter-proximal pausing is used by the cell for regulatory purposes and why this is a major leap forward in our understanding of transcription regulation.

In order to introduce the reader to the relevant information needed to understand this work, I first review selected milestones in the transcription field. For this purpose, I focus on the transcription cycle, its regulation, and the concept of a gene (chapter 1). Then, I focus on the details of promoter-proximal pausing which links transcription initiation to productive elongation (chapter 2). Finally, I will summarize the consensus methods (chapter 3) and present results and contributions that expand the current understanding of global transcription regulation in human cells (chapter 4).

## 1. Chromatin transcription and regulation in human cells

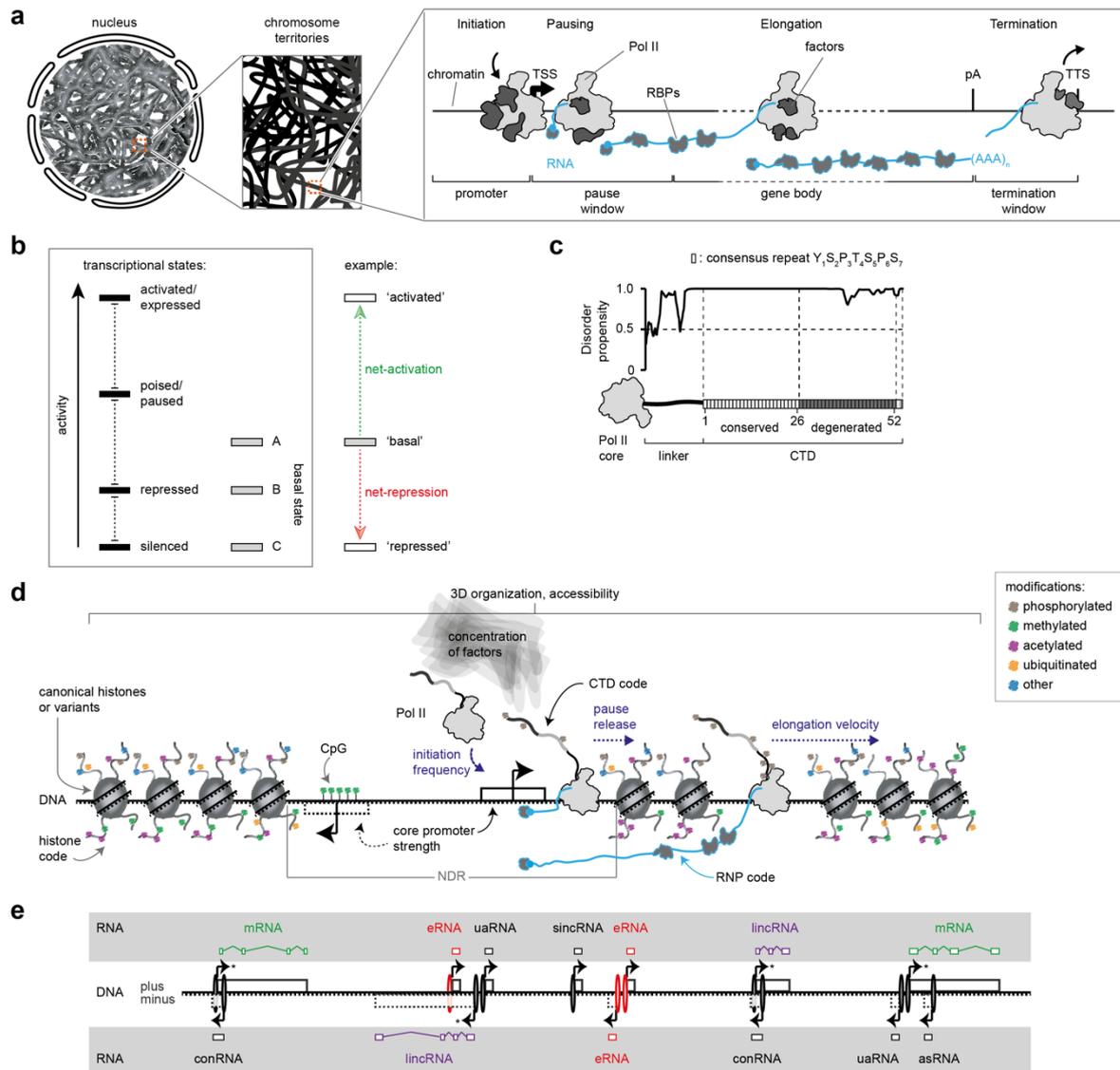
The nucleus of eukaryotic cells provides a nanoenvironment for the tightly compacted but highly organized 3D architecture of the genome<sup>20-23</sup> (**Figure 1 a**, left). The building blocks of this complex structure are canonical histones or histone variants that assemble into octamers<sup>24-26</sup>, which then form nucleosome core particles with DNA<sup>27-29</sup>. These particles are then organized into higher ordered 3D structures with additional architectural factors<sup>30</sup>.

Different ‘phases’, such as hetero- and euchromatin<sup>31</sup>, or the A- and B-compartment<sup>5</sup>, as well as subordinate entities (0.2-1.0 Mbp range) of the chromatin architecture have been identified, including chromosome territories<sup>32</sup>, topologically associating domains (TADs)<sup>33,34</sup>, lamina-associated domains (LADs)<sup>35-37</sup> or nucleolar-associated domains (NADs)<sup>38,39</sup>. Numerous chromatin remodeling mechanisms operate on these structures to ensure accessible surface area for transcription initiation<sup>40,41</sup>. Once transcription is initiated, polymerases are capable of reading through nucleosomes to unravel the stored genetic information<sup>42-46</sup>.

Transcription starts at a defined position: the 5’ end of the transcribed unit which is also referred to as transcription start site (TSS)<sup>47,48</sup>. The TSS is flanked by 50 bp sequences (core promoter) which serve as binding platform for the pre-initiation complex (PIC)<sup>49,50</sup>. The PIC consists of RNA polymerase II (Pol II) and general transcription factors (TFs) TFIID, TFIIA, TFIIB, TFIIF, TFIIIE and TFIIH<sup>51-53</sup>. The molecular details of transcription initiation have been extensively studied and are reviewed elsewhere<sup>48,54</sup>. Together with upstream regulatory sequences<sup>55</sup> (the so-called extended promoter) and more distant located regulatory elements (enhancers)<sup>56-58</sup>, PIC recruitment allows for a highly regulated initiation of transcription. For most transcribed units, transcription initiation occurs bidirectional on two promoters sharing a nucleosome depleted region (NDR)<sup>59-62</sup>. After transcribing 20-100 nucleotides (nt), Pol II pauses<sup>18,19</sup> (see chapter 2). By the action of several pause release factors, transcription proceeds into productive elongation where other elongation factors join Pol II to complete copying the transcribed unit<sup>19</sup>. The entire process of transcribing a human gene can take between a few seconds up to several hours (average time of ~30 min)<sup>63</sup>. Once Pol II passes the polyadenylation (pA) signal, the nascent RNA is cleaved. Pol II is terminated at a transcription termination site (TTS)<sup>64,65</sup>, and eventually recycled<sup>17,66</sup> (**Figure 1 a**, right). Co-transcriptionally, several RNA processing events<sup>67</sup> are initiated or completed, including 5’ end capping (the addition of a methylated guanoside to the 5’ end of the nascent RNA)<sup>68,69</sup>, splicing (the removal of noncoding introns)<sup>70,71</sup>, editing (modification of ribonucleosides)<sup>72-74</sup> and polyadenylation (the addition of a poly(A) stretch to the 3’ end of the RNA) or 3’ end cleavage at a stem-loop structure<sup>64,65</sup>.

## 1.1 Key concepts of Pol II transcription regulation

As outlined above, Pol II undergoes an elaborate and repetitive cycle from initiation to recycling which is enabled and regulated by factors acting in *trans* with Pol II and the nascent RNA backbone<sup>75,76</sup>. Multiple interconnected steps allow for control of when and to which extend transcriptional output is generated. These regulatory steps entail changes in chromatin accessibility<sup>77,78</sup>, (co-) regulator recruitment<sup>79</sup>, as well as allosteric changes<sup>80</sup> and post-translational modifications of factors involved in initiation, promoter-proximal pausing, elongation, termination and recycling<sup>81</sup>. Post-translational modifications are reversible and allow for a dynamic code using regulation by writers, readers and erasers<sup>82</sup>. Since the characterization of Pol II was initiated 50 years ago by Pierre Chambon and Robert Roeder<sup>83-86</sup>, substantial advances have been made towards a mechanistic understanding of transcription and its regulation.



**Figure 1. Key regulatory concepts of chromatin transcription in human cells.**

(a) Genome organization. Left: 3D organization of chromatin in the cell nucleus. Middle: zoom to chromosome territories. Right: representation of the transcription cycle. Pol II is depicted in silver (the CTD of Pol II is not shown) and additional factors in dark grey. RNA and RNA-binding proteins (RBPs) are depicted in blue. Solid line represents DNA and nucleosomes (chromatin). Right schematic is adapted from Hantsche and Cramer<sup>87</sup>.

(b) Transcriptional states ordered by activity. The basal state varies between segments of the genome. The ground state of the majority of the genome is repressed (basal state 'B'), while certain regions have a higher intrinsic activity (basal state 'A') such as promoters. Parts of the genome are strongly repressed (basal state 'C') such as pericentric heterochromatin. Right: negative co-factors repress transcription, while activators and positive co-factors increase transcriptional activity. Schematic is adapted from Burley and Roeder<sup>88</sup>.

(c) Pol II has an unstructured CTD. Disorder analysis (top) and schematic view (bottom) was kindly provided by Marc Böhning (MPI-bpc, Dept. of Molecular Biology)<sup>89</sup>.

(d) Simplified representation shows the layers of transcription regulation. For details refer to main text.

(e) Diagram illustrating the classes of genes (as boxes on the plus or minus DNA strand) encoding protein-coding and long noncoding RNAs: messenger (m) RNA in green; long intergenic noncoding (linc) RNA in purple; enhancer (e) RNA in red; antisense (as) RNA, upstream antisense (ua) RNA, convergent (con) RNA, and short intergenic noncoding (sinc) RNA in black. Sense TSS is marked by asterisk. Promoter states (grey ovals) are associated with multiple gene classes, whereas enhancer states (red ovals) are only associated with genes encoding eRNAs. Top and bottom panels: boxes represent transcribed exons, solid lines represent introns.

For a detailed state-of-the-art picture of the individual steps of transcription the reader is referred to several excellent reviews with recent structure-function<sup>87,90,91</sup>, biochemical, single-molecule imaging<sup>92,93</sup>, or functional genomics insights<sup>25,48,64,71,94,95</sup>. The following paragraphs highlight multiple layers and the dynamics of transcription regulation relevant to this work (**Figure 1**).

*3D organization and chromatin accessibility.* Transcriptional activity of a genomic region depends on its accessibility to pioneering factors, remodelers, and transcription factors<sup>96</sup>. This is determined by chromatin compaction<sup>41,97</sup>, DNA sequence<sup>98</sup>, topology<sup>99</sup> and its modifications (of CpG islands)<sup>100</sup> either of which might restrict access to underlying DNA elements. It is further influenced by the composition and post-translational modifications of the histone octamers<sup>78</sup>. Histones, particularly their accessible N-terminal tail region, can be methylated (me), acetylated (ac), phosphorylated, ubiquitinated, sumoylated, ADP ribosylated, propionylated, butyrylated, deaminated<sup>78,101,102</sup> and serotonylated<sup>103</sup> (histone code). The addition or removal of modifications, or deposition of specialized histone variants can reduce chromatin compaction, act as scaffolds to recruit transcription activators or repressors and thus, associate dynamically with certain transcription states (active, poised, repressed, silenced)<sup>101</sup> (**Figure 1 b**). On top of this, nucleosome positioning<sup>41</sup> and spacing is actively regulated by ATP-dependent chromatin-remodeling complexes which slide, exchange, and evict nucleosomes<sup>40</sup>.

*Repetitive carboxy-terminal domain (CTD) of POLR2A.* Pol II is a 514 kDa enzyme consisting of 12 subunits<sup>104,105</sup>. The largest subunit of Pol II, POLR2A (alias human Rpb1) has a large unstructured CTD which consists of a linker and 52 repeats with the consensus sequence YSPTSPS (Y: tyrosine, S: serine, P: proline, and T: threonine)<sup>82,106,107</sup> (**Figure 1 c**). It serves as landing platform for transcription factors. The CTD of Pol II is hypophosphorylated when it is not bound to its template DNA<sup>108,109</sup>. Beginning from initiation, cyclin-dependent kinases (CDKs) decorate the CTD linker<sup>110,111</sup> and repeats with phosphorylations<sup>82</sup> (**Figure 1 d**). These dynamic modifications recruit factors specific to each step of the transcription cycle<sup>81,82</sup> (CTD code). Upon recycling, the RNA is released, the CTD modifications are reset and Pol II is available for another round of transcription<sup>112</sup>.

*Concentration of factors.* Binding of any factor to Pol II or nascent RNA backbone is dictated by both the concentration of the factor itself and the number of competing binding sites on the target<sup>113</sup>. Each step of the transcription cycle requires a distinct set of factors which have to be provided and organized in space and time<sup>17,114</sup>. An attractive model for local ‘caging’ and organization of multiple factors is liquid-liquid phase separation (LLPS)<sup>115-117</sup>. Furthermore, LLPS might be a crucial for 3D nuclear organization<sup>95</sup>. As it stands today, it is not entirely clear what the components (DNA<sup>118</sup>, RNA<sup>119</sup>, proteins) of transcription condensates are. However, recent reports show that proteins with low complexity intrinsically disordered regions like the CTD of POLR2A<sup>89,120</sup> or the histidine-rich domain in the cyclin subunit of P-TEFb<sup>121</sup> (chapter 2.1) have the potential to phase separate. Switches of the phosphorylation status of the Pol II CTD allow to drive or prevent LLPS dynamically<sup>89,120</sup>.

*(Co-) regulators such as the Mediator complex.* Regulators might act as activators or repressors to modulate transcription<sup>122</sup>. Transcription initiation is stimulated by co-

activators<sup>123</sup> and the Mediator complex to direct Pol II to the correct genomic loci<sup>124</sup>. Activators are characterized by a bipartite organization consisting of a sequence-specific DNA-binding domain and an activation domain<sup>125</sup>. The activation domain contacts and recruits additional multi-protein complexes which are referred to as co-activators<sup>126,127</sup>. A crucial co-activator is the Mediator complex which bridges dynamically between both activators and Pol II as part of the PIC<sup>124</sup>.

*Transition to productive elongation.* After Pol II escapes the promoter, the RNA 5' end is capped and Pol II enters a promoter-proximal window<sup>68,69</sup>. Pol II traverses this window at a rate controlled by accessibility, concentration of factors, and regulators before its release into productive elongation<sup>19,92</sup> (chapter 2). Elongating Pol II is highly processive<sup>128</sup>. Elongation occurs at different velocities along a gene<sup>129,130</sup> (chapter 3.1.1) and thus, coordinates co-transcriptional mechanisms such as RNA processing<sup>71,131,132</sup>, or chromatin modifications by recruitment of modifiers<sup>133</sup>. The process of co-transcriptional mechanisms might also affect the elongation velocity<sup>63</sup>. In addition, dynamically formed RNA secondary structures might modulate elongation velocity due to co-transcriptional folding<sup>134</sup>.

*Termination and recycling.* Efficient termination is important for maintaining the pool of free polymerases for re-initiation<sup>112</sup>. Termination is coupled to RNA 3' end processing and depends on the dissociation of elongation factors in concert with recruitment of termination factors<sup>135</sup>. For recycling and re-initiation of Pol II and the transcription machinery, two nonexclusive models have been proposed: chromatin loop formation to bring the 3' end to the vicinity of the TSS<sup>136-138</sup>, or chromatin compartmentalization<sup>139</sup> for local caging of transcription factors<sup>112</sup>.

If factors involved in the transcription cycle act independently, the informational output would be a simple summation of the effects of each individual factor listed above. However, fine-tuning of transcriptional output suggests an intensive crosstalk between molecular processes which remains to be characterized (**Figure 1 d**). Factors may synergize or antagonize the functions of each other, leading to a complex output dependent on the specific composition.

## 1.2 The definition of a gene

By the late 1970s, a gene was defined as a continuous segment (transcribed unit) within the DNA molecule coding for a protein<sup>140</sup>. This definition was challenged when Richard Roberts and Phillip Sharp discovered that a gene might be discontinuous, containing noncoding (introns) and coding segments (exons)<sup>141-143</sup>. After characterization of the human genome by various high-throughput experimental and computational techniques<sup>144</sup>, a gene was redefined as 'a union of genomic sequences encoding a coherent set of potentially overlapping functional products'<sup>145</sup>. To date, deep sequencing has uncovered the existence of many novel noncoding RNAs with or without an evident functional role of the RNA itself. The latter may still be considered functional due to local effects of the transcription process (incl. proteins related to transcription) on the 3D nuclear architecture e.g. impacting the gene expression of nearby gene loci<sup>146-148</sup>, or as an evolutionary driver in the generation of new functional RNAs<sup>149</sup>.

However, the current challenge is to distinguish transcripts with yet unknown functionality against transcriptional noise. Given that uncertainty, we refer to a gene as the continuous segment (transcribed unit) within the DNA molecule beginning at the TSS of its 5' end, including intronic and exonic sequences<sup>150</sup> and excluding the requirement of validated functionality.

The exon-intron architecture of a gene is highly species-dependent<sup>151</sup>. Most human genes are multi-exonic with an average number of eight exons per gene<sup>3,152</sup>. The maximum number of exons in a human gene is 312 coding for titin – whose name originates from the giants (Titans) of Greek mythology<sup>153,154</sup>. Most exons are short (< 0.2 kbp)<sup>155</sup> while introns are 10- to 100-fold longer (< 3 kbp, only 10 % of introns are > 10 kbp)<sup>154</sup>. An accurate annotation of genes is crucial for the study of molecular biology, and annotation errors propagate into downstream analyses<sup>156</sup>.

### 1.2.1 Long noncoding RNA

Long noncoding (lnc) RNA refers to autonomously transcribed RNA of >200 nt in length (to exclude small RNAs such as miRNAs and snRNAs) that does not encode a protein<sup>157</sup>. By current estimates from the GENCODE reference annotation for the human genome (version 29, GRCH38.p12)<sup>158,159</sup>, 27 % of the total number of genes encode lncRNAs (34 % encode protein-coding RNAs). Similar to protein-coding (m) RNAs, they are often capped<sup>160</sup>, spliced<sup>161</sup> and polyadenylated<sup>162</sup> or form a triple-helix at the 3' end<sup>163</sup>. LncRNAs can be nuclear, cytoplasmic or both<sup>164</sup>. Despite recent insights, many questions concerning the roles of lncRNAs remain<sup>146,165</sup>. In the following paragraphs, I discuss the classification of lncRNA according to genomic location and states (promoter versus enhancer) (**Figure 1 e**), and if applicable, some recent examples of each class.

Genes encoding long intergenic noncoding RNA (lincRNA) do not overlap with protein-coding genes<sup>166,167</sup>. Due to their stability, the annotation of genes encoding lincRNAs is equally well established as for protein-coding genes, and can serve as reference (sense TSS) for the annotation of antisense lncRNAs<sup>168</sup>. Prominent examples of functional lincRNAs in humans are the Xist lincRNA in X-chromosome inactivation<sup>169,170</sup>, or the MALAT1 lincRNA in modulation of splicing efficiency<sup>171,172</sup>. Less stable transcripts of genes that do not overlap with protein-coding genes, are referred to as short intergenic noncoding RNA (sincRNA)<sup>168</sup>. In humans, sincRNAs are on average five-times shorter than lincRNAs<sup>168</sup>.

Genes encoding antisense RNA (asRNA) overlap with annotated genes on the opposite DNA strand<sup>173</sup>. They often have the potential to regulate their overlapping genes<sup>174</sup>. For example, the Sirt1 asRNA increases the stability of Sirt1 mRNA by masking miRNA binding sites<sup>175</sup>, or erbA asRNA inhibits splicing of the erbA pre-mRNA<sup>176</sup>. If the TSS of a gene encoding an asRNA is in close proximity to the sense TSS (<1 kbp downstream) it is referred to as convergent RNA (conRNA)<sup>177</sup>. Elongating Pol II molecules stop upon head-to-head collision<sup>178</sup>. However, in living cells, transcription of sense and antisense transcripts might be lagged and therefore rarely occurring simultaneously. Upstream antisense RNA (uaRNA) is encoded by genes upstream to annotated genes on the opposite DNA strand<sup>174</sup>. It is unclear whether these transcripts have a functional role or are byproducts of the intrinsic bidirectional nature of mammalian promoters and enhancers<sup>59-61,179</sup>.

All of the above listed lncRNA classes might fall in either promoter or enhancer states<sup>48</sup>. Enhancers are regulatory sequences which when in contact with a promoter can

activate gene expression, independently of position, orientation and distance relative to their target gene<sup>56-58</sup>. How to accurately annotate sequences as promoter or enhancer is a significant challenge, especially since enhancer definitions strongly vary<sup>180</sup>. Generally, promoter states are marked by higher levels of trimethylation of lysine 4 of histone 3 (H3K4me3) compared to monomethylation of the same residue (H3K4me) at their promoter<sup>166</sup>. The opposite marks (low H3K4me3, high H3K4me) in combination with H3K27ac were found for enhancers<sup>181</sup>. However, due to the emerging evidence of highly active enhancers displaying a H3K4me3 promoter mark<sup>182</sup>, and promoters serving as enhancers<sup>183,184</sup>, it has to be acknowledged that classification purely based on epigenetic marks is incomplete. Thus, to better distinguish enhancers from promoters, additional characteristics of enhancer functionality have to be included such as chromatin accessibility, the scaffold function of enhancer sequences for transcription factors and chromatin-modifying enzymes<sup>185</sup>, and a functional dissection by reporter gene analysis and CRISPR/Cas9 deletion experiments<sup>186</sup>. If genes fall in enhancer states, their instable product is referred to as enhancer RNA (eRNA)<sup>56</sup>. To date, it remains unclear how eRNAs participate in enhancer function<sup>187,188</sup>.

### 1.3 Case study for transcription regulation: heat shock

At any given time, only a fraction of genes is active while the majority of the genome is repressed (basal state)<sup>88,189</sup>. Transcription regulation allows cells to react in a rapid and coordinated manner to temporary signaling cues (activated or repressed states), or to adapt to permanent changes in their environments<sup>190</sup> (**Figure 1 b**). The regulated changes in transcription can be grouped in three categories: (i) 3D architecture level, (ii) transcription activator level, (iii) transcription machinery and RNA level (**Figure 1 d**). Here, the heat shock response (HSR) was chosen as a case study of transcription regulation because it provides a well-established model system, and it involves global transcriptional mechanisms that are conserved across species<sup>191-195</sup>. The adaptive HSR is an emergency response hallmarked by a disruption in protein homeostasis, and subsequent high induction of heat shock (HS) genes<sup>196</sup>. HS refers to a temperature above the optimum growth temperature ( $T$ ) of an organism representing a challenge for its survival<sup>197</sup>. Humans have a core body temperature ( $T_{\text{core}}$ ) of  $\sim 37$  °C upon homeostasis.  $T_{\text{core}}$  increases upon fever to 39-41 °C<sup>198</sup>, and  $T_{\text{core}}$  above 43 °C is defined as upper limit for survival<sup>199</sup>. To date, a substantial amount of our understanding regarding the HSR has come from studies in yeast<sup>200,201</sup> and *Drosophila*<sup>139,193,202</sup>. The human HSR is less well understood and molecular details are just beginning to emerge. In humans, physiological HS inducing conditions are environmental stress (e.g. heat, energy depletion, ultraviolet radiation), cell development and differentiation, or physical exercise<sup>196,203-205</sup>. Pathological inducers are inflammation and fever, aging, neurodegeneration, or tissue injuries and repair<sup>196,204,206</sup>. Furthermore, a diverse spectrum of human cancers shows a strong correlation of increased gene expression of heat shock factors with reduced survival rates<sup>207,208</sup> (**Figure 36**). After sensing of HS, the cell's survival is not only accomplished on the transcriptional level but further supported by complex post-transcriptional<sup>209-211</sup>, and translational processes<sup>212,213</sup> which are reviewed elsewhere. The aim of this subsection is to briefly summarize what is known for human cells about the various factors that modulate gene expression in response to HS, with focus on the above defined layers (i to iii) of global transcriptional regulation<sup>214</sup>.

*3D architectural changes upon HS.* The 3D architecture of the genome can be measured by high-throughput chromosome conformation capture assays such as Hi-C at ~50 kbp resolution<sup>215</sup>. In human cells, no changes in global compartments (TADs) or looping interactions were observed upon short exposure to heat (30 min)<sup>216</sup>. These data suggest that the 3D architecture (>50 kbp) is already established prior to HS to enable a rapid response. On the chromatin accessibility level, however, dramatic changes were observed upon HS<sup>217</sup>. These changes occur due to nucleosome loss at activated loci<sup>218</sup> and histone modifications, i.e. an increased acetylation of H4<sup>219</sup>.

*Activation.* Unlike *Drosophila* or yeast, plants<sup>220</sup> and animals<sup>204</sup> have multiple heat shock factor (HSF) isoforms. HSFs are the primary stress sensors for HS gene upregulation<sup>206</sup>. In humans, studies have analyzed the genome-wide distribution of HSFs<sup>207,221</sup> and Pol II<sup>219,222</sup> in steady state and upon HS. These studies have identified HSF target genes which are marked by promoter-proximal paused Pol II in steady state<sup>219</sup>. HSF1 is considered the master regulator of the HSR. Upon HS, the monomeric HSF1 is modified by acetylation, phosphorylation and sumoylation, it trimerizes, and translocates to the nucleus<sup>223</sup>. There it binds to certain sequence elements (inverted repeats of nGAAn pentamers) at promoters and enhancers of HS-inducible genes<sup>224</sup>. For selected model genes, it has been shown that transcription is upregulated by recruiting various chromatin remodelers and co-activators<sup>217,225</sup>. Furthermore, HS induces polyADP-ribosylation (PARylation) of chromatin which results in nucleosome destabilization and mimics the effects of linker histone 1 depletion<sup>226</sup>. PARylation also supports local ‘caging’ of transcription factors and Pol II at the activated HSP70 gene locus<sup>139,227</sup>.

*Downregulation.* Upon HS, cells shift all resources from growth to combat stress and ensure survival<sup>197</sup>. This involves the global shut-down of transcription and re-distribution of the transcription machinery to loci with stress-related functions as was shown in yeast during HS<sup>228</sup>. This immediate downregulation of transcription could so far not be investigated quantitatively owing to limitations in extracting transcriptional changes of mRNA in microarray analyses<sup>224,229,230</sup> which measure stable mRNAs and lack the temporal resolution for rapid changes. In the context of global transcriptional shut-down, it is still unknown how HS is sensed and what causes the downregulation at the molecular level. In mice, it was previously reported that HSFs are not required for downregulation<sup>231</sup>. Recent data in human cells suggest that Pol II<sup>219</sup> and negative elongation factors such as NELF increase their residence time on promoters of downregulated genes<sup>232</sup>. This supports a model in which increased promoter-proximal Pol II pausing sustains downregulation upon HS<sup>194,219,231</sup>.

## 2. Pol II promoter-proximal pausing

Widespread peaks of elongating Pol II at 20 to 100 nt near promoters in eukaryotes were initially described for several model genes, including the mammalian  $\beta$ -globin locus<sup>233</sup>, *Drosophila* heat shock genes<sup>234</sup>, human c-myc and c-fos genes<sup>235-238</sup>. Over the last decades<sup>239,240</sup>, promoter-proximal pausing of Pol II has emerged as a general phenomenon across protein-coding genes (including up to 90 % of active promoters<sup>241</sup>), and has blossomed into a rich research field. More recently, Pol II pausing was reported at noncoding genes that produce long noncoding RNAs<sup>242,243</sup>, including enhancer RNAs<sup>182</sup>, and upstream antisense RNAs<sup>244</sup>. In this chapter, I briefly summarize the biological impact of Pol II pausing in human cells, its regulation – focusing on the prominent release factor P-TEFb –, and an emerging model of the molecular mechanism of Pol II pausing.

*What biological functions might Pol II pausing serve in human cells?* Interestingly, Pol II pausing occurs at both, genes actively producing RNA and genes without efficient transcript completion<sup>19</sup>. At genes that are fully transcribed, pausing facilitates the assembly of RNA processing factors<sup>245</sup>. This pausing checkpoint ensures that 5' ends of nascent RNA molecules are capped and protected from degradation prior to elongation<sup>246,247</sup>. Genes that experience initiation but not elongation are often in an uninduced state<sup>248</sup>. As a result, the Pol II pausing associated transcription machinery alters chromatin and maintains an open promoter structure enabling regulatory factors to access underlying DNA elements<sup>18</sup>. Upon developmental regulation<sup>249,250</sup>, or in response to other stimuli<sup>251,252</sup> pausing is reduced and responsive genes are fully transcribed. Furthermore, pausing was shown to synchronize gene activation events in *Drosophila*<sup>253</sup>. Negative consequences of pausing may be transcription-replication conflicts during S phase, especially when paused Pol II encounters the replication machinery head-to-head, leading to DNA damage and genomic instability<sup>254</sup>. Paused Pol II could also interfere with transcription of other genes in its vicinity by steric hindrance<sup>255</sup>.

*What factors affect Pol II pausing in human cells?* At the outset of my graduate work, it has been reported in *Drosophila*<sup>256</sup> and *Escherichia coli*<sup>257</sup> that the DNA sequence composition is affecting Pol II pausing. However, the underlying sequence determinants at human genes were unknown. Similar to other transcription steps (section 1.1), pausing is stabilized by several factors, including DSIF composed of Spt4 and Spt5<sup>258</sup> and the NELF complex with A, B, C/D and E subunits<sup>259</sup>. Paused Pol II has been shown to relocate nucleosomes<sup>260</sup>, and vice versa, nucleosomes seem to enhance pausing<sup>261</sup>. Additional factors influence the stability of paused Pol II, such as GDOWN1 and TFIIF<sup>262</sup>. Most prominent among the pause release factors is the CDK9-containing kinase complex P-TEFb<sup>263-265</sup>. Other factors involved in the transition to productive elongation are the PAF1 complex<sup>110,266-268</sup>, the elongation factors SPT6<sup>269-272</sup>, and TFIIS<sup>273</sup>, as well as several other factors recruiting P-TEFb.

## 2.1 Positive transcription elongation factor b (P-TEFb)

The positive transcription elongation factor b (P-TEFb)<sup>263</sup> is a heterodimer which is constitutively expressed throughout the cell cycle<sup>274</sup> (recently reviewed in<sup>275,276</sup>). It consists of the cyclin-dependent kinase CDK9 originally termed PITALRE<sup>264,277</sup> and a T-type cyclin (CCNT1 or CCNT2)<sup>278-280</sup>. In human cells, P-TEFb is either active or inactive<sup>281,282</sup>. In its inactive form, 7SK RNA serves as a scaffold for interacting proteins (LARP7, MePCE, HEXIM1, HEXIM2<sup>283</sup>), ultimately sequestering P-TEFb in a 7SK small nuclear ribonucleoprotein complex (snRNP)<sup>284</sup>. The ratio of both forms varies between different cell lines but the majority is inactive in steady state<sup>281,282,285</sup>.

*Recruitment and activation of P-TEFb.* Both forms of P-TEFb can be recruited to its target genes via several different recruitment complexes including (co-)activators or chromatin-associated factors (reviewed in<sup>276</sup>). Activators as C-MYC<sup>240,286</sup>, NF-kappaB<sup>287</sup> or the viral HIV Tat<sup>288</sup> transcription factors can directly recruit active P-TEFb to target genes by physically interacting with its CCNT subunit. The subunit MED26 of the Mediator complex binds the super elongation complex (SEC) which contains active P-TEFb, ELL/EAF family members and other factors<sup>289</sup>. Another recruitment mechanism of active P-TEFb utilizes histone tail binders such as BRD4 or MePCE. BRD4 interacts with promoter-proximal histone 4 acetylated lysine (H4K16ac) via its bromodomain<sup>290,291</sup>. MePCE acts independently during P-TEFb recruitment by interacting with histone H4 of the +1 nucleosome<sup>292</sup>. The inactive, 7SK snRNP-bound P-TEFb can be nuclear or chromatin-bound by TRIM28/KAP1<sup>293</sup>. Nuclear 7SK snRNP-bound P-TEFb is activated by T-loop phosphorylation of CDK9 by CDK7 which frees P-TEFb from the 7SK snRNP complex<sup>294,295</sup>. However, it is not known which factor releases P-TEFb from the chromatin-associated 7SK snRNP complexes. To date, our understanding of P-TEFb recruitment remains incomplete and needs to be established at a genome-wide scale. The timing of P-TEFb activation and the localization of its recruitment complexes at a certain gene might define the length of the pause duration.

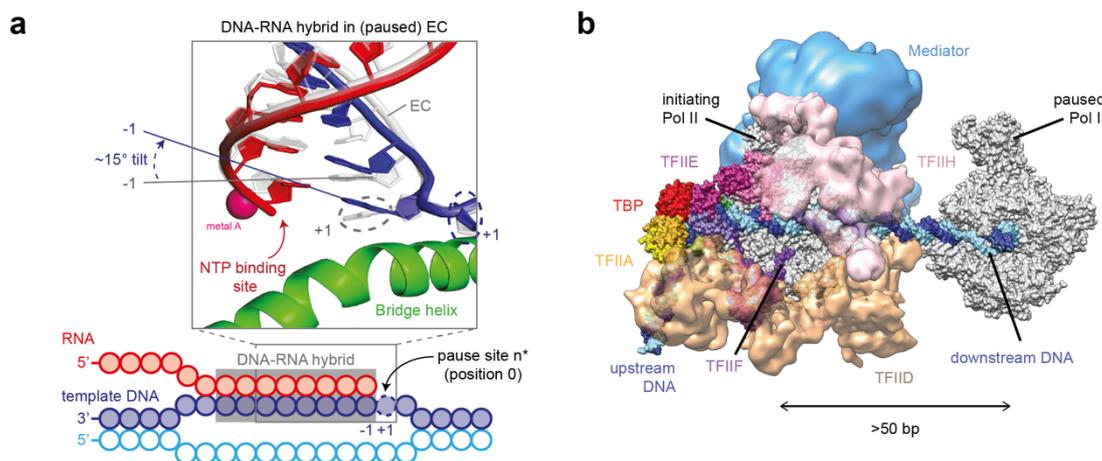
*P-TEFb controlled checkpoints: promoter-proximal pausing and pA site.* After recruitment to its target and activation, the CDK9 kinase phosphorylates the Spt5 subunit of DSIF<sup>296</sup>, NELF<sup>297</sup> and the CTD of POLR2A<sup>298</sup>. For the latter, CDK9 has been shown to phosphorylate the POLR2A linker region of the CTD<sup>110,111</sup>, as well as serine 2<sup>299,300</sup> and serine 5<sup>301</sup> of the CTD heptad repeats (**Figure 1 c**). CDK9 was recently shown to be implicated in a second checkpoint around the polyadenylation site<sup>302</sup>. The second checkpoint is still poorly understood and might implicate CDK9 in 3'-end RNA processing and transcription termination<sup>302,303</sup>. This implication is supported by the observation that P-TEFb enhances the activity of the transcription termination factor XRN2 by phosphorylation<sup>303</sup>.

## 2.2 Integration of pausing models

Today we have a better knowledge of the biochemical composition of paused and elongating Pol II complexes, and many pause and release factors have been identified<sup>304</sup>. But our knowledge about the timing and molecular mechanisms of assembly and composition of Pol II pause and release complexes remains limited (**Supplementary Note 1**). Three models for

establishing promoter-proximal pausing were described in the literature: the kinetic model, the barrier model and the interaction model (reviewed in <sup>304</sup>). The kinetic model describes pausing as a combination of the slow elongation rate of Pol II (see section 3.1.1), reversible sliding of Pol II with low processivity along DNA and RNA (backtracking) <sup>305</sup> and the delayed recruitment rate of Pol II release (P-TEFb complex) and elongation factors prior to productive elongation. The barrier model (also referred to as ‘ubiquitous pausing’ <sup>46</sup>) builds on observations that nucleosomes, especially the first downstream of the TSS, hinder Pol II’s transition to productive elongation <sup>261</sup>. The interaction model suggests that factors stabilizing the paused Pol II complex such as NELF, DSIF, or other factors (of which the activity still needs to be established) determine pausing. Depending on the gene’s architecture and nanoenvironment, the energy landscape and thus, the transition rate to productive elongation will be variable and each of the three site models might contribute differently to establishing a paused Pol II <sup>306</sup> (see also **Supplementary Note 1**).

Recent studies revealed structures of the Pol II elongation complex in the paused and activated state, and provided the first mechanistic insights into the P-TEFb dependent switch to active elongation <sup>110,307</sup> (**Figure 2 a**). The RNA-DNA hybrid within the paused polymerase is in the tilted state that hinders nucleotide addition at the active site <sup>307</sup>. Thus, the subsequent nucleotide is not added yet. We defined the pause site (position 0) to be the position in line with the ‘post-translocated’ RNA rather than with the ‘pre-translocated’ DNA (see Methods II.2.2.4). A tilted hybrid might be the hallmark of a paused state and was also observed in bacterial elongation complexes <sup>308</sup>.



**Figure 2. Structural modeling of promoter-proximal pausing and initiation.**

(a) Top: close-up view of the DNA-RNA hybrid in the paused transcription elongation complex (paused EC) (in dark blue and red) (PDB-code 6GML <sup>307</sup>) compared to the elongation complex (EC) (in silver) (PDB-code 5OIK <sup>105</sup>). In the paused EC, the DNA-RNA hybrid is in an offline transcription state. The template DNA strand (in dark blue) passes over the Pol II bridge helix (in green). Structural view (top) was kindly provided by Dr. Seychelle Vos (MPI-bpc, Dept. of Molecular Biology) <sup>307</sup>. Bottom: schematic of nucleic acid residues (DNA in blue, RNA in red). Shaded area highlights the DNA-RNA hybrid. The pause site ( $n^*$ ) (throughout this work referred to as ‘position 0’) is denoted at +1 of the template DNA. Bottom schematic is adapted from Armache *et al.* <sup>309</sup>. (b) Modeling shows that paused Pol II (silver, right) positioned 50 bp downstream of the TSS allows for formation of the Pol II initiation complex (different colors, left). Modeling is based on the latest structural information (Mediator EMD-8307 <sup>310</sup>, TFIID EMD-3305 <sup>311</sup>, TFIIF EMD-3307 <sup>312</sup>, closed complex PDB-code 5FZ5 <sup>313</sup>, EC PDB-code 1WCM <sup>314</sup>). Structural modeling was performed by Dr. Merle Hantsche (MPI-bpc, Department of Molecular Biology).

Jesper Svejstrup and co-workers explored footprints of initiating and paused Pol II molecules and proposed a theoretic model (referred to as ‘Ehrensberger theory’) in which a paused polymerase interferes with the binding of the Pol II initiation complex, or a newly initiating polymerase triggers the release of a paused polymerase to productive elongation<sup>315</sup>. As a consequence, the authors suggest that initiation might still be the rate limiting step of transcription and pausing might serve as a window of opportunity to collect necessary transcription elongation factors for full processivity during elongation<sup>315</sup>. Structural modeling shows that Pol II positioned 50 bp downstream of the TSS allows for formation of the Pol II initiation complex while shorter distances between the active sites of paused and initiating Pol II are predicted to lead to steric clashes (**Figure 2 b**). Even if a paused Pol II is located further downstream, it may still restrict initiation events if additional polymerases line up behind it.

### 3. Studying transcription kinetics genome-wide

To arrive at a mechanistic understanding of transcription it must be studied at multiple levels: by structural analysis of the components, their functionality *in vitro* (biochemical and single-molecule assays) and *in vivo* (systems biology, imaging). This work contributes to the field of systems biology which combines experimental data and theory, with many researchers contributing to both. In particular, this work focusses on Pol II transcription kinetics in human cells during steady state and changes of transcription kinetics upon perturbation in a genome-wide manner.

#### 3.1 Background reading on transcription kinetics

Transcription kinetics refer to all relevant parameters involved in transcribing RNA in units of molecules or nucleotides and time. In human cells, RNA synthesis depends on multiple (potentially) rate-limiting steps (see chapter 1.1).

Here, I briefly discuss the progress that has been made in recent years in a quantitative description of transcription kinetics, specifically initiation frequency, pause duration, and elongation rate. At the onset of this work, elongation rates were the only rate which had been studied quantitatively genome-wide.

##### 3.1.1 Elongation rates

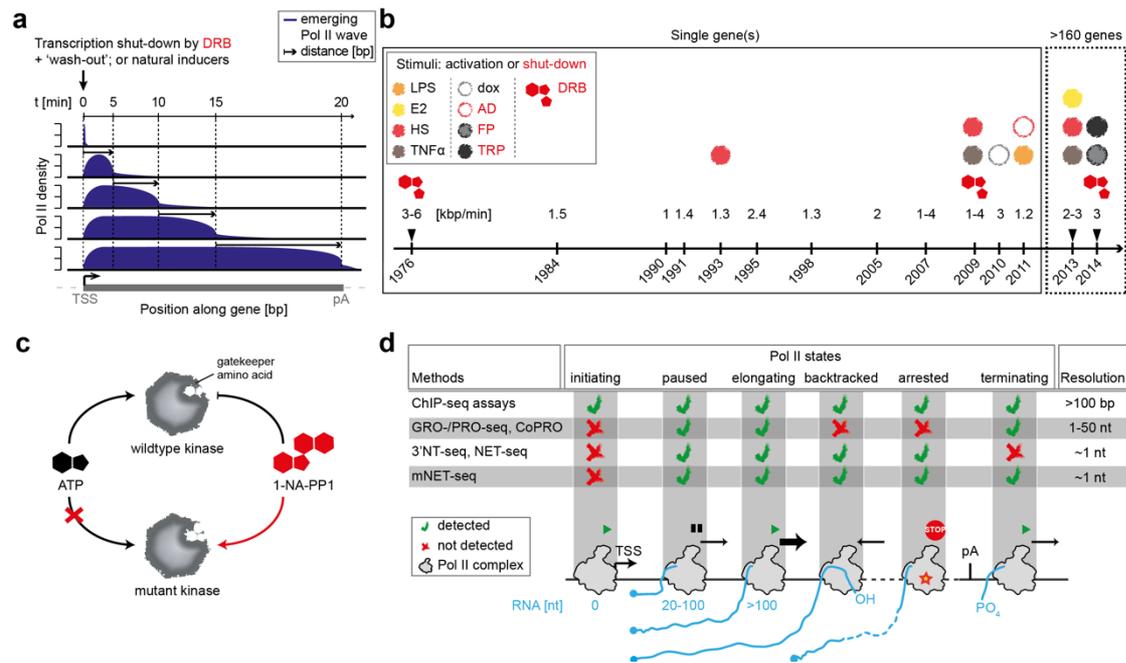
The elongation velocity of Pol II is influenced by the translocation rate of Pol II as well as the frequency and duration of pauses along a gene<sup>316</sup>. The translocation rate is governed by the dynamics of the trigger loop structure of Pol II<sup>317</sup>. The translocation rate of Pol II’s movement per base depends on the DNA sequence and other processivity factors joining Pol II during elongation<sup>92</sup>. At certain positions, Pol II moves slowly (e.g. promoter-proximal pause window) while it accelerates within other regions (e.g. intronic sequences)<sup>129,130</sup>. The promoter-proximal pause duration can also be influenced by several factors (see chapter 2).

By the late 1950s, Igor Tamm and colleagues suggested that 5,6-Dichloro-1- $\beta$ -D-ribofuranosylbenzimidazole (DRB) reversibly reduces the rate of transcription<sup>318</sup>. Two decades later, they were the first measuring elongation rates (3-6 kbp/min) by following the declining

wave of elongating polymerases after DRB induced transcriptional shut-down by radioisotope pulse labeling<sup>319</sup>. Several groups joined to delineate the mode of action of DRB<sup>320,321</sup>, and identified the DRB-sensitive component as the kinase component CDK9 of P-TEFb<sup>264</sup>. Thus, DRB inhibits promoter-proximal pause release resulting in a failure of promoter-proximal Pol II to progress further, while elongating Pol II complexes are able to complete transcription<sup>264</sup>. Following DRB drug removal (wash-out), promoter-proximal Pol II enters productive elongation and nascent transcription can be followed. By measuring the width of the emerging (wash-out) or retreating (shut-down) transcription wave, elongation rates could be assessed (distance/time) (**Figure 3 a**). DRB shut-down or wash-out experiments in combination with a variety of techniques, incl. RT-qPCR<sup>322-324</sup>, tiling microarrays<sup>325</sup>, fluorescent labeling of Pol II<sup>272,326-328</sup> or RNA<sup>329-333</sup>, became a consensus approach for estimating elongation rates of individual model genes (**Figure 3 b**).

Until 2013, studies were limited to a single or small number of genes and reported a wide range of elongation rates (1 to 6 kbp/min) (**Table 28**). This was initially explained by variations between cell types, treatment systems, and technical differences, however, systematic studies of elongation rates were missing. The application of GRO-seq (see subsection 3.1.2) in combination with DRB<sup>129</sup>, or with physiological inducers (17 $\beta$ -estradiol<sup>334</sup> or the cytokine TNF $\alpha$ <sup>334</sup>), was a breakthrough, soon to be followed by a number of genome-wide studies<sup>130,335,336</sup>. These studies addressed elongation rate changes along the gene and between different gene classes. They found that Pol II elongation rates are the slowest close to the beginning of the gene (5' end) and increase within the first 15 kbp for human genes<sup>129,334</sup>. However, the local resolution of these studies was limited to windows of >5 kbp, and thus, offers insufficient resolution on short genes<sup>19</sup>. In addition, DRB does not have high specificity and targets other kinases involved in transcription. Thus, in order to minimize possible physiological perturbations caused by drug treatment, future experiments would benefit from substitution of DRB with a more specific and potent CDK9 kinase inhibitor. To date, commercially available compounds targeting CDK9 suffer from multiple off-target effects (**Figure 35**) and the field lacks tools to do precise molecular perturbations of CDK9 (for review see<sup>337</sup>).

An elegant chemical-genetic approach to inhibit kinases *in vivo* highly specific and within minutes was developed by Kevan Shokat<sup>338-340</sup>. By mutating the so-called gatekeeper residue in the active site of the kinase from the natural amino acid to an amino acid with a smaller side chain (glycine or alanine), a unique interface of the kinase's ATP binding pocket to a small molecule is engineered<sup>339</sup>. The mutant (analog-sensitive) kinase retains its function and cellular distribution *in vivo*<sup>338,341,342</sup>. At the same time, the enlarged binding pocket allows for an unnatural bulky ATP analog (e.g. 1-NA-PP1) to bind which is non-hydrolysable<sup>339</sup> (**Figure 3 c**). It has to be acknowledged that any type of reversible inhibitory ligand spends part of its time 'off' its binding site. However, once bound, the dissociation constant and thus, a background kinase activity, are minimal for bulky ATP analogs<sup>338</sup>. With the development of the CRISPR/Cas9 system<sup>343-345</sup>, a tool was in place to engineer endogenous analog-sensitive CDK9 in human cells.



**Figure 3. Studying transcription kinetics genome-wide.**

(a) Pol II density (arbitrary unit) across a gene. DRB was used to shut-down transcription. Waves of emerging Pol II (blue) after DRB (red) wash-out, or after induction by natural stimuli were studied by time-course. Elongation velocity [bp/min] was assessed by the ratio of distance [bp] and time [min]. (b) Historic overview of studies measuring elongation rates. Single gene studies dominated the transcription field until 2013. A detailed overview of the conducted studies including applied methods and references can be found in **Table 28**. Abbreviations: endotoxin/lipopolysaccharides (LPS) (orange), 17 $\beta$ -estradiol (E2) (yellow), heat shock (HS) (red), tumor necrosis factor alpha (TNF $\alpha$ ) (brown), doxycycline (dox) (white), actinomycin D (AD) (red), flavopiridol (FP) (grey), triptolide (TRP) (black), 5,6-Dichloro-1- $\beta$ -D-ribofuranosylbenzimidazole (DRB) (red). Bottom: elongation rates [kbp/min] for single gene(s) (box with solid line) or average of >160 genes (dotted line) are displayed. (c) Schematic of chemical-genetic approach (Shokat mutation): mutant analog-sensitive kinase (bottom, grey) binds the bulky non-hydrolysable ATP analog (1-NA-PP1, red) which blocks binding of ATP (black). Wildtype kinase (grey, top) can only bind ATP (black) due to gatekeeper amino acid. (d) Comparison of methods used to detect human Pol II complexes. Detected Pol II states and resolution are depicted for each method. Except ChIP-seq assays, all methods (GRO-seq<sup>346</sup>, PRO-seq<sup>347</sup>, CoPRO<sup>247</sup>, 3'NT-seq<sup>348</sup>, NET-seq<sup>349</sup>, mNET-seq<sup>350</sup>) are strand-specific. For comparison see<sup>168,351</sup>. 3'NT-seq<sup>348</sup> and NET-seq without IP<sup>349</sup> are not covered in the main text. Both methods purify chromatin-associated RNAs and are not specific to Pol II<sup>351</sup>. CoPRO is a combination of PRO-seq with a subsequent selection of (un-)capped transcripts which allows accurate mapping of the TSS<sup>247</sup>.

### 3.1.2 Occupancy profiles as a consensus measure of Pol II half-lives

Genes can be distinguished on the basis of their Pol II distribution: no Pol II across the entire gene, evenly distributed Pol II across the gene, or Pol II enriched on the 5'-end of genes<sup>352,353</sup>. Polymerase occupancy can be mapped along DNA or RNA<sup>351</sup>. The respective techniques to measure polymerase occupancy are described in chapter 3.2.2.

Most available studies estimated the degree of Pol II pausing as the relative ratio of Pol II occupancy in the promoter-proximal region and the gene body, which has been termed the traveling ratio<sup>354</sup>, or the pausing index<sup>352,353</sup>. Rapid transition of promoter-proximal Pol II to elongation would yield a low ratio. Promoter-proximal enrichment of Pol II and thus, a slow transition to productive elongation would yield a high ratio. Measuring changes in Pol II occupancy upon chemically inhibiting initiation was the only measure to determine the half-life of Pol II at paused genes. The decay of Pol II occupancy was mostly studied at selected

model genes by ChIP-qPCR<sup>355</sup>, run-on assay<sup>356</sup>, or imaging<sup>357</sup> upon initiation inhibition by triptolide (TRP). TRP blocks the ATPase activity of the TFIIH helicase XPB, thus, opening of the DNA template is not possible<sup>358,359</sup>. However, reported half-lives of paused Pol II studied at individual model genes varied strongly among studies. For example, Pol II half-lives at hsp70 in steady state were reported to be of 2.5 min<sup>360</sup>, 5 min<sup>357</sup>, 10 min<sup>355</sup>, 25 min<sup>356</sup> in *Drosophila*, or even up to 60 min in human cells<sup>361</sup>. Furthermore, half-lives of paused Pol II varied strongly between genes (at 13 genes in *Drosophila*: 1 to 25 min<sup>355</sup>). A single study addressed half-lives genome-wide: in mouse embryonic stem cells, an average Pol II half-life of 6.9 min was measured at 3,181 genes by GRO-seq with TRP<sup>130</sup>. Systematic studies of half-lives in human cells were missing at the outset of this work.

It is critical to point out that although Pol II occupancy depends on pausing, it does not directly relate to the kinetics of pausing<sup>315</sup>. This is because the Pol II occupancy signal at a given time depends on the number of polymerases and their speed, and cannot be used in isolation to distinguish between these two. Indeed, when a Pol II occupancy peak increases, this can be due to an increase in the number of pausing polymerases or due to an increase in the duration of pausing, or both. Taken together, conclusions about the pause duration starting from occupancy profiles can only be drawn by factoring in the number of polymerases released into elongation (productive initiation frequency).

## 3.2 A novel multi-omics approach to quantify transcription kinetics

Kinetic modeling minimally requires: (i) a measure of transcriptional output per time and (ii) a measure of Pol II occupancy. At the outset of my graduate work, existing evidence was primarily limited to measures of Pol II occupancy providing only snapshots but no kinetic insights to promoter-proximal pausing (see section 3.1).

We used TT-seq (transient transcriptome sequencing) in combination with mNET-seq (mammalian native elongating transcript sequencing) analysis, hereafter called multi-omics, as a new method to analyze elongation velocity, pause duration and productive initiation frequency genome-wide in living human cells.

### 3.2.1 Transient transcriptome sequencing (TT-seq)

Natural derivatives of nucleosides such as 4-thiouridine (4sU) are rapidly taken up by a broad range of cells, undergo phosphorylation by cellular uridine kinases, and get incorporated into endogenous RNA<sup>362</sup>. Metabolic labeling with 4sU was first introduced 40 years ago<sup>362</sup>. It was later used in combination with affinity purification and microarray analysis<sup>363-367</sup>. More recently, it was combined with next-generation sequencing (4sU-seq)<sup>368-370</sup> (for review see<sup>371</sup>). Gene expression profiling using 4sU-seq has higher sensitivity than standard transcriptomics, as 4sU-seq separates newly transcribed RNA from the pre-existing RNA pool<sup>366,368,372</sup>. However, if the labeling pulse is shorter than 30 min, these protocols are less sensitive for organisms with larger transcriptomes<sup>373</sup>. This is due to the long unlabeled 5' region of transcripts, which were produced in cells before labeling, compared to the short, labeled newly transcribed 3' region<sup>373</sup>.

The development of transient transcriptome sequencing (TT-seq) solved this problem by introducing an additional fragmentation step after total RNA isolation, followed by sequencing only newly synthesized RNA fragments that are formed during the 5 min labeling

time<sup>168</sup> (see **Figure 6 a** for a detailed experimental workflow). TT-seq can quantify transcriptional activity and RNA metabolism in steady state<sup>168</sup> and during transcription responses such as T cell activation<sup>374</sup>. TT-seq quantifies even short-lived noncoding RNAs, including enhancer RNAs (eRNA)<sup>375</sup> and their changes<sup>374</sup>. TT-seq also detects transient RNAs between the polyadenylation site (pA) and the site for transcription termination (TTS)<sup>168</sup>. Compared to RNA-seq, TT-seq is particularly powerful in monitoring downregulation of transcription because it is not biased by stable transcripts which mask rapid changes. Downregulation of transcription was studied by TT-seq for example during X chromosome inactivation in mouse embryonic stem cells<sup>376</sup>.

Other recently developed methods involve chemical conversion of 4sU into cytosine analogs which is detected as T-to-C point mutation in sequencing<sup>377-379</sup>. The benefit is that biochemical separation of 4sU labeled RNAs is not necessary. However, since these assays preserve labeled and unlabeled molecules, a much higher sequencing depth is required. This is especially problematic if the focus is on fast transcription kinetics using short labeling times. Since the relevant labeled RNA fragments are strongly diluted in unlabeled molecules, these methods are outperformed by TT-seq (for a comparison see<sup>380</sup>).

TT-seq provides a measure of transcriptional output per time (criterion i) as it measures local synthesis rates at nucleotide resolution. Thus, it allows to monitor and quantify productive transcription initiation frequency globally<sup>381</sup> which could not be done before since no alternative method was available.

### 3.2.2 Mammalian native elongating transcript sequencing (mNET-seq)

Polymerase distribution can be mapped with DNA or RNA. Chromatin immunoprecipitation (ChIP) assays such as ChIP-seq<sup>382-384</sup>, ChIP-exo<sup>385</sup> or ChIP-nexus<sup>386</sup> detect the distribution of Pol II by crosslinking to DNA, fragmentation and immunoprecipitation using a Pol II specific antibody. ChIP assays lack strand specificity and are limited in their resolution to >100 bp. To date, ChIP assays are the only method covering all stages of transcription including initiating Pol II<sup>351</sup> (**Figure 3 d**).

RNA-based methods are powerful tools to locate polymerases in a strand-specific manner and at high resolution. To uniquely align an RNA sequence back to the human genome, RNA fragments of >20 nt length are required<sup>346</sup>. Thus, RNA-based methods cannot map initiating Pol II. Widely used RNA-based techniques are nuclear run-on (NRO) assays<sup>346,347</sup>, and native elongating transcript sequencing (NET-seq)<sup>350,387</sup>. NRO assays such as global run-on sequencing (GRO-seq)<sup>346</sup> or precision nuclear run-on sequencing (PRO-seq)<sup>347</sup> can detect actively elongating and paused polymerases in isolated nuclei, but not backtracked and arrested polymerases. To this end, isolated nuclei are treated with sarkosyl which blocks new initiation but releases elongation-competent polymerases for *in vitro* polymerization (NRO) of approximately 1 to 50 nt<sup>346,347</sup>. During NRO, 5-bromouridine 5'-triphosphate (BrUTP) (GRO-seq)<sup>346</sup>, or biotin-labeled ribonucleotide triphosphate analogs (biotin-rNTP) (PRO-seq)<sup>347</sup> are incorporated, allowing the affinity purification of nascent RNAs. Thus, these protocols are specific to NRO-RNAs but not to Pol II-associated RNAs.

Mammalian NET-seq (mNET-seq) maps Pol II by tracking the RNA in the active site of Pol II without crosslinking or *in vitro* run-on<sup>350,387</sup>. mNET-seq in adherent human cells (HeLa S3)<sup>350,388</sup> was published as an updated version of the original NET-seq in *Saccharomyces cerevisiae*<sup>387</sup>. The mNET-seq protocol makes use of cellular fractionation<sup>350</sup>. The solubilized chromatin fraction is subjected to a Pol II-specific immunoprecipitation (IP)<sup>350,387</sup>.

Crosslinking is not required because the chromatin-associated Pol II-RNA complex is very stable and remains bound even in the presence of high salt, urea, polyanions and detergents<sup>389,390</sup>. Next, RNA fragments are selected by size (25 to 110 nt), purified and subjected to cDNA library preparation<sup>350</sup>. The strand-specific distribution of Pol II is characterized by sequencing the 5' and 3' ends of each RNA molecule. The 3' end sites of individual RNA molecules are then used to identify Pol II occupancy at single-nucleotide resolution<sup>350,387</sup> (see **Figure 6 b** for a detailed experimental workflow). With the transfer of the protocol from yeast to human cells, a bias of (m)NET-seq became visible: non nascent RNA intermediates with 3'-OH groups remain bound to Pol II after IP and lead to an overestimation of Pol II molecules in certain areas along the gene<sup>391</sup>. These non nascent RNAs arise due to stable protein-protein interactions of RNA processing complexes with Pol II e.g. spliceosome or microprocessor complex associated RNAs<sup>391,392</sup>. The addition of the strong detergent Empigen BB during the IP (0.1-2 % v/v) overcomes these limitations by weakening protein-protein interactions<sup>393,394</sup>. Thus, non nascent RNAs are lost and only nascent RNAs associated with the active center of Pol II remain<sup>391,392</sup>.

To date, mNET-seq (with Empigen BB treatment) is the superior method catching all post-initiation states of Pol II including backtracked and arrested polymerases (which is critical to estimate the total number of Pol II molecules on a gene, criterium ii) at single-nucleotide resolution and without crosslinking or *in vitro* NRO (for a comparison see<sup>351</sup>, **Table 27**, and **Figure 3 d**).

## 4. Motivation and aims of this thesis

At the outset of my graduate work the Ehrensberger theory<sup>315</sup> (chapter 2.2) served as an attractive explanation for how promoter-proximal pausing might regulate the amount of RNA synthesized per time. However, a fundamental component of this hypothesis remained untested *in vivo*: how does the productive initiation frequency change upon pause duration changes? We reasoned that addressing this central question could enable a mechanistic understanding of how genes are regulated in steady state cellular conditions as well as in response to natural stress, and might reveal which of the steps is rate limiting (**Figure 4 a**).

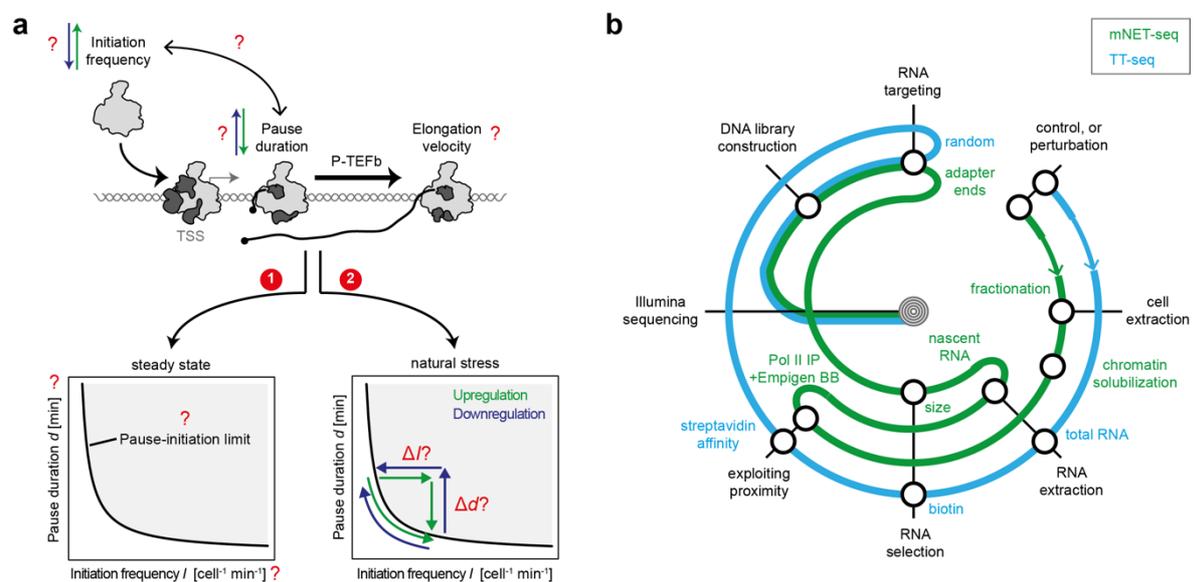
As pointed out above (chapter 3) current consensus methods to determine transcription rates have several shortcomings. Limitations are low resolution (>5 kbp) and unspecific inhibitors for elongation rate estimation, ambiguous occupancy read-out of Pol II for pause duration estimation, or the lack of a method to quantify rates *in vivo* to estimate the initiation frequency. The major obstacle to addressing transcription kinetics was the lack of biochemical techniques to measure newly synthesized nucleotides per time. The development of TT-seq<sup>168</sup> enabled us to solve this issue and to measure productive initiation frequencies *in vivo*. The second obstacle was the lack of a strand-specific Pol II occupancy profiling protocol in human suspension cells (**Figure 4 b**). Therefore, I set out to establish the recently reported method mNET-seq (with Empigen BB treatment) in our lab. A third development (CRISPR/Cas9 system) allowed us to engineer a human cell line for highly specific chemical inhibition of the P-TEFb kinase CDK9, the key factor that triggers promoter-proximal pause release of Pol II. By combining the specific pause-release kinase inhibition with our novel multi-omics approach,

we measured global transcription kinetics and their changes at human genes. In specific, the following questions were investigated:

- Question 1. Can we measure the kinetic parameters initiation frequency, pause duration, and elongation velocity globally?
- Question 2. How can Pol II promoter-proximal pausing control the frequency of productive initiation and by that the amount of RNA synthesized per time?
- Question 3. What are the determinants of long pause duration in human cells?

Furthermore, our multi-omics approach allows to refine our understanding of (non)coding transcription regulation. Thus, in a second study, we annotated (non)coding transcript classes in human hematopoietic cell lines. To this end, we combined the segmentation algorithm GenoSTAN<sup>177</sup> with GRO-cap<sup>50</sup> which recovers nascent RNAs with capped 5' ends of transcriptionally engaged polymerases. In addition, we used published cell line specific epigenomic datasets<sup>177</sup> to define coding and noncoding transcript classes including putative enhancer RNAs. We used perturbation experiments (CDK9 inhibition or heat shock) to stimulate cells in a controlled manner and measured their response with the aim to more fully expose the underlying transcriptional program of different transcript classes. We focused on the following questions:

- Question 4. Are there differences with respect to average Pol II pausing and/or productive initiation behavior of coding compared to noncoding transcript classes?
- Question 5. How are rapid and reversible changes in transcription upon stress achieved at a molecular level? Specifically, are changes refined by the pause-initiation limit?



**Figure 4. Multi-omics analysis of transcription kinetics in human cells.**

(a) Schematic of questions to be addressed in this work using a novel multi-omics approach to derive productive initiation frequency  $I$ , pause duration  $d$  and elongation velocity  $v$  genome-wide. Parameters and their relationship unknown at the outset of this work are highlighted with red question marks. (b) Structure of our multi-omics experiments. mNET-seq is depicted in green, TT-seq in blue. Individual steps of the protocol are highlighted by black circles. For a detailed experimental overview see also **Figure 6**. The subway map is inspired by Shendure and Liebermann Aiden<sup>395</sup>.

## II. Materials and Methods

### 1. Materials

This section lists chemicals, enzymes, kits and consumables used as part of this thesis. All materials, consumables and chemicals used for (RT-qPCR, TT-seq, RNA-seq, mNET-seq) were RNase-free, and molecular biology grade. If not stated otherwise, materials were obtained from Sigma-Aldrich <sup>[A]</sup>, Thermo Fisher Scientific <sup>[B]</sup>, Merck Millipore <sup>[C]</sup>, PanReac AppliChem <sup>[D]</sup>, Jena Bioscience <sup>[E]</sup>, Carl Roth <sup>[F]</sup>, bioworld <sup>[G]</sup>, Bio-Rad <sup>[H]</sup>, QIAgen <sup>[I]</sup>, New England Biolabs (NEB) <sup>[J]</sup>, Epicentre <sup>[K]</sup>, Eppendorf <sup>[L]</sup>, Miltenyi Biotec <sup>[M]</sup>, Agilent <sup>[N]</sup>, Illumina <sup>[O]</sup>, NuGEN <sup>[P]</sup>. Instruments and laboratory equipment are listed in part V (**Table 29**).

#### 1.1 Chemicals, enzymes, kits and other consumables

**Table 1. Chemicals and reagents used in this study in alphabetical order.**

Chemicals	Supplier	Identifier	Application(s)
1-NA-PP1	[C]	529605	CDK9 <sup>as</sup> inhibition
4-thio-uridine, 4-thio-UTP	[E]	NU-1156S	spike-ins
4-thiouracil, 4sU	[A]	T4509	TT-seq
Ammonium persulfate, APS	[D]	A00304	mNET-seq, spike-ins
Bis-acrylamide 19:1 (40 %)	[A]	A9926-100	mNET-seq, spike-ins
$\beta$ -mercaptoethanol	[F]	4227.1	Western blot
Bovine serum albumin, BSA (1 %) in PBS	[G]	40120706-1	mNET-seq
Bromophenol blue (tracking dye)	[A]	B0126	mNET-seq
Chloroform	[A]	372978	mNET-seq, TT-seq
DL-Dithiothreitol, DTT	[A]	43815	TT-seq
Dimethylsulfoxid, DMSO	[A]	D8418	mNET-seq, TT-seq
dNTPs mix (10 mM each)	[B]	R0191	RT-qPCR
Empigen BB (~30 %)	[A]	30326	mNET-seq
Ethanol	[C]	159010	mNET-seq, TT-seq
EDTA (0.5 M, pH 8.0)	[B]	15575020	mNET-seq, TT-seq
EGTA (0.5 M, pH 8.0)	[G]	40520008-1	mNET-seq
EZ-link HPDH Biotin	[B]	21341	TT-seq
Glycerol	[A]	G5516	mNET-seq
GlycoBlue co-precipitant	[B]	AM9516	mNET-seq

InstantBlue gel staining solution	Expedeon	ISB1L	Western blot
Isopropanol	[C]	109634	TT-seq
Laemmli Sample Buffer (4 x)	[H]	#1610747	Western blot
Magnesium chloride solution, MgCl <sub>2</sub> (1 M)	[A]	M1028	mNET-seq
Milk powder	[F]	T145.2	Western blot
MOPS buffer (20 x)	[B]	NP0001-02	Western blot
N,N-Dimethylformamide, DMF	[A]	D4551	TT-seq
Sodium chloride, NaCl (5 M)	[A]	S5150	TT-seq
Nonident™ P40 Solution (NP-40)	[A]	74385	mNET-seq
NuPAGE Transfer Buffer (20 x)	[B]	NP0006-1	Western blot
Phosphatase Inhibitor Cocktail Tablets	[A]	4906837001	mNET-seq
Ponceau S	[A]	P3504	Western blot
Protease inhibitor cocktail (100 x)	[A]	P8340	mNET-seq
QIAzol lysis reagent	[I]	79306	TT-seq
QuickExtract™ DNA Extraction Solution	[K]	QE09050	CDK9 <sup>as</sup> validation
Random Hexamer primers	[B]	48190-011	RT-qPCR
RNase-free AMPure XP beads	Beckman	A63987	TT-seq, spike-ins
Sodium acetate, NAOAc (3 M, pH 5.2)	[B]	R1181	mNET-seq, TT-seq
Sucrose	[A]	S0389	mNET-seq
SYBR Safe DNA Gel Stain	[B]	S33102	CDK9 <sup>as</sup>
SYBR™ Gold Nucleic Acid Gel Stain	[B]	S-11494	mNET-seq, spike-ins
Tetramethylethylenediamine, TEMED	[A]	A00115	mNET-seq, spike-ins
TRIS borate EDTA, TBE (10 x)	[A]	93290-1L	mNET-seq, spike-ins
Tris-HCl (1 M, pH 7.0)	[B]	AM9851	mNET-seq
Tris-HCl (1 M, pH 7.5)	[B]	15567-027	mNET-seq
Tris-HCl (1 M, pH 8.0)	[B]	15568-025	mNET-seq
Trypan blue solution (0.4 %)	[A]	T8154	Cell viability
Tween-20	[A]	P9416	TT-seq, Western blot
UltraPure Agarose	[B]	16500500	CDK9 <sup>as</sup>
Urea	[D]	A1049	mNET-seq, spike-ins
Water, H <sub>2</sub> O	[B]	10977049	all
Xylene cyanol (tracking dye)	[A]	X4126	mNET-seq

**Table 2. Enzymes used in this study in alphabetical order.**

Enzymes	Supplier(s)	Identifier	Application
AccuPrime GC-Rich DNA Polymerase	[B]	12337-016	CDK9 <sup>as</sup>
BstUI enzyme (1,000 units)	[J]	R0518S	CDK9 <sup>as</sup>
HL-dsDNase	ArcticZymes	70800-201	TT-seq
Maxima H Minus Reverse Transcriptase Mix	[B]	EP0742	RT-qPCR
RNase-free DNase Set	[I]	79254	TT-seq
SuperScript II Reverse Transcriptase	[B]	18080-044	mNET-seq
T4 Polynucleotide Kinase (3' phosphatase minus)	[J]	M0236S	mNET-seq
T4 RNA ligase, deletion mutant 2	[K]	LR2D1132K	mNET-seq

**Table 3. Kits used in this study in alphabetical order.**

Kits	Supplier	Identifier	Application
ERCC RNA Spike-in Mix	[B]	4456740	spike-ins
KAPA HIFI Library Amp Real Time	peqlab	07-KK2701-01	TT-seq
MEGAscript T7 kit	[B]	AM1334	spike-ins
miRNeasy Micro Kit	[I]	217084	TT-seq
Ovation Universal RNA-Seq System	[P]	0343-32	TT-seq
QIAquick MinElute Gel Extraction Kit	[I]	28704	mNET-seq
Qubit dsDNA HS Assay Kit	[B]	Q32854	mNET-seq
Qubit RNA HS Assay Kit	[B]	Q32855	TT-seq
RNA PCR Primer Indices A, D	[O]	RS-200-0012; RS-200-0048	mNET-seq
Strand Selection Reagent, SS5 ver 3 (human rRNA depletion)	[P]	S01859	TT-seq
Super Signal West Pico Plus Chemiluminescent Substrate	[B]	34580	Western blot
SYBR Select Master Mix	[B]	4472919	RT-qPCR
TapeStation D1000 Reagents	[N]	5067-5583	mNET-seq
TapeStation D1000 ScreenTape	[N]	5067-5582	mNET-seq
TapeStation HS RNA ScreenTape	[N]	5067-5579	TT-seq
TapeStation HS RNA ScreenTape Sample Buffer	[N]	5067-5580	TT-seq
TruSeq Small RNA Library Kit	[O]	RS-200-0012	mNET-seq
TURBO DNA-free Kit	[B]	AM1907	RT-qPCR
μMACS Streptavidin Kit	[M]	130-074-101	TT-seq

**Table 4. Consumables used in this study.**

Consumables	Supplier	Identifier	Application
96 well plate, round bottom	VWR	734-0955	RNA, cDNA
96 well plates, FrameStar® 96 non-skirted	4titude	4ti-0711	RT-qPCR
Cassette (1 mm)	[B]	NC2010	mNET-seq
E-Gel High-Resolution (4 %) agarose gels	[B]	G501804	mNET-seq
SurPhob SafeSeal LOW BINDING tips, polyethylen (PE), Biozym	Sarstedt	70.1186.210	RNA
MaXtract High Density tubes (1.5 mL)	[I]	129046	TT-seq
MaXtract High Density tubes (2 mL)	[I]	129056	TT-seq
microTUBE AFA Fiber Snap-Cap (130 µL)	Covaris Ltd	520045	TT-seq
Optical adhesive cover, Microseal 'B' seal	[H]	MSB1001	RT-qPCR
Polypropylene (PP) low bind filter tips 10 µL	Biozym	VT0200	TT-seq
Polypropylene (PP) low bind filter tips 20 µL	Biozym	VT0220	TT-seq
Polypropylene (PP) low bind filter tips 30 µL	Biozym	VT0250	TT-seq
Polypropylene (PP) low bind filter tips 1250 µL	Biozym	VT0270	TT-seq
Polypropylene (PP) lowbind tubes (1.5 mL)	[L]	022431021	TT-seq
Polypropylene (PP) lowbind tubes (2 mL)	[L]	022431048	TT-seq
Precast polyacrylamide gel, NuPAGE (4-12 %, 15-well)	[B]	NP0323BOX	Western blot
PVDF Transfer Membrane (0.45 µm)	[B]	88518	Western blot
Qubit Assay Tubes	[B]	Q32856	TT-seq, mNET-seq
Scalpels no. 23, sterile disposable	Swann-Morton	0510	mNET-seq
Spin-X column with cellulose acetate membrane	Costar	8160	mNET-seq
µ columns	[M]	130-042-701	TT-seq
µMACS Streptavidin kit	[M]	130-074-101	TT-seq

**Table 5. Protein, DNA and RNA ladder used in this study.**

Ladder	Range	Supplier	Identifier	Application
TrackIt 25 bp DNA	10-1,200 bp	[B]	10488022	mNET-seq
HRL	100-500 bp	[O]	-	mNET-seq
CRL	145, 160, 500 bp	[O]	-	mNET-seq
Precision Plus Protein™ All Blue Prestained Protein standard	10-250 kDa	[H]	1610373	Western blot
Small RNA marker	20-100 nt	Abnova	R0007	mNET-seq
High Sensitivity RNA ScreenTape ladder	25-6,000 nt	[N]	5067-5581	TapeStation
D1000 Ladder	35-1,000 bp	[N]	5067-5586	TapeStation

## 1.2 Plasmids and oligonucleotides

**Table 6. Plasmids used as IVT template of RNA spike-ins in this study.**

The cloning vector for all spike-in inserts was pUC19<sup>396</sup>. See Appendix for plasmid maps and subsection 2.2.1 and **Table 20** for experimental details. Cloning of the selected spike-in sequence into a pUC19 cloning vector and verification was done by Dr. Anna Sawicka and Dr. Kristina Žumer (MPI-bpc, Dept. of Molecular Biology).

Name	Insert IVT template	Backbone	Source
2-4sU-pUC19	Spike-in 2 (derived from ERCC-00043)	pUC 19 (e.g. [J], N3041)	Cramer lab, this study
12-pUC19	Spike-in 12 (derived from ERCC-00170)	pUC 19 (e.g. [J], N3041)	Cramer lab, this study
4-4sU-pUC19	Spike-in 4 (derived from ERCC-00136)	pUC 19 (e.g. [J], N3041)	Cramer lab, this study
5--pUC19	Spike-in 5 (derived from ERCC-00145)	pUC 19 (e.g. [J], N3041)	Cramer lab, this study
8-4sU-pUC19	Spike-in 8 (derived from ERCC-00092)	pUC 19 (e.g. [J], N3041)	Cramer lab, this study
9-pUC19	Spike-in 9 (derived from ERCC-00002)	pUC 19 (e.g. [J], N3041)	Cramer lab, this study

**Table 7. Plasmid used for CRISPR/Cas9 engineering in this study.**

This plasmid was a gift from Feng Zhang (Addgene plasmid #48138, RRID:Addgene\_48138)<sup>397</sup>. See Appendix for plasmid map, **Table 10** for sgRNA primers and subsection 2.3.1 for experimental details. Cloning of the 20 nt guide RNA sequence for the canonical CDK9 into the Addgene plasmid and verification was done by Dr. Weihua Qin (LMU Munich, Dept. of Biology II, group of Prof. Dr. Heinrich Leonhardt).

Name	Insert	Source
pSpCas9(BB)-2A-GFP (PX458)	20nt guide RNA sequence targeting CDK9	Dr. Weihua Qin, this study <sup>381</sup>

## 1.3 Primers

Primers were synthesized by Integrated DNA Technologies (IDT) or Biomers.

**Table 8. RT-qPCR primers used in this study.**

See Appendix for synthetic spike-in sequences. Primer sequences for Spike-in 4-sU and Spike-in 12 were designed by Dr. Margaux Michel (MPI-bpc, Dept. of Molecular Biology).

Target name	Direction	Primer sequences 5'>3'	Length (nt)	Tm (°C)	Amplicon size (bp)	Genomic Sequence	Amplified Region
Spike-in 4-4sU	forward	CCGAGTTCGCCT TACTGCTC	20	60	95	Synthetic, ERCC-00136	438 ... 532
	reverse	AATCGATCGGAA TCACGCCG	20	60			
Spike-in 5	forward	CATAAGCGGAGA AAGAGGGAATGA C	25	59	103	Synthetic, ERCC-00145	15 ... 117
	reverse	GCTAAATAGAGA GCATCCACACCT C	25	58			

Spike-in 12	forward	AGACTGGCATTC CCGTGATA	20	57	97	Synthetic, ERCC-00170	222 ... 318
	reverse	GCTAAAACCCCT GCCTGCAA	20	60			
ACTB	forward	ACTCTTCCAGCC TTCCTTCC	20	62	102	Gene ID: 60, NC_000007.14 Homo sapiens chromosome 7 Reference GRCh38.p12	2,299 ... 2,495
	reverse	TACAGGTCTTTG CGGATGTC	20	60			
HSPA1A	forward	GATCTTACCAC CTACTCCGACA	23	59	86	Gene ID: 3303, NC_000006.12 Homo sapiens chromosome 6 Reference GRCh38.p12	31,817,03 4 ... 31,817,11 9
	reverse	GATTGTTGTCTT TCGTCATGGCCC	24	60			

**Table 9. CDK9<sup>as</sup> screening: PCR and sequencing primers.**

Sequences which do not map to the target locus are underlined. CDK9<sup>as</sup> screen (WQ) was designed by Dr. Weihua Qin (LMU Munich, Dept. of Biology II, group of Prof. Dr. Heinrich Leonhardt).

Target name	Direction	Primer sequences 5'>3'	Length (nt)	Tm (°C)	Amplicon size (bp)	Genomic Sequence	Amplified Region
CDK9 <sup>as</sup> screen (WQ)	forward	<u>CCCCG</u> TAGCT GGTGCTTCCTC G	22	61	500	Gene ID: 1025, NC_000009 Homo sapiens chromosome 9, GRCh38.p12	127,787,74 8 ... 127,788,24 0
	reverse	<u>CCCCA</u> GCAGCC TTCATGTCCCT AT	24	59			
CDK9 <sup>as</sup> screen (SG-1)	forward	CCTCTTCTTAA CTCAGATGGA CCC	24	58	412	Gene ID: 1025, NC_000009 Homo sapiens chromosome 9, GRCh38.p12	127,787,69 3 ... 127,788,10 4
	reverse	GTGGATGTAG TAGAGGCCGT TAAG	24	59			
CDK9 <sup>as</sup> screen (SG-2)	forward	GTGAAGGAAG GAACAGACAG ATGC	24	59	388	Gene ID: 1025, NC_000009 Homo sapiens chromosome 9, GRCh38.p12	127,787,81 2 ... 127,788,19 9
	reverse	AGAGTGGGTC CTTTAGTGACA TCC	24	60			

**Table 10. Other oligonucleotides used in this study.**

The listed oligonucleotides were designed Dr. Weihua Qin (LMU Munich, Dept. of Biology II, group of Prof. Dr. Heinrich Leonhardt). See **Table 7** and subsection 2.3.1 for experimental details.

Name	Sequence 5'>3'
HDR template: A103 is underlined, BstUI cutting site in small letters	AAAGTGTGTTGGGTGTGGTTTTCTTGACTTTTTCTTCTTTCTAT TCCTGCCTCAGCTTCCCCCTATAACCGCTGCAAGGGTAGTATAT ACCTGGT <sub>cgcg</sub> GACTTCTGCGAGCATGACCTTGCTGGGCTGTTGA GCAATGTTTTGGTCAAGTTCACGCTGTCTGAGATCAAGAGGGT GATGCAGATGCTGCTTAACGGCCT
Primers for sgRNA generation and screening: CDK9-sgRNA-F	<u>CACCGCTCGCAGAAGTCGAACACC</u>
Primers for sgRNA generation and screening: CDK9-sgRNA	<u>AAACGGTGTTCGACTTCTGCGAGCC</u>

## 1.4 Antibodies for Western blot and IP

**Table 11. Primary antibodies used in this study.**

If available, Research Resource Identifiers (RRIDs) are listed. The serum of 3E10 and 3E8 was a kind gift of Prof. Dr. Dirk Eick (Helmholtz Center Munich, Dept. of Molecular Epigenetics).

Target	Host	Dilution	Reference	Lot. #
GAPDH (C 71.1)	mouse	1:20,000	[A], G8795; RRID: AB_1078991	067M4785V
Histone H2B (C.362.2)	mouse	1:2,000	[B], MA5-14835; RRID: AB_10982286	RI2272615
Histone H3, HRP-coupled	rabbit	1:5,000	Abcam, ab21054; RRID: AB_880437	GR261138-1
POLR2A (hRPB1), C-terminus unphosphorylated + phosphorylated (CMA601)	mouse	3 µg/10 <sup>7</sup> cells (IP)	Biozol, MABI0601; RRID: AB_2728735	15013
POLR2A (hRPB1), C-terminus Ser2P (3E10)	rat	1:60	Gift of Dirk Eick	Serum
POLR2A (hRPB1), C-terminus Ser2P (CMA602)	mouse	5 µg/10 <sup>7</sup> cells (IP)	Biozol, MABI0602; RRID: AB_2747403	15013
POLR2A (hRPB1), C-terminus Ser5P (CMA602)	mouse	5 µg/10 <sup>7</sup> cells (IP)	Biozol, MABI0603; RRID: AB_2728736	16003
POLR2A (hRPB1), C-terminus Ser5P (3E8)	rat	1:60	Gift of Dirk Eick	Serum
POLR2A (hRPB1), hRPB1, N-terminus (F-12)	mouse	1:100	Santa Cruz Biotechnology, sc-55492; RRID: AB_630203	E2913
POLR2C (hRPB3)	rabbit	1:2,000	Bethyl Laboratories, A303-771A; RRID: AB_11218388	A303-771A-1
U1 snRNP70, C-terminus (C-3)	mouse	1:200	Santa Cruz Biotechnology, sc-390899	D0114
mouse IgG (mock control)	mouse	3 µg/10 <sup>7</sup> cells (IP)	Abcam, ab37355; RRID: AB_2665484	NA

**Table 12. HRP-coupled secondary antibodies used for Western blotting in this study.**

Target	Host	Dilution	Reference
Rabbit IgG	donkey	WB: 1:5,000	GE Healthcare, NA934
Mouse IgG	goat	WB: 1:3,000	Abcam, ab5870
Rat IgG	goat	WB: 1:5,000	[A], A9037

## 1.5 Cell culture: cell lines, medium and consumables

**Table 13. Human cell lines used in this study.**

If available, Research Resource Identifiers (RRIDs) are listed.

Species	Cell line	Reference	Identifiers	Additional information
Homo sapiens; male	Raji B lymphocyte	DSMZ	Cat# ACC-319; RRID:CVCL_0511	Epstein-Barr virus-related Burkitt lymphoma
Homo sapiens; male	Raji B (CDK9 <sup>as</sup> ) lymphocyte	This study <sup>381</sup>	NA	Homozygous mutation of F103A at the CDK9 gene
Homo sapiens; female	K562 erythro- leukemia	DSMZ	Cat# ACC-10; RRID:CVCL_0004	Chronic myelogenous leukemia, BCR-ABL1 positive

**Table 14. Growth and freeze medium composition used in this study.**

Cell lines	Medium	Components	Supplier, identifier	Final concentration	Used for
Raji B, Raji B (CDK9 <sup>as</sup> )	Growth	RPMI-1640	[B], 31870-074	90 % v/v	This study <sup>381</sup>
		Heat inactivated FBS	bio-sell	10 % v/v	
		L-glutamine	[B]	200 $\mu$ M	
		(Pen/Strep	[B]	1 x	
Raji B (CDK9 <sup>as</sup> ), K562	Growth	RPMI-1640	[B], 31870-074	90 % v/v	This study
		Heat inactivated FBS	[B], 10500-064	10 % v/v	
		GlutaMAX	[B], 35050061	1 x	
Raji B, Raji B (CDK9 <sup>as</sup> ), K562	Freeze	Heat inactivated FBS	[B], 10500-064	90 % v/v	This study
		DMSO (cryoprotectant)	[A], 2438	10 % v/v	
Raji B, Raji B (CDK9 <sup>as</sup> ), K562	Wash	DPBS, no calcium, no magnesium	[A]	NA	This study

**Table 15. Cell culture consumables used in this study.**

<b>Instrument</b>	<b>Supplier</b>	<b>Identifier</b>
Fungizide	WAK-Chemie	WAK-ZF-1
Acryl Aqua Clean, Microbiocidal Additive for Water Soluble Heating Bath Fluids	WAK-Chemie	WAK-AQA-250-50l
Plasmo Test Mycoplasma Detection Kit	InvivoGen	rep-pt1
Serological pipettes (50 mL, 25 mL, 10 mL, 5 mL, 1 mL)	Sarstedt	Ref 86.1254.001
Tissue culture plates 6-well, 12-well, 24-well, 96-well	Sarstedt	83.1836
Trypan blue solution (0.4 %)	[A]	T8154
Cell culture T25/T75/T175/T225 flask for suspension cells (standard TC surface treatment)	Greiner Bio-One	NA
Sterile cryovial (1.2 mL)	Kisker	CV11-1
Sterile cryovial (2 mL)	Kisker	CV11-2
Tubes 50 mL, 15 mL, 2 mL, 1.5 mL	Greiner Bio-One	NA

## 1.6 Buffers

**Table 16. Western blot: buffers used in this study.**

<b>Buffer</b>	<b>Component</b>	<b>Final concentration</b>	<b>Step</b>
4 x LB	Laemmli Sample Buffer (4 x)	4 x	Western blot
	$\beta$ -mercaptoethanol	10 % v/v	
Ponceau S staining	Ponceau S	0.1 % w/v	Western blot
	Acetic acid	5 % v/v	
10 x PBS stock	NaCl	1.37 M	Western blot
	KCl	27 mM	
	Na <sub>2</sub> HPO <sub>4</sub>	100 mM	
	KH <sub>2</sub> PO <sub>4</sub> pH 7.4	18 mM	
MOPS running buffer	MOPS buffer (20 x)	1 x	Western blot
Transfer buffer	NuPAGE Transfer Buffer (20x)	1 x	Western blot
	Ethanol	10 % v/v	
PBS-T	PBS	1 x	Western blot
	Tween20	0.05 % v/v	
Blocking solution	Milk powder	5 % w/v	Western blot
	PBS	1 x	
	Tween20	0.05 % v/v	
Binding solution	Milk powder	2 % w/v	Western blot
	PBS	1 x	
	Tween20	0.05 % v/v	

**Table 17. TT-seq: solutions and buffers used in this study.**

Solution / buffer	Component	Final concentration	Stored at	Step
50 mM 4sU	4-thiouridine	50 mM	-20 °C	Labeling
	RNase-free H <sub>2</sub> O	-		
EZ-Link Biotin-HPDP	EZ-Link Biotin-HPDP	1 mg/ml	4 °C, -20 °C	Biotinylation
	DMF	-		
10 x biotinylation buffer	Tris HCl pH 7.4	100 mM	4 °C	Biotinylation
	EDTA pH 8.0	10 mM		
µMACS washing buffer (RT)	Tris HCl pH 7.4	100 mM	Prepare freshly, degas for >10min, RT	Pull-down of labeled RNAs
	EDTA pH 8.0	10 mM		
	NaCl	1 M		
	Tween20	0.1 % v/v		
µMACS washing buffer (65 °C)	Tris HCl pH 7.4	100 mM	Prepare freshly, degas for >10min, 65 °C	Pull-down of labeled RNAs
	EDTA pH 8.0	10 mM		
	NaCl	1 M		
	Tween20	0.1 % v/v		

**Table 18. mNET-seq: solutions and buffers used in this study.**

Solution / buffer	Component	Final concentration	Stored at	Step
10 x PhosSTOP	PhosSTOP tablet	10 x	-20 °C	Buffers
1 x MNase buffer (1 mL/10 samples)	10 x MNase reaction buffer	1 x	Prepare freshly	Chromatin wash
	100 x BSA	1 x		
Chromatin digest buffer (1 mL/10 samples)	10 x MNase reaction buffer	1 x	Prepare freshly	Chromatin digest
	100 x BSA	1 x		
	MNase	50 U		
PNKT buffer (460 µL/replicate)	T4 PNK buffer	1 x	-20 °C	RNA phosphorylation
	Tween	0.1 % v/v		
Cell lysis buffer (Ørom) (4 mL/10 samples)	Tris-HCl pH 7.4	10 mM	Prepare freshly	Fractionation
	NaCl	150 mM		
	NP-40	0.15 % v/v		
	PhosSTOP	1 x		
	Protease Inhibitor	1 x		
Sucrose buffer (Ørom) (10 mL/10 samples)	Tris-HCl pH 7.4	10 mM	Prepare freshly	Fractionation
	NaCl	150 mM		
	Sucrose	24 % w/v		
	PhosSTOP	1 x		
	Protease Inhibitor	1 x		
Glycerol buffer (Ørom) (2.5 mL/10 samples)	Tris-HCl pH 7.4	20 mM	Prepare freshly	Fractionation
	NaCl	75 mM		
	EDTA	0.5 mM		
	Glycerol	50 % v/v		

	PhosSTOP	1 x		
	Protease Inhibitor	1 x		
Nuclear lysis buffer (Ørom) (2.5 mL/10 samples)	Tris-HCl pH 7.4	10 mM	Prepare freshly	Fractionation
	NaCl	300 mM		
	EDTA	0.2 mM		
	MgCl <sub>2</sub>	7.5 mM		
	NP-40	1 % v/v		
	Urea	1 M		
	PhosSTOP	1 x		
	Protease Inhibitor	1 x		
Dynabeads wash (DW) buffer	DPBS, no Ca <sup>2+</sup> , no Mg <sup>2+</sup>		Prepare freshly	IP
	BSA	0.1 % v/v		
	EDTA	2 mM		
IP buffer (15 mL/replicate)	Tris-HCl pH 7.4	50 mM	Prepare freshly	IP
	NaCl	150 mM		
	NP-40	0.05 % v/v		
	Empigen BB (~30 %)	1 % v/v		
	PhosSTOP	1 x		
	Protease Inhibitor	1 x		
7 M urea LB (10 µL/replicate)	Urea	7 M	RT	RNA size selection
7 M urea LB with tracking dyes	Urea	7 M	RT	RNA size selection
	Bromophenol blue	0.05 % w/v		
	Xylene cyanol	0.05 % w/v		
Elution buffer (1 mL/replicate)	Sodium acetate pH 5.2	1 M	Prepare freshly	RNA size selection
	EDTA	1 mM		
6 % PAGE (10 mL)	Urea	7 M	Prepare freshly	RNA size selection
	40 % bis-acrylamide (19:1)	6 % v/v		
	10 x TRIS borate - EDTA (TBE) buffer solution	1 x		
	TEMED	10 µL		
	10 % w/v ammonium persulfate (APS)	50 µL		

## 1.7 Software and databases

**Table 19. Overview of software and databases used in this thesis.**

Software	Developer	Version	Ref
Mac OS High Sierra	Apple, USA	10.13.6	
Adobe Illustrator CS6	Adobe Systems Inc., USA	16.0.0	
Microsoft Word and Excel for Mac	Microsoft Corp., USA	16.16.8	
Fiji / Image J	Open source	2.0.0	<sup>398</sup>
qPCRsoft	Analytik Jena, Germany	3.4.6.0	
ChemoStar Imager	Intas	v.0.3.39	
NanoDrop Software	Peqlab		
4200 TapeStation Analysis Software	Agilent Technologies Inc.	A.02.01	
PROSize Data Analysis Software (Fragment Analyzer)	Agilent Technologies Inc.	3.0	
SonoLab 7 (Covaris S220)	Covaris, USA	7.2.1.0	
Primer3	HHMI, open source software for high-throughput primer design, USA	v.0.4.0	<sup>399</sup>
NCBI	U.S. National Library of Medicine, USA	NA	
Integrative Genomics Viewer, IGV desktop application	Broad Institute and the Regents of the University of California	2.4.10	<sup>400</sup>
Endnote X8	Clarivate Analytics	X8.2	
Bcl2fastq	Illumina	2.20.0.422	
fastQC	Babraham Institute Bioinformatics	0.11.5	<sup>401</sup>
Cutadapt	Marcel Martin	1.9.1	<sup>402</sup>
STAR	Open source	2.3.0	<sup>403</sup>
Samtools	John Marshall and Petr Danecek	1.3.1	<sup>404</sup>
HTSeq	EMBL Heidelberg, Genome Biology Unit	0.6.1	<sup>405</sup>
R/Bioconductor	<a href="https://www.r-project.org/">https://www.r-project.org/</a>	3.1.2	<sup>406</sup>
Cyberduck	David Kocher, Yves Langisch	6.6.2	
Clustal Omega	EMBL-EBI, Wellcome Genome Campus	1.2.4	<sup>407</sup>
Optimized CRISPR design	Web tool, Feng Zhang <a href="http://crispr.mit.edu/">http://crispr.mit.edu/</a>	NA	
SnapGene	GSL Biotech LLC	3.3.4	
KInhibition, kinase inhibitor selection tool	Fred Hutch online portal to search publicly available datasets	NA	<sup>408</sup>

---

## 2. Methods

This section lists methods used in this thesis. If applicable, a detailed list of contributions can be found at the beginning of each subsection or paragraph (*in italic*).

### 2.1 General methods

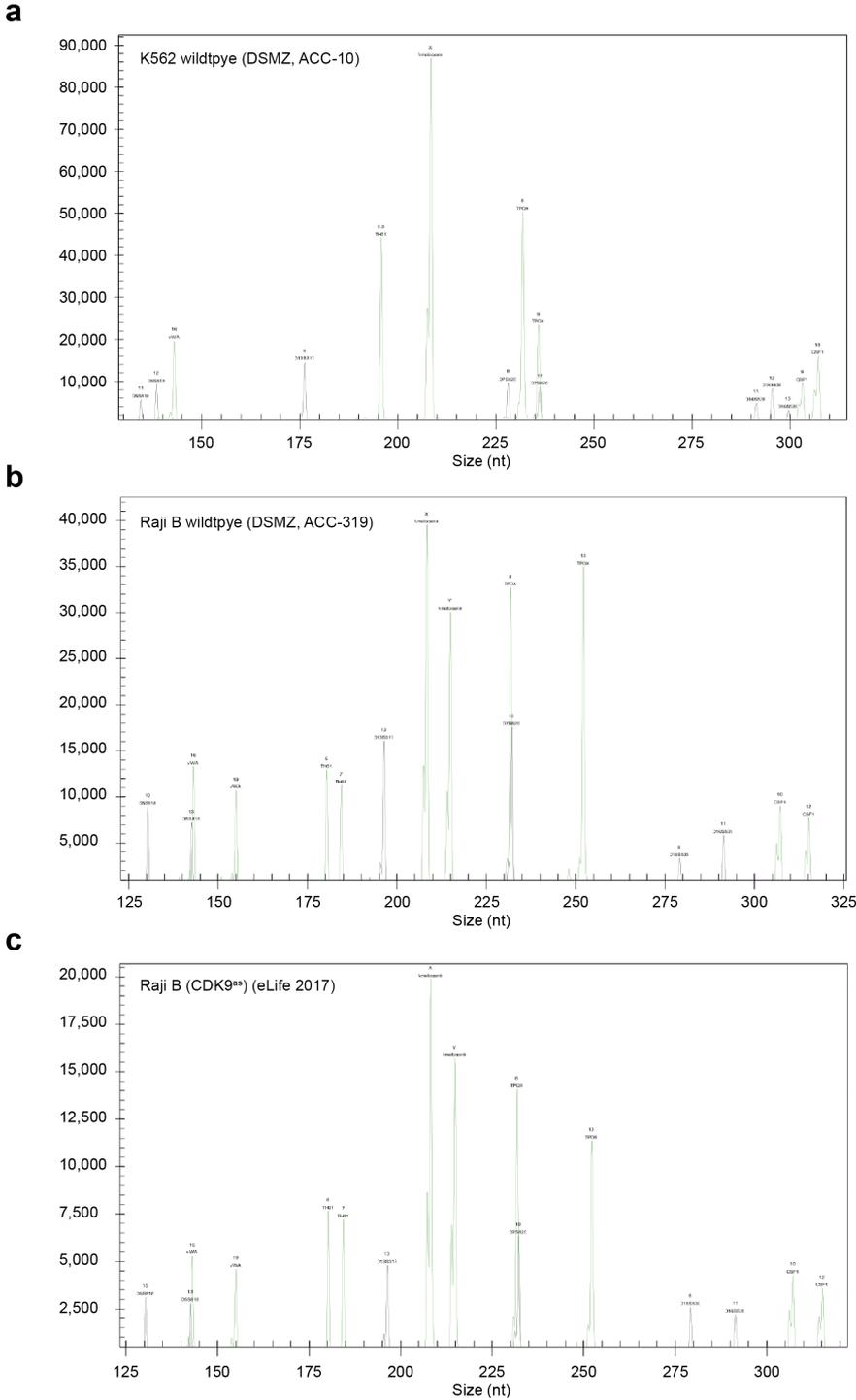
#### 2.1.1 Maintenance of human hematopoietic cell lines

**Growth conditions.** I cultured K562 and Raji B (wildtype or CDK9<sup>as</sup>) cells (**Table 13**) in accordance with the DSMZ Cell Culture standards in RPMI 1640 medium containing 10 % heat inactivated fetal bovine serum (FBS), and L-glutamine or 1 x GlutaMAX supplement at 37 °C in a humidified 5 % CO<sub>2</sub> incubator (**Table 14**). For short term storage, cells were stored at -80 °C. For long term storage, cells were stored in liquid nitrogen. The freezing medium is described in **Table 14**. For each working stock, a growth curve was generated. Cell morphology was frequently controlled by light microscopy. All three cell lines used in this study displayed the phenotypic properties, including morphology and proliferation rate, that have been described in literature.

**Authentication.** *K562, Raji B wildtype and Raji B (CDK9<sup>as</sup>) cell lines were verified by Dr. Wilhelm Dirks (Leibniz Institute DSMZ, Molecular Biology group).* DNA profiling was carried out using 18 different and highly polymorphic short tandem repeat (STR) loci. All three cell lines were derived of pure human cell cultures and had full-matching STR profiles, as indicated by a search of databases of cell banks ATCC (USA), JCRB (Japan), RIKEN (Japan), KCLB (Korea), and DSMZ (Germany) (international STR reference database)<sup>409</sup>. The respective STR electropherograms are shown in **Figure 5**.

**Mycoplasma testing.** *Tests were performed in-house by Dr. Anna Sawicka (MPI-bpc, Dept. of Molecular Biology).* K562, Raji B wildtype and Raji B (CDK9<sup>as</sup>) cells were verified to be free of mycoplasma contamination using Plasco Test Mycoplasma Detection Kit according to manufacturer's instructions. The kit provides a cellular assay for colorimetric detection using HEK-Blue<sup>TM</sup>-2 cells in detection medium as sensor.

**Cell counting and biological replicates.** I counted cells using a hemocytometer (Neubauer counting chamber) in 3 technical replicates, or by a Scepter<sup>TM</sup> 2.0 Cell Counter (Merck KGaA) according to manufacturer's instructions. Biological replicates were cultured independently.



**Figure 5. STR electropherograms of human hematopoietic cell lines used in this study.** Shown are the STR profiles of human K562 cells (a), Raji B wildtype (b), and Raji B (CDK9<sup>as</sup>) (c) cells. Authentication of cell lines was performed by Dr. Wilhelm Dirks (Leibniz Institute DSMZ, Molecular Biology group). See main text for additional information.

### 2.1.2 Quantitative reverse transcription PCR of RNA in a two-step assay

*I established a qPCR assay to screen for optimal inhibitor concentrations and determine the best time points for subsequent TT-seq, RNA-seq and mNET-seq analysis. This approach makes use of RNA spike-ins for normalization<sup>410</sup> and targets introns of highly expressed and rapidly degraded transcripts (the selection is based on half-lives measured in K562 cells<sup>168</sup>), or treatment-specific up- or downregulated transcripts (based on literature).*

**Total RNA isolation and DNase treatment.** I harvested  $3-4 \times 10^6$  cells at  $3,000 \times g$  for 2 min for each time point and biological replicate ( $n > 2$ ). Total RNA was isolated with QIAzol according to manufacturer's instructions except for the addition of 10 ng RNA spike-in mix<sup>168</sup> together with QIAzol. To remove possible genomic DNA contamination, isolated RNA (10  $\mu$ g) was treated with TURBO DNase according to manufacturer's instructions.

**Reverse transcription (RT).** For RT, I used random hexamer priming (5'-d(NNNNNN)-3', N = G, A, T, or C) according to manufacturer's instructions. Briefly, 1  $\mu$ g of DNase-treated RNA, Random Hexamer primers (final concentration of 5 ng/ $\mu$ L), dNTPs mix (final concentration of 0.5 mM) were mixed and incubated at 65 °C for 5 min. Subsequently, Maxima H Minus Reverse Transcriptase (final concentration of 200 U) and 5 x Maxima RT buffer were added (+RT reaction). For DNA contamination control, cDNA synthesis was performed without enzyme (-RT reaction; substituted with water). The (-/+ ) RT reactions were incubated in a PCR cycler at 25 °C for 10 min, 50 °C for 30 min, and 85 °C for 5 min.

**Primer design.** I designed primers for quantitative PCR (qPCR) by using the online primer design software Primer3 v.0.4.0<sup>399</sup>. Criteria that differ to default settings were the following: product size ranges '80-150 100-200'; primer size optimum '25'; primer Tm minimum '60', optimum '64', maximum '70'; maximum Tm difference '2', table of thermodynamic parameters 'SantaLucia 1998'; product Tm optimum '50'; primer GC minimum '35', optimum '65', maximum '80'; salt correction formula 'SantaLucia 1998'; concentration of divalent cations '3.5 mM'; concentration of dNTPs '0.2 mM'. Primer specificity (single product peak) was validated by melting profiles. Primer sequences, length, annealing temperature, amplicon length and position on target are reported in **Table 8**.

**Quantitative PCR (qPCR).** cDNAs (50 ng) were amplified with SYBR® Select Master Mix according to manufacturer's instruction with a final primer concentration of 400 nM. PCR reactions were run in 96-well optical plates sealed with optical adhesive cover on a qTOWER 2.0/2.2 instrument. The following thermal cycling conditions were used (SYBR Select Master Mix reference, standard cycling mode): 50 °C for 2 min, 95 °C for 2 min, 40 cycles of 95 °C for 15 sec and 60 °C for 1 min. Three synthetic RNA spike-ins were used for normalization. The  $2(-\Delta\Delta Ct)$  method was applied to calculate the normalized target gene expression fold change, with the amplification efficiency (E) for each target gene, slope of standard curve (S) and mean threshold cycle (Ct)<sup>411</sup>.

### 2.1.3 Western blotting

*I performed Western blotting to determine the quality of cellular fractionation by enrichment of marker proteins in cytoplasm (GAPDH), nucleoplasm (U1 snRNP0), and chromatin (H3/H2B, phosphorylated POLR2A).*

The protocol consists of the following steps: sample preparation, protein separation, protein electrotransfer to a membrane, detection of marker proteins including blocking, probing with primary antibody and subsequently, with secondary horseradish peroxidase (HRP) conjugated antibody (subsection 1.4), chemiluminescence detection, and band quantification. Buffer compositions are listed in **Table 16**.

Protein samples were denatured in 1 x Laemmli buffer at 70 °C for 10 min. Samples were stored at -20 °C, or directly applied to polyacrylamide gel electrophoresis (PAGE). NuPAGE™ 4-12 % Bis-Tris Protein Gels and MOPS buffer<sup>412</sup> were used according to manufacturer's instructions. Briefly, proteins were separated at 200 V for 50 min. A PVDF membrane was cut to the size of the gel, activated in 100 % ethanol for 30 sec, and subsequently equilibrated in transfer buffer. After 50 min, NuPAGE was removed from the cassette, placed onto a filter paper and assembled in a XCell II Blot Module (semi-wet transfer unit) according to manufacturer's instructions. Transfer of a single NuPAGE onto PVDF membrane was performed in transfer buffer at 30 V for 1 h.

Membrane was stained with Ponceau S staining solution to visualize successful protein transfer<sup>413</sup>. Membrane blocking was performed in 5 % milk PBS-T on a rocking surface for at least 1 h. Primary antibody was added overnight. A list of primary antibodies used in this study can be found in **Table 11**. Three washes of the membrane were performed using PBS-T, each with incubation on a rocking surface for 5 min. Secondary antibody was added on a rocking surface for 1 h. A list of HRP-coupled secondary antibodies used in this study can be found in **Table 12**. Three washes were performed using PBS-T, each with incubation on a rocking surface for 5 min.

ECL working solution of SuperSignal West Pico PLUS Chemiluminescent Substrate was prepared by mixing equal parts of the Stable Peroxide Solution and the Luminol/Enhancer Solution. After the last PBS-T wash, the membrane was rinsed several times with ECL working solution. Marker proteins were visualized by chemiluminescence detection on INTAS. Band quantification was performed by Fiji / Image J.

### 2.1.4 Figures

All figures in this work were generated in Adobe Illustrator. Genome browser views were generated in R/Bioconductor or using the Integrative Genomics Viewer. Metagene plots were generated in R/Bioconductor. Bar plots were generated in Excel.

All plots were subsequently modified for visual purposes in Adobe Illustrator. The respective programs are listed in **Table 19**.

## 2.2 Multi-omics

### 2.2.1 *In vitro* transcription (IVT) of RNA spike-ins

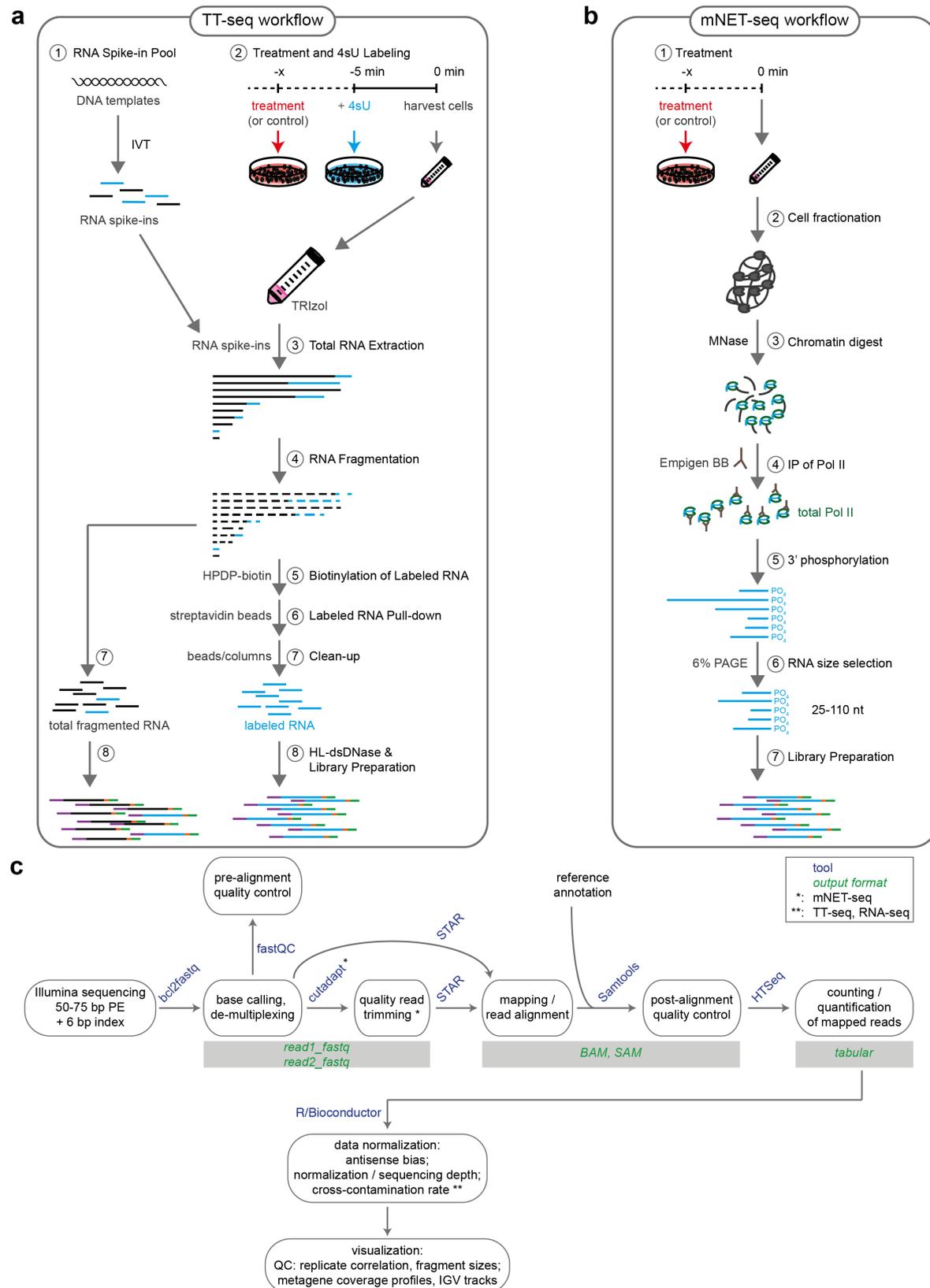
Six synthetic RNA spike-in controls (**Table 20**) were selected and established by Dr. Katja Lidschreiber (MPI-bpc, Dept. of Molecular Biology) as previously published<sup>168</sup>. Cloning of the selected spike-in sequence into a pUC19 cloning vector and verification was done by Dr. Anna Sawicka and Dr. Kristina Žumer (MPI-bpc, Dept. of Molecular Biology).

**Table 20. Characteristics of 4sU-labeled and unlabeled RNA spike-ins.**

Full-length sequences of the respective spike-in and plasmid maps are listed in the Appendix. Molecular mass (M) is calculated as described in section 2.2.4, and rounded to first decimal place. For 4sU-labeled spike-ins, 10 % of U is substituted with 4sU ( $\tau = 0.1$ ).

Name	Derived from ERCC ID	Length $L_i$ [nt]	Nucleotide composition				%GC	$\tau$	M [g/mol]
			A	U	C	G			
2-4sU	ERCC-00043	985	354	297	127	207	34	0.1	318,331.0
12	ERCC-00170	949	323	296	114	216	35	0	306,481.8
4-4sU	ERCC-00136	1,014	314	268	189	243	43	0.1	327,586.2
5	ERCC-00145	1,015	286	264	207	258	46	0	327,385.0
8-4sU	ERCC-00092	1,079	233	287	265	294	52	0.1	347,569.7
9	ERCC-00002	1,037	226	263	272	276	53	0	333,378.4

For IVT template generation, 3  $\mu$ g of plasmid (**Table 6**, **Table 20**) was linearized using EcoRV-HF (blunt end cut) digestion mix containing 1 x CutSmart buffer and 1  $\mu$ L of EcoRV-HF enzyme (final volume 50  $\mu$ L). I incubated the digestion mix at 37 °C for 1 h and terminated the reaction by adding 1/20 volume of 0.5 M EDTA. Subsequently, DNA was precipitated in 1/10 volume of 3 M sodium acetate pH 5.2, and 2 volumes of 100 % ethanol at -20 °C for 15 min. DNA was collected by centrifugation at 4 °C and 16,000 x g for 15 min. The pellet was washed twice using 75 % ethanol. DNA was air-dried and resuspended in 5  $\mu$ L of H<sub>2</sub>O at a concentration of 0.1-1.0  $\mu$ g/ $\mu$ L (quantified by NanoDrop). *In vitro* transcription (IVT) was performed using the MEGAscript T7 kit following manufacturer’s instruction. For IVT of 4sU-labeled spike-ins, 10 % of UTP was substituted with 4-thio-UTP. This is to ensure at least similar 4sU incorporation rates in the IVT as has been observed in human cell lines<sup>414-416</sup>. IVT reactions were incubated at 37 °C. After 4 h, reaction volume was filled up with H<sub>2</sub>O to 40  $\mu$ L, then 2  $\mu$ L of TURBO DNase was added and incubated at 37 °C for additional 15 min. RNA spike-ins were purified with RNAClean XP beads following manufacturer’s instructions. RNA was quantified using NanoDrop and 5 % denaturing PAGE (see **Figure 37** for representative PAGE). The final RNA spike-in pool contained equal amounts (1 ng/ $\mu$ L per spike-in) of all six RNA spike-ins and was stored in aliquots at -80 °C. The same batch of spike-in pool was used for all of my TT-seq and RNA-seq experiments (**Table 21**, **Table 23**).



**Figure 6. Detailed experimental workflow of TT-seq and mNET-seq experiments.**

See text of section 2.2.2 (TT-seq, RNA-seq) (a) and 2.2.3 (mNET-seq) (b) for details. (c) This diagram depicts the major (pre-)processing steps and corresponding tools (in dark blue). Steps unique to either mNET-seq, or TT-seq and RNA-seq are highlighted with asterisk symbol(s). The output format is highlighted in green. For details see Supplementary Information V.2.1. Abbreviations: quality controls (QC), Integrative Genomics Viewer (IGV).

### 2.2.2 Transient transcriptome sequencing (TT-seq)

*I optimized the experimental conditions for labeling of hematopoietic (suspension) cell lines (K562, Raji B wildtype, Raji B CDK9<sup>ts</sup>) upon different treatments (see also subsection 2.1.2) based on a protocol published by the Cramer lab<sup>168</sup>. I performed all TT-seq experiments (listed in **Table 21** and **Table 23**).*

The methods presented in subsection 2.2.2 are currently under peer review for publication:

**S. Gressel** \*, K. Lidschreiber \* *et al.* Transient transcriptome sequencing: experimental protocol to monitor genome-wide RNA synthesis including enhancer transcription.

*Methods Mol Biol, in revision*

(\* joint first authorship)

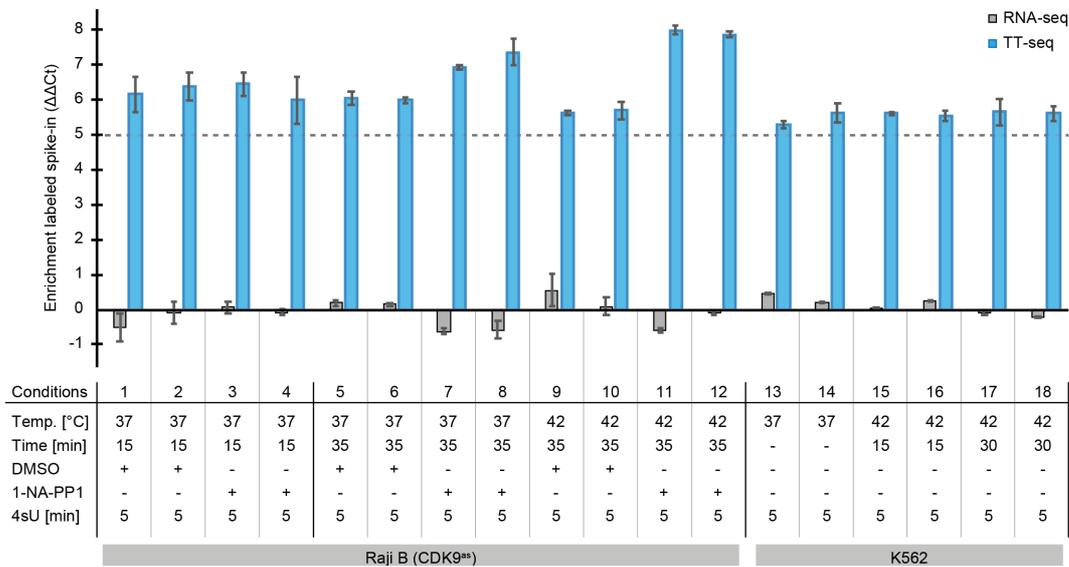
The TT-seq protocol<sup>168</sup> (**Figure 6 a**) consists of the following steps: (1) RNA spike-in pool preparation, (2) cell treatment and 4sU labeling of cellular RNA, (3) RNA spike-in addition and total RNA extraction (protocol based on<sup>417</sup>), (4) mild RNA fragmentation, (5) biotinylation of labeled RNA using HPDP-biotin (protocol based on<sup>365</sup>), (6) pull-down of labeled RNA using magnetic streptavidin beads, (7) labeled and total fragmented RNA clean-up including DNase I digest, (8) next-generation sequencing library preparation including random priming, cDNA fragmentation, rRNA depletion, and determination of PCR amplification cycle number, (9) quality control and Illumina sequencing. Buffer compositions are listed in **Table 17**.

- (1) **RNA spike-ins.** Six synthetic RNA spike-in controls (three unlabeled and three 4sU labeled spike-ins) were established for validation of labeled RNA enrichment, global normalization and for estimating possible cross-contamination from unlabeled RNA fragments<sup>168</sup>. The spike-ins are derived from selected RNAs of the ERCC RNA Spike-in Mix. For a detailed protocol see subsection 2.2.1 and **Table 20**.
- (2) **Cell treatment and 4sU labeling.** Cells were diluted 48 h before the labeling experiment was performed. Growth medium was exchanged 24 h before labeling. Prior to 4-thiouridine (4sU) labeling, cells were counted and treatments were performed (subsection 2.3.2 or 2.4.1). The exact cell count is required to calculate the corresponding amount of RNA spike-ins. For labeling, 4sU was added to  $\sim 3.5 \times 10^7$  cells in 50 mL of growth medium to a final concentration of 500  $\mu$ M, and incubated at 37 °C, 5 % CO<sub>2</sub> for 5 min. Exactly after 5 min of labeling, cells were centrifuged at 37 °C and 3,000 x g for 2 min. The cell pellet was vortexed with 5 mL of QIAzol lysis reagent, incubated for 5 min, and stored at -80 °C.
- (3) **Total RNA extraction.** Lysates were thawed and 150 ng of RNA spike-in pool ( $\sim 43$  ng/10<sup>7</sup> cells) was added, mixed and incubated for 5 min. Per 1 mL of Qiazol 0.2 mL of chloroform (1 mL) was added, mixed for 15 sec and incubated for 5 min. Lysates were centrifuged at

4 °C and 13,000 x g for 15 min. Aqueous phase was transferred to new 15 mL centrifuge tube, an equal volume of isopropanol was added, mixed, and incubated for 10 min. To precipitate total RNA, samples were centrifuged at 4 °C and 13,000 x g for 10 min. The RNA pellet was washed twice using 75 % ethanol at 4 °C and 13,000 x g for 10 min. The pellet was dissolved in 600 to 1,000  $\mu$ L H<sub>2</sub>O (aiming for a final concentration of  $\geq 750$  ng/ $\mu$ L). Concentration and purity of total RNA was determined by NanoDrop spectrophotometer. Total RNA was stored at -80 °C.

- (4) **Mild RNA fragmentation.** Total fragmented RNA should be in a range between 0.2 to 15 kbp. Two-times 150  $\mu$ g of total RNA in a volume of 130  $\mu$ L were transferred to 130  $\mu$ L microTUBE AFA Fiber Snap-Cap tubes. The following settings were used on Covaris S220: water level 12, temperature 7 °C, peak incident power 100 W, duty factor 1 %, cycles per burst 200, treatment time 10 sec.
- (5) **Biotinylation.** EZ-link HPDP-biotin is pyridyldithiol-activated and allows thiol-specific biotinylation of labeled RNA. Total fragmented RNA was incubated at 65 °C for 10 min, then placed on ice for 5 min. If RNA-seq was performed, 5  $\mu$ L of total fragmented RNA were saved for DNase I digest. RNA was splitted into two 2 mL PP microtubes (150  $\mu$ g/reaction) and processed in parallel. Water was added up to 700  $\mu$ L. Subsequently, 100  $\mu$ L of 10 x biotinylation buffer, 200  $\mu$ L of EZ-link HPDP-Biotin stock solution were added. Samples were incubated in the dark with rotation for 2 h. RNA was extracted and unbound biotin was removed using MaXtract High Density tubes following manufacturer's instructions. For RNA precipitation, 1/10 volume of 5 M NaCl and an equal volume of isopropanol was added, mixed and centrifuged at 4 °C and 13,000 x g for 30 min. The RNA pellet was washed twice using 75 % ethanol at 4 °C and 13,000 x g for 10 min. The pellet was dissolved in 100  $\mu$ L H<sub>2</sub>O per biotinylation reaction. The two RNA samples processed in parallel were pooled.
- (6) **Pull-down of labeled RNA.**  $\mu$ MACS wash buffer (3 mL per sample) was heated to 65 °C, or kept at room temperature (3 mL per sample).  $\mu$ MACS streptavidin beads were added to RNA sample (200  $\mu$ L) in a ratio of 1:2, and incubated at 4 °C for 15 min. In the meantime,  $\mu$ MACS columns were equilibrated with 100  $\mu$ L of nucleic acids equilibration buffer. Elution buffer (0.1 M DTT) was prepared freshly. RNA-bead-mix was transferred to the  $\mu$ MACS column, the flow-through was collected in LoBind tubes and reapplied twice. Columns were washed 3-times with 900  $\mu$ L of wash buffer (65 °C), and 3-times with 900  $\mu$ L of wash buffer (RT). Labeled RNA was eluted twice in 100  $\mu$ L of 0.1 M DTT (final volume of 200  $\mu$ L).
- (7) **RNA clean-up.** Labeled RNA (TT-seq) and 1  $\mu$ g of total fragmented RNA (RNA-seq) were processed in parallel. RNA was purified using either RNAClean XP beads or miRNeasy Micro Kit according to manufacturer's instructions (as recommended, buffer RWT was prepared with isopropanol instead of ethanol). RNA was quantified by Qubit High Sensitivity RNA kit. Labeled RNA enrichment was controlled by RT-qPCR (see subsection 2.1.2) using primer pairs for RNA spike-ins (**Table 8**) to estimate labeled enrichment. For total fragmented RNA, the Ct values of labeled and unlabeled RNA spike-ins were very similar. Labeled RNA showed a Ct value difference of  $>5$  for labeled

and unlabeled RNA spike-ins (**Figure 7**).

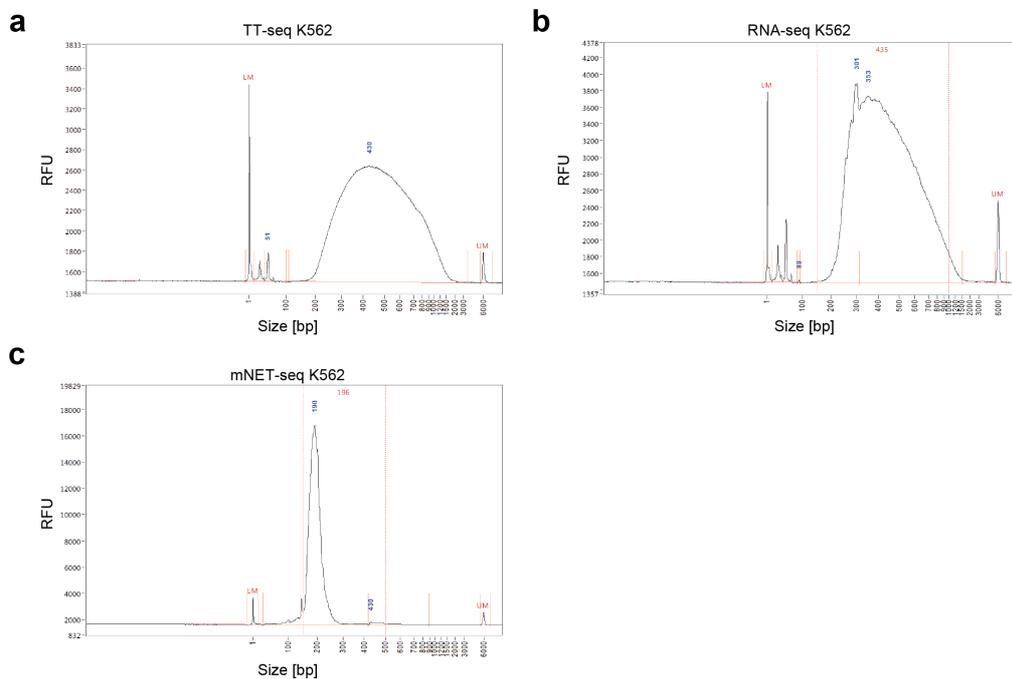


**Figure 7. Labeled spike-ins are enriched in TT-seq compared to RNA-seq samples.**

Bar plot depicts enrichment ( $\Delta\Delta\text{Ct}$ ) of labeled spike-in for TT-seq (5 min of 4sU labeling; blue) and total RNA-seq samples (grey) collected in Raji B (CDK9<sup>as</sup>) or K562 cells. Conditions are listed in the table at the bottom. RT-qPCR was performed as described in subsection 2.1.2 using primers for spike-ins 12 and 4 (**Table 8**). Corresponding normalized Ct values ( $\Delta\text{Ct}$ ) and enrichment values ( $\Delta\Delta\text{Ct}$ ) can be found in **Table 30**.

(8) **Library preparation.** Input RNA (100 ng) was treated with HL-dsDNase and used for strand-specific library preparation according to the Ovation Universal RNA-Seq System with minor modifications. For ‘First Strand Primer Pre-mix Preparation’, only Random primer mix was used (poly-d(T) primers were substituted with 0.4  $\mu\text{L}$  H<sub>2</sub>O). The User guide’s instructions were followed from ‘First Strand Synthesis Using DNase-treated RNA’ to ‘Second Strand cDNA Synthesis’. For fragmentation, Covaris S220 and 130  $\mu\text{L}$  microTUBE AFA Fiber Snap-Cap tubes were used with the following settings: water level 12, temperature 7 °C, peak incident power 145 W, duty factor 10 %, cycles per burst 200, treatment time 50 sec. The User guide’s instructions were followed from ‘cDNA Concentration After Fragmentation’ to ‘Adaptor Cleavage’. During ‘Ligation’ unique indexes (barcodes) were used per library (**Table 21**, **Table 23**). To avoid over-amplification, the precise number of PCR cycles was determined by the KAPA HiFi Library Amp Real Time kit following manufacturer’s instructions. As input, 4.2  $\mu\text{L}$  (~10 %) of the library, 5  $\mu\text{L}$  of 2x Kapa HiFi Hot Start Master Mix and 0.8  $\mu\text{L}$  Amplification Primer Mix P2 were used. For ‘PCR amplification’ of the remaining ~45  $\mu\text{L}$  of library the determined number of PCR cycles (usually 10-15 cycles) were used following the User guide’s instructions for amplification as described in the Appendix ‘Using qPCR to determine the Number of PCR Cycles’. The User guide’s instructions were followed from ‘Bead Purification of the Amplified Material’ to ‘Quantitative and Qualitative Assessment of the Library’. Library cDNA concentration was measured by Qubit HS RNA kit.

- (9) **Quality control and Illumina sequencing.** The size of the purified cDNA libraries was determined on a Fragment Analyzer. A representative size distribution of TT-seq and RNA-seq libraries prior to Illumina sequencing is depicted in **Figure 8 a-b**. Clustering and sequencing were performed on a HiSeq 1,500 or 2,500 (Illumina) in paired-end mode with 50 bp read length<sup>381</sup>, or on a NextSeq 550 (Illumina) in paired-end mode with 75 bp read length. An overview of generated libraries and sequencing statistics can be found in **Table 22** and **Table 25**.
- (10) **Data processing.** *Data processing was performed by Dr. Björn Schwab (MPI-bpc, Dept. of Molecular Biology) as previously published<sup>168,381</sup>. For details see Supplementary Information V.2.1.1 and **Figure 6 c**.*



**Figure 8. Size distribution of cDNA libraries generated in this study.**

Digital electropherograms show distribution of cDNA fragments size (library size includes both, adapters and insert). Purified cDNA libraries were analyzed by monitoring the relative fluorescence unit (RFU) intensity during capillary electrophoresis separation (size range 0.1-6 kbp) using Fragment Analyzer. The Fragment Analyzer run was performed in-house by Sigurd Hille, Kerstin Maier or Petra Rus (MPI-bpc, Dept. of Molecular Biology). (a) TT-seq cDNA library. (b) RNA-seq cDNA library. (c) mNET-seq cDNA library.

### 2.2.3 Mammalian native elongating transcript sequencing (mNET-seq)

*I established a protocol for human hematopoietic (suspension) cell lines (K562, Raji B CDK9<sup>as</sup>) based on protocols published by the Proudfoot lab<sup>350,388</sup> and performed all mNET-seq experiments (listed in **Table 21** and **Table 23**).*

The final mNET-seq protocol (**Figure 6 b**) consists of the following steps: (1) cell fractionation (protocol based on<sup>418</sup>) and chromatin digestion, (2) immunoprecipitation (IP) of total Pol II including Empigen BB detergent (protocol based on<sup>391</sup>), (3) RNA phosphorylation, (4) RNA precipitation and size selection (protocol based on<sup>419</sup>), (5) next-generation sequencing library preparation including 3' and 5' RNA adapter ligation, reverse transcription, cDNA library amplification, size selection and purification, and (6) quality control and Illumina sequencing. Buffer compositions are listed in **Table 18**.

- (1) **Cell fractionation.** Cells were diluted 48 h before the experiment was performed, and growth medium was exchanged 24 h before the experiment. Cells were counted and treatments were added (see subsection 2.3.2 or 2.4.1). Cells were collected in a swing-bucket centrifuge at 4 °C and 200 x g for 5 min. Cells were washed using 10 mL of DPBS and centrifuged at 4 °C and 200 x g for 5 min. The cell pellet was resuspended at 10<sup>7</sup> cells per mL in DPBS and transferred to 1.5 mL tubes (i.e. for 10<sup>8</sup> cells, ten-times 1 mL). Per replicate, samples were processed in parallel for fractionation and MNase digest, and pooled before the IP step. Cells were centrifuged in a fixed-angle benchtop centrifuge at 4 °C and 200 x g for 2 min. The supernatant was removed, the cell pellet was gently resuspended in 400 µL of cell lysis buffer, and incubated on ice for 5 min. The cell lysate was overlaid on top of 1 mL sucrose buffer, and centrifuged at 4 °C and 2,000-3,500 x g for 10 min. The resulting supernatant contains the cytoplasm (kept for fractionation control), and the pellet contains cell nuclei. During the centrifugation step, MNase was added to the chromatin digest buffer to a final concentration of 50 U, and warmed up to 37 °C. Cell nuclei were resuspended in 250 µL of glycerol buffer and 250 µL of nuclear lysis buffer, and incubated on ice for 2 min. To precipitate the chromatin-Pol II-RNA complex, samples were centrifuged at 4 °C and 13,000 x g for 2 min. The resulting supernatant contains the nucleoplasm (kept for fractionation control). The pellet was washed once with 100 µL of 1 x MNase buffer (without MNase). To digest the chromatin-RNA pellet, 100 µL of pre-warmed chromatin digest buffer as added at 37 °C and 1,400 rpm for 90-120 sec (K562: 120 sec; Raji B CDK9<sup>as</sup>: 90 sec). For a reproducible digest it is critical to control temperature of the thermomixer. The digest was stopped by adding 11 µL of 250 mM EGTA (final concentration 25 mM) and transfer to ice. Samples were centrifuged at 4 °C and 13,000 x g for 5 min. Supernatant of all samples per replicate were pooled (i.e. for 10<sup>8</sup> cells, ~1 mL). The quality of cell fractionation was controlled by Western blot of marker proteins.
- (2) **IP of total Pol II with 1 % v/v Empigen BB (~30 % active substance).** Chromatin-Pol II-RNA complexes were 8-fold diluted with IP buffer. Antibody-conjugated Dynabeads M-280 Sheep anti-mouse IgGs were prepared based on manufacturer's instructions. Briefly, for 10<sup>8</sup> cells 250 µL of Dynabeads (~10 mg/mL) were washed three-times with DW buffer. 30 µg of Pol II POLR2A total CTD (all isoforms) antibody (CMA601) was

added, and coupled by slow rotation at 4 °C and 8 rpm for 2 h or overnight. Antibody-conjugated Dynabeads were washed three-times with DW buffer, resuspended in IP buffer and added to diluted chromatin-Pol II-RNA complexes. IP was incubated on a rotating wheel at 4 °C and 12 rpm for 1 h. The tube was transferred to a big magnet for 5 min, the supernatant removed by decanting, and the remaining beads were transferred to a new 1.5 mL LoBind tube using 1 mL of IP buffer. Beads were washed six times with 1 mL IP buffer. For each wash, solution was transferred to the magnet for 1 min before supernatant was removed. Beads were resuspended in 300  $\mu$ L of PNKT buffer.

- (3) **RNA phosphorylation.** To control RNA size distribution, 1/10 volume of IP (corresponding to  $10^7$  cells) was subjected to  $^{32}\text{P}$ - $\gamma$ -ATP labeling in the hot lab according to MPI-bpc safety regulations. For  $^{32}\text{P}$ - $\gamma$ -ATP labeling, beads were resuspended in 10  $\mu$ L PNK hot reaction mix containing 8.5  $\mu$ L of PNKT, 0.5  $\mu$ L of T4 polynucleotide kinase (PNK) 3' phosphatase minus (20 U/ $\mu$ L) and 1  $\mu$ L of fresh  $^{32}\text{P}$ - $\gamma$ -ATP. For ATP labeling, beads were resuspended in 200  $\mu$ L of PNK reaction mix containing 160  $\mu$ L PNKT, 10  $\mu$ L of T4 PNK 3' phosphatase minus (20 U/ $\mu$ L) and 30  $\mu$ L of 10 mM ATP. Reactions were incubated at 37 °C and 800 rpm for 10 min. Beads were washed with 1 mL IP buffer, resuspended in 1 mL of TRIzol reagent and incubated for 5 min. Samples were stored at -20 °C.
- (4) **RNA precipitation and size selection (25-110 nt).** Samples were thawed, 200  $\mu$ L of chloroform was added, mixed by vortexing and incubated for 2 min. Subsequently, samples were centrifuged at 4 °C and 16,000 x g for 15 min. The upper aqueous phase (~0.5 mL) was transferred to a new 1.5 mL tube. 1  $\mu$ L of GlycoBlue co-precipitant and equal volume (~0.5 mL) of isopropanol were added, mixed and incubated for 10 min. RNA was precipitated by centrifugation at 4 °C and 16,000 x g for 10 min. RNA pellets were air-dried for ~3 min and dissolved in 10  $\mu$ L of 7 M urea LB. Denaturing 6 % polyacrylamide gel solution (PAGE) was prepared, poured into a 1 mm cassette with a 15-well comb and polymerized for at least 30 min. The comb was removed and PAGE was pre-run in electrophoresis tank with 1 x TBE buffer at 200 V for 30 min. Wells were washed with a syringe to remove urea before sample loading. 7 M urea LB containing 0.05 % bromophenol blue and 0.05 % xylene cyanol dyes was used to follow sample migration. In a 6 % PAGE, bromophenol blue migrates at ~25 nt and xylene cyanol at ~110 nt (**Figure 39**). Samples were run at 200 V for 30 min. For  $^{32}\text{P}$ - $\gamma$ -ATP labeled RNA, size distribution was visualized by phosphorimaging on a Typhoon FLA 9500 instrument. For ATP labeled RNA, the gel was cut above the bromophenol blue band and above the xylene cyanol band. RNA was extracted by crush-and-soak. Briefly, 2 mL tubes were prepared for RNA recovery by inserting a 0.5 mL tube which was pierced several times with a 22 gauge syringe needle. Gel slices were transferred to 0.5 mL tube and centrifuged at 16,000 x g for 2 min. The 0.5 mL tube was removed, 500  $\mu$ L of elution buffer was added to the gel fragments and rotated at 12 rpm for 3 h. Samples were centrifuged at 16,000 x g for 1 min, and the elution was repeated from the gel fragments. Supernatant was transferred to a spin-X column with cellulose acetate membrane and centrifuged at 16,000 x g for 1 min. Flow-through was transferred to new 2 mL tube, 0.5  $\mu$ L of GlycoBlue co-precipitant and 2.5 volumes of 100 % ethanol were added and incubated at -20 °C for 30 min. Samples were centrifuged at 4 °C and 16,000 x g for 10 min. RNA pellets were air-

---

dried for ~5 min. The first pellet was dissolved in 5  $\mu\text{L}$  of  $\text{H}_2\text{O}$ , and then transferred to the RNA pellet of the second elution (final volume 5.5  $\mu\text{L}$ ). RNA was stored at  $-80\text{ }^\circ\text{C}$ .

- (5) **Library preparation.** Input RNA (5  $\mu\text{L}$ ) was used for strand-specific library preparation according to the TruSeq Small RNA Library Prep Kit (Illumina) with modifications described below. The Reference Guide's instructions were followed from 'Ligate 3' Adapter' to 'Amplify Libraries'. During amplification (15 cycles) unique PCR primer barcodes of Indices A or D Box (sequences 1-12, or 37-48) were used per library according to Illumina's Pooling Guidelines. cDNA size selection (150-230 nt) was performed using 4 % E-Gel High-Resolution agarose gels, loading a maximum of two libraries per gel. Gels were loaded with TrackIt 25 bp DNA ladder, HRL ladder, CRL ladder, 25  $\mu\text{L}$  of library in two wells, and 20  $\mu\text{L}$  of  $\text{H}_2\text{O}$  in all empty wells. Gels were run in iBase Power system using program 2 at maximum run time (40 min). Bands were visualized on a transilluminator by exposure to UV light. Libraries were cut using sterile disposable scalpels no. 23. Libraries were purified using QIAquick MinElute Gel Extraction Kit according to manufacturer's instructions. The kit recovers ~80 % of DNA (70 bp – 4 kbp). cDNA was eluted in 11  $\mu\text{L}$ .
- (6) **Quality control and Illumina sequencing.** The size of the purified cDNA libraries was determined on a Fragment Analyzer. A representative size distribution of mNET-seq libraries is depicted in **Figure 8 c**. Clustering and sequencing were performed on a HiSeq 1,500 or 2,500 (Illumina) in paired-end mode with 50 bp read length <sup>381</sup>, or on a NextSeq 550 (Illumina) in paired-end mode with 75 bp read length. An overview of generated libraries and sequencing statistics can be found in **Table 22** and **Table 25**.
- (7) **Data (pre-)processing.** *Data processing was performed by Dr. Björn Schwalb (MPI-bpc, Dept. of Molecular Biology) as published <sup>381</sup>.* For details see Supplementary Information V.2.1.2 and **Figure 6 c**.

### 2.2.4 Kinetic modeling ( $I$ , $d$ ), pause site $n^*$ and the ‘pause-initiation limit’

*Kinetic modeling was designed in collaboration with Dr. Björn Schwalb (MPI-bpc, Dept. of Molecular Biology) who also performed the final analysis.*

The methods presented in subsection 2.2.4 have been published:

**S. Gressel**\*, B. Schwalb\* *et al.* CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* 6, doi:10.7554/eLife.29736 (2017).

(\* joint first authorship)

**Detection of pause sites.** For all expressed (c)TUs or RefSeq-TUs (exceeding 10 kbp in length with one unique TSS given all RefSeq annotated isoforms (UCSC RefSeq GRCh38))  $i$  the pause site  $n^*$  was calculated for all bases  $m$  in a window of 350 bp downstream of the GRO-cap refined TSS, or in a window from the TSS to the end of the first exon (excluding the last 5 bp) via maximizing the function

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

where  $\rho_i$  needed to exceed 5 times the median of the signal strength  $p_{im}$  for all non-negative antisense bias corrected mNET-seq coverage values. In order to maximize the chances of finding the most likely pause site, two replicate tracks were constructed by taking the maximum of each nucleotide over the first and second replicates, respectively, regardless of treatment. We defined the pause site at the active site based on structural information<sup>307</sup> (**Figure 2 a**). The RNA-DNA hybrid within the paused polymerase is in the so-called tilted state that hinders nucleotide addition at the active site. Thus, the subsequent nucleotide is not added yet. We defined the pause site to be the position in line with the 'post-translocated' RNA rather than with the 'pre-translocated' DNA. In conclusion, the pause site was calculated as  $n^* = m^* + 1$ , where  $m^*$  is the argument that maximizes  $\rho_i$ .

**Molecular weight conversions.** The known sequence and mixture of the utilized spike-ins allows to calculate a conversion factor to RNA amount per cell [cell<sup>-1</sup>] given their molecular weight assuming perfect RNA extraction. The number of spike-in molecules per cell  $N$  [cell<sup>-1</sup>] was calculated as

$$N = \frac{m}{Mn} N_A$$

with the mass  $m$   $25 \cdot 10^{-9}$  [g] per spike-in (in this study, pool with equal mass per spike-in), the number of cells  $n$  (first study:  $3.27 \cdot 10^7$  Raji B CDK9<sup>as</sup> cells; second study:  $3.8 \cdot 10^7$  for K562,  $3.4 \cdot 10^7$  for RajiB CDK9<sup>as</sup>) the Avogadro constant  $N_A$   $6.02214085774 \cdot 10^{23}$  [mol<sup>-1</sup>] and molar-mass (molecular weight) of the spike-ins  $M$  [g mol<sup>-1</sup>] calculated as

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

where  $A_n$ ,  $U_n$ ,  $C_n$ , and  $G_n$  are the number of each respective nucleotide within each spike-in

polynucleotide.  $\tau$  is set to 0.1 in case of a labeled spike-in and 0 otherwise. The addition of 159 to the molecular weight considers the molecular weight of a 5' triphosphate (see **Table 20**). Provided the above, the conversion factor to RNA amount per cell  $\kappa$  [cell<sup>-1</sup>] can be calculated as

$$\kappa = \text{mean} \left( \text{median}_i \left( \frac{tb_i}{L_i \cdot N} \right) \right)$$

for all labeled spike-in species  $i$  with length  $L_i$ . Note that imperfect RNA extraction efficiency would lead to an underestimation of cellular labeled RNA in comparison to the amount of added spike-ins and thus to an underestimation of initiation frequencies. In case of a strong underestimation however the real initiation frequencies would lie above the pause-initiation limit, which is theoretically impossible. Thus, we assume this effect to be insignificant.

**Estimation of the productive initiation frequency  $I$ .** The antisense bias corrected number of transcribed bases  $tb_i^{control}$  was calculated on all expressed (c)TUs or expressed RefSeq-TUs (exceeding 10 kbp in length with one unique TSS given all RefSeq annotated isoforms (UCSC RefSeq GRCh38))  $i$ . For each (c)TU or RefSeq TU  $i$  the productive initiation frequency  $I_i$  [cell<sup>-1</sup> min<sup>-1</sup>], which corresponds to the pause release rate, was calculated as

$$I_i = \frac{1}{\kappa} \cdot \frac{tb_i^{control}}{t \cdot L_i}$$

with labeling duration  $t = 5$  [min] and length  $L_i$ . Note that for RefSeq-TUs,  $tb_i^{control}$  and  $L_i$  were restricted to regions of non-first constitutive exons (exonic bases common to all isoforms) located in the first 25 kbp. Given a 30 min treatment, we expect only the first 35 kbp to be significantly affected by changes in initiation frequency assuming an average elongation velocity of 2.4 kbp/min. Additionally, changes in splicing rate upon heat shock treatment<sup>209,420</sup> should not influence constitutive exonic regions. This is to ensure, not to be biased by alternative splicing.

**Estimation of pause duration  $d$ .** For all expressed (c)TUs or expressed RefSeq-TUs (exceeding 10 kbp in length with one unique TSS given all RefSeq annotated isoforms (UCSC RefSeq GRCh38))  $i$  the pause duration  $d_i$  [min] was calculated as the residing time of the polymerase in a window +/- 100 bases  $m$  around the pause site (see above) as

$$d_i = \frac{\sum_{\pm 100} p_{im}}{I_i} \cdot \text{median}_i \left( \frac{v_i}{I_i v_i (t^* - t) / \sum_{\text{response window}} p_{im}} \right)$$

with pause release rate  $I_i$  and the number of polymerases  $p_{im}$  (antisense bias corrected mNET-seq coverage values in a window +/- 100 bases around the pause site. For pause sites below 100 bp downstream of the TSS the first 200 bp of the TU were considered. Note that the right part of the formula is restricted to mNETseq instances above the 50% quantile for robustness and adjusts  $d_i$  to an absolute scale by comparing the CDK9 derived elongation velocities  $v_i$  with those derived from combining mNET-seq and TT-seq data in the response window [200,  $v_i(t^* - t)$ ]. The productive initiation frequency represents the ‘true’ initiation frequency if the fraction of Pol II terminating within the pause window is insignificant

(unknown fraction of early termination). The pause duration  $d$  obtained in this way reflects the effective pause between two initiation events that successfully lead to productive elongation of a transcript and thus the relevant transcriptional outcome.

**Pause-initiation limit.** The previously derived inequality from <sup>315</sup>

$$\frac{v}{l} \geq 50 [bp]$$

states that new initiation events into productive elongation are limited by the velocity of the polymerase in the promoter-proximal region and that steric hindrance occurs below a distance of 50 bp between the active sites of the initiating Pol II and the paused Pol II. Given the calculations of pause duration  $d$  and (productive) initiation frequency  $I$  above, we can reformulate this inequality to

$$\frac{200 [bp]}{d \cdot I} \geq 50 [bp]$$

with 200 [bp] being the above defined pause window.

---

## 2.3 CDK9-dependent Pol II pausing controls transcription initiation

---

The methods presented subsections 2.3.1 (first paragraph), 2.3.4 and 2.3.5 have been published:

**S. Gressel \***, B. Schwalb \* *et al.* CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* 6, doi:10.7554/eLife.29736 (2017).

(\* joint first authorship)

---

### 2.3.1 CRISPR/Cas9 engineering of human Raji B analog-sensitive CDK9 cell line

*The human Raji B (CDK9<sup>as</sup>) cell line was generated and validated by Dr. Weihua Qin (LMU Munich, Dept. of Biology II, group of Prof. Dr. Heinrich Leonhardt).*

**CRISPR/Cas9 engineering.** Analog sensitive CDK9 (CDK9<sup>as</sup>) contains a point mutation of the so-called gatekeeper residue that enables the kinase active site to accept bulky ATP analogs as 1-NA-PP1 (**Figure 3 d-e**). To identify the gatekeeper residue <sup>339</sup>, the amino acid sequence of the human CDK9 kinase (UniProt, P50750-1) was aligned with sequences of previously characterized kinases carrying analog sensitive mutations. Multiple sequence alignment was performed with the web tool Clustal Omega 1.2.4 <sup>407</sup>. For the canonical isoform of CDK9, phenylalanine (F) 103 was identified as the gatekeeper residue and selected for mutation to alanine (A). Mutation of F103 at the CDK9 gene loci in Raji B cells was performed using the CRISPR/Cas9 system <sup>344,345</sup> as described <sup>421</sup> with minor modifications (**Figure 9 a**). Briefly, the single guide RNA (sgRNA) for editing CDK9 was designed by using the web tool Optimized CRISPR design (<http://crispr.mit.edu/>), and was incorporated to pSpCas9(BB)-2A-GFP (PX458) vector by BpiI restriction sites (Addgene plasmid #48138) <sup>397</sup>

(**Table 7**, see Appendix for plasmid map). For nucleotide replacement (gttc to cgcg), 200 nt single-stranded DNA oligonucleotides (ssODNs) were synthesized by Integrated DNA Technologies and used as homology-directed repair (HDR) template (**Table 10**). A BstUI cutting site was incorporated into the HDR template for screening. The vector and HDR template were introduced into human Raji B cells by using Amaxa Mouse ES Cell Nucleofector® Kit according to the manufacturer's instructions. Two days after transfection, GFP positive cells were single cell sorted into 96 well plates using FACS Aria II instrument. After two weeks, individual colonies were expanded for genomic DNA isolation.

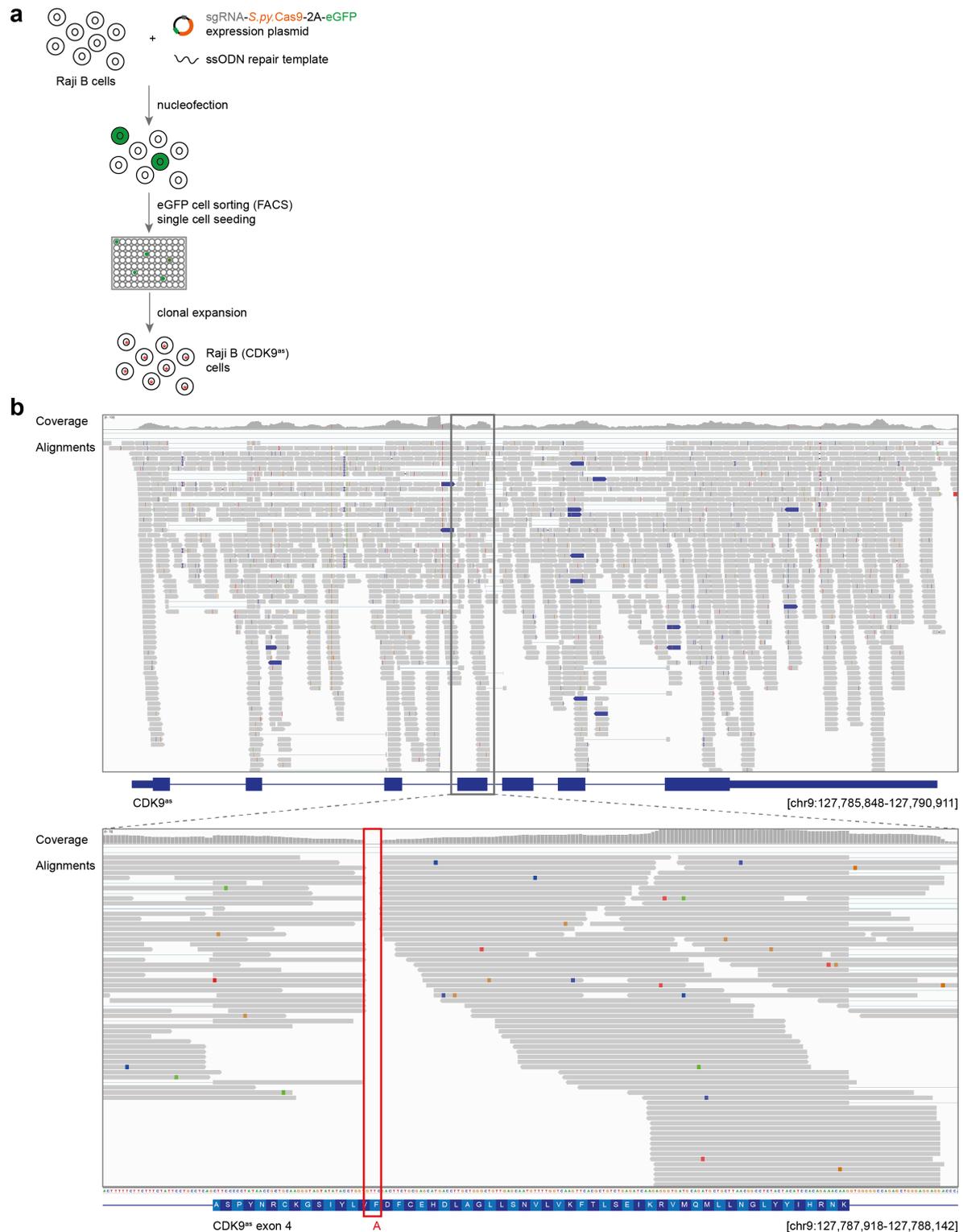
**Genomic DNA (gDNA) extraction.**  $10^5$  Raji B wildtype or Raji B (CDK9<sup>as</sup>) cells were used for gDNA extraction. Medium was aspirated and cells washed once with DPBS. Cell lysis and gDNA extraction was performed using 100  $\mu$ L of Quick Extraction solution per  $10^5$  cells according to manufacturer's instructions. Briefly, the lysate was incubated at 65 °C for 15 min and subsequently at 95 °C for 10 min. The final DNA concentration was measured by NanoDrop.

**Target DNA PCR amplification and purification.** Per reaction, 100 ng of gDNA template were used. PCR amplification was performed using AccuPrime GC-Rich DNA Polymerase according to manufacturer's instructions. Briefly, forward and reverse primers (final 200 nM each) (**Table 9**), buffer A (GC-rich template, 2 mM MgSO<sub>4</sub>), AccuPrime DNA polymerase (1 U) were added to gDNA template. Reactions were run using the following program in a thermal cycler with a lid temperature of 99 °C. The PCR product was analysed on a 2 % agarose gel containing SYBRsafe. Bands were visualized on a transilluminator by exposure to UV light. DNA band was cut using sterile disposable scalpels no. 23 (Swann-Morton) and purified using QIAquick MinElute Gel Extraction Kit according to manufacturer's instructions. DNA was eluted in 11  $\mu$ L. DNA concentration was measured by NanoDrop.

Step	Temperature	Time	Cycle
1	95 °C	3 min	1 x
2	95 °C	30 sec	25-30 x
3	(T <sub>m</sub> of primers – 5 °C)	30 sec	
4	72 °C	30 sec	
5	72 °C	10 min	1 x
6	4 °C	hold	

**BstUI digest.** For restriction digest, purified PCR product (200 ng) was digested using 49  $\mu$ L of BstUI digestion mix containing 1 x CutSmart buffer and 1  $\mu$ L of BstUI enzyme. The digestion mix was incubated at 60 °C for 15 min and terminated by adding 1 volume of gel loading dye at 70 °C for 10 min. Samples (wildtype, CDK9<sup>as</sup>) were analyzed on a 2 % agarose gel containing SYBRsafe. Bands were visualized on a transilluminator by exposure to UV light.

**Sanger cycle sequencing using capillary electrophoresis.** 1  $\mu$ g of purified PCR product was submitted for Sanger sequencing. Sequencing primers are listed in **Table 9**.



**Figure 9. CRISPR/Cas9 engineering of analog-sensitive CDK9 in human Raji B cells.**

(a) Detailed experimental workflow of CRISPR/Cas9 engineering of the CDK9<sup>as</sup> cell line. See Method 2.3.1 for details. (b) IGV<sup>400</sup> browser shot of TT-seq data (**Table 21**) at the CDK9 locus (top: entire gene; bottom: zoomed to exon 4 carrying F103A mutation in CDK9<sup>as</sup>). Reads containing mutated residues (gttc to cgcg) (>2 mismatches per 50 bp read) cannot be mapped to the reference genome (highlighted by red box).

### 2.3.2 Cell treatment: CDK9<sup>as</sup> inhibition

*Optimization of 1-NA-PP1 concentration was performed by Dr. Tim M. Decker (Helmholtz Center Munich, Dept. of Molecular Epigenetics, group of Prof. Dr. Dirk Eick) as published<sup>381</sup>.*

Based on the cell proliferation (MTS) assay and Pol II CTD phosphorylation data we chose 5  $\mu$ M of 1-NA-PP1 and 15 min as treatment conditions for the TT-seq and mNET-seq analysis of Raji B (CDK9<sup>as</sup>) cells.

### 2.3.3 Multi-omics experiments

*TT-seq and mNET-seq experiments were performed as described (section 2.2). I performed all TT-seq and mNET-seq experiments. For four TT-seq experiments (Table 21, no. 1-4), Dr. Tim M. Decker (Helmholtz Center Munich, Dept. of Molecular Epigenetics, group of Prof. Dr. Dirk Eick) contributed to steps (2) to (7) (see section 2.2.2, steps 2-7).*

**Table 21. Information on experimental conditions used in this study.**

No.	Assay	Cell type	Condition name	Replicate no.	Treatment
1	TT-seq	Raji B (CDK9 <sup>as</sup> )	Ctrl	1	0.05 % v/v DMSO, 15 min.
2	TT-seq	Raji B (CDK9 <sup>as</sup> )	Ctrl	2	0.05 % v/v DMSO, 15 min.
3	TT-seq	Raji B (CDK9 <sup>as</sup> )	CDK9 inhibited	1	5 $\mu$ M 1-NA-PP1, 15 min.
4	TT-seq	Raji B (CDK9 <sup>as</sup> )	CDK9 inhibited	2	5 $\mu$ M 1-NA-PP1, 15 min.
5	TT-seq	Raji B	Ctrl	1	0.05 % v/v DMSO, 15 min.
6	TT-seq	Raji B	Ctrl	2	0.05 % v/v DMSO, 15 min.
7	TT-seq	Raji B	CDK9 inhibited	1	5 $\mu$ M 1-NA-PP1, 15 min.
8	TT-seq	Raji B	CDK9 inhibited	2	5 $\mu$ M 1-NA-PP1, 15 min.
9	mNET-seq	Raji B (CDK9 <sup>as</sup> )	Ctrl	1	0.05 % v/v DMSO, 15 min.
10	mNET-seq	Raji B (CDK9 <sup>as</sup> )	Ctrl	2	0.05 % v/v DMSO, 15 min.
11	mNET-seq	Raji B (CDK9 <sup>as</sup> )	CDK9 inhibited	1	5 $\mu$ M 1-NA-PP1, 15 min.
12	mNET-seq	Raji B (CDK9 <sup>as</sup> )	CDK9 inhibited	2	5 $\mu$ M 1-NA-PP1, 15 min.

For TT-seq pre-processing and global spike-in normalization, as well as mNET-seq pre-processing see Supplementary Information V.2.1.

### 2.3.4 Analysis of CDK9 response ratio

*Data analysis was performed by Dr. Björn Schwalb (MPI-bpc, Dept. Molecular Biology).*

**Size factor normalization.** We first checked that no significant global shifts were detected in a comparison of two TT-seq replicates with 1-NA-PP1 (CDK9<sup>as</sup> inhibitor) treatment against two TT-seq replicates with DMSO treatment (control) in the described spike-ins normalization strategy (V.2.1.1). Then all samples were subjected to an alternative, more robust normalization procedure. For each sample  $j$  the antisense bias corrected number of transcribed bases  $tb_j$  was calculated on all expressed TUs  $i$  exceeding 125 kbp in length. 50 kbp were truncated from each side of the selected TUs to avoid influence of the response to CDK9<sup>as</sup> inhibition<sup>302</sup>. On the resulting intervals, size factors for each sample  $j$  were determined as

$$\sigma_j = \text{median}_i \left( \frac{tb_{ij}}{(\prod_{v=1}^m tb_{ij})^{1/m}} \right)$$

where  $m$  denotes the number of samples. This formula has been adapted<sup>422</sup> and was used to correct for library size and sequencing depth variations.

**Calculation of response ratios.** For each condition  $j$  (control or CDK9<sup>as</sup> inhibited) the antisense bias corrected number of transcribed bases  $tb_i^j$  was calculated on all expressed TUs  $i$  exceeding 10 kbp in length. Of all remaining TUs only those were kept harboring one unique TSS given all Refseq annotated isoforms (UCSC RefSeq GRCh38). Response ratios were calculated for a window from the TSS to 10 kbp downstream (excluding the first 200 bp) for each TU  $i$  as

$$r_i = 1 - \frac{tb_i^{CDK9^{as} \text{ inhibited}}_{[0.2, 10 \text{ kbp}]}}{tb_i^{Control}_{[0.2, 10 \text{ kbp}]}}$$

where negative values were set to 0.

**Estimation of robust common elongation velocity.** For each condition  $j$  (control or CDK9<sup>as</sup> inhibited) the antisense bias corrected number of transcribed bases  $tb_i^j$  was calculated on all expressed TUs  $i$  with a given response ratio  $r_i$ , excluding the first 200 bp. All TUs were truncated by 5 kbp in length from the 3' end prior to calculation to avoid influence of some alterations in signal around the pA site after CDK9<sup>as</sup> inhibition<sup>302</sup>. A robust common elongation velocity estimate was calculated by finding an optimal fit for all TUs  $i$  between 25 to 200 kbp in length  $L_i$ , i.e. minimizing the function

$$\text{loss} = \text{median}_i \left( \left| 1 - \frac{tb_i^{CDK9^{as} \text{ inhibited}}}{tb_i^{Control}} - \frac{r_i v(t^* - t)}{L_i} \right| \right)$$

on the interval  $[0, 10000]$  with inhibitor treatment duration  $t^* = 15$  [min] and labeling duration  $t = 5$  [min], given that

$$tb_i^{CDK9^{as} \text{ inhibited}} - tb_i^{Control} = r_i \frac{tb_i^{Control}}{L_i} v_i(t^* - t)$$

, i.e. the difference of transcribed bases obtained by the CDK9<sup>as</sup> inhibitor treatment equals the number of transcribed bases per nucleotide  $tb_i^{Control}/L_i$  times the number of nucleotides traveled  $v_i(t^* - t)$  in  $t^* - t$  minutes corrected by the amount of the response  $r_i$ .

**Estimation of gene-wise elongation velocity.** For each condition  $j$  (control or CDK9<sup>as</sup> inhibited) the antisense bias corrected number of transcribed bases  $tb_i^j$  was calculated on all expressed TUs  $i$  exceeding 35 kbp in length, excluding the first 200 bp. All TUs were truncated by 5 kbp in length from the 3' end prior to calculation to avoid influence of some alterations in signal around the pA site after CDK9<sup>as</sup> inhibition<sup>302</sup>. Of all remaining TUs only those were kept harboring one unique TSS given all Refseq annotated isoforms (UCSC RefSeq GRCh38). For each TU  $i$  with  $r_i > 0.25$  the elongation velocity  $v_i$  [kbp/min] was calculated as

$$v_i = \frac{tb_i^{Control} - tb_i^{CDK9^{as} \text{ inhibited}}}{tb_i^{Control} \cdot \frac{r_i}{L_i} (t^* - t)}$$

with inhibitor treatment duration  $t^* = 15$  [min] and labeling duration  $t = 5$  [min].

### 2.3.5 Analysis of human promoter-proximal pause sites

*Data analysis was performed by Dr. Björn Schwalb (MPI-bpc, Dept. Molecular Biology).*

**mNET-seq data normalization.** mNET-seq coverage tracks were size factor normalized on 260 TUs that showed a response of less than 5% ( $r_i < 0.05$ ) in the TT-seq signal upon 1-NA-PP1 (CDK9<sup>as</sup> inhibitor) treatment. The response ratio  $r_i$  was determined as described (2.3.4) including also TUs with multiple TSS to extend the number of TUs for normalization. Note that variation of the response ratio cutoff and thereby the number of TUs available for normalization does virtually not change the normalization parameters. Coverage tracks for further analysis were restricted to the last nucleotide incorporated by the polymerase in the aligned mNET-seq reads.

**DNA-RNA and DNA-DNA melting temperature calculation.** The gene-wise mean melting temperature of the DNA-RNA and DNA-DNA hybrid was calculated from subsequent melting temperature estimates of 8-base pair DNA-RNA<sup>423</sup> and DNA-DNA<sup>424</sup> duplexes tiling the respective area.

***In vivo* RNA secondary structure (DMS-seq).** The gene-wise DMS-seq coverage<sup>425</sup> for a window of [-15, -65] bp upstream of the pause site was normalized by subtraction from the respective DMS-seq coverage (denatured) allowing for maximal 5% negative values which were set to 0 (sequencing depth adjustment). The gene-wise mean values were subsequently normalized by dividing with the initiation frequency. Note that the latter normalization has an insignificant effect.

**Prediction of RNA secondary structure (*in silico*).** The gene-wise mean minimum free energy for a window of [-15,-65] bp upstream of the pause site was calculated from subsequent minimum free energy estimates of 13-base pair RNA fragments tiling the respective area using RNAfold from the ViennaRNA package<sup>426</sup>.

### 2.3.6 Illumina sequencing and data availability

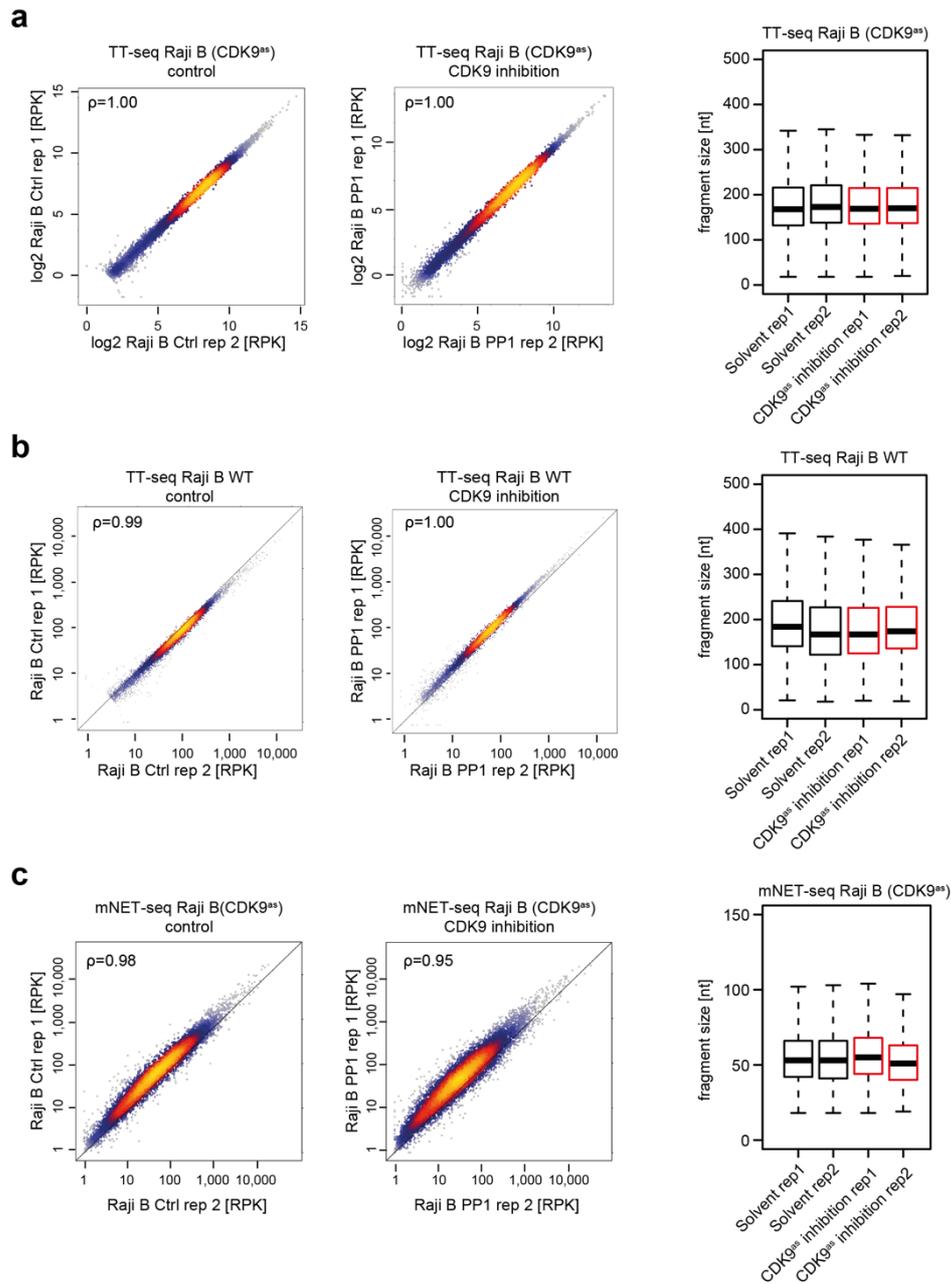
*For this study, Illumina sequencing was performed at the Genomics research unit of Dr. Helmut Blum at Gene Center Munich (LAFUGA) by Dr. Stefan Krebs, and at the Deep-Sequencing Facility at UMG Göttingen (TAL).*

The sequencing data and processed files were deposited in the GEO database under accession code GSE96056<sup>381</sup>. A detailed overview of sequencing data and processed files is depicted in **Table 22**. Experimental conditions are listed in **Table 21**.

**Table 22. Sequencing statistics of 12 libraries generated in this study.**

All libraries were sequenced on an Illumina HiSeq 1,500 or 2,500 sequencing platform in 50 bp paired-end mode. Numbers refer to experimental conditions listed in **Table 21**. Correlation of replicates is calculated as Spearman's rho rounded to second decimal place. Sequenced fragments were determined by FastQC, uniquely mapped fragments were counted after mapping using STAR 2.3.0<sup>403</sup>.

No.	Barcode	GEO no.	Fragment numbers			Duplicates (%)	Replicate correlation (RPKs)
			Sequenced	Mapped uniquely	Replicates combined		
1	GGAGAA	GSM2528067	314,918,545	268,871,103		12.1	
2	AGCATG	GSM2528068	309,157,897	263,833,832	532,704,935	12.7	1.00
3	GAGTCA	GSM2528066	306,381,221	252,242,142		12.2	
4	CGTAGA	GSM2528069	297,142,421	238,445,837	490,687,979	9.7	1.00
5	AAGAGG	GSM2728731	47,534,035	44,476,779		16.7	
6	GGAGAA	GSM2728732	54,910,285	50,449,470	94,926,249	25.2	0.99
7	AGCATG	GSM2728733	53,643,687	49,779,461		49.2	
8	GAGTCA	GSM2728734	41,709,549	38,712,626	88,492,087	44.9	1.00
9	GCCAAT	GSM2728735	115,716,948	44,550,953		35.1	
		GSM2728736	155,873,045	96,105,747		46.9	
10	TGACCA	GSM2728737	249,818,334	28,847,051	169,503,751	36.6	0.98
11	CTTGTA	GSM2728738	86,286,005	38,611,242		37.0	
		GSM2728739	135,015,202	82,905,554		48.4	
12	ACTTGA	GSM2728740	185,224,474	26,036,620	147,553,416	50.7	0.95



**Figure 10. Correlations and fragment sizes of cDNA libraries (Raji B cells).**

(a) TT-seq experiments were performed from two independent biological replicates for control (DMSO, 15 min), or CDK9<sup>as</sup> inhibition (5  $\mu$ M 1-NA-PP1, 15 min) in Raji B (CDK9<sup>as</sup>) cells (**Table 21**). Scatter plots show the density of reads of 11,384 RefSeq GRCh38 annotated genes. Both axes depict reads per kilobase (RPK). The color scale corresponds to the density of points. Correlation of replicates is calculated as Spearman's rank correlation coefficient ( $\rho$ ) rounded to second decimal place (shown in each plot). (b) TT-seq experiments were performed from two independent biological replicates for control (DMSO, 15 min), or CDK9<sup>as</sup> inhibition (5  $\mu$ M 1-NA-PP1, 15 min) in Raji B wildtype (WT) cells (**Table 21**). Scatter plots show the density of reads of 10,363 RefSeq GRCh38 annotated genes. (c) mNET-seq (with Empigen BB treatment) experiments were performed from two independent biological replicates for control (DMSO, 15 min), or CDK9<sup>as</sup> inhibition (5  $\mu$ M 1-NA-PP1, 15 min) in Raji B (CDK9<sup>as</sup>) cells (**Table 21**). Scatter plots show the density of reads of 10,040 RefSeq GRCh38 annotated genes. (a-c) Right: size distributions [nt] of library inserts.

## 2.4 The pause-initiation limit restricts transcription activation

---

The methods presented in section 2.4 are currently under peer review for publication:

**S. Gressel \***, B. Schwalb \* *et al.* The pause-initiation limit restricts transcription activation in human cells. *Nature communications, in revision.*

(\* joint first authorship)

---

### 2.4.1 Cell treatment: heat shock

*To determine the optimal time-points for the multi-omics analysis, I performed a time-course in which K562 or Raji B (CDK9<sup>as</sup>) cells were exposed to 42 °C in a water bath. Transcriptional changes were analyzed by RT-qPCR (see subsection 2.1.2), and cell viability was checked by trypan blue exclusion assay (see below) and MTS assay (data not shown).*

**Heat shock treatment.** To avoid transcriptional changes by freshly added growth medium <sup>427</sup>, fresh growth medium was added ~24 h prior to heat shock treatments. Heat shock treatments of K562 or Raji B (CDK9<sup>as</sup>) cells were performed in T175 flasks in a volume of 50 mL at 0.6 x 10<sup>6</sup> cells/mL in a water bath at 42 °C. Temperature was monitored by thermometer. It took 5 min until the cell suspension reached 42 °C. For RT-qPCR and cell viability assessment, cells were treated for a time-course of 0 to 75 min. For TT-seq, RNA-seq and mNET-seq experiments, cells were treated for 0 (Ctrl), 15 (HS15) or 30 min (HS30).

**Trypan blue exclusion.** Cell viability levels were evaluated by the trypan blue exclusion method as described by <sup>428</sup>. Cell viability assessment was performed with two biological replicates. Briefly, 5 x 10<sup>5</sup> cells were treated as above for 0, 15, 30, 45, 60, or 75 min. Cells were pelleted at 200 x g for 5 min and resuspended in 1 mL DPBS prior to counting. Equal volumes of cell suspension and 0.4 % trypan blue solution were mixed and incubated for 2 min. The solution was applied to a hemacytometer and viable cells were counted using light microscopy. For each treatment time, the cell count was duplicated (dilution factor for trypan blue) and the average value was obtained. Cell viability was calculated as the ratio of viable cells upon heat shock (15-75 min) to viable cells of control (0 min).

### 2.4.2 Multi-omics experiments

*TT-seq and mNET-seq experiments were performed as described (section 2.2). I performed all TT-seq, RNA-seq and mNET-seq experiments in this study (Table 23).*

For TT-seq pre-processing and global spike-in normalization, as well as mNET-seq pre-processing see Supplementary Information V.2.1.

**Table 23. Information on experimental conditions used in this study.**Abbreviations used: control (Ctrl), heat shock (HS), analog sensitive CDK9 (CDK9<sup>as</sup>).

No.	Assay	Cell type	Condition name	Replicate no.	Treatment
1	mNET-seq	K562	Ctrl	1	-
2	mNET-seq	K562	Ctrl	2	-
3	mNET-seq	K562	HS15	1	42 °C, 15 min.
4	mNET-seq	K562	HS15	2	42 °C, 15 min.
5	mNET-seq	K562	HS30	1	42 °C, 30 min.
6	mNET-seq	K562	HS30	2	42 °C, 30 min.
7	TT-seq	K562	Ctrl	1	-
8	TT-seq	K562	Ctrl	2	-
9	TT-seq	K562	HS15	1	42 °C, 15 min.
10	TT-seq	K562	HS15	2	42 °C, 15 min.
11	TT-seq	K562	HS30	1	42 °C, 30 min.
12	TT-seq	K562	HS30	2	42 °C, 30 min.
13	RNA-seq	K562	Ctrl	1	-
14	RNA-seq	K562	Ctrl	2	-
15	RNA-seq	K562	HS15	1	42 °C, 15 min.
16	RNA-seq	K562	HS15	2	42 °C, 15 min.
17	RNA-seq	K562	HS30	1	42 °C, 30 min.
18	RNA-seq	K562	HS30	2	42 °C, 30 min.
19	TT-seq	Raji B (CDK9 <sup>as</sup> )	Ctrl Solvent	1	0.05 % v/v DMSO, 35 min.
20	TT-seq	Raji B (CDK9 <sup>as</sup> )	Ctrl Solvent	2	0.05 % v/v DMSO, 35 min.
21	TT-seq	Raji B (CDK9 <sup>as</sup> )	Ctrl CDK9 inhibited	1	5 μM 1-NA-PP1, 35 min.
22	TT-seq	Raji B (CDK9 <sup>as</sup> )	Ctrl CDK9 inhibited	2	5 μM 1-NA-PP1, 35 min.
23	TT-seq	Raji B (CDK9 <sup>as</sup> )	HS Solvent	1	0.05 % v/v DMSO, 35 min; after 5 min: 42 °C, 30 min.
24	TT-seq	Raji B (CDK9 <sup>as</sup> )	HS Solvent	2	0.05 % v/v DMSO, 35 min; after 5 min: 42 °C, 30 min.
25	TT-seq	Raji B (CDK9 <sup>as</sup> )	HS CDK9 inhibited	1	5 μM 1-NA-PP1, 35 min; after 5 min: 42 °C, 30 min.
26	TT-seq	Raji B (CDK9 <sup>as</sup> )	HS CDK9 inhibited	2	5 μM 1-NA-PP1, 35 min; after 5 min: 42 °C, 30 min.

**mNET-seq data pre-processing and normalization.** We first identified a subgroup of RefSeq-TUs with unchanged behavior over the response to heat shock in the spike-ins normalized TT-seq data via k-means clustering. On the resulting 3,416 RefSeq-TUs  $i$ , size factors for each sample  $j$  were determined as

$$\sigma_j = \operatorname{median}_i \left( \frac{p_{ij}}{\left(\prod_{v=1}^m p_{ij}\right)^{1/m}} \right)$$

where  $m$  denotes the total number of antisense corrected mNET-seq samples ( $p_{ij}$ ). This formula has been adapted<sup>422</sup> and was used to correct for library size and sequencing depth variations.

### 2.4.3 Heat shock treatment response ratio

**Calculation of response ratios (for calibration of pause duration).** For each condition  $j$  (control or heat shock 15 minutes) the antisense bias corrected number of transcribed bases  $tb_i^j$  was calculated on all expressed RefSeq-TUs  $i$  (exceeding 35 kbp in length with one unique TSS given all Refseq annotated isoforms (UCSC RefSeq GRCh38)). Response ratios were calculated for a window from the TSS to 10 kbp downstream (excluding the first 200 bp) for each RefSeq-TU  $i$  as

$$r_i = 1 - \frac{tb_i^{heat\ shock}_{[0.2, 10\ kbp]}}{tb_i^{control}_{[0.2, 10\ kbp]}}$$

where negative values were set to 0 and values above 1 were set to 1.

**Estimation of gene-wise elongation velocity (for calibration of pause duration).** For each condition  $j$  (control or heat shock 15 minutes) the antisense bias corrected number of transcribed bases  $tb_i^j$  was calculated on all expressed RefSeq-TUs  $i$  (exceeding 35 kbp in length with one unique TSS given all Refseq annotated isoforms (UCSC RefSeq GRCh38)), excluding the first 200 bp. All TUs were truncated by 5 kbp in length from the 3' end prior to calculation to avoid influence of some alterations in signal around the pA site after heat shock. For each TU  $i$  with  $r_i > 0.1$  the elongation velocity  $v_i$  [kbp/min] was calculated as

$$v_i = \frac{tb_i^{control} - tb_i^{heat\ shock}}{tb_i^{control} \cdot \frac{r_i}{L_i} (t^* - t)}$$

with heat shock treatment duration  $t^* = 15$  [min] and labeling duration  $t = 5$  [min].

#### 2.4.4 Major transcript classes annotation

*Annotation was performed by Dr. Björn Schwalb (MPI-bpc, Dept. Molecular Biology) based on <sup>168,177</sup>.*

**Transcription units based on UCSC RefSeq genome assembly.** For each annotated gene, transcription units were defined as the union of all existing inherent transcript isoforms (UCSC RefSeq GRCh38).

**Definition of isoform-independent exonic regions (constitutive exons).** Isoform-independent exonic regions were determined using a model for constitutive exons <sup>429</sup> based on UCSC RefSeq annotation (GRCh38).

**Transcription units based on GenoSTAN (TUs, K562 and Raji).** Annotation of different transcript classes was done as in <sup>168</sup> with minor differences. In brief, genome-wide coverage was calculated from all TT-seq fragment midpoints in consecutive 200 bp bins throughout the genome. In order to create a comprehensive annotation independent of heat shock induced length differences, two replicate tracks were constructed by taking the maximum of each bin over the first and second replicates, respectively, regardless of treatment. A two-state hidden Markov model with a Poisson-Log-Normal emission distribution was 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 a validated GENCODE mRNA or lincRNA (version 22) or showed uninterrupted coverage supported by all TT-seq samples. Subsequently, TU start and end sites were refined to nucleotide precision by finding borders of abrupt coverage increase or decrease between two consecutive segments in the four 200 bp bins located around the initially assigned start and stop sites via fitting a piecewise constant curve to the TT-seq coverage profiles for both replicates using the segmentation method from the R/Bioconductor package "tilingArray" <sup>430</sup>.

**GRO-cap TSS refinement of TUs (cTUs, K562).** For all TUs  $i$ , the GRO-cap refined transcription start site  $tss^*$  was determined as the closest non-zero GRO-cap signal <sup>50</sup> in a window of 500 bp around the start of the TUs. Note that all TUs without an assigned GRO-cap site were not used. It was recently shown that upon 1 h heat shock TSS architecture remains mostly unchanged <sup>247</sup>. Thus, we assume changes of TSS architecture upon 30 min of heat shock to be insignificant.

**Transcript sorting (K562 and Raji).** We sorted each gene (cTU for K562, TU for Raji) into one of the following seven classes: eRNA, sincRNA, asRNA, conRNA, uaRNA, lincRNA and mRNA. First, (c)TUs reciprocally overlapping by at least 50% with a validated GENCODE mRNA or lincRNA (version 22) on the same strand were classified as mRNAs and lincRNAs. (c)TUs reciprocally overlapping by less than 50% with a validated GENCODE mRNA or lincRNA (version 22) on the same strand were not classified. Next, (c)TUs located on the opposite strand of either a mRNA or lincRNA were classified as asRNA if the TSS was located  $> 1$  kbp downstream of the sense TSS on the opposite strand, as uaRNA if its TSS was located  $< 1$  kbp upstream of the sense TSS, and as conRNA if its TSS was located  $< 1$  kbp downstream of the TSS on the opposite strand. For K562, each of the remaining cTUs was

classified as sincRNA. Every ncRNA (sincRNA, asRNA, conRNA or uaRNA) was re-classified as eRNA if its TSS fell into a K562 enhancer state<sup>177</sup>. This resulted in 11,324 non-ambiguously classified RNAs in K562 cells (**Table 24**). For Raji, each of the remaining TUs was classified into eRNA – if its TSS exhibited a high (>1) ratio H3K4me1/H3K4me3 – or as sincRNA – if its TSS exhibited a low (<1) ratio of H3K4me1/H3K4me3. This resulted in 16,452 non-ambiguously classified RNAs (**Table 24**).

**Table 24. Annotation of RNA classes in human hematopoietic cell lines (K562, Raji B).**

For details see Methods and **Figure 1 e**. Abbreviations: capped (c) transcription unit (TU).

Class	Class	Definition	cTUs in K562	TUs in Raji B
<b>mRNA</b>	messenger RNA	GENCODE annotation (>50 % overlap)	6,549	6,990
<b>lincRNA</b>	intergenic RNA	GENCODE annotation (>50 % overlap); promoter state	239	243
<b>asRNA</b>	antisense RNA	located on opposite strand of mRNA or lincRNA, TSS located >1 kbp downstream of sense TSS; promoter state	564	1,398
<b>uaRNA</b>	upstream antisense RNA	located on opposite strand of mRNA or lincRNA, TSS located <1 kbp upstream of sense TSS; promoter state	1,064	565
<b>conRNA</b>	convergent RNA	located on opposite strand of mRNA or lincRNA, TSS located <1 kbp downstream of sense TSS; promoter state	502	326
<b>sincRNA</b>	short intergenic noncoding RNA	promoter state	1,581	3,479
<b>eRNA</b>	enhancer RNA	location as asRNA, uaRNA, or conRNA; enhancer state	825	3,451
<b>Total</b>			11,324	16,452

#### 2.4.5 Illumina sequencing and data availability

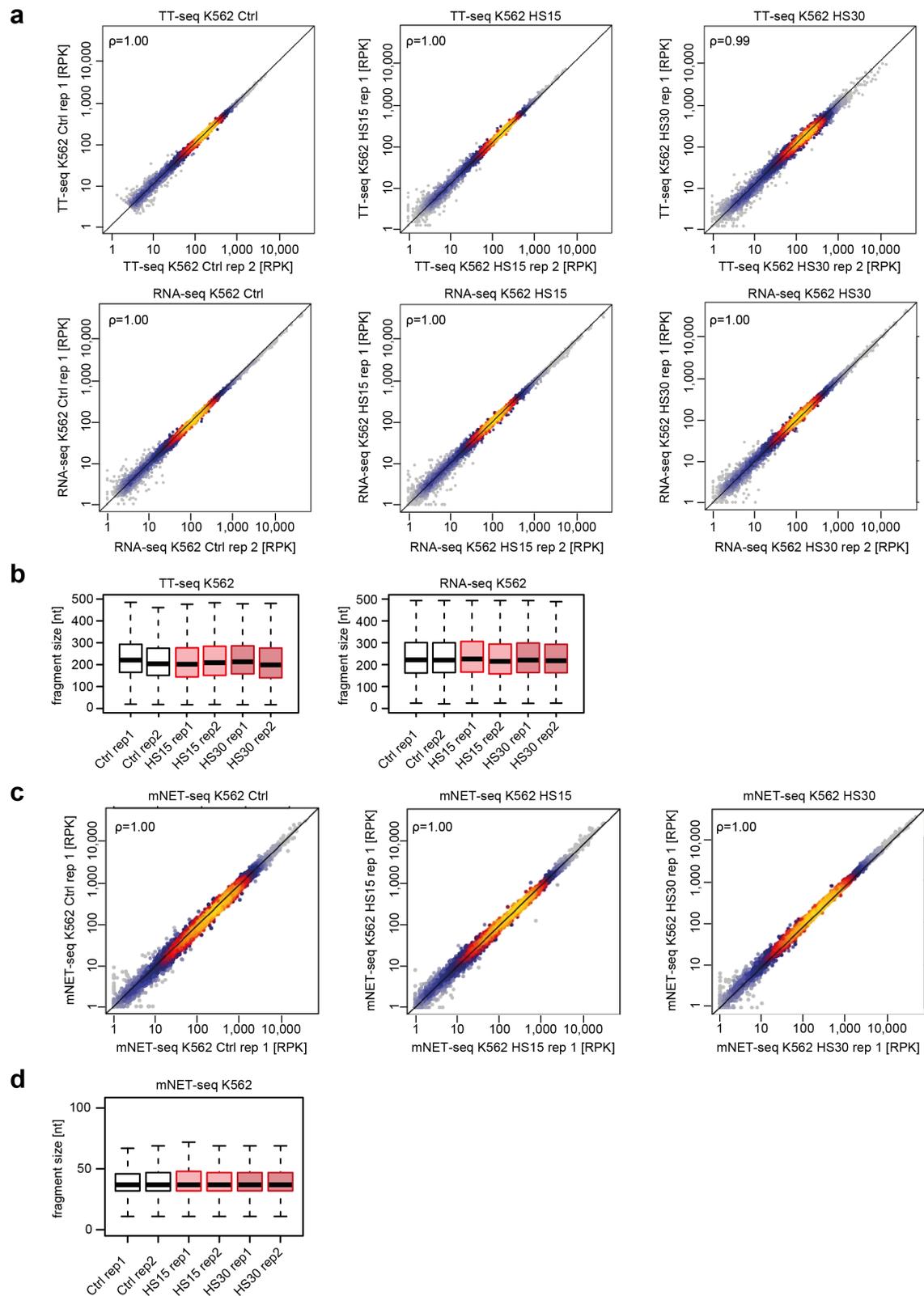
*For this study, all Illumina sequencing was performed in-house by Kerstin Maier and Petra Rus (MPI-bpc, Dept. of Molecular Biology).*

The sequencing data and processed files were deposited in the GEO database under accession code GSE123980. Transcript annotations were deposited as supplementary GTF file (GSE123980\_transcript.annotation). A detailed overview of sequencing data and processed files is depicted in **Table 25**. Experimental conditions are listed in **Table 23**.

**Table 25. Sequencing statistics of 26 libraries generated in this study.**

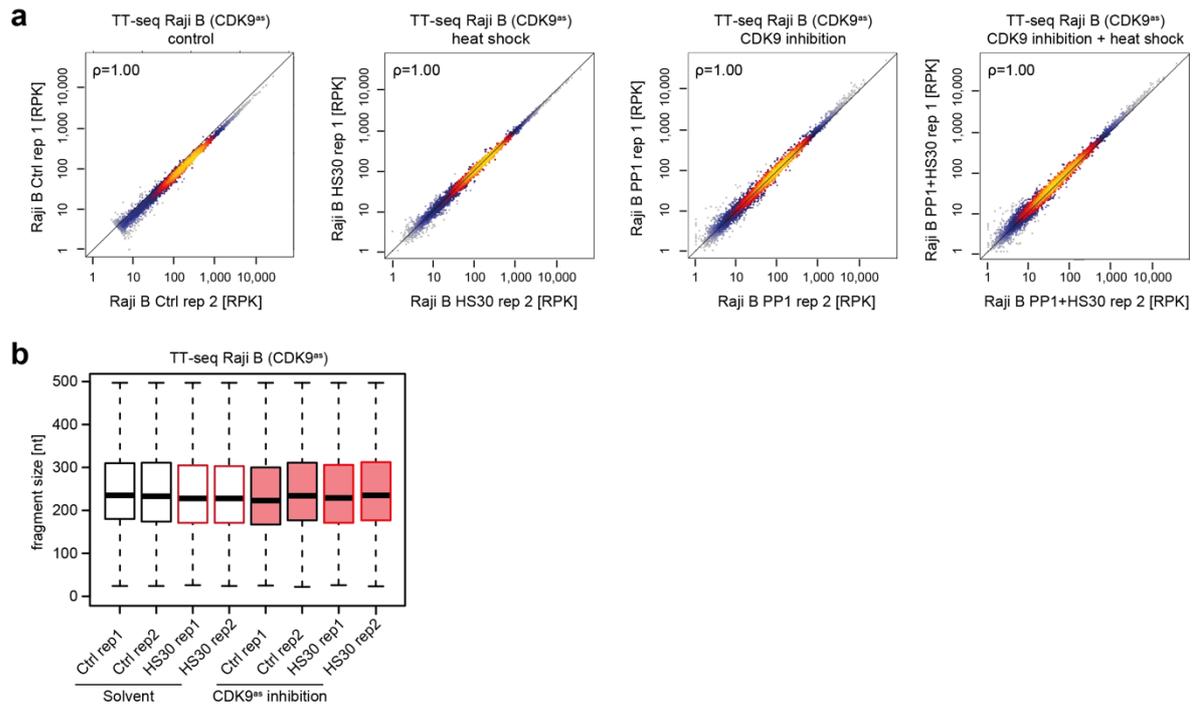
All libraries were sequenced on a NEXTseq 550 sequencing platform in 75 bp paired-end mode. Numbers refer to experimental conditions listed in **Table 23**. Correlation of replicates is calculated as Spearman's rho rounded to second decimal place. Sequenced fragments were determined by FastQC, uniquely mapped fragments were counted after mapping using STAR 2.3.0 <sup>403</sup>.

No.	Barcode	GEO no.	Fragment numbers			Duplicates (%)	Replicate correlation (RPKs)
			Sequenced	Mapped uniquely	Replicates combined		
1	CGGAAT	GSM3518117	159,190,513	118,817,488		54.1	
2	CTAGCT	GSM3518118	173,587,084	130,654,902	249,472,390	54.5	1.00
3	GACGAC	GSM3518119	164,964,729	117,933,814		48.0	
4	TAATCG	GSM3518120	151,647,527	111,634,549	229,568,363	44.3	1.00
5	TCATTC	GSM3518121	202,842,209	151,118,454		52.9	
6	TCCCGA	GSM3518122	180,695,911	134,847,488	285,965,942	47.5	1.00
7	AAGCCT	GSM3518105	149,673,334	134,367,932		7.2	
8	GTCGTA	GSM3518106	151,253,343	136,351,777	270,719,709	11.5	1.00
9	AAGAGG	GSM3518107	161,455,492	145,506,316		8.4	
10	GGAGAA	GSM3518108	157,619,872	141,621,200	287,127,516	11.3	1.00
11	AGCATG	GSM3518109	147,638,717	132,928,558		8.1	
12	GAGTCA	GSM3518110	170,970,259	153,960,400	286,888,958	16.2	0.99
13	AACCAG	GSM3518111	102,213,680	84,822,615		35.6	
14	TGGTGA	GSM3518112	107,355,413	89,846,989	174,669,604	33.7	1.00
15	AGTGAG	GSM3518113	100,709,060	84,896,024		35.4	
16	GCACTA	GSM3518114	105,407,471	88,182,473	173,078,497	41.8	1.00
17	CACAGT	GSM3518115	108,465,109	90,694,343		35.6	
18	TTGGCA	GSM3518116	105,327,756	88,593,759	179,288,102	41.2	1.00
19	AACCAG	GSM3518123	57,249,175	52,426,514		5.3	
20	TGGTGA	GSM3518124	76,191,286	69,811,550	122,238,064	6.3	1.00
21	AGTGAG	GSM3518125	70,801,390	64,421,316		15.1	
22	GCACTA	GSM3518126	67,661,931	61,817,282	126,238,598	10.2	1.00
23	ACCTCA	GSM3518127	70,180,248	64,629,409		6.5	
24	GTGCTT	GSM3518128	74,373,094	68,428,032	133,057,441	8.5	1.00
25	AAGCCT	GSM3518129	82,625,021	75,042,304		21.7	
26	GTCGTA	GSM3518130	77,366,563	70,613,758	145,656,062	14.4	1.00



**Figure 11. Correlations and fragment sizes of cDNA libraries (K562 cells).**

TT-seq (a, top; b, left), RNA-seq (a, bottom; b, right), and mNET-seq (with Empigen BB treatment) (c, d) experiments were performed from two independent biological replicates for control (37 °C), or heat shock (42 °C) conditions for 15 min (HS15) and 30 min (HS30) (Table 23). (a, c) Scatter plots with color-coded density of reads for 12,315 RefSeq GRCh38 annotated genes. Both axes depict reads per kilobase (RPK). The color scale corresponds to the density of points. Correlation of replicates is calculated as Spearman's rank correlation coefficient ( $\rho$ ) rounded to second decimal place (shown in each plot). (b, d) Size distributions of library inserts of TT-seq, RNA-seq (b) and mNET-seq (d) in K562 cells treated as described in (a).



**Figure 12. Correlations and fragment sizes of TT-seq libraries (Raji B).**

(a) TT-seq experiments were performed from two independent biological replicates for control (37 °C), or heat shock conditions for 30 min (42 °C) with solvent (DMSO, 35 min) or CDK9 inhibition (5  $\mu$ M 1-NA-PP1, 35 min) (**Table 23**). Scatter plots show the density of reads of 10,294 RefSeq GRCh38 annotated genes. Both axes depict reads per kilobase (RPK). The color scale corresponds to the density of points. Correlation of replicates is calculated as Spearman's rank correlation coefficient ( $\rho$ ) rounded to second decimal place (shown in each plot). (b) Size distributions of library inserts of TT-seq in Raji B (CDK9<sup>as</sup>) treated as described in (a).

## III. Results

### 1. Establishing mNET-seq in human suspension cells

The mNET-seq protocol can be divided into five main steps: (i) cellular fractionation, (ii) chromatin solubilization, (iii) immunoprecipitation of polymerase (Pol) II, (iv) RNA 3' end phosphorylation and size selection, (v) preparation of cDNA libraries and Illumina sequencing. In an effort to establish a reproducible mNET-seq protocol in human hematopoietic suspension cells (K562 and Raji B), I revisited most steps of the protocol published by the Proudfoot lab <sup>350,388</sup> (**Figure 13 a**).

Here, I briefly summarize the critical steps (i-iv) which have been optimized for mNET-seq application in human hematopoietic suspension cells.

**Table 26. Comparison of three cellular fractionation protocols.**

Step	Mayer's protocol <sup>431</sup>	Ørom's protocol <sup>418</sup>	Nojima's protocol <sup>388</sup>
<b>Mild cell lysis</b>	10 mM Tris-HCl pH 7.0 150 mM NaCl 0.15 % v/v NP-40	10 mM Tris-HCl pH 7.4 150 mM NaCl 0.15 % v/v NP-40	10 mM Tris-HCl pH 7.5 10 mM NaCl 0.5 % v/v NP-40 2.5 mM MgCl <sub>2</sub>
<b>Sucrose cushion</b>	10 mM Tris-HCl pH 7.0 150 mM NaCl 25 % w/v sucrose	10 mM Tris-HCl pH 7.4 150 mM NaCl 24 % w/v sucrose	10 mM Tris-HCl pH 7.4 10 mM NaCl 10 % w/v sucrose 0.5 % NP-40 2.5 mM MgCl <sub>2</sub>
<b>Glycerol</b>	20 mM Tris-HCl pH 8.0 75 mM NaCl 0.5 mM EDTA 50 % v/v glycerol	20 mM Tris-HCl pH 7.4 75 mM NaCl 0.5 mM EDTA 50 % v/v glycerol	20 mM Tris-HCl pH 7.9 75 mM NaCl 0.5 mM EDTA 50 % v/v glycerol
<b>Nuclear lysis</b>	20 mM HEPES pH 7.5 300 mM NaCl 1 M urea 0.2 mM EDTA 1 % v/v NP-40	10 mM Tris-HCl pH 7.4 300 mM NaCl 1 M urea 0.2 mM EDTA 1 % v/v NP-40 7.5 mM MgCl <sub>2</sub>	20 mM HEPES-KOH pH 7.6 300 mM NaCl 1 M urea 0.2 mM EDTA 1 % v/v NP-40 7.5 mM MgCl <sub>2</sub>
<b>Processing time</b>	~45 min	~45 min	~60 min

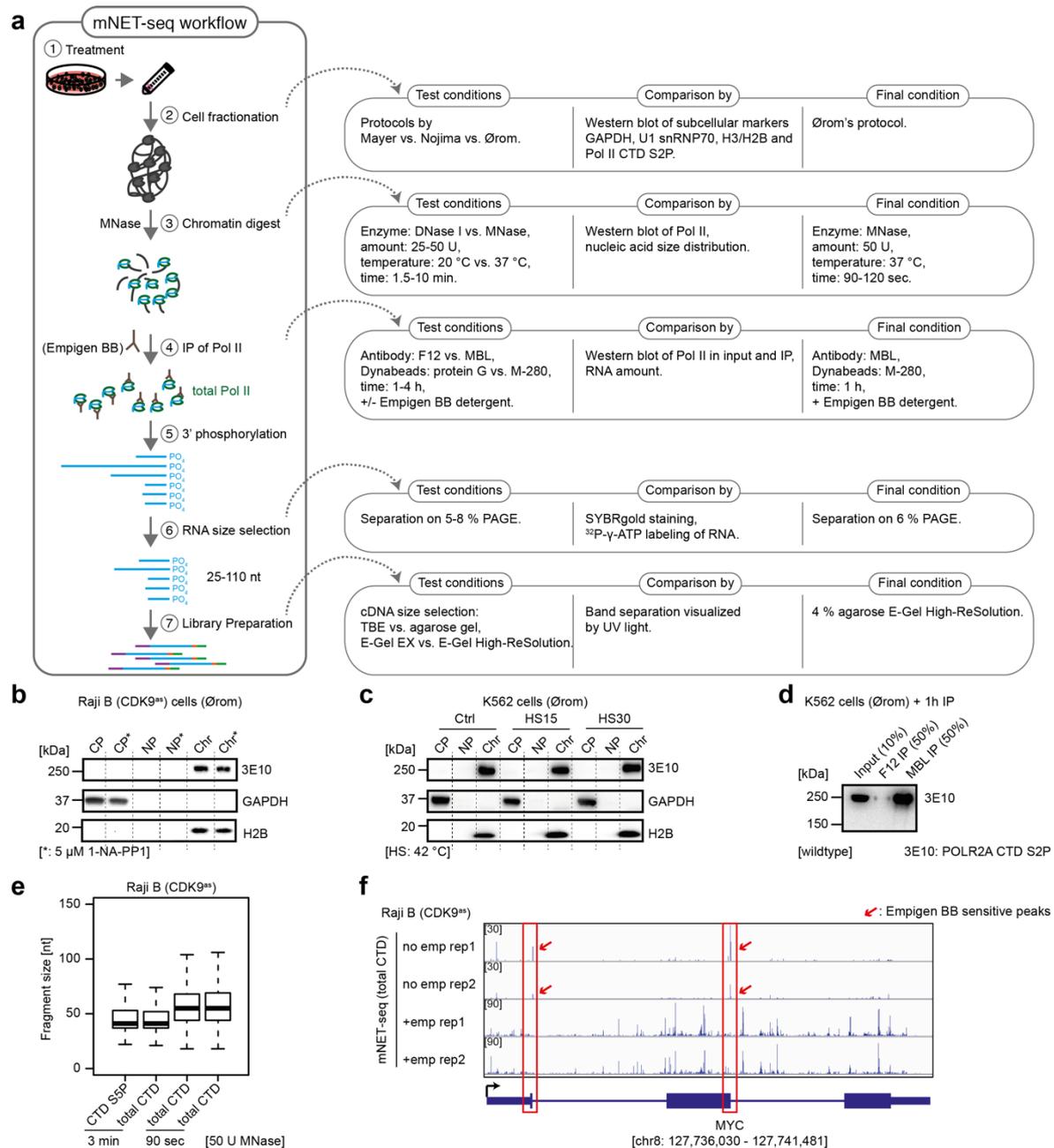
*Cellular fractionation.* Due to differences in cell size and chromatin compaction of adherent and suspension cells, the efficiency of a cellular fractionation protocol differs. In general, cells are lysed with mild detergent (NP-40), and centrifuged through a sucrose cushion to yield purified nuclei. Then, the nuclear envelop is disrupted by NP-40 and urea, and the chromatin is precipitated by centrifugation<sup>390</sup>. I compared three fractionation protocols (**Table 26**) with respect to processing time, quality and reproducibility of the fractionation. The quality and reproducibility were verified by Western blotting of marker proteins (**Figure 38 a**). One protocol failed to resolve clean fractions of hematopoietic cells (**Figure 38 b**) while the other two performed equally well. Since short processing time was also an important aspect, Ørom's protocol<sup>418</sup> worked best in steady state and after perturbations (CDK9 inhibition, or heat shock) (**Figure 13 b-c**).

*Chromatin solubilization.* Cellular fractionation is followed by solubilization of the chromatin fraction by a DNase. Micrococcal nuclease (MNase) has the propensity to cleave double stranded DNA, within internucleosomal linker regions, and single stranded RNA<sup>432-434</sup> resulting in solubilization of Pol II-RNA complexes from chromatin. However, overly harsh MNase digestion affects nascent RNA length. I found that short incubation times (<180 sec) with 50 U MNase yield sufficient solubilization of Pol II-RNA complexes, and RNA fragment lengths corresponding to the desired RNA size range of 25 to 110 nt (**Figure 13 e**).

*Immunoprecipitation (IP) of POLR2A with Empigen BB.* I tested two commercial antibodies for the subsequent IP step. Both immunoprecipitate total Pol II by targeting the largest subunit of human Pol II, POLR2A. Specifically, one binds to the N-terminus of POLR2A (Santa Cruz, clone F12), and the other targets the unphosphorylated and phosphorylated CTD of POLR2A (MBL, clone CMA601) (**Table 11**). By comparing the IP efficiency, the MBL antibody performed with ~10-fold higher yield compared to the F12 antibody (**Figure 13 d**). Thus, the MBL antibody was used for IP. A crucial improvement for the IP step was the addition of the strong detergent Empigen BB to the IP reaction and subsequent washing steps to weaken protein-protein interactions of Pol II with RNA processing complexes. Consistent with observations by recent studies<sup>391,392</sup>, strong peaks at the end of exons were lost upon Empigen BB addition (**Figure 13 f**). These peaks originate from non nascent RNA intermediates with 3'-OH groups such as spliceosome or microprocessor complex associated RNAs.

*RNA size selection.* I tested different acrylamide percentages (5-8 %) for RNA size selection with the aim to monitor RNA size distribution by the migration of two blue dyes (bromophenol blue and xylene cyanol blue), without radioactive labeling. The use of 6 % denaturing urea polyacrylamide gel electrophoresis (PAGE) allowed to select the target size (25-110 nt) by cutting between the two dyes (**Figure 39**). A few month ago the Proudfoot lab also decreased the acrylamide concentration from 8 % (used in 2015 to 2017<sup>350,388,391</sup>) to 6 %<sup>392,435</sup> in order to monitor the RNA size by the two tracking dyes.

The optimized mNET-seq protocol for suspension cells is described in detail in the methods section II.2.2.3.



**Figure 13. Establishing mNET-seq of total Pol II in human hematopoietic cell lines.**

(a) Left: experimental workflow of mNET-seq. Right: several steps were optimized for application in hematopoietic cells. (b-c) Western blot analysis (INTAS imager) of cytoplasm (CP), nucleoplasm (NP) and solubilized chromatin (Chr) fractions (Ørom's protocol<sup>418</sup>) from Raji B (CDK9<sup>as</sup>) (b) or K562 cells (c) using indicated antibodies. Antibodies are listed in Materials II.1.4. (b) Cells were treated with solvent control (DMSO) or CDK9 inhibition (5  $\mu$ M 1-NA-PP1) for 15 min prior to fractionation (Table 21, no. 9, 11). Chromatin of Raji B (CDK9<sup>as</sup>) cells was digested for 90 sec. (c) Cells were subjected to standard growth conditions (Ctrl), 15 or 30 min of heat shock (HS) at 42 °C prior to fractionation (Table 23, no. 1, 3, 5). Chromatin of K562 cells was digested for 2 min. (d) Comparison of two POLR2A antibodies after 1 h of IP using digested chromatin of 10<sup>7</sup> K562 cells as input. Uncropped images can be found in Figure 40. (e) Library insert size after sequencing of mNET-seq libraries generated in Raji B (CDK9<sup>as</sup>) cells. Left: trials using 3 min digest time. Right: final conditions using 90 sec of MNase digest. (f) Representative genome browser view of mNET-seq data at the MYC gene locus on chromosome 8 (plus strand) (IGV, version 2.4.10; human hg38)<sup>400</sup>. Shown is the strand-specific Pol II occupancy with single nucleotide resolution. Top panels: two biological replicates of mNET-seq in Raji B (CDK9<sup>as</sup>) cells without Empigen BB. Bottom panels: two biological replicates of mNET-seq in Raji B (CDK9<sup>as</sup>) cells with 1 % v/v Empigen BB (~30 % active substance) treatment. Empigen BB sensitive peaks are marked by red arrows.

---

## 2. CDK9-dependent Pol II pausing controls transcription initiation

---

The results presented in chapter 2 have been published:

**S. Gressel** \*, B. Schwalb \* *et al.* CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* 6, doi:10.7554/eLife.29736 (2017).

(\* joint first authorship)

A detailed list of author contributions can be found on page VII.

---

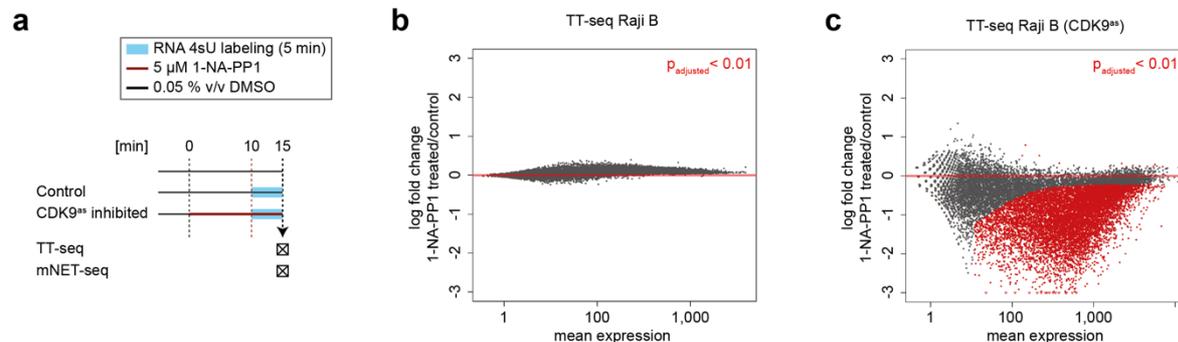
### Chapter summary

Gene transcription can be activated by decreasing the duration of RNA polymerase II pausing in the promoter-proximal region, but how this is achieved remains unclear. Here we use a ‘multi-omics’ approach to demonstrate that the duration of polymerase pausing generally limits the productive initiation frequency of transcription initiation in human cells (‘pause-initiation limit’). We further engineer a human cell line to allow for specific and rapid inhibition of the P-TEFb kinase CDK9, which is implicated in polymerase pause release. CDK9 activity decreases the pause duration but also increases the productive initiation frequency. This shows that CDK9 stimulates release of paused polymerase and activates transcription by increasing the number of transcribing polymerases and thus the amount of mRNA synthesized per time. CDK9 activity is also associated with long-range chromatin interactions, suggesting that enhancers can influence the pause-initiation limit to regulate transcription.

### Acknowledgements

We would like to thank Helmut Blum and Stefan Krebs (LAFUGA, LMU Munich) for sequencing. We also thank Merle Hantsche (Max Planck Institute for Biophysical Chemistry) for structural modeling. We thank Julien Gagneur (Technical University of Munich) for initial discussions. HL was funded by SFB 1064 TP A17. DE was funded by SFB1064 (Chromatin Dynamics). PC was funded by Advanced Grant TRANSREGULON of the European Research Council and the Volkswagen Foundation.

## 2.1 CRISPR/Cas9 engineered mutation allows for specific CDK9 inhibition



**Figure 14. Inhibition of the human CDK9<sup>as</sup> kinase activity specific to CRISPR/Cas9 engineered cells.**

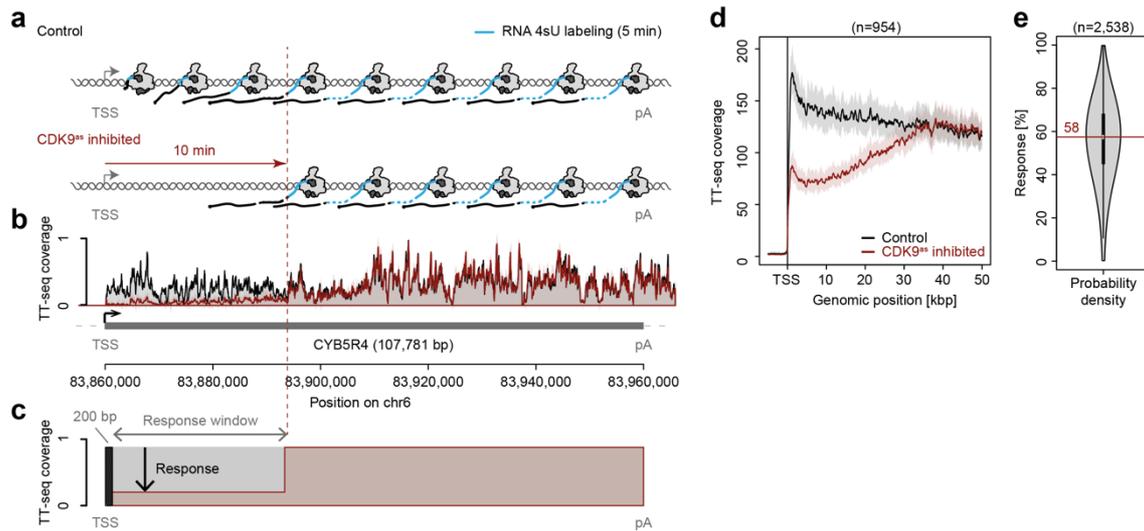
(a) Experimental design. TT-seq was carried out in Raji B (wildtype, or CDK9<sup>as</sup>) cells after treatment with 0.05 % v/v solvent DMSO (control) or 5  $\mu$ M 1-NA-PP1 (CDK9<sup>as</sup> inhibited) for 15 min (including 5 min of 4sU labeling). (b-c) Log fold change upon 1-NA-PP1 treatment (5  $\mu$ M for 15 min) versus normalized mean read count across two biological replicates and conditions for wildtype Raji B cells (b) and Raji B (CDK9<sup>as</sup>) cells (c). Significantly up- or downregulated genes (adjusted p-value < 0.01) are marked in red.

To specifically inhibit CDK9, we used a chemical biology approach<sup>338,339</sup> that circumvents off-target effects of standard CDK9 inhibitors<sup>337</sup>. We introduced a CDK9 analog sensitive mutation (CDK9<sup>as</sup>) into human Raji B cells by CRISPR/Cas9 (Methods II.2.3.1; **Figure 9**). This allows for rapid and highly specific CDK9 inhibition with the adenine analog 1-NA-PP1<sup>338,339</sup>, which does not have any effect on wild type cells (**Figure 14 b-c**). CDK9 protein levels were unchanged in CDK9<sup>as</sup> mutant cells compared to wild type cells (**Figure 41 a**). After 72 h of incubation with 1-NA-PP1, growth of CDK9<sup>as</sup> cells ceased, whereas wild type cells grew normally (**Figure 41 b**).

## 2.2 TT-seq monitors immediate response to CDK9 inhibition

We treated CDK9<sup>as</sup> cells with 5  $\mu$ M of 1-NA-PP1 for 10 min and monitored changes in RNA synthesis by TT-seq<sup>168</sup> (Methods II.2.2.2; **Figure 6 a**), using an RNA labeling time of 5 min (**Figure 15 a-b**). TT-seq data were highly reproducible (**Figure 10 a**; Spearman correlation coefficient 1) and monitored transcription activity before and after CDK9 inhibition (**Figure 15**). CDK9 inhibition resulted in reduced TT-seq signals at the beginning of genes, indicating that less Pol II was released into gene bodies (**Figure 15 c-e**). This gave rise to a ‘response window’ revealing the distance traveled by Pol II during 10 min inhibitor treatment (**Figure 15 c-d**). Downstream of the response window, the TT-seq signal was largely unchanged, indicating continued RNA synthesis from Pol II elongation complexes that had been released before CDK9 inhibition. To determine the relative response of genes to CDK9 inhibition, we calculated response ratios for those transcribed units based on the UCSC RefSeq genome assembly GRCh38 (RefSeq-TUs; referred to as ‘genes’ in chapter 2) (Methods II.2.3.4) that synthesized RNA, harbored a single TSS, and exceeded 10 kbp in length (2,538 genes). The response ratio of genes varied between 0 % to 100 % (fully responding genes) with a median

of 58 % (Figure 15 f). A remaining TT-seq signal in the response window likely reflects the proportion of polymerases that move to productive elongation without CDK9 kinase activity, but we cannot exclude that it stems from incomplete CDK9 inhibition. However, based on the assumption that the inhibitor is evenly distributed across cells and within, the portion of CDK9 that has not been fully inhibited must be very low (see also Figure 42).

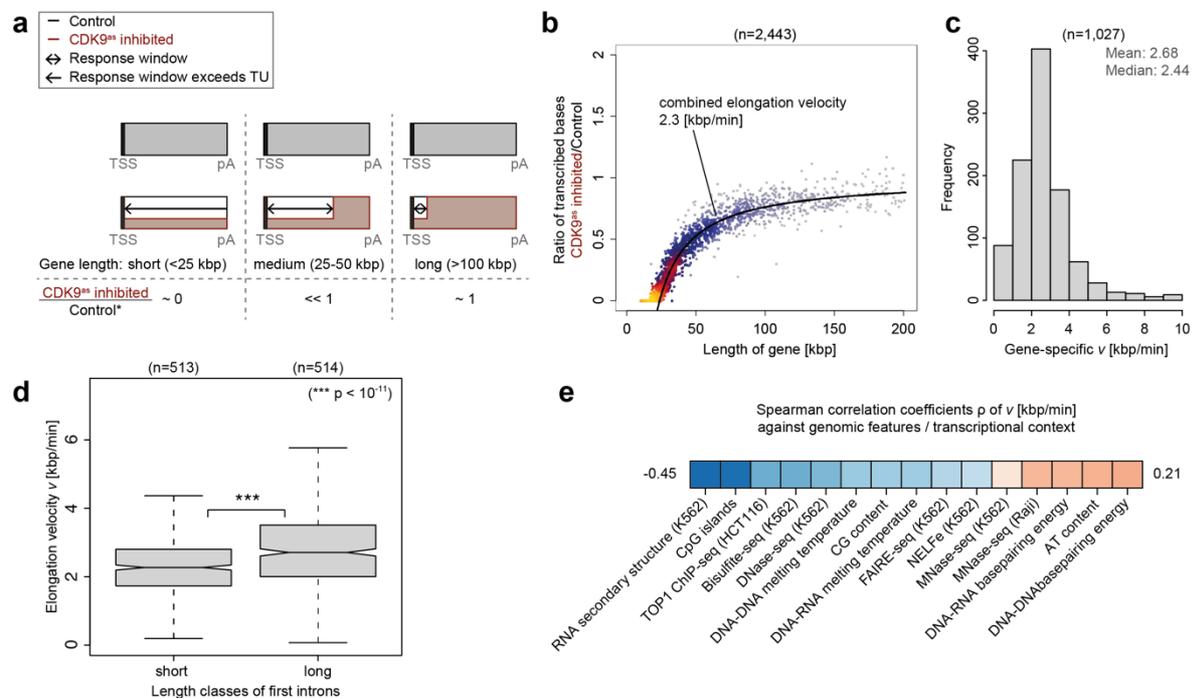


**Figure 15. Decrease of RNA synthesis in the 5'-region of genes upon CDK9 inhibition is gene-specific.**

(a) Experimental design. TT-seq was carried out in Raji B (CDK9<sup>as</sup>) cells after treatment with 0.05 % v/v solvent DMSO (control) or 5  $\mu$ M 1-NA-PP1 (CDK9<sup>as</sup> inhibited) for 15 min (including 5 min of 4sU labeling) (Table 21, Table 22, Figure 10). (b) TT-seq signal before (black) and after (red) CDK9 inhibition at the CYB5R4 gene locus (107,781 [bp]) on chromosome 6. Two biological replicates were averaged. The grey box depicts the transcript body from the transcription start site (TSS, black arrow) to the polyA site (pA). (c) Schematic representation of changes in TT-seq signal showing the definition of the response window. The decrease of RNA synthesis in the 5' region of a gene corresponds to the response [%]. Colors are as in (b). (d) Metagene analysis comparing the average TT-seq signal before and after CDK9 inhibition. The TT-seq coverage was averaged for 954 out of 2,538 investigated genes that exceed 50 [kbp] in length. Genes were aligned with their TSS. Shaded areas around the average signal (solid lines) indicate confidential intervals. (e) Violin plot showing the relative response to CDK9 inhibition for 2,538 investigated genes defined as  $1 - (\text{CDK9}^{\text{as}} \text{ inhibited} / \text{Control}) \cdot 100$  [%] for a window from the TSS to 10 [kbp] downstream, excluding the first 200 [bp] (c). The red line indicates the median response (58 %).

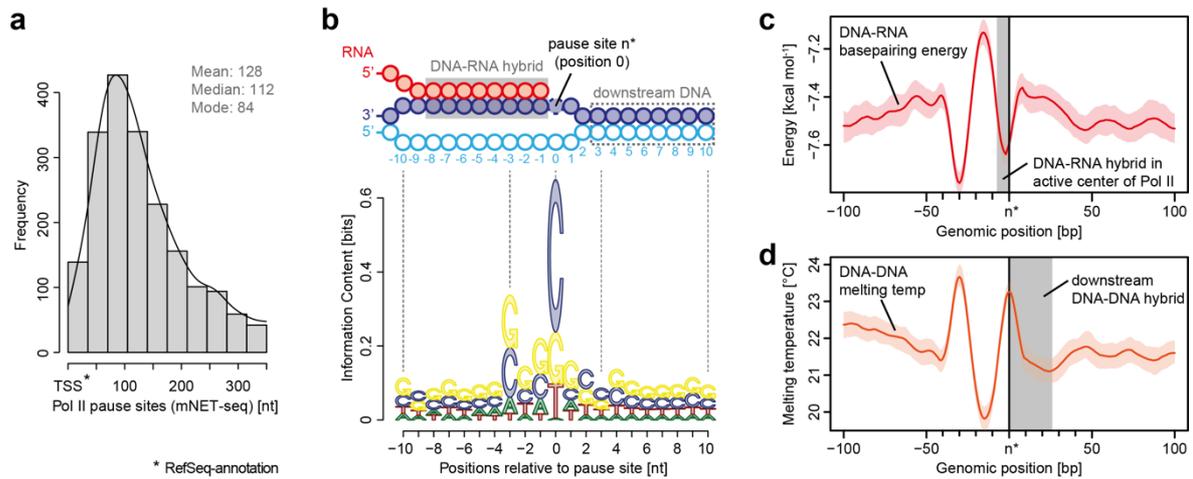
## 2.3 Pol II elongation velocity is gene-specific

The width of the response window differs between genes (**Figure 15 c-d**) and informs on Pol II elongation velocity (Methods II.2.2.4). The average width of the response window was 23 kbp which would correspond to an average elongation velocity of 2.3 kbp/min (**Figure 16 a**). This is in agreement with gene-specific elongation velocity measurements (**Figure 16 b**, combined 2.3 kbp/min; **Figure 16 c**, mean 2.7 kb/min, median 2.4 kb/min). This also agrees with previous estimates<sup>129,130,335,336</sup>. Gene-specific elongation velocities (**Figure 16 c**) were significantly higher in genes with longer first introns (**Figure 16 d**; Wilcoxon rank sum test,  $p$ -value  $< 1.916 \cdot 10^{-11}$ ), consistent with faster transcription of introns<sup>130</sup>. Elongation velocity correlated positively with nucleosome density, and negatively with the stability of the DNA-RNA hybrid, CpG density and topoisomerase occupancy (**Figure 16 e**, **Table 31**).



**Figure 16. Pol II elongation velocity is gene-specific.**

(a) Schematic representation of observed response window of TT-seq signal upon CDK9 inhibition (red) or control (black) for genes of three different length classes (short  $< 25$  [kbp], medium-length 25-50 [kbp] and long  $> 100$  [kbp]). (b) Scatter plot of the ratio of transcribed bases (CDK9<sup>as</sup> inhibited/control) against the length of the genes in nucleotides [kbp] revealed that the schematic representation in (a) holds true for 2,443 investigated genes. The ratio is corrected for response of individual genes. Modeling of the observed relation allows estimation of a robust combined elongation velocity of 2.3 [kbp/min] (solid black line). (c) Distribution of gene-wise elongation velocity (1,027 genes) depicted as a histogram (mean 2.7 [kbp/min], median 2.4 [kbp/min]). (d) Distribution of elongation velocity [kbp/min] depicted for 513 genes with short first intron ( $< 50$  % quantile, left) and 514 genes with long first intron ( $> 50$  % quantile, right). (e) Color encoded Spearman correlation coefficients (-0.45 in blue to 0.21 in red) of elongation velocity [kbp/min] against genomic features and measures of transcriptional context (all data sets are listed in **Table 31**). Note that the conclusions we draw across different cell lines are all based on metagene analysis, involving 500 up to more than 2,000 genes. Thus, we assume cell-line specific differences to have an insignificant influence and that the tendencies we observe rather suggest strong conservation.



**Figure 17. Characteristics of the promoter-proximal pause site**

(a) Histogram depicts the distribution of pause sites ( $n^*$ ) with respect to RefSeq annotated TSS for 2,135 investigated genes (mean 128 [bp], median 112 [bp], mode 84 [bp]). Two biological mNET-seq replicates were averaged (Table 21, Table 22, Figure 10). (b) Position weight matrix (PWM) logo representation of bases at positions  $-10$  to  $+10$  [bp] around the pause site (position 0,  $n^*$ ). (c) Mean basepairing energy [kcal mol<sup>-1</sup>] (right y-axis, in red) of the DNA-RNA hybrid aligned at the pause site  $n^*$ . The DNA-RNA hybrid in the active center of Pol II is highlighted in grey. (d) Mean melting temperature [°C] (left y-axis, in orange) of the DNA duplex aligned at the pause site  $n^*$ . The downstream DNA duplex is highlighted in grey. Shaded areas around the average signal (solid lines) indicate confidence intervals.

## 2.4 Characteristics of the promoter-proximal pause site

To study the kinetics of CDK9-dependent Pol II pause release, we generated mNET-seq data that map the RNA 3' end of engaged Pol II and extracted the position of paused polymerases (Methods II.2.2.3; Figure 6 b). mNET-seq data were highly reproducible (Figure 10 c; Spearman correlation coefficient  $>0.95$ ). Of the above genes, 2,135 (84 %) showed mNET-seq signal peaks above background. The called pause sites  $n^*$  were distributed around a maximum located 84 bp downstream of the RefSeq-annotated TSS (Figure 17 a). We asked how sequence content (DNA sequence at the pause site) and energetic properties (RNA-DNA hybrid in the active site of Pol II, downstream DNA duplex) contribute to pausing at human genes. At pause sites, we detected an enrichment for G/C-C/G dinucleotides with a strongly conserved cytosine at the RNA 3'-end (Figure 17 b). We observed a minimum of the predicted basepairing energy of the DNA-RNA hybrid immediately upstream of the pause site in the active site of Pol II (Figure 17 c, highlighted in grey). A weak DNA-RNA hybrid in the active center of Pol II is known to destabilize the elongation complex<sup>436</sup>, and could be a major determinant for establishing the paused state. The DNA duplex downstream of the pause site shows a local minimum of the DNA-DNA melting temperature but appears to have a minor impact on the pause (Figure 17 d, highlighted in grey).

## 2.5 Multi-omics analysis provides pause duration $d$ and initiation frequency $I$

To quantify pausing, we defined the pause duration  $d$  as the time a polymerase needs to pass through a 200 bp ‘pause window’ located  $\pm$  100 bp around the pause site. The pause duration  $d$  can now be derived from a combination of mNET-seq and TT-seq data. In particular, the mNET-seq signal corresponds to the number of polymerases in the pause window, which is determined by  $d$  and by the initiation frequency  $I$  (**Figure 18 a**)<sup>315</sup>. Thus,  $d$  is proportional to the ratio of the mNET-seq signal over  $I$ . To calculate  $I$  we integrated TT-seq signals over exons, excluding the first exon (Methods II.2.2.4). This provides the ‘productive initiation frequency’, i.e. the number of polymerases that initiate and successfully exit from the pause window. We use the term ‘productive’ because we do not know whether there is a small fraction of polymerases terminating within the pause window. Finally, to derive absolute values of  $d$ , we scaled the reciprocal of  $d$  (the elongation velocity in the pause window) according to the elongation velocity obtained from CDK9 inhibition.

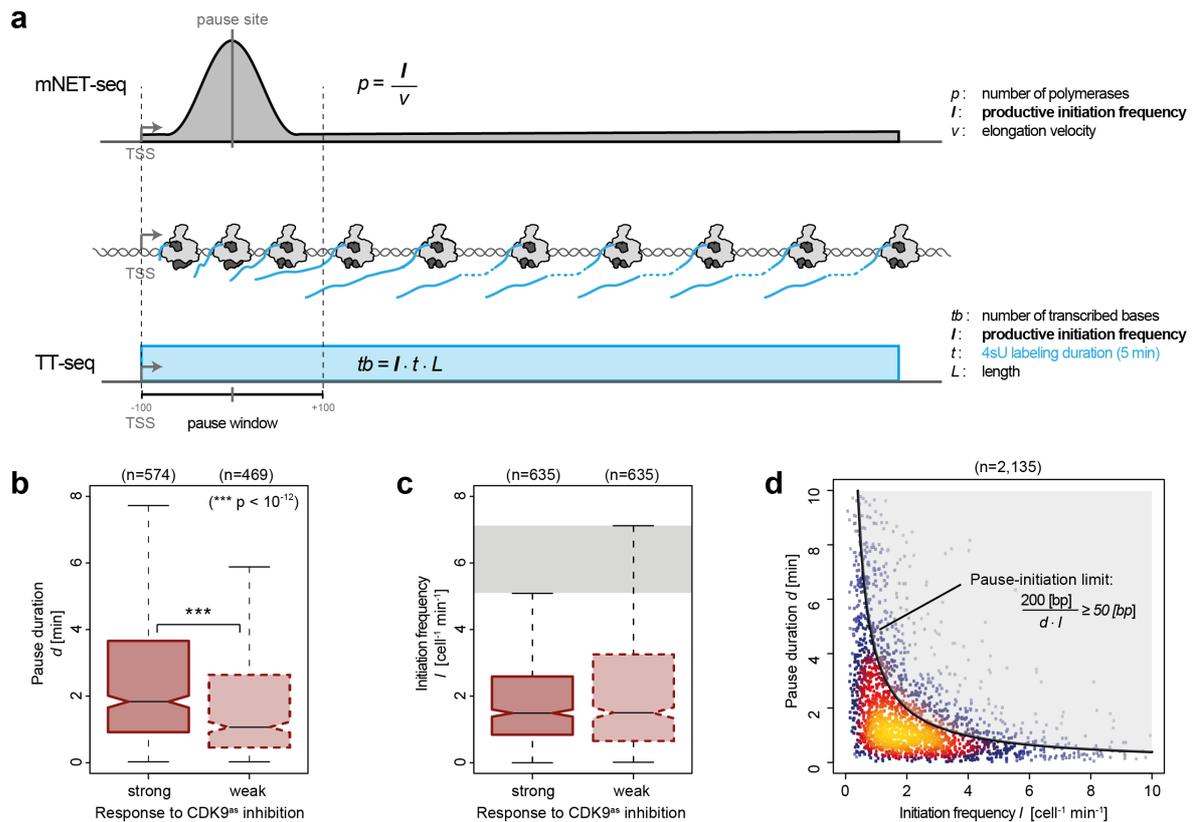
We obtained a mean productive initiation frequency of 2.7 polymerases cell<sup>-1</sup>min<sup>-1</sup>, and pause durations in the range of minutes, with strong variations between genes. The pause durations are generally consistent with reported half-lives of paused Pol II in mouse<sup>130</sup> and *Drosophila* cells<sup>355,357</sup> but slightly shorter. Pause durations were also consistent with kinetic modeling of TT-seq data alone. At genes with long pause durations we observed less labeled RNA in the short region between the TSS and the pause site. This confirms that indeed initiation frequencies are altered. It also indicates that the fraction of Pol II enzymes that terminate within the pause window is low, in agreement with previous findings<sup>355</sup> (see **Supplementary Note 1**). For strongly CDK9-responding genes, we obtained a significantly longer pause duration (**Figure 18 b**; Wilcoxon rank sum test, p-value < 10<sup>-12</sup>) and lower initiation frequencies (**Figure 18 c**, highlighted in grey).

## 2.6 Human genes have a ‘pause-initiation limit’

These results prompted us to ask whether the pause duration is generally related to the initiation frequency. We indeed found a robust anti-correlation between  $I$  and  $d$  in normally growing cells, and an upper boundary for combinations of  $I$  and  $d$  which we call ‘pause-initiation limit’ (**Figure 18 d**; Methods II.2.2.4). Thus, genes with shorter pausing show higher initiation frequencies and more RNA synthesis. In conclusion, independent mNET-seq and TT-seq data led to independent measures of pause duration and productive initiation frequency for each gene, which were then observed to be globally anti-correlated.

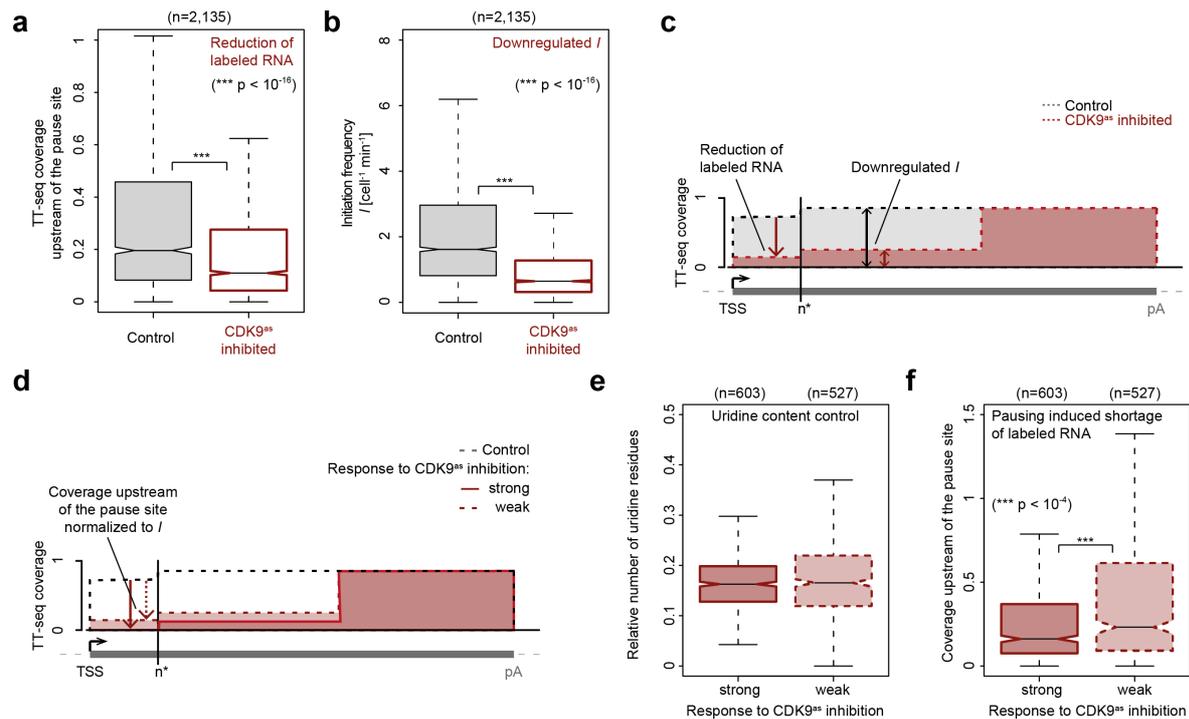
These findings now allowed us now to test directly whether longer pause durations lead to lower initiation frequencies, by analyzing TT-seq data after CDK9 inhibition. CDK9 inhibition resulted in significantly reduced labeled RNA in the short region between the TSS and the pause site (**Figure 19 a, c**; Wilcoxon rank sum test, p-value < 10<sup>-16</sup>). Productive initiation frequencies were significantly downregulated after CDK9 inhibition (**Figure 19 b-c**; Wilcoxon rank sum test, p-value < 10<sup>-16</sup>). Because CDK9 specifically targets paused Pol II, and not initiating polymerase, these results show that pausing limits initiation, and not the other way around. Thus, human genes have a ‘pause-initiation limit’.

To monitor the occupancy of engaged Pol II we generated mNET-seq data before and after CDK9 inhibition (Methods II.2.2.3). CDK9 inhibition resulted in increased mNET-seq signal at the beginning of genes and decreased signal in the gene body, indicating that less Pol II was released from the pause site (**Figure 20 a**). Indeed, calculation of pause durations from mNET-seq and TT-seq data after CDK9 inhibition showed that Pol II resides significantly longer at the pause site after CDK9 inhibition (**Figure 20 b**; Wilcoxon rank sum test,  $p$ -value  $< 10^{-16}$ ). Taken together, CDK9 inhibition increases the pause duration and decreases the initiation frequency at human genes (**Figure 20 c**).



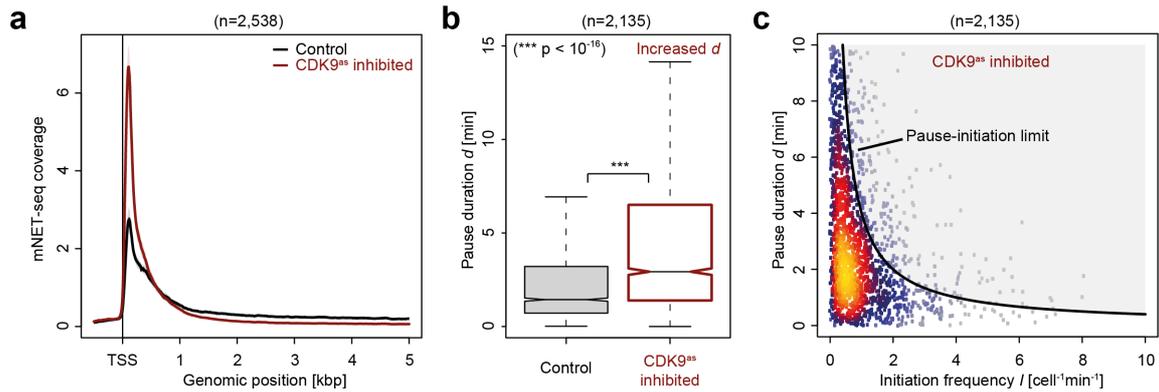
**Figure 18. Pol II pausing generally limits transcription initiation ('pause-initiation limit').**

(a) Schematic representation of polymerase flow in the promoter-proximal region. The mNET-seq signal (top) is the ratio of the productive initiation frequency  $I$  over the elongation velocity  $v$ . The TT-seq signal (bottom) corresponds to the productive initiation frequency  $I$ . Thus,  $v$  can be derived from the ratio of the TT-seq over the mNET-seq signal, and the reciprocal of  $v$  in the pause window corresponds to the pause duration  $d$ . (b) Distributions of gene-wise pause duration  $d$  [min] for genes with a CDK9 response ratio  $> 75$  % quantile (574 genes) and genes with a response ratio  $< 25$  % quantile (469 genes). (c) Distributions of gene-wise initiation frequency  $I$  [ $\text{cell}^{-1} \text{min}^{-1}$ ] for genes with a CDK9 response ratio  $> 75$  % quantile (635 genes) and genes with a response ratio  $< 25$  % quantile (635 genes). (d) Scatter plot between the initiation frequency  $I$  [ $\text{cell}^{-1} \text{min}^{-1}$ ] and the pause duration  $d$  [min] for 2,135 common genes with color-coded density estimation. The grey shaded area depicts impossible combinations of  $I$  and  $d$  according to published kinetic theory<sup>315</sup> and assuming that steric hindrance occurs below a distance of 50 bp between the active sites of the initiating Pol II and the paused Pol II (**Figure 2 b**).



**Figure 19. Increasing Pol II pause duration decreases the frequency of transcription initiation.**

(a) Boxplot comparing distributions of gene-wise mean TT-seq signals (labeled RNA) in the region between the TSS and the pause site, before (control) and after CDK9 inhibition for 2,135 genes, normalized to the initiation frequency *I* before CDK9 inhibition. (b) Boxplot comparing distributions of gene-wise initiation frequencies before (control) and after CDK9 inhibition. (c) Schematic summary of observed decrease in TT-seq signal upon CDK9 inhibition, upstream (analyzed in (a)) and downstream (analyzed in (b)) of the pause site ( $n^*$ ) (Figure 2 a). The grey box depicts the transcript body from the transcription start site (TSS, black arrow) to the polyA site (pA). (d) Schematic representation of labeled RNA coverage ratio calculation for strong (red solid line) and weak (red dashed line) CDK9 response. (e) Distributions of gene-wise uridine content in the region between the TSS and the pause site for genes with a strong response ratio (>75% quantile, 603 genes) and genes with a weak response ratio (<25% quantile, 527 genes). (f) Distributions of gene-wise mean real TT-seq signal in the region between the TSS and the pause site normalized to initiation frequency for subsets as in (e).



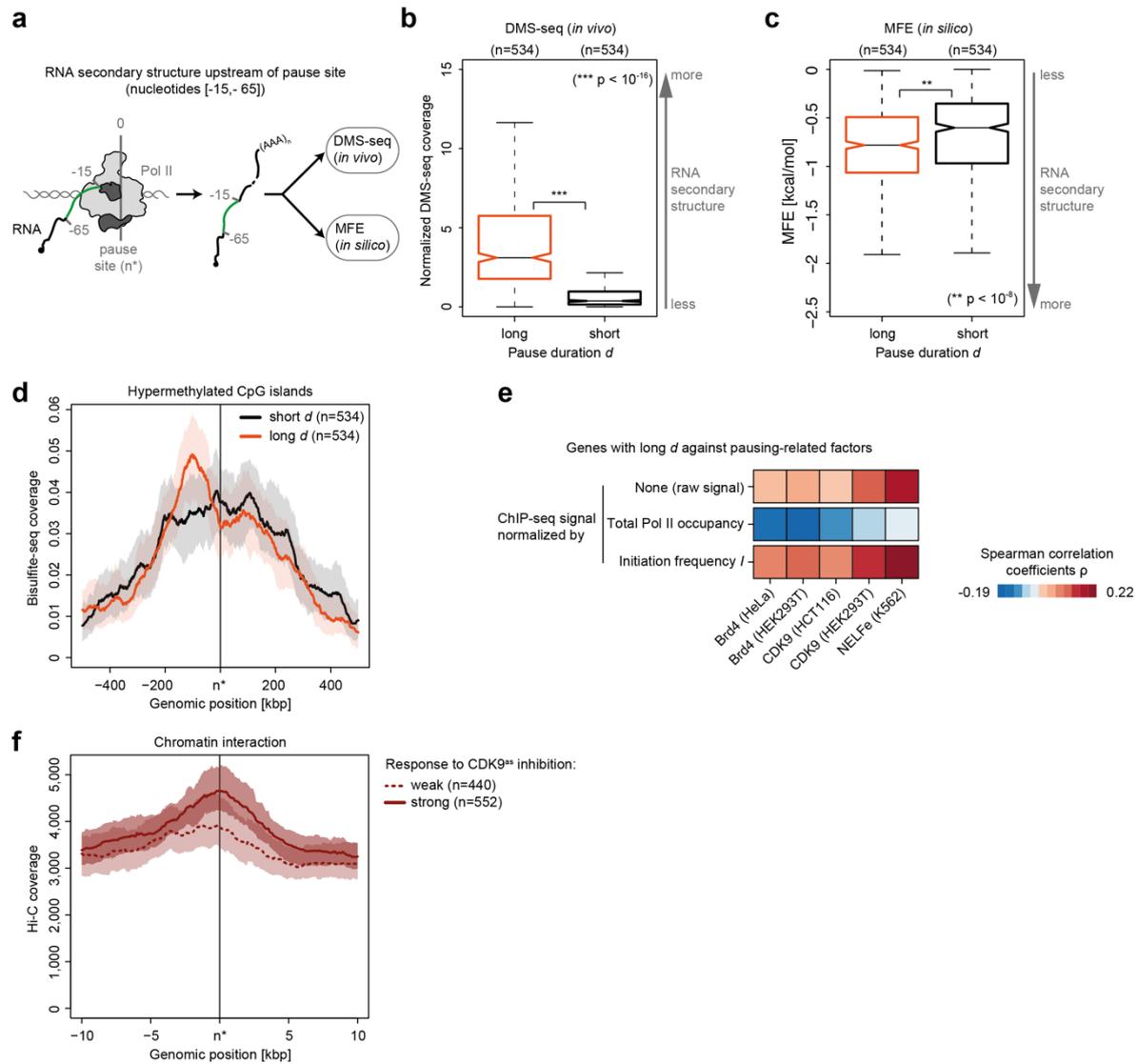
**Figure 20. CDK9 inhibition leads to increased pause duration.**

(a) Metagenome analysis comparing the average mNET-seq signal before and after CDK9 inhibition. Two biological replicates were averaged (Table 21, Table 22, Figure 10). The mNET-seq coverage was averaged for 2,538 investigated genes. Genes were aligned with their TSS. Shaded areas around the average signal (solid lines) indicate confidentiality intervals. (b) Distributions of gene-wise pause duration  $d$  [min] before (control) and after CDK9 inhibition for 2,135 genes. (c) Scatter plot between the initiation frequency  $I$  [ $\text{cell}^{-1} \text{min}^{-1}$ ] and the pause duration  $d$  [min] after CDK9 inhibition for 2,135 genes with color-coded density estimation. The grey shaded area depicts impossible combinations of  $I$  and  $d$ <sup>315</sup> assuming that steric hindrance occurs below a distance of 50 [bp] between the active sites of the initiating Pol II and the paused Pol II (Figure 2 b).

## 2.7 Determinants of promoter-proximal pausing

To investigate possible reasons for polymerase pausing and its consequences, we compared different properties of genes with long and short pause durations. For the 5' region of genes with longer pause durations, the transcript adopts more RNA secondary structure *in vivo*<sup>425</sup> (Figure 21 a-b; Wilcoxon rank sum test, p-value  $< 10^{-16}$ ) and *in silico*<sup>426</sup> (Figure 21 c Wilcoxon rank sum test, p-value  $< 10^{-8}$ ). Genes with longer pause durations were also enriched for hyper-methylated CpG islands<sup>144</sup> upstream of the pause site (Figure 21 d), consistent with a previous report<sup>256</sup>.

Comparison of strongly and weakly CDK9-responding genes around the pause site showed that genes that responded strongly to CDK9 inhibition showed a higher tendency to establish long-range chromatin interactions (Figure 21 f) as observed by Hi-C<sup>437</sup>. This is consistent with the idea that interactions of an enhancer with its target promoter can stimulate Pol II pause release<sup>240,438</sup>. This tendency however seems to be independent of the pause duration as comparing genes with long and short pause durations leads to no observable difference in Hi-C signal. Finally, we investigated which factors preferentially occupy pause windows with longer pause durations. This is now possible because ChIP-seq signals can be normalized with the productive initiation frequency. Without such normalization, ChIP-seq derived factor occupancies are artificially high in pause windows with long pause durations<sup>315</sup>. Correlation of such normalized ChIP-seq signals in the pause window with pause durations (Figure 21 e) resulted in a positive correlation for Pol II phosphorylation at sites that are associated with elongation, and also for NELF-E, CDK9, and Brd4, which are all factors involved in Pol II pausing and release.



**Figure 21. Determinants of long promoter-proximal pause duration  $d$ .**

(a) Secondary structure of a 50 nt long RNA fragment (-15 to -65 bp) upstream of pause site (position 0,  $n^*$ ) is compared for genes with long  $d$  to genes with short  $d$ . RNA secondary structure is read out by *in vivo* DMS-seq, or *in silico* (minum free energy, MFE [kcal mol<sup>-1</sup>]). (b) Distribution of gene-wise mean *in vivo* DMS-seq signals<sup>425</sup> for a window defined in (a) for genes with long  $d$  ( $d > 75\%$  quantile, 534 genes) and with short  $d$  ( $d < 25\%$  quantile, 534 genes) normalized to denatured DMS-seq coverage<sup>425</sup>. (c) Distributions of gene-wise MFE<sup>426</sup> as in (b). (d) Metagene analysis comparing the average Bisulfite-seq signal for subsets as in (b) aligned at the pause site ( $n^*$ ) (red, long  $d$ , and black, short  $d$ ). Shaded areas around the average signal (solid lines) indicate confidence intervals. (e) Heatmap showing the pairwise Spearman correlation (color encoded,  $-0.19$  in blue to  $0.22$  in red) of 534 genes with long  $d$  against ChIP-seq measurements in the pause window ( $n^* \pm 100$  bp, or TSS + 200 bp) of factors involved in pausing: Brd4<sup>439</sup>, CDK9<sup>439,440</sup>, and NELF-E<sup>144</sup> in K562 cells. ChIP-seq signal is shown as raw (top), or normalized by total Pol II occupancy<sup>440</sup> (middle), or by initiation frequency  $I$  (bottom). (f) Metagene analysis comparing the average Hi-C signal<sup>437</sup> (detecting long-range chromatin interactions) for strongly CDK9-responsive genes (red, response ratio  $> 75\%$  quantile, 552 genes) and weakly CDK9-responsive genes (black, response ratio  $< 25\%$  quantile, 440 genes) aligned at the pause site ( $n^*$ ). All data sets are listed in **Table 31**.

---

### 3. The pause-initiation limit restricts transcription activation

---

The results presented in chapter 3 are currently under peer review for publication:

S. Gressel \*, B. Schwalb \* *et al.* The pause-initiation limit restricts transcription activation in human cells. *Nature Communications, in revision.*

(\* joint first authorship)

A detailed list of author contributions can be found on page VII.

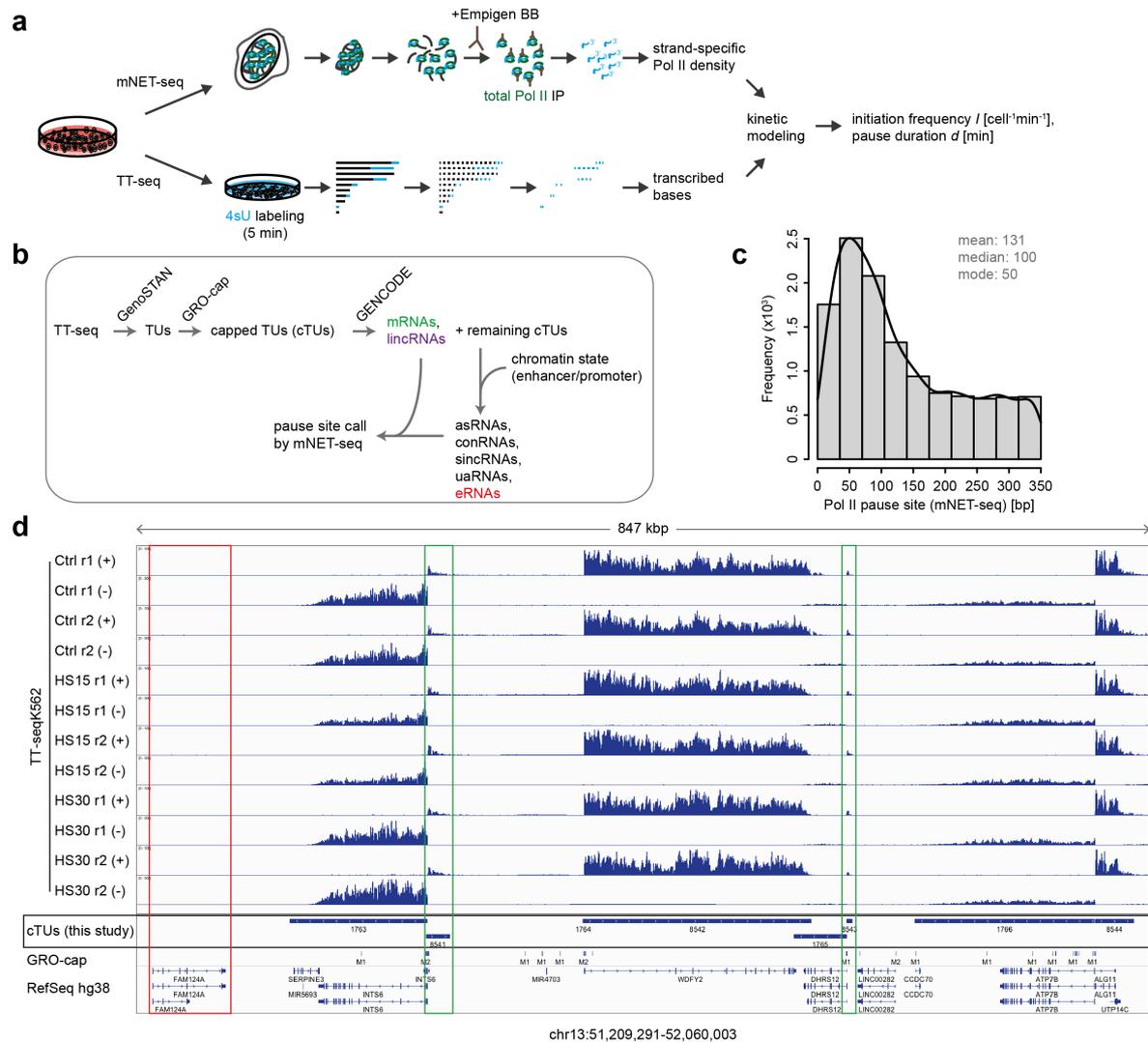
---

#### Chapter summary

Eukaryotic gene transcription is often controlled at the level of RNA polymerase II (Pol II) pausing in the promoter-proximal region. Pausing Pol II limits the frequency of transcription initiation ('pause-initiation limit'), predicting that the pause duration must be decreased for transcriptional activation. To test this prediction, we conducted a genome-wide kinetic analysis of the heat shock response in human cells. We show that the pause-initiation limit restricts transcriptional activation at most genes. Gene activation requires the activity of the P-TEFb kinase CDK9, which decreases the duration of Pol II pausing and thereby enables an increase in the productive initiation frequency. The transcription of enhancer elements is generally not pause-limited and can be activated without CDK9 activity. Our results define the kinetics of Pol II transcriptional regulation in human cells at all gene classes during a natural transcription response.

#### Acknowledgements

We would like to thank Kerstin Maier and Petra Rus (Max Planck Institute for Biophysical Chemistry) for sequencing. We would like to thank Anna Sawicka and Kristina Žumer (Max Planck Institute for Biophysical Chemistry) for sharing the spike-in plasmids. PC was supported by the Deutsche Forschungsgemeinschaft (SFB860, SPP1935), the European Research Council Advanced Investigator Grant TRANSREGULON (grant agreement No 693023), and the Volkswagen Foundation.



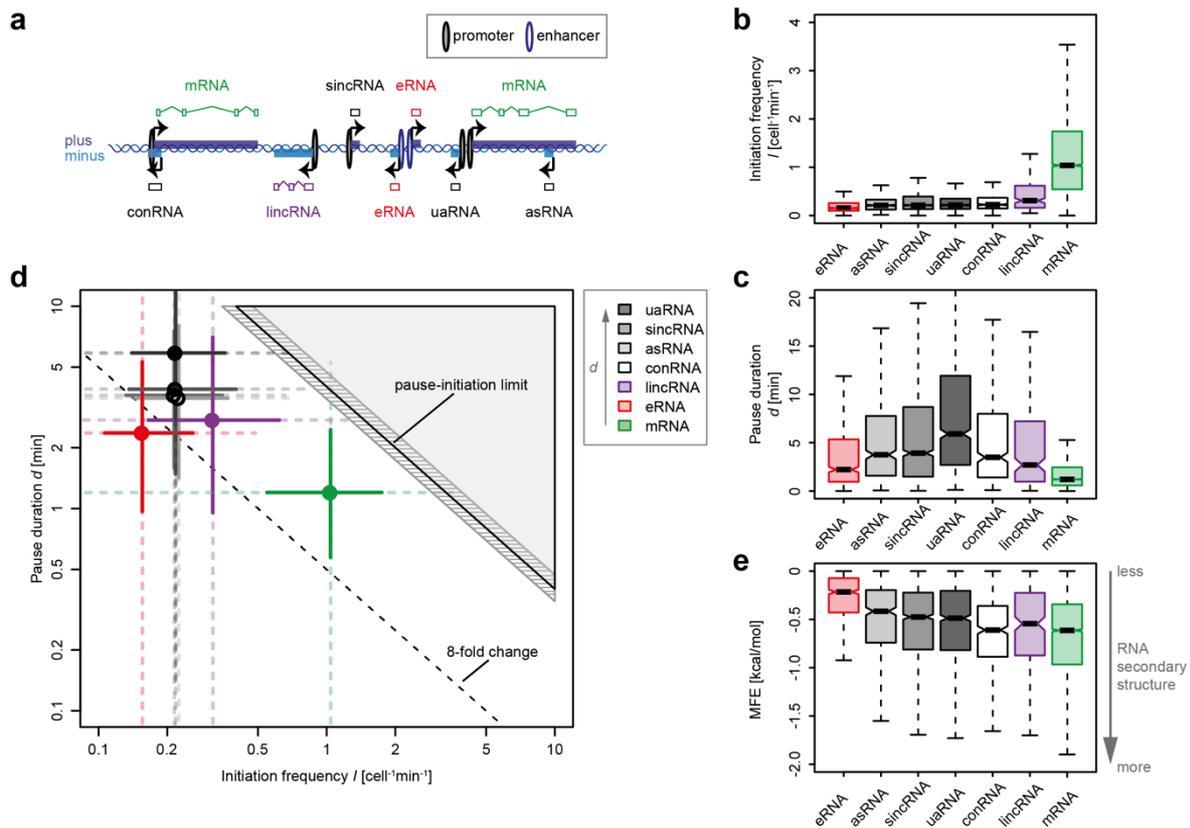
**Figure 22. Multi-omics analysis and annotation of different gene classes in human K562 cells.**

(a) Experimental schematic of multi-omics approach. Human K562 cells were subjected to TT-seq to estimate the productive initiation frequency  $I$  [ $\text{cell}^{-1} \text{min}^{-1}$ ]. By combining TT-seq data with mNET-seq the pause duration  $d$  [min] is calculated<sup>381</sup>. (b) Schematic overview of capped transcription unit (cTU) annotation in K562 cells. First, TT-seq data ( $n=6$ ) were subjected to the genomic segmentation algorithm GenoSTAN<sup>177</sup>. The transcription start sites (TSSs) were further refined by published GRO-cap data generated in K562 cells<sup>50</sup>. To call promoter-proximal pause sites mNET-seq data ( $n=6$ ) generated in K562 cells were used (Table 23, Table 25, Figure 11). (c) Distribution of pause site distance [bp] for 12,160 investigated cTUs measured by mNET-seq depicted as a histogram with respect to GRO-cap refined TSS (mean 131 [bp], median 100 [bp], mode 50 [bp]). (d) Genome browser view of a 847 kbp region located in chromosome 13 (chr13:51,209,291-52,060,003) visualized with the Integrative Genomics Viewer (IGV, version 2.4.10; human hg38)<sup>400</sup>. From top to bottom, tracks represent: TT-seq coverage in K562 cells ( $n=6$ ) on respective plus (+) or minus (-) strand, GRO-cap data<sup>50</sup>, new cTU annotation, RefSeq GRCh38 annotation. cTUs missing in the RefSeq GRCh38 annotation are highlighted in a green. RefSeq transcript not present in K562 cells is highlighted in red.

### 3.1 Multi-omics analysis and annotation of transcription units

We first aimed at defining genome-wide Pol II kinetic parameters during steady state conditions. To obtain the productive initiation frequency  $I$  and the promoter-proximal pause duration  $d$ <sup>381</sup>, we carried out TT-seq and mNET-seq of total Pol II (with Empigen BB) in human K562 cells (**Figure 6**, **Figure 22 a**). We generated TT-seq data for two independent biological replicates after 5 min of metabolic labeling with 4sU (**Figure 11 a**; Spearman correlation  $\rho = 1.00$ ) as well as mNET-seq data for two independent biological replicates (**Figure 11 c**; Spearman correlation  $\rho = 0.99$ ) (**Table 23**). We then used the TT-seq data to create a genome-wide transcription unit (TU) annotation with the segmentation algorithm GenoSTAN<sup>177</sup> (**Figure 22 b**, **Figure 45**). As a strategy for identifying accurate transcription start sites (TSSs) for the TT-seq derived TUs, we used published GRO-cap data<sup>50</sup>, which recovers nascent RNAs with 5' caps of transcriptionally engaged RNA Pol II. To be eligible for further analysis, annotated TUs needed a GRO-cap signal in a window of 250 bp around the GenoSTAN-derived start site, and an expression of  $>5$  reads per base in the sum of both replicates of the TT-seq signal. We then sorted each capped TU (cTU) into one of the following seven classes using GENCODE annotation, respective location and chromatin state annotation<sup>177</sup> for enhancer and promoter classification (**Figure 23 a**; Methods II.2.4.4): protein-coding (m), long intergenic noncoding (linc), antisense (as), convergent (con), upstream antisense (ua), short intergenic noncoding (sinc), and putative enhancer (e) RNAs.

Subsequently, we used the mNET-seq data to extract the position of paused polymerases for all cTUs in each class that showed mNET-seq signal peaks above background (Methods II.2.4.4). The called pause sites were distributed around a maximum located  $\sim 50$  bp downstream of the TSS (**Figure 22 c**), in contrast to the pause sites that were previously derived based on the TSS annotation from RefSeq, which were located  $\sim 30$  bp further downstream (**Figure 2 a**; **Figure 48 c**). This agrees well with recently published data in K562 cells<sup>247</sup>. We did not observe any substantial differences in the positions of called pause sites among the different classes. This resulted in a total of 10,363 expressed cTUs, for which a pause site call was successful, encoding 604 eRNAs, 471 asRNAs, 1,314 sincRNAs, 965 uaRNAs, 445 conRNAs, 209 lincRNAs and 6,355 mRNAs (**Figure 22 b-d**, **Figure 23 a**). Below we will refer to these cTUs simply as 'genes' or with the respective RNA transcript class they give rise to.



**Figure 23. Transcription kinetics of different gene classes in steady state.**

(a) Schematic representation of seven major transcript classes annotated in this study: 604 eRNAs (in light red), 471 asRNAs, 1,314 sincRNAs, 965 uaRNAs, 445 conRNAs (in different shades of grey), 209 lincRNAs (in purple), 6,355 mRNAs (in green), depicted on plus (dark blue) or minus (light blue) strand (Methods II.2.4.4). (b-c) Boxplots of productive initiation frequency  $I$  [ $\text{cell}^{-1} \text{min}^{-1}$ ] (b) or pause duration  $d$  [min] (c) for transcript classes defined in (a). Black bars represent medians, boxes span from upper to lower quartiles, whiskers represent 1.5 times the inter-quartile range. (d) Plot shows the median productive initiation frequency  $I$  [ $\text{cell}^{-1} \text{min}^{-1}$ ] depicted against the median pause duration  $d$  [min] for all transcript classes (circles) in log scale. The two solid perpendicular lines define the inter-quartile range, the dotted whiskers represent 1.5 times the inter-quartile range of the respective estimates for the entire transcript class. The grey shaded area depicts impossible combinations of  $I$  and  $d$ <sup>315,381</sup>. Striped area shows confidence intervals of the pause-initiation limit. The dotted line defines an 8-fold possible fold change until a gene would be restricted by the pause-initiation limit. (e) Boxplots of minimum free energy MFE [kcal/mol]. MFE was calculated in a window of [-15,-65] bp upstream of the pause site to predict RNA secondary structure<sup>426</sup> (Figure 21 a) for each transcript in a transcript class.

### 3.2 Transcription kinetics differ between gene classes

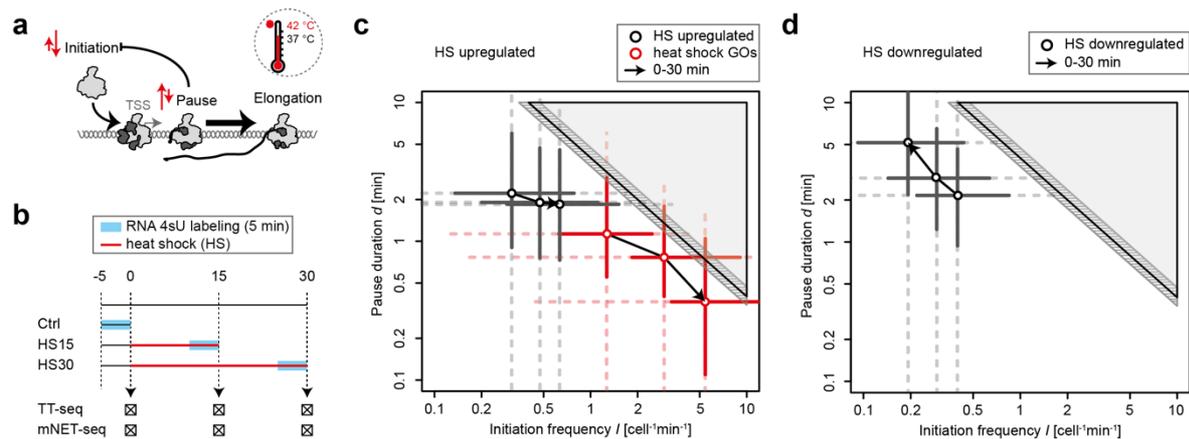
We next used our TT-seq and mNET-seq data in combination with our previously described kinetic modeling<sup>381</sup> to derive estimates of  $I$  and  $d$  for all annotated genes (**Figure 23 a**). We observed a reciprocal behavior of  $I$  and  $d$  for all classes of genes except enhancers (**Figure 23 b-d**). When  $I$  was high,  $d$  was generally low, and the other way around, consistent with an anti-correlation between these two parameters<sup>381</sup>. Protein-coding and lincRNAs showed the highest  $I$  value (with median initiation events of 1 and 0.3 cell<sup>-1</sup> min<sup>-1</sup>) and the lowest  $d$  values (median 1 and 2.7 min), consistent with high expression levels (**Figure 23 b, d; Figure 46**). On average, lincRNAs show significantly longer pausing compared to mRNAs (**Figure 23 c**; Wilcoxon rank sum test, p-value  $< 2.2 \cdot 10^{-16}$ ), in contrast to a recent study<sup>391</sup>. Thus, genes encoding lincRNAs initiate on average about half as frequently as protein-coding genes, and smaller mNET-seq peaks for these genes indicate longer pause durations, which is counterintuitive but clearly revealed by the kinetic analysis (**Supplementary Note 2**). Generally, all noncoding (nc) RNA classes (except eRNAs) showed high  $d$  values, with a median ranging around 4-6 minutes, explaining their low levels of expression. The low pause durations at enhancers might be explained by the fact that eRNAs do generally not adopt stable secondary structure in the nascent RNA exiting from the polymerase (**Figure 23 e**), which is associated with pausing<sup>381</sup>. However, other chromatin features and factor availability at different loci might contribute to pausing as well. A particularly long pause duration was observed for upstream antisense RNAs (median 6 min), and this could impair initiation events in the noncoding direction of bidirectional promoters (**Figure 23 c**).

### 3.3 Monitoring transcription kinetics during a natural transcription response

We next investigated how transcription activation kinetics change upon response to heat shock (**Figure 24**). The optimal time points were determined by quantitative reverse transcription PCR (RT-qPCR) of the major cell stress protein HSPA1A (human Hsp70) and cell viability assays (**Figure 47**; Methods II.2.1.2 and II.2.4.1). We performed TT-seq, RNA-seq and mNET-seq in K562 cells that were maintained under optimal growth conditions at 37 °C (control), or placed in a 42 °C water bath for 15 or 30 min (heat shock) (**Figure 24 b**). TT-seq, RNA-seq and mNET-seq libraries were prepared in two biological replicates that were highly reproducible (TT-seq: rho  $> 0.99$  for all time points; RNA-seq: rho  $> 0.99$  for all time points; mNET-seq: rho  $> 0.99$  for all time points) (**Table 23, Table 25, Figure 11**). In order to capture global changes in transcription profiles, we used a spike-in normalization strategy<sup>168</sup>. This revealed that 899 genes were significantly upregulated, whereas 2,614 genes were downregulated after 30 min of heat shock (**Figure 25 a**). To normalize the respective mNET-seq signals in the heat shock conditions versus the control, we identified 3,416 genes that were unchanged in their TT-seq signal, and globally calibrated the mNET-seq data to show no change on these genes during the heat shock response (Methods V.2.1.2).

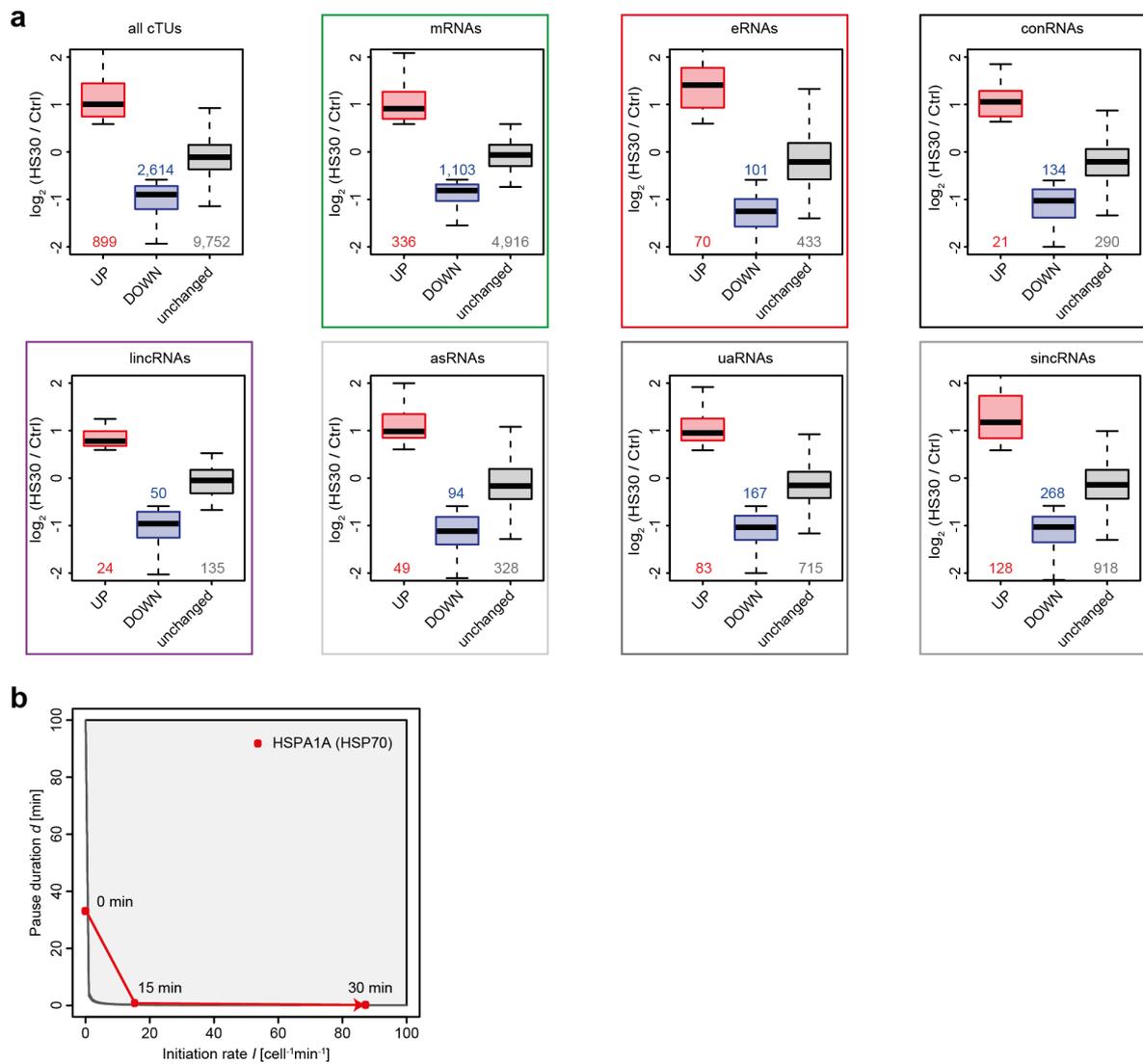
### 3.4 The pause-initiation limit restricts transcription activation

Next, we calculated the productive initiation frequency  $I$  and the pause duration  $d$  for all genes after 15 min and after 30 min of heat shock. Our results show that upon heat shock, activated genes generally show increased  $I$  and decreased  $d$  (**Figure 24 c**, **Figure 48 f**), suggesting that gene activation requires a decrease in pause duration, which in turn allows for higher productive initiation frequencies. This behavior is even more evident at protein-coding genes linked to heat shock (**Figure 24 c**). This can be exemplified at the HSPA1A (HSP70) gene which is a prominent model gene for the heat shock response (Supplementary Fig. 4 b). The pause duration  $d$  of HSPA1A changes from  $\sim 30$  min at steady state to  $\sim 0$  min upon 30 min of heat shock which agrees well with literature estimates<sup>356</sup>. Consistent with an anti-correlation between  $d$  and  $I$ ,  $I$  increased upon heat shock to 87 productive initiation events ( $\text{cell}^{-1} \text{min}^{-1}$ ). The results also revealed that downregulated genes exhibited decreased  $I$  and increased  $d$  (**Figure 24 d**, **Figure 48 e**). Thus, this multi-omics analysis reveals that the pause-initiation limit restricts gene regulation at genes which are located close to the limit in steady state. The use of TT-seq in this respect is critical because it directly monitors RNA synthesis activity and productive initiation frequency (see also **Supplementary Note 2**, **Figure 43**).



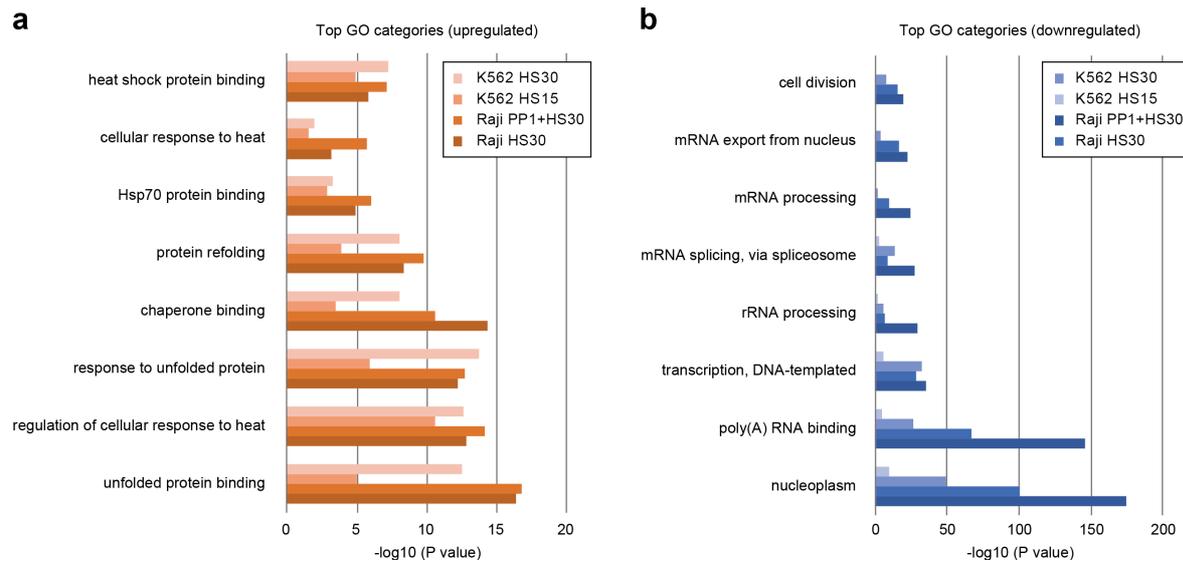
**Figure 24. Regulation of transcription upon heat shock time-course in human K562 cells.**

(a) Illustration of the regulatory changes (red arrows) at a model gene upon heat shock. Pol II is depicted in silver, decorated with initiation, pausing or elongation factors (dark grey). (b) Experimental set-up of the heat shock (HS) time-course (0 to 30 min at 42 °C) in human K562 cells. TT-seq and mNET-seq experiments were performed in two independent biological samples (**Table 23**, **Table 25**, **Figure 11**). (c) Pause-initiation trajectories upon heat shock time-course. Shown are 706 significantly upregulated cTUs (in grey), and 33 significantly upregulated RefSeq-TUs (in red) of overrepresented GO categories linked to heat shock (unfolded protein binding, regulation of cellular response to heat, response to unfolded protein, chaperone binding) (**Figure 26**). Median estimates per time point (0, 15 and 30 min) are depicted as circles. The two solid perpendicular lines define the inter-quartile range, the dotted whiskers represent 1.5 times the inter-quartile range of the respective estimates for the entire transcript class. The grey shaded area depicts impossible combinations of  $I$  and  $d$ <sup>381</sup>. Striped area shows confidence intervals of the pause-initiation limit. (d) Pause-initiation trajectories as in (c) upon heat shock time-course of 1,907 significantly downregulated cTUs (in grey).



**Figure 25. DE analysis of transcript classes and kinetics of HSPA1A upon heat shock.**

(a) Transcript classes as annotated in this study (Figure 22; Methods II.2.4.4). Boxplots show differential expression (DE) analysis of upregulated (light red), downregulated (dark blue), and unchanged genes (grey) in human K562 cells upon 30 min of heat shock (42 °C). Heat shock samples (HS30) were compared to respective control (Ctrl). Spike-in normalized TT-seq data were analyzed for DE of all genes (cTUs), mRNAs (green frame), eRNAs (light red frame), conRNAs (black frame), lincRNAs (purple frame), asRNAs, uaRNAs, sincRNAs (frame in different shades of grey). A minimum fold change of 1.5 in all conditions and a maximum adjusted p-value of 0.1 was set for calling a significant expression change. (b) Plot shows the median productive initiation frequency  $I$  [cell<sup>-1</sup> min<sup>-1</sup>] depicted against the median pause duration  $d$  [min] for hsp70 (HSPA1A) (red) for 0, 15, and 30 min at 42 °C (in red). The grey shaded area depicts impossible combinations of  $I$  and  $d$ .



**Figure 26. GO analysis of differentially expressed RefSeq-TUs upon heat shock.**

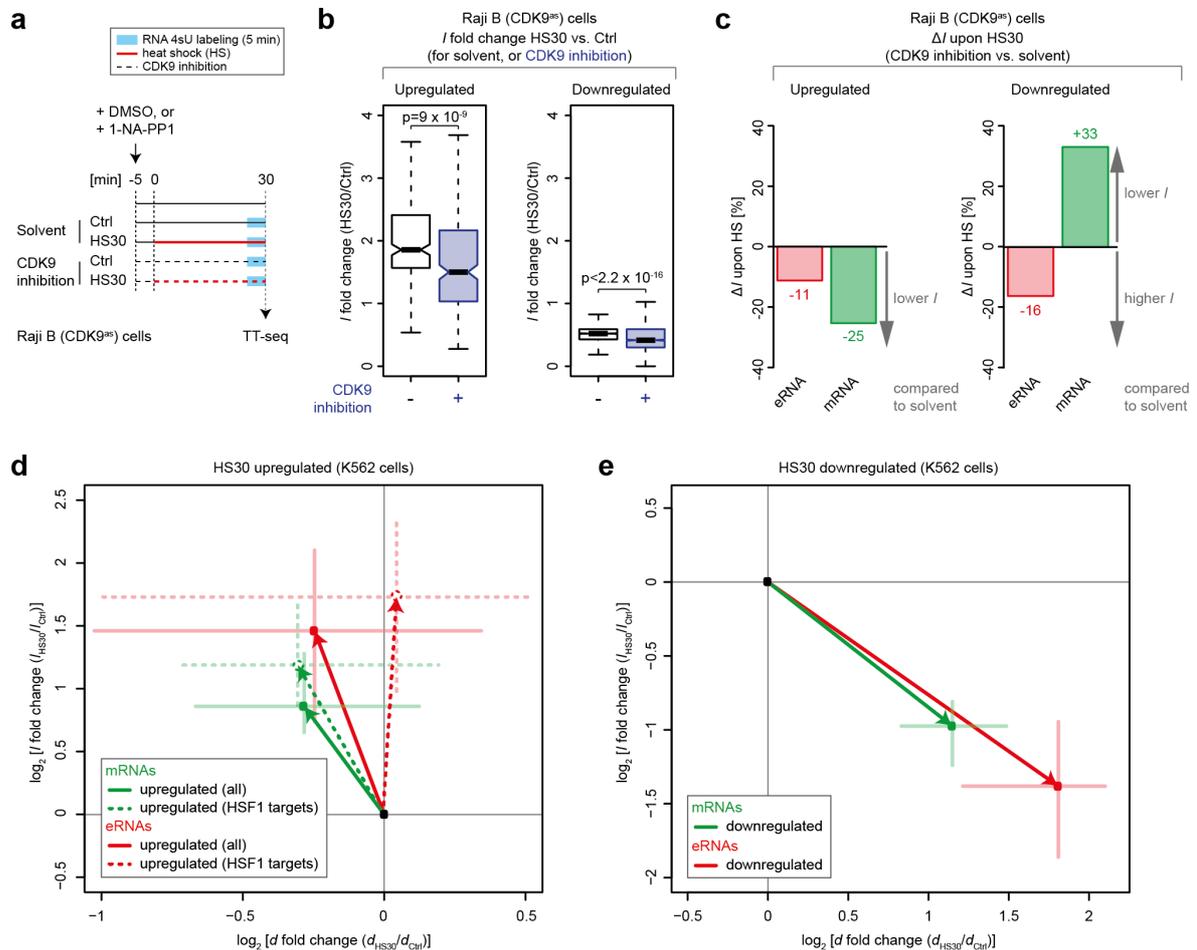
Gene Ontology (GO) analysis<sup>441</sup> of significantly overrepresented categories linked to upregulated (a) or downregulated (b) categories in heat shock for human K562 or Raji B (CDK9<sup>as</sup>) cells. For K562 cells, heat shock samples (HS15, or HS30) were compared to respective control (Ctrl). For Raji B (CDK9<sup>as</sup>) samples, solvent heat shock sample (HS30) was compared to solvent control, while CDK9 inhibited heat shock sample (PP1+HS30) was compared to CDK9 inhibited control. Two biological replicates were averaged per condition and time-point.

### 3.5 CDK9 activity lowers the pause-initiation limit for gene activation

Our kinetic modeling revealed that the activation of genes restricted by the pause-initiation limit requires a decrease in the pause duration. To corroborate this, we utilized the highly specific and rapid inhibition of an analog-sensitive CDK9 Raji B cell line using the bulky ATP analog 1-NA-PP1<sup>381</sup> (**Figure 27 a**). Raji B and K562 cell lines are predicted to show a conserved response to heat shock with respect to timing of HSPA1A upregulation, cell viability (**Figure 26**) and GO terms of up- or downregulated TUs (**Figure 47**).

We generated TT-seq data for two independent biological replicates after 5 min of metabolic labeling with 4sU to measure changes in *I* upon specific inhibition of the P-TEFb kinase CDK9 prior and during heat shock (**Table 23**, **Table 25**). TT-seq data were highly reproducible (**Figure 12 a**; Spearman correlation  $\rho = 1.00$ ) and CDK9 kinase inhibition was very rapid (**Figure 44**). We again annotated TUs genome-wide with GenoSTAN (**Figure 49**), and this resulted in 6,990 mRNAs, 3,451 eRNAs, 243 lincRNAs, 1,398 asRNAs, 326 conRNAs, 565 uaRNAs, and 3,479 sincRNAs. We now tested our hypothesis by investigating the changes in *I* upon heat shock after CDK9 inhibition (**Figure 27**). We derived estimates of the productive initiation frequency *I* for all 12,958 expressed, non-ambiguously classified genes after spike-in normalization. Changes in *I* for upregulated genes confirmed a strong dependence of transcription activation on CDK9 kinase activity. Genes were significantly less inducible during heat shock when CDK9 was inhibited (**Figure 27 b**, left; Wilcoxon rank sum test, p-value =  $9 \cdot 10^{-9}$ ), confirming an obligatory decrease in pause duration for upregulation of the productive initiation frequency. Productive initiation events of genes encoding mRNAs decreased to 75 % after CDK9 inhibition (**Figure 27 c**, left). Downregulation of genes was overall stronger upon CDK9 inhibition (**Figure 27 b**, right; Wilcoxon rank sum test, p-value

$< 2.2 \cdot 10^{-16}$ ), indicating less possible initiation events (33 % for mRNAs) due to even longer pause durations compared to downregulation upon heat shock alone (**Figure 27 c**, right). Thus, CDK9 activity lowers the pause duration to allow for high gene activation that is restricted by the pause-initiation limit.



**Figure 27. Enhancer transcription is generally not pause limited and less dependent on CDK9.**

(a) Experimental set-up. The adenine analog 1-NA-PP1 allows for rapid and highly specific inhibition of analog sensitive CDK9<sup>338,339</sup> in CRISPR/Cas9 engineered human Raji B (CDK9<sup>as</sup>) cells<sup>381</sup> (**Figure 12 b**). CDK9 was inhibited using 5 μM 1-NA-PP1 (CDK9 inhibition) in combination with heat shock at 42 °C for 30 min in human Raji B (CDK9<sup>as</sup>) cells. DMSO was used as solvent control. (b) Boxplots of productive initiation frequency  $I$  fold change before (in white) and after CDK9 kinase inhibition (in dark blue). Shown are 241 significantly upregulated genes (left boxplot), and 2,795 significantly downregulated genes (right boxplot) annotated in Raji B (CDK9<sup>as</sup>) cells (**Figure 49 b**). (c) Left: bar plot comparing productive initiation frequency  $I$  change ( $\Delta I$ ) with and without CDK9 inhibition for 92 significantly upregulated mRNAs (in green), and 54 significantly upregulated eRNAs (in red) annotated in Raji B (CDK9<sup>as</sup>) cells.  $\Delta I$  upon upregulation between heat shock with CDK9 inhibition (CDK9 inhibited HS30) and heat shock with solvent control (Solvent HS30) is shown in percentage [%]. Grey arrow indicates that less productive initiation events occur compared to solvent control. Right: bar plot comparing productive initiation frequency  $I$  change ( $\Delta I$ ) with and without CDK9 inhibition for 2,210 significantly downregulated mRNAs (in green), and 223 significantly downregulated eRNAs annotated in Raji B (CDK9<sup>as</sup>) cells.  $\Delta I$  upon downregulation between heat shock with CDK9 inhibition (CDK9 inhibited HS30) and heat shock with solvent control (Solvent HS30) is shown in percentage [%]. Grey arrows indicate that less productive initiation events occur for mRNAs, or more productive initiation events for eRNAs compared to solvent control. (d) Log<sub>2</sub> fold change of pause duration  $d$  and initiation frequency  $I$  for 336 significantly upregulated mRNAs (in green), and 67 significantly upregulated eRNAs (in light red) in K562 cells upon 30 min of heat shock (HS30) (**Figure 25 a**). Dashed lines represent HSF1 driven subsets of 91 mRNAs and 20 eRNAs. (e) Log<sub>2</sub> fold change of pause duration

$d$  and initiation frequency  $I$  for 1,101 significantly downregulated mRNAs (in green), and 99 significantly downregulated eRNAs (in light red) in K562 cells upon 30 min of heat shock (HS30) (**Figure 25 a**).

### 3.6 Enhancer transcription is generally not pause limited

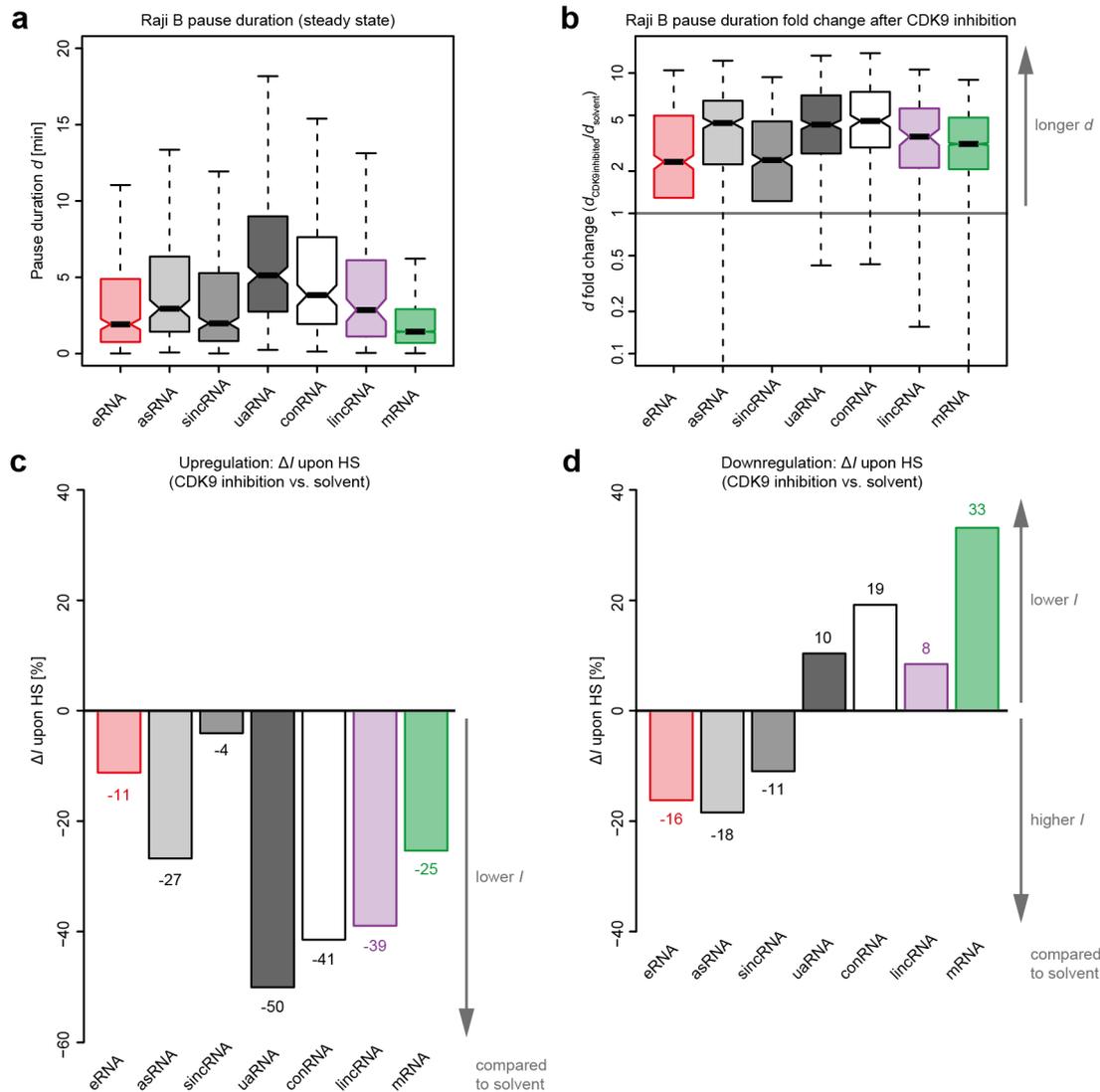
Amongst all seven gene classes, enhancers showed the greatest distance from the pause-initiation limit at steady state in K562 cells, thus forming a notable exception from other gene classes (**Figure 23 d**). Enhancers showed a median pause duration of only 2.2 min and a median productive initiation frequency of only one initiation event every 6-7 min. This indicated that enhancers were generally not restricted by the pause-initiation limit, i.e.  $I$  could increase several-fold without any change in pause duration, until the pause-initiation limit would be reached. Indeed, upon heat shock, the productive initiation frequency  $I$  for enhancers was increased 1.5-fold more than for mRNAs given the same change in pause duration  $d$  (**Figure 27 d**). This difference is even stronger when comparing heat shock factor 1 (HSF1) targeted mRNAs and eRNAs. HSF1 is a major activator in heat shock induced transcription upregulation. HSF1 driven eRNAs can be activated without a change in pause duration, while HSF1 driven mRNAs still require a shortening of the pause duration.

Another implication of this exceptional kinetic behavior is that enhancer transcription can generally only be reduced by a strong increase (>8-fold, **Figure 23 d**) of the pause duration  $d$ . However, downregulation of enhancer transcription is in line with the general observation of prolonged pausing to inhibit new initiation events, as is the case for mRNA synthesis (**Figure 27 e**). In conclusion, enhancers differ from protein-coding genes, because their productive initiation frequency appears generally not to be restricted by the pause duration.

### 3.7 Enhancer transcription is less dependent of CDK9

Although enhancer transcription is generally not limited by pausing in K562 cells at steady state (**Figure 23 d**), it remained unclear whether enhancer transcription is controlled by P-TEFb. Using mNET-seq data in Raji B (CDK9<sup>as</sup>) cells<sup>381</sup> we found that the median pause duration of all transcript classes and the exceptional role of eRNAs is conserved in unperturbed Raji B cells (**Figure 28 a**). We confirmed this also in the context of the upregulation of enhancer transcription upon heat shock in K562 cells. Enhancers showed higher initiation frequency fold-changes provided the same fold-change in pause duration as found for protein-coding genes (**Figure 27 d**). Consistent with the results in K562, activation of enhancer transcription in Raji cells was only reduced by 11% upon CDK9 inhibition (**Figure 27 c**, **Figure 28 c**).

This shows that enhancer transcription can be activated even when CDK9 is inhibited. The overall behavior of impaired activation after CDK9 inhibition for all gene classes resembles the pause durations calculated for Raji cells<sup>381</sup> and strongly supports our estimates (**Figure 23 d**, **Figure 28 b**). Surprisingly, downregulated enhancers were not repressed by inhibition of CDK9, consistent with our assumption that higher pause durations do not cause lower productive initiation frequencies at enhancers (**Figure 27 c**, **Figure 28 d**). Taken together, enhancer transcription and thus, eRNA synthesis can be up- and down-regulated to a large extent without changes in pause duration.



**Figure 28. Gene classes vary in CDK9-induced pause duration changes upon heat shock.**

(a) Boxplots of pause duration  $d$  [min] for 498 eRNAs (in light red), 319 asRNAs, 504 sincRNAs, 462 uaRNAs (in different shades of grey), 254 conRNAs, 144 lincRNAs (in purple), and 5,967 mRNAs (in green) annotated in Raji B (CDK9<sup>as</sup>) cells (Methods II.2.4.4). By combining TT-seq data with published mNET-seq in Raji B (CDK9<sup>as</sup>) cells<sup>381</sup>,  $d$  was calculated. Black bars represent medians, boxes span from upper to lower quartiles, whiskers represent 1.5 times the inter-quartile range. Note that the median pause duration (black bars) per transcript class in Raji B (CDK9<sup>as</sup>) cells resembles the median pause duration per transcript class in K562 cells (**Figure 23 c**). Thus, our data suggest strong conservation across human hematopoietic cell lines. (b) Fold change of pause duration  $d$  for different transcript classes (as in (a)) in Raji B (CDK9<sup>as</sup>) cells upon CDK9 inhibition using published TT-seq and mNET-seq data<sup>381</sup>. CDK9 inhibited samples (5  $\mu$ M 1-NA-PP1, 15 min) were compared to respective solvent control (DMSO, 15 min). Grey arrow indicates shift to longer pause durations upon CDK9 inhibition for all genes. (c) Bar plot comparing productive initiation frequency  $I$  change ( $\Delta I$ ) with and without CDK9 inhibition for significantly upregulated genes in Raji B (CDK9<sup>as</sup>) cells. From left to right: 55 eRNAs, 13 asRNAs, 51 sincRNAs, 16 uaRNAs, 11 conRNAs, 3 lincRNAs, and 92 mRNAs annotated in this study (colors as in (a)) (**Figure 49 b**).  $\Delta I$  upon upregulation between heat shock with CDK9 inhibition (CDK9 inhibited HS30) and heat shock with solvent control (Solvent HS30) is shown in percentage [%]. Grey arrow indicates that less productive initiation events occur compared to solvent control. (d) Bar plot comparing productive initiation frequency  $I$  change ( $\Delta I$ ) with and without CDK9 inhibition for significantly downregulated genes in Raji B (CDK9<sup>as</sup>) cells. From left to right: 223 eRNAs, 45 asRNAs, 227 sincRNAs, 38 uaRNAs, 20 conRNAs, 35 lincRNAs, and 2,211 mRNAs annotated in this study (colors as in (a)) (**Figure 49 b**).  $\Delta I$  upon downregulation between heat shock with CDK9 inhibition (CDK9 inhibited HS30) and heat shock with solvent control (Solvent HS30) is shown in percentage [%]. Grey arrows indicate that less productive initiation events occur for mRNAs, or more productive initiation events for eRNAs (less downregulation) compared to solvent control.

## IV. Discussion and Outlook

In this part, I will review the results of my work with respect to the five questions (Introduction I.4) raised at the beginning of my graduate work (chapter 1). Furthermore, I included a chapter on current technical limitations of our multi-omics approach. I hope that this work will encourage further development of the methodology (TT-seq and mNET-seq) to address the technical limitations discussed in chapter 2. In the future, multiple challenges such as the general mechanism of Pol II pausing and (in)active P-TEFb recruitment have to be addressed. Thus, in a closing chapter, I will look beyond the current state to anticipate what might be the next exciting questions to tackle in the field of transcription regulation (chapter 3).

### 1. Our contributions to the field

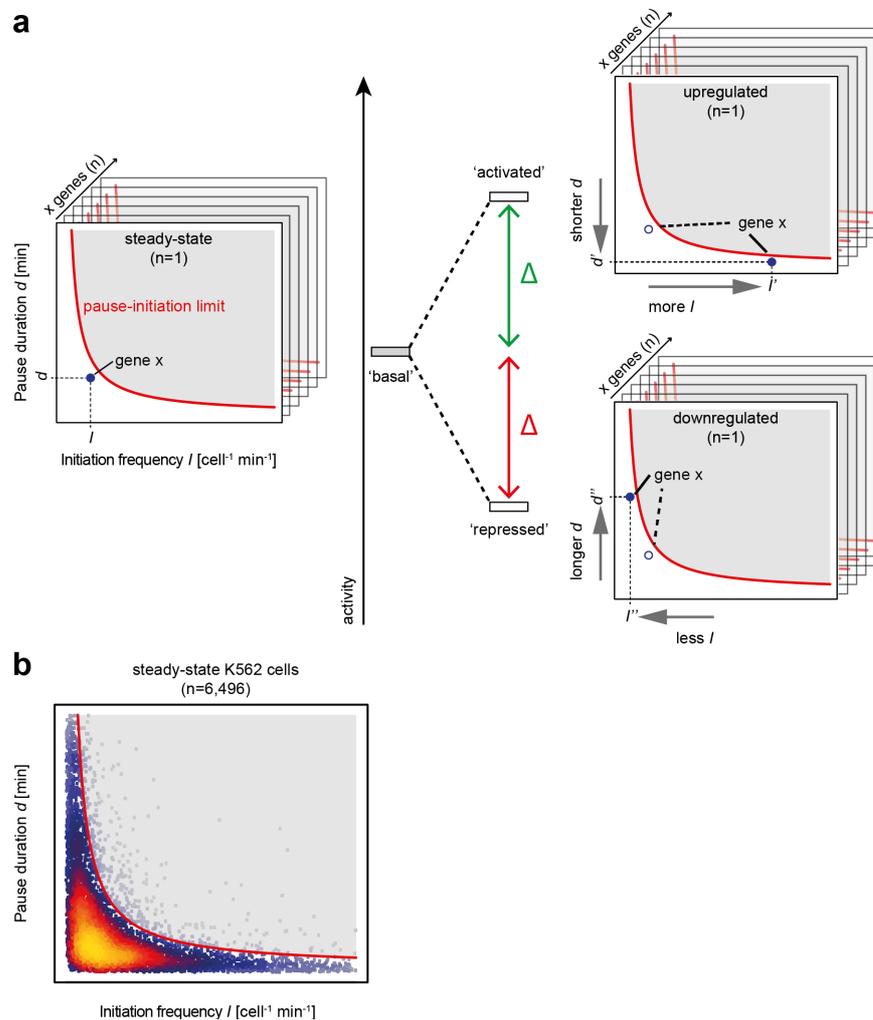
#### 1.1 A novel multi-omics approach

*Question 1. Can we measure the kinetic parameters initiation frequency, pause duration, and elongation velocity globally?*

Major open questions in systems biology relate to how cells control when and how much of the molecular information contained in their genome is transcribed into RNA. To capture the complexity of this regulation, a quantitative description of the transcriptional output and the kinetics of the transcription process itself is crucial. To date, methods are available to quantify RNA synthesis and degradation rates<sup>368,442</sup> (recently reviewed in<sup>443</sup>). However, at the outset of this work, efforts to capture transcription kinetics revealed limitations: the number of initiating polymerases (initiation frequency) and the time a polymerase needs to pass through the promoter-proximal window (pause duration) could not be quantified. Consensus methods in the field measured elongation velocity (in units of distance [bp] per time [min]) for genes longer than 5 kbp, and estimated the degree of Pol II pausing as relative ratio of promoter-proximal region and gene body occupancies of Pol II (see Introduction I.3.1).

Here, we solve this challenge with the combination of two state-of-the-art sequencing methods (TT-seq, mNET-seq) and novel kinetic modeling (multi-omics). Our approach enables genome-wide measurements of the kinetic parameters elongation velocity  $v$ , productive initiation frequency  $I$  and pause duration  $d$  in human cells (**Figure 18 a**, **Figure 22 a**). Specifically, the TT-seq signal (transcribed bases,  $tb$ ) in the exons of the gene body is proportional to the productive initiation frequency  $I$  (in units of molecules per cell per time [ $\text{min}^{-1}$ ]). Synthetic spike-ins allow for calculation of a conversion factor to RNA amount per

cell (see Methods II.2.2.4). The promoter-proximal pause site (PS) which is marked by a peak in Pol II occupancy can be determined by mNET-seq. The number of polymerases released into elongation (productive initiation frequency  $I$ ) measured by TT-seq can be combined with Pol II occupancy profiles measured by mNET-seq. This allows for quantitative conclusions to be drawn about the pause duration  $d$  (in units of time [min]) which is the time a polymerase resides in the pause window (PS +/- 100 bp) (**Figure 18 a**). These efforts have led to a kinetic description of genes in steady state, upon pause release inhibition and natural stress (**Figure 29**). Given that sufficient input material (number of cells) can be generated, our multi-omics approach enables for the calculation of kinetics in any cell line with minimal optimization (chapter 2). Ultimately, transcription kinetics are a critical building block in our understanding of transcriptional regulation which may enable us to develop novel treatments for diseases with transcriptional malignancies (chapter 3).



**Figure 29. Summary of the novel multi-omics analysis.**

(a) For simplicity, the kinetic parameters ( $I$ ,  $d$ ) are depicted for a model gene  $x$  in steady state (left) and upon upregulation (right, top) or downregulation (right, bottom). The multi-omics approach, allows us to follow the kinetic parameters of a specific gene, e.g. the HSPA1A (human HSP70) gene over the heat shock time-course (**Figure 25 b**). Furthermore, our multi-omics approach enables us to globally extract kinetic parameters of expressed genes (symbolized by shifted frames, see also **Figure 18**), or sets of genes (see **Figure 23**, **Figure 24**). Plots are ordered by transcription activity (see also **Figure 1 b**). (b) Kinetic analysis of 6,496 genes in steady state K562 cells.

## 1.2 The pause-initiation limit

---

Parts of the discussion presented in this subsection 1.2 has been published as:

**S. Gressel \***, B. Schwalb \* *et al.* CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* 6, doi:10.7554/eLife.29736 (2017).

(\* joint first authorship)

---

*Question 2. How can Pol II promoter-proximal pausing control the frequency of productive initiation and by that the amount of RNA synthesized per time?*

Taken together, our results show that Pol II pausing can control transcription initiation and demonstrate the central role of CDK9 in controlling pause duration and thereby the productive initiation frequency. Our results have implications for understanding gene regulation. Genes that show initiation frequencies below the pause-initiation limit may be activated by increasing the initiation frequency without changing the pause duration. However, activation of genes that are transcribed at the pause-initiation limit requires a decrease in pause duration, i.e. stimulation of pause release, to enable higher initiation frequencies. We suggest that pause-controlled initiation evolved because mutations in the promoter-proximal region can change pause duration, and thereby limit initiation, but do not compromise a high initiation capacity of the core promoter around the TSS. This may have enabled the evolution of genes that remain highly inducible but can be efficiently downregulated.

After our work had been completed, a publication appeared in *Nature Genetics* that concluded that polymerase pausing inhibits new transcription initiation<sup>444</sup>. The conclusion in this paper is consistent with our general finding of an interdependency of Pol II pausing and transcription initiation, but the two studies differ in three aspects.

First, we used human cells whereas the published work was conducted in *Drosophila* cells. Second, our work uses a multi-omics approach to enable a kinetic description, whereas the published work is based on changes in factor occupancy. Third, we selectively inhibited CDK9 using CRISPR/Cas9-based engineering and chemical biology, whereas the published work used small molecule inhibitors that may target multiple kinases (DRB, flavopiridol; see Introduction I.3.1.1). Despite these differences, the general conclusion that promoter-proximal pausing of Pol II sets a limit to the frequency of transcription initiation holds for both human and *Drosophila* cells and is likely a general feature of metazoan gene regulation.

To understand genome function, the regulatory steps of gene transcription have to be defined and it is crucial to analyze under which conditions they become rate limiting. A rate-limiting step may be defined as the slowest molecular transition in the process that limits the overall progression and the transcriptional output (reviewed in<sup>445,446</sup>). As outlined above, we and another group concluded that prolonged promoter-proximal pausing of Pol II impairs new initiation, and thus, reduces the amount of mRNA synthesized per time (‘pause-initiation limit’)<sup>381,444</sup>. Mechanistically, however, it is not fully understood if steric hindrance by the

paused polymerase (**Figure 2 b**) is the only way to restrict initiation, or if promoter-proximal pausing additionally represses Pol II recruitment to the promoter or PIC assembly. In the future, the interdependence of Pol II pausing and initiation should be investigated further, e.g. by structural analysis.

### 1.3 DNA sequence determines pause site but not pause duration

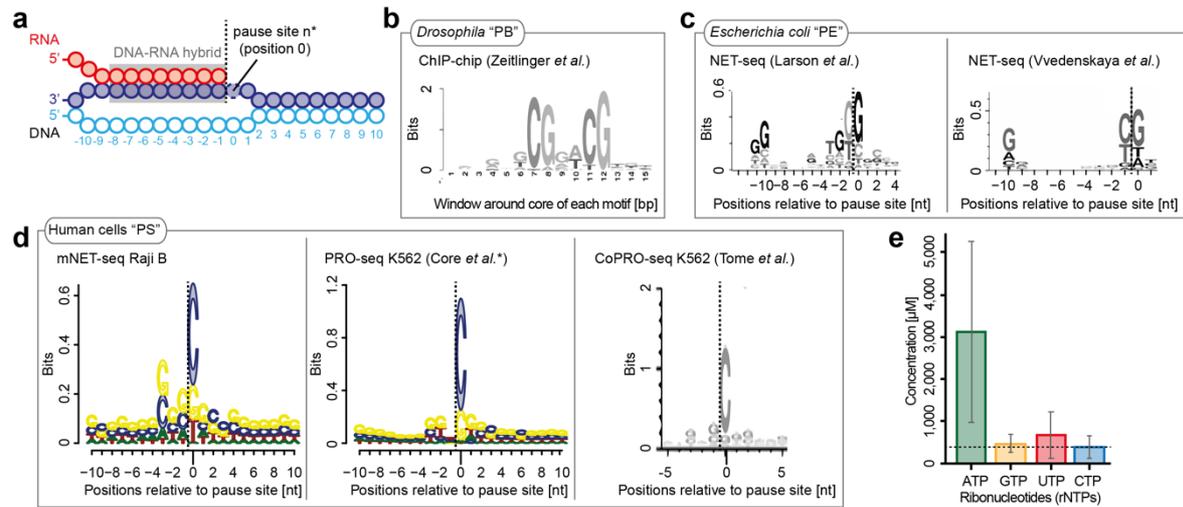
*Question 3. What are the determinants of long pause duration in human cells?*

Polymerase pausing is characterized by two parameters: the pause position or pause signal to enter the paused state, and the pause duration which is the dwell time until a Pol II molecule transitions from the promoter-proximal paused state to productive elongation.

Across different organisms, promoter-proximal polymerase pause positions can be called from polymerase peaks downstream of the TSS using (m)NET-seq, PRO-seq, or CoPRO data (**Figure 30 a**). The first DNA sequence analysis of genes with promoter-proximal Pol II peaks was performed on ChIP-chip data generated in *Drosophila melanogaster* embryos. In this study an enrichment of cytosine (C) and guanine (G) was observed for a window of [-5, + 50] around the TSS which the authors referred to as the pause button (PB)<sup>256</sup>(**Figure 30 b**). Technical advances in resolution by the development of PRO-seq defined the position of the *Drosophila* PB at ~40 bp downstream of the TSS<sup>347</sup>. At the *Drosophila* Hsp70 promoter, it was shown that insertion of sequences upstream of the PB sequence shifted the pause site downstream of its original position (e.g. the addition of 5 bp shifted the pause site ~5 bp downstream)<sup>347</sup>. This indicates that the PB defines the pause position in *Drosophila*. NET-seq analysis of pause site sequences in *Escherichia coli* (*E.coli*) showed the following features: G/C immediately upstream of the DNA-RNA hybrid in the active site of the polymerase, C/T at the 3' end of the RNA and G/C as the first incoming nucleotide (pause element, PE)<sup>257,447</sup> (**Figure 30 c**). Our work was the first study analyzing the pause sequence (PS) at human pause sites by mNET-seq. We found an enrichment for G/C-C/G dinucleotides with a strongly conserved C as the first incoming nucleotide (**Figure 17 b**). The strongly conserved C was also found in other mammalian cell lines – by re-analysis of published PRO-seq data (human K562 cells)<sup>50</sup> and by another recent study using CoPRO experiments (human K562 cells and mouse embryonic fibroblasts)<sup>247</sup> (**Figure 17 c**).

In summary, high-throughput sequencing of nascent RNAs found that C/G nucleotides are frequently enriched at polymerase pause sites in *Drosophila*, *E.coli* and mammalian cells. The conservation of DNA sequence elements at pause sites (PB, PE, or PS) across diverse species suggests that pausing might have evolved early and supports its key regulatory roles in prokaryotes and eukaryotes (reviewed in<sup>46</sup>). Another contributing factor might be that CTP and GTP are the least frequent ribonucleotides<sup>448</sup> (**Figure 17 d**). However, pause durations in bacteria (~1 second<sup>449</sup>) are much shorter than in *Drosophila* and mammalian cells (range of minutes, ~2.5 min in this study). This indicates that the pause duration must depend on stabilizing factors acting in *trans* such as NELF and DSIF which are not present in bacteria. Interestingly, the dependence on the P-TEFb kinase CDK9 for pause release varies significantly between human genes (**Figure 15 e**). Some genes (low responding) are capable of escaping the pause window independently of CDK9 while other

genes (high responding) are highly dependent on CDK9 (mean response  $\sim 60\%$ ). Several important questions about the factors ‘timing’ the pause duration remain to be answered.



**Figure 30. Pause site specification by sequence in flies, bacteria and humans.**

(a) Pause site  $n^*$  is defined at position 0 (see also **Figure 2 a**). The RNA 3'-end is depicted as dotted line. (b) Logos depict the level of sequence conservation (bits of information) at each position. In *Drosophila melanogaster* embryos, the pause button (PB) is enriched from [-5 to +55 bp]. *Ab initio motifsearch*<sup>256</sup> was performed on ChIP-chip data<sup>353</sup> (logo modified from figure 2 a of<sup>256</sup>). (c) In *Escherichia coli*, a consensus pause element (PE) was identified by NET-seq. Top: Larson *et al.*<sup>257</sup> aligned pause sequences at their 3' end and performed motif analysis by weblogo.berkeley.edu<sup>450</sup> (logo modified from figure 2 a of<sup>257</sup>). Bottom: Vvedenskaya *et al.*<sup>447</sup> analyzed pause sites by WebLogo<sup>451</sup> (logo modified from figure 1 a of<sup>447</sup>). (d) In human cells, the pause sequence (PS) was identified by mapping the active site of Pol II along nascent RNA at the pause site  $n^*$ . Left to right: mNET-seq of Raji B cells<sup>381</sup>, PRO-seq of K562 cells<sup>(50)</sup>, reanalyzed in this study), and CoPRO-seq of K562 cells<sup>247</sup>. Position weight matrix (PWM) logo representation of bases -10 to +10 [nt] around the pause site  $n^*$  (position 0). Tome *et al.*<sup>247</sup> analyzed sequences of the pause region (logo modified from figure 2 b of<sup>247</sup>). (e) Average physiological concentration of ribonucleotides across mammalian tumor cell cultures (ATP in green, GTP in yellow, UTP in red, CTP in blue). The values for average concentration [ $\mu\text{M}$ ] and the standard deviation are derived from table 4, column ‘tumor cells’ of<sup>448</sup>.

---

## 1.4 Transcription of enhancer elements is generally not pause-limited

---

The discussion presented in subsection 1.4 and 1.5 is currently under peer review for publication:

**S. Gressel** \*, B. Schwab \* *et al.* The pause-initiation limit restricts transcription activation in human cells. *Nature Communications*, in revision.

(\* joint first authorship)

---

*Question 4. Are there differences with respect to average Pol II pausing and/or initiation behavior of coding compared to noncoding transcript classes?*

Changes in the kinetics of Pol II initiation and pausing had not been quantified genome-wide during a transcription response, which is required as definitive evidence that natural gene regulation is controlled by pausing kinetics. To test this assumption, we conducted a genome-wide kinetic analysis of the heat shock response in human cells.

We quantified transcription kinetics of protein-coding and noncoding genes in steady state and during the dynamic transcriptional response of human cells to heat shock. To this end, we annotated protein-coding RNAs (mRNAs), and six major long noncoding transcript classes, i.e. lincRNAs, asRNAs, eRNAs, uaRNAs, conRNAs, and sincRNAs in human hematopoietic cell lines (K562 and Raji B) (**Figure 22 b-e**, **Figure 49 a**). We then used our multi-omics approach to follow changes in productive initiation frequency  $I$  and pause duration  $d$  in a quantitative manner. In steady state, we observed a reciprocal behavior of  $I$  and  $d$  for genes encoding all transcript classes except eRNAs. Protein-coding and lincRNAs were among the classes with the shortest pause durations, consistent with their high transcription levels. The longest median pause duration was observed for upstream antisense RNAs (uaRNAs), presumably impairing initiation events in the noncoding direction of bidirectional promoters. Enhancers showed a median pause duration of 2.2 min and a median productive initiation frequency of only one initiation event every 6-7 min (**Figure 23 c-d**). Thus, enhancer elements are generally not pause-limited. We found that in contrast to the other transcript classes, the nascent eRNA upstream of the pause site does not adopt a stable secondary structure (**Figure 23 e**). This could be an explanation for the short pause durations. Furthermore, it might be that the absence of stabilizing factors at enhancer elements leads to shorter pause durations.

## 1.5 Towards a model for gene activation

*Question 5. How are rapid and reversible changes in transcription upon stress achieved at a molecular level? Specifically, are changes refined by the pause-initiation limit?*

So far, a model was proposed where transcriptional regulation upon heat shock is coordinated at the single step of promoter-proximal pause release<sup>219,231</sup>. Here we could show that this holds true for genes that are close to the pause-initiation limit such as protein-coding genes (**Figure**

**24 c).** Using the transcriptional response to heat shock, we show that upregulation of the productive initiation frequency is restricted by the P-TEFb kinase CDK9. Enhancers form a notable exception to this rule, because changes in pause duration do not cause changes in eRNA production during heat shock (**Figure 27**, **Figure 28 c-d**).

More generally, upregulation of a gene requires an increase in initiation frequency, which leads to a higher number of polymerases loaded onto the gene, and a higher amount of RNA synthesis over time (**Figure 29**, top). At genes that are at the pause-initiation limit, pausing limits initiation, and a decrease in pause duration is required for upregulation of transcription by allowing for higher initiation frequencies. This is apparently often the case at protein-coding genes. In contrast, upregulation of enhancer transcription is often possible without changes in pause duration because enhancers are generally not pause-limited. These observations lead to a simple model for gene activation, which involves upregulation of enhancer transcription by an increase in productive initiation frequency, followed by a decrease in pause duration at the protein-coding target genes, which in turn allows for an increase of productive initiation frequency at the target genes. These mechanisms however rely on the availability of polymerases and transcription factors, and assume that transcription is generally processive.

Considering the critical role of factors decreasing the pause duration for gene activation, the mechanisms of polymerase pause release should be studied further in the future. Despite the P-TEFb kinase CDK9, a role in pause release upon heat shock has been reported for other factors: the TFIIF-associated kinase CDK7<sup>452-454</sup>, the elongation factor TFIIS<sup>273</sup>, the DNA-PK kinase, the ATM kinase, the 7SK snRNP recruitment factor TRIM28/KAP1<sup>455</sup>, the pause stabilizing factor GRINL1A/GDOWN1<sup>262</sup>, and the CTD phosphatase FCP1<sup>456</sup>. These factors were mainly studied at selected model genes. It is however unclear if these factors reduce the pause duration directly, or if they operate indirectly, e.g. by recruitment of P-TEFb to activated genes. Methods developed here and elsewhere can now be used to dissect the involvement of these factors as well as the kinetics underlying the mechanisms of P-TEFb delivery and activation in a quantitative and genome-wide manner, ultimately unravelling the nature of gene regulation in mammalian cells.

## 2. Technical improvements

Our multi-omics approach constitutes a systematic tool for measuring transcription rates in living cells. The current input requirements for TT-seq are 300 to 600  $\mu\text{g}$  of total RNA to yield  $>100$  ng of labeled RNA which is the optimal input for the cDNA library preparation. If sufficient input material is available, TT-seq can be applied in a broad range of cells including *Drosophila* and mammalian cells<sup>168,374,457</sup>. For mNET-seq,  $\sim 1 \times 10^8$  of mammalian cells (K562, Raji B, HeLa S3, murine lymphoid cells) are required for a good coverage across the genome. Transferring the mNET-seq protocol to other cell lines may require optimization of two steps, cell fractionation and chromatin solubilization. If the number of cells is not available for mNET-seq experiments, mNET-seq can be replaced by another occupancy profiling technique which requires less input material (e.g. PRO-seq). However, it should be noted that PRO-seq does not monitor backtracked or arrested polymerases and may thus underestimate the total number of polymerases at a gene.

As for every experimental approach, the two underlying techniques (TT-seq and mNET-seq) can be improved further to increase the efficiency (TT-seq, mNET-seq) and quantitative accuracy (mNET-seq) of the current protocols.

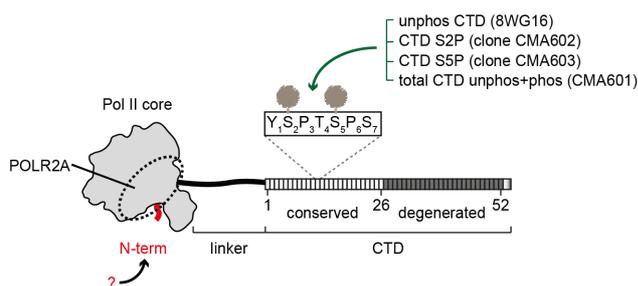
### 2.1 TT-seq

To date, the application of TT-seq is primarily limited by its need for large amounts of input material. This is required to perform metabolic labeling and subsequently, biochemical separation of newly synthesized from pre-existing RNAs. Depending on cell type specific parameters (doubling time, transcriptional activity, and amount of cellular RNA), between 300 and 600  $\mu\text{g}$  of total RNA is needed to isolate sufficient amounts of labeled RNA for sequencing. Typically, this corresponds to  $1.5$  to  $5 \times 10^7$  cells. Overcoming these limitations would allow application of TT-seq in cases of rare cell populations and limited input material as those commonly obtained in clinical settings. To do so, either the metabolic labeling efficiency (not discussed here), or the efficiency of the biochemical separation step of 4sU-labeled RNA fragments from unlabeled fragments must be increased. The separation step is essential to focus the sequencing capacity on newly synthesized RNA fragments which were transcribed during the 5 min labeling pulse (see Introduction I.3.2.1). The current protocol uses EZ-link HPDP-biotin which allows thiol-specific biotinylation of 4sU-labeled RNA. It was suggested that methylthiosulfonate-activated biotin (MTS-biotin) can be used as a more efficient alternative to HPDP-biotin<sup>458</sup>. Our group performed extensive control experiments to evaluate the described use of MTS-biotin for our TT-seq protocol. To assess the specificity of both biotins, we used 4sU-free total cellular RNA spiked-in with 0.1 % of 4sU-labeled spike-ins. When using MTS-biotin we found high levels of cross-contamination of the labeled RNA with unlabeled RNA (Katja Lidschreiber, *unpublished*). This is in agreement with Marzi *et al.* who detected 10-fold higher background levels for MTS-biotin compared to HPDP-biotin<sup>459</sup>. Recently, an improved MTS-biotin protocol including optimized biotin chemistry (MTS-resin) was developed<sup>460</sup>. We have not tested this improved biotinylation protocol yet, but, if the new protocol overcomes the specificity issues it might offer an alternative to the currently used HPDP-biotin and could be a step towards overcoming the limitation of input material.

## 2.2 mNET-seq

The three main technical challenges in mNET-seq are (i) the low efficiency (requirement of large amounts of input material) and imperfect specificity (heterogeneity of the target epitope) of the POLR2A antibody, (ii) the amplification bias (which might reduce quantitative accuracy), and (iii) the global normalization strategy (crucial for perturbation experiments).

One of the critical steps in mNET-seq is the immunoprecipitation of the Pol II-RNA complex. Several antibodies are available for the largest subunit of Pol II, POLR2A. They target the C-terminal domain (CTD) of POLR2A, specifically, unphosphorylated CTD (8WG16), CTD with phosphorylation(s) at serine 2 (S2P) (MBL, clone CMA602), CTD with phosphorylation(s) at serine 5 (S5P) (MBL, clone CMA603), or total, unphosphorylated and phosphorylated CTD (MBL, clone CMA601) (**Figure 31**). Since the total number of Pol II molecules is required for kinetic modeling, the monoclonal MBL antibody (clone CMA601)<sup>350,461</sup> was used to immunoprecipitate total POLR2A. During the transcription cycle, the CTD is heavily phosphorylated and interacts with multiple transcription factors (see CTD code, Introduction I.1.1). The MBL antibody was validated by overall immunoreactivity in ELISA (enzyme-linked immunosorbent assay)<sup>462</sup> against purified peptides with and without phosphorylations (antigens)<sup>350</sup>. ELISAs are limited by the provided antigen information and *in vivo* target epitopes of certain Pol II states might be masked by transcription factors leading to an underrepresentation in the mNET-seq data. However, comparison of Pol II occupancies measured by mNET-seq to published PRO-seq data which do not rely on Pol II antibodies (Introduction I.3.2.2) indicates that the fraction of Pol II that cannot be captured with a total CTD antibody must be very low. Two technical developments could circumvent the aforementioned problems: epitope tagging of POLR2A by CRISPR/Cas9<sup>343</sup>, or the development of a novel N-terminal POLR2A antibody. Introducing an epitope tag may influence stability or interactions of Pol II with transcription factors and thus, differ from the wildtype Pol II and would affect the outcome of the mNET-seq experiment. As an alternative, the N-terminus of POLR2A would provide a less heterogenous target compared to the CTD, but to date no efficient antibody for IP is commercially available (**Figure 31**; see **Figure 13 d** for Santa Cruz’s antibody F12 raised against the N-terminus of POLR2A). The generation of a novel N-terminal POLR2A antibody would be advantageous<sup>463</sup>, e.g. by collaboration with in-house Nanobodies experts (MPI-bpc, Dept. of Cellular Logistics, Prof. Dr. Dirk Görlich). Nanobodies are single-domain antibodies of camelid origin<sup>464,465</sup> which can be produced renewable as recombinant fusion proteins in bacteria without the need of immunizing and killing of animals<sup>466</sup>. Epitope tagging or development of a novel antibody with higher efficiency would also allow to use less input material for the IP.



**Figure 31. Target regions of human POLR2A antibodies.**

The largest subunit of human Pol II, POLR2A (position highlighted by circle) is used as target for commercial antibodies. For the N-terminus, no efficient antibody for IP applications is available to date and the field would benefit from a novel N-terminal antibody. Boxes show consensus repeat of YSPTSPS<sup>107</sup> and recognized phosphorylation sites (brown dots).

All current RNA-based Pol II profiling methods use amplification by PCR in their library strategy (i.e. GRO-seq 12-20 PCR cycles<sup>467,468</sup>, PRO-seq 12-18 PCR cycles<sup>347</sup>, mNET-seq 10-15 PCR cycles<sup>388</sup>). To correct for amplification bias, namely to distinguish seemingly identical fragments arising from distinct RNA molecules from copies arising through PCR amplification of the same RNA molecule, Unique Molecular Identifiers (UMIs) have been described for RNA-protein crosslinking methods or single-cell RNA-seq<sup>469,470</sup>. UMIs are random oligonucleotide barcodes which make each RNA molecule unique and allow quantitative read counting after correction for amplification bias. For mNET-seq (and all current small RNA-seq methods), overlapping or identical reads arising from distinct RNA molecules are expected in regions with a high number of polymerases or regions with slowly elongating polymerases. Thus, the majority of reads mapping to identical coordinates might not be due to PCR duplication but rather from biological distinct RNA molecules and should also be considered for quantification. It was recently shown by Fu *et al.* that for small RNA-seq data the majority of sequencing reads (~60 %) are counted as PCR duplicates. However, if UMIs were used to distinguish identical copies from unique RNAs from PCR duplicates only ~5 % were actual duplicates<sup>471</sup>. Thus, removing PCR duplicates without UMIs would lead to a biased, unbalanced removal of identical but biological meaningful mNET-seq reads. The addition of UMIs in the cDNA library would improve the quantitative accuracy of mNET-seq Pol II occupancy changes.

To compare Pol II occupancies reproducibly among experiments (control versus perturbation), normalization is required. There are a number of different normalization strategies, including normalization to sequencing depth, quantile normalization and other methods. In this study, we employed a set of ‘neutral’ genes which were defined by TT-seq (see Methods II.2.3.5 and II.2.4.2) for global normalization. However, certain perturbations might not allow for this normalization strategy. To address the problem of global normalization in these cases, IP spike-ins have been described<sup>472,473</sup>. To this end, a constant number of Pol II-RNA complexes from another organism (e.g. *Drosophila*) is spiked-in prior to IP. The spike-in controls serve as internal calibrators for technical variations during IP and subsequent steps of the protocol. A crucial prerequisite of this approach is that IP antibodies are used that target conserved peptides across organisms. Unfortunately, the currently used MBL antibody does not recognize *Drosophila* or yeast Pol II due to sequence differences between their CTDs<sup>474</sup>. Thus, a novel N-terminal antibody would allow immediate application of spike-ins during IP since the N-terminus of POLR2A is conserved across organisms.

### 3. Future explorations

**Box. Outstanding challenges in the pausing field.**

General pausing mechanism

- (1) What influences Pol II pause duration at pause sites? Is it a specific factor or the combinatorial action of multiple factors (regulatory network)? \*
- (2) What transcription factors and chromatin regulators are involved in changing the pause duration in response to cellular signaling pathways? \*
- (3) How is pause duration regulated in a gene or domain (e.g. TAD) specific manner? \*
- (4) What are the consequences of pausing for development? How is paused Pol II established *de novo* at previously closed promoters during development (special pioneering factors)?
- (5) How do enhancers (and eRNAs) facilitate pause-release? \*
- (6) Which steps are disrupted and lead to aberrant release of pausing in disease? \*
- (7) How much non-productive initiation (promoter-proximal attenuation) occurs genome-wide upon steady state? \*\*

P-TEF recruitment

- (8) How are the various recruitment mechanisms for P-TEFb coordinated and regulated?
- (9) How is the ratio of anchored Brd4-P-TEFb and anchored KAP1-7SK-snRNP-P-TEFb on chromatin? Are there sets of genes which have related and highly similar compositions? \*
- (10) How is active P-TEFb recruited to specific promoters/enhancers? \*
- (11) How is inactive, chromatin-anchored KAP1-7SK-snRNP-P-TEFb released on-site at specific promoters/enhancers?
- (12) How does P-TEFb after its release become part of SEC, and what is the link for SEC assembly (suggested for AFF1 and AFF4<sup>475</sup>)?

\* covered in the main text (chapter 3); \*\* requires development of a novel method

This work represents a detailed genome-scale picture of kinetic parameters and serves as a template for future studies measuring transcription kinetics. However, to arrive at a mechanistic understanding of the peculiar step of Pol II pausing *in vivo* it must be studied further by systems biology approaches (and as mentioned in the introduction, by other methods as well to arrive at a holistic understanding of the process). In systems biology, a four-step framework can be applied to most challenges<sup>476</sup>. This requires (i) the knowledge of all components involved in the pausing step, (ii) systematic perturbations of pausing-related factors and monitoring changes in transcription kinetics or changes of components (e.g. by post-translational modifications), (iii) the development of one or several pausing model hypotheses, and (iv) to repeat the previous steps to expand, refine and test the pausing model(s) by experimental data which might depend on the development of novel techniques.

In this perspective, I focus first on a strategy to identify the set of pausing factors at specific promoters (step i, subsection 3.1), and state-of-the-art tools to rapidly perturb pausing-related factors (step ii, subsection 3.2). Combining rapid perturbations with our multi-omics approach will provide answers to many of the outstanding challenges in the pausing field (listed in the **Box** above). At the end, I will briefly cover the requirements of a novel methodology to address promoter-proximal attenuation (step iv, subsection 3.3), and

the application of our multi-omics approach to study elongation velocity changes along the gene in the context of co-transcriptional processes such as splicing (subsection 3.4).

Future studies of pausing promise many important discoveries: since de-regulation of pausing is involved in multiple diseases (e.g. MLL-fusion leukemia, HIV infection, cardiac hypertrophy; summarized in <sup>250</sup>), an understanding of Pol II pausing regulation holds great promise for novel targeted therapies. Furthermore, cancers are an attractive target for multi-omics studies of transcription kinetics because cancer is a disease of gene regulation (as discussed in an excellent recent review <sup>477</sup>).

### 3.1 Gene-specific identification of (un)known pausing factors

Engineered DNA-binding molecule mediated chromatin immunoprecipitation (enChIP)<sup>478,479</sup> in combination with mass spectrometry (MS) offers a strategy to isolate physiologically relevant complexes within a specific genomic region. The protocol consists of a cell fractionation step <sup>418</sup>, crosslinking of DNA and proteins within isolated nuclei <sup>480</sup>, chromatin isolation and fragmentation by sonication (1-2 kbp fragments), as well as MS analysis (**Figure 32 a**). Challenges for the development of a reproducible enChIP-MS protocol will be the identification of a suitable crosslinker, testing the efficiency and background of epitope tagged nuclease dead Cas9 protein with guide RNA <sup>481</sup> for the IP of the DNA at a specific locus, and identifying a suitable MS normalization strategy between samples. The characterization of factors enriched at pause sites of selected genes with long pause duration versus genes with short pause duration, and changes of the repertoire upon perturbation will provide new insights (**Box**, questions no. 1-3, 9).

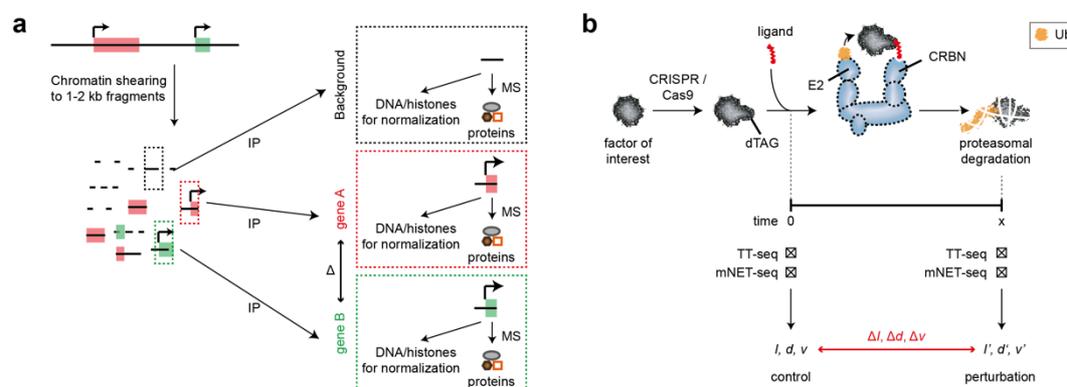
### 3.2 Perturbation tools with high kinetic resolution

Combining the read-out of transcription kinetics by our multi-omics approach with a specific and rapid perturbation of pausing and elongation factors, or of chromatin architectural factors will be highly informative (**Box**, questions 1-2, 6, 8). To date, multiple perturbation strategies are available including genetic approaches (CRISPR/Cas9 system <sup>482</sup> and RNA interference <sup>483</sup>), pharmacologic inhibitors (small molecules), and target(ed) protein degradation (TPD) (direct: Protacs <sup>484</sup>, molecular glues <sup>485</sup>; indirect: dTAG strategy <sup>486</sup>, AID system <sup>487</sup>). The strength and caveats of the different strategies were recently compared in an excellent technology review <sup>488</sup>.

As mentioned above, a specific and rapid perturbation is critical for our application to dissect the primary role of the respective factor. For this purpose, the indirect chemical-genetic TPD approach performs best on a wide range of target proteins especially if structural information is available <sup>488</sup>. Two strategies are commonly used, the degradation tag (dTAG) and the auxin-inducible degron (AID) system. Both strategies require engineering (addition of a tag for ligand binding) of the endogenous gene which encodes the protein of interest by CRISPR/Cas9 mediated knock-in. Once the tagged protein of interest is expressed and a standardized ligand is added to the cells, the ligand binds to the tag and re-directs the ubiquitination activity of ubiquitin E3 ligases to the target protein. Ubiquitination leads to a target-specific proteasomal degradation. The advantage of the dTAG system is that it employs

an endogenous E3 ligase complex (CRBN) while AID requires ectopic expression of a ligase complex (TIR1). The dTAG-ligand is a small molecule which can rapidly enter cells<sup>486</sup>. The top five targets for perturbation in combination with multi-omics experiments (**Figure 32 a**) are the pausing factor NELF, the insulator CTCF, the PAF1 complex, the cap-binding complex (CBC) and the cyclin subunit of P-TEFb (CCNT1) (see also **Table 35**).

Genome architecture and transcriptional activity are intimately linked: the 3D nuclear architecture governs when and to what extent the linear information might be accessed. In this study, we found that genes with high CDK9 responsiveness are also associated with long-range chromatin interactions (**Figure 21 f**), suggesting that enhancers can influence the pause-initiation limit at target genes to regulate transcription. Thus, it would be very interesting to dissect pause duration changes after perturbation of 3D chromatin loops. Hereby established enhancer-promoter interactions would be disrupted. These loop structures are formed by the interaction of DNA with two CTCF insulator proteins, and cohesin<sup>489</sup>. Using the above introduced dTAG system, CTCF and cohesin might be interesting targets to disrupt the genome architecture. Knockdown (KD) experiments of CTCF suggested that CTCF is required for P-TEFb delivery at the model gene C-MYC<sup>490</sup> and a viral transcript<sup>491</sup>. This supports our model that certain chromatin loops established by CTCF and cohesin allow for enhanced P-TEFb delivery to release paused Pol II, especially at genes with long pause durations.



**Figure 32. Future developments and applications of our multi-omics approach.**

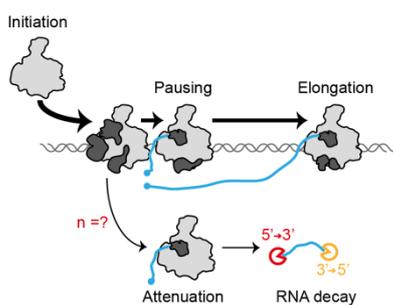
(a) Identification of novel or gene-specific pausing factors in human cells. For details see main text. The experimental strategy was developed after a discussion with Prof. Dr. Henning Urlaub (MPI-bpc, Dept. Bioanalytical Mass Spectrometry). (b) Targeted protein degradation (TPD) using dTAG strategy in combination with our multi-omics approach.

### 3.3 The ratio of non-productive to productive initiation frequency

The TT-seq<sup>168</sup> protocol enables us to measure productive initiation frequencies *in vivo*. However, since TT-seq requires the incorporation of 4sU into newly synthesized RNA, shorter RNAs (<100 nt) have a higher probability to escape the labeled RNA purification. As a consequence, the coverage upstream of the pause site cannot be recovered fully (depending on its sequence composition and 4sU content). Our comparison of genes with long pause durations

to genes with short pause durations (both groups had similar U content) showed that less labeled RNA was recovered for genes with long pause durations (**Figure 19**). This suggests that the fraction of Pol II that might terminate in the pause window (promoter-proximal attenuation, see **Supplementary Note 1**) is generally low. However, it is technically challenging to quantify the extent of promoter-proximal attenuation (non-productive initiation frequency) (**Figure 33**). The combination of productive initiation frequency with the unknown fraction of non-productive initiation frequency yields the ‘true’ initiation frequency. As described in **Supplementary Note 1**, occupancy measurements are insufficient to define the fate of paused Pol II. Below, two ideas for novel approaches to tackle this challenge are listed:

- Estimate the ‘true’ initiation frequency by capping events per time.  
Technical challenges: develop an *in vivo* cap-labeling approach, e.g. using cap analogs (NAD<sup>+</sup> etc.). Optimize efficiency of metabolic uptake and incorporation of cap analog.
- *In vitro* transcription assay to reconstitute the transition from initiation via promoter-proximal pausing/attenuation to productive elongation.  
Technical challenge: identification, purification and complex assembly of essential pausing and release factors and Pol II.



**Figure 33. The challenge to quantify promoter-proximal attenuation.**  
The figure is adapted from <sup>357</sup>.

### 3.4 Elongation velocity at high resolution across the gene

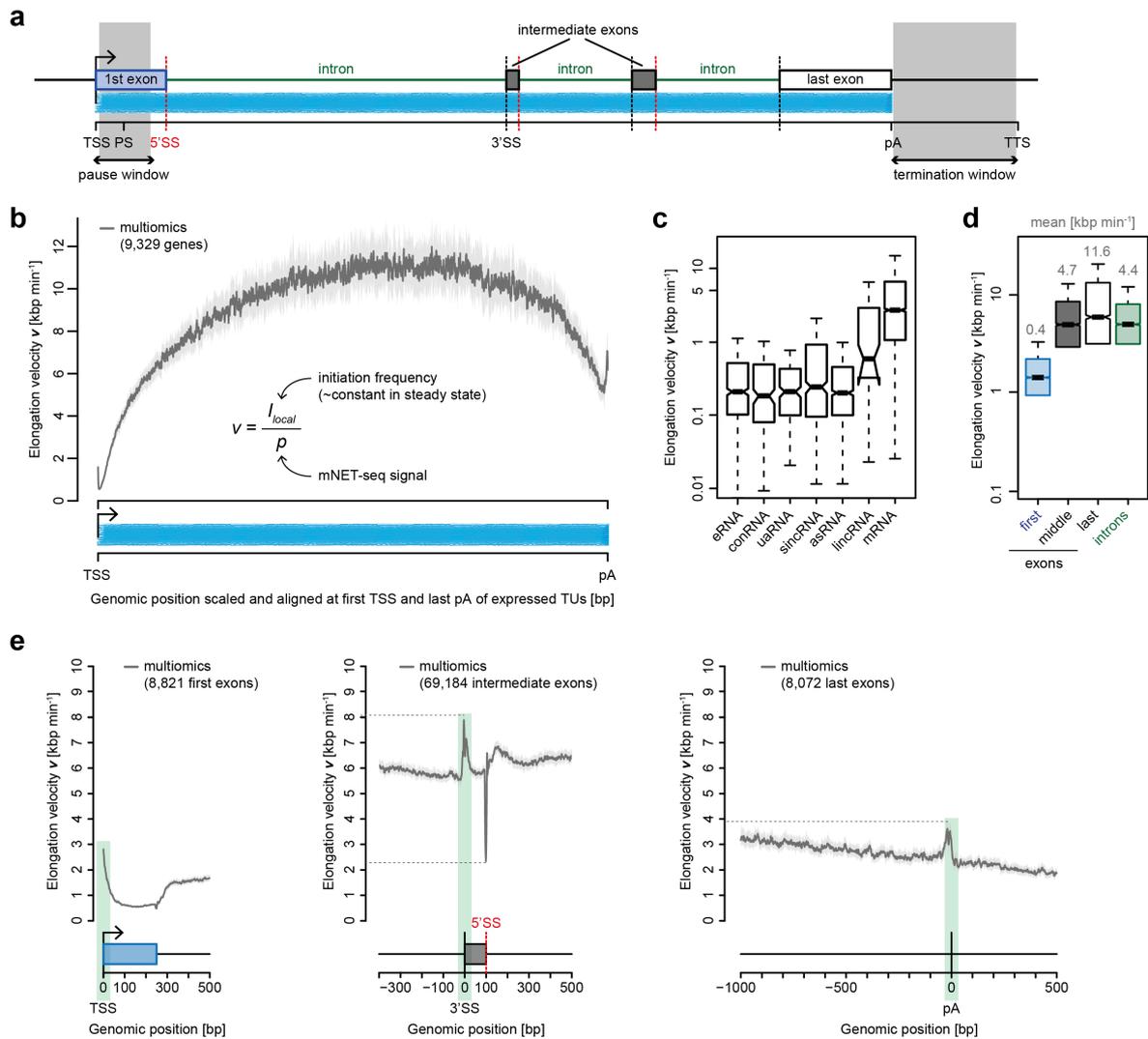
In order to introduce the reader to the relevant information needed to understand the next paragraphs I will briefly revisit the human gene architecture (**Figure 34 a**). A gene consists of exons which code for the mature transcript, and introns which are removed by splicing (see also Introduction I.1.2). Introns of nascent RNAs are mostly spliced co-transcriptionally <sup>392,492,493</sup>. For each intron the spliceosome needs to assemble, thus, multiple spliceosomes are needed per transcript (recently reviewed in <sup>71</sup>). The 5' end of the first exon is defined as transcription start site (TSS). Downstream of the TSS is the Pol II pause site (PS) which marks the center of the promoter-proximal pause window (PS +/- 100 nt). The 3' end of the last exon is referred to as poly(A) (pA) site. The RNA is cleaved at the pA signal and most RNA 3' ends get a poly(A) tail <sup>65</sup>. Pol II continues transcription downstream of the pA site until it reaches a transcription termination site (TTS). The region between pA site and TSS is defined as termination window. Slowly elongating polymerases are not only observed in the promoter-proximal window but can also be found at intron-exon junctions (splice sites) and over the termination window <sup>19</sup>. Our multi-omics approach quantifies

elongation velocities at different genes (**Figure 34 b**, *unpublished*), at genes encoding different transcript classes (**Figure 34 c**, *unpublished*), or at specific gene segments (**Figure 34 d**, *unpublished*). Multiple genomic features and factors modulate the Pol II elongation velocity  $v$  (**Figure 16 e**). Here, I focus on the interplay of  $v$  with co-transcriptional splicing and the dynamic histone code (see also Introduction I.1.1).

*The complex interplay of elongation velocity and splicing.* Slow Pol II elongation velocity can promote splicing and inclusion of alternative exons<sup>494-496</sup>, but it can also lead to exon skipping by extending the window of opportunity to recruit splice repressors<sup>497</sup>. It was recently shown in plants that light can accelerate Pol II elongation rates along photosynthesis genes, and as result of the changed velocity, the corresponding transcripts are subject to alternative splicing<sup>498</sup>. High resolution (close to single-nucleotide precision) is required to distinguish  $v$  at exon-intron borders. As discussed, occupancy profiling is ambiguous (Introduction I.3.1.2, **Supplementary Note 2**). But, it can provide clues if Pol II might elongate more (local minimum) or less rapidly (accumulation) in specific regions of the gene, assuming that Pol II transcription is generally processive and the drop-off rate within the gene body is low. This was indeed suggested by published PRO-seq data<sup>347</sup>: alignment at 3' splice sites (3'SS, intron-exon border) showed a drop in Pol II occupancy at the end of the intron, immediately followed by Pol II accumulation at the beginning of the exon. Alignment of PRO-seq density at 5' SS (exon-intron border) showed Pol II accumulation at the end of the exon, and decreasing Pol II occupancy within the first 50 bp of the intron after which it returns to baseline levels<sup>347</sup>. Another study found that the 3' ends of ~17 nt long splice-site RNAs (spliRNAs) align perfectly with the 3' end of exons and are conserved across species (human, mouse, *Drosophila*, *C.elegans* and marine sponge)<sup>499</sup>. This might indicate a longer residence time of the polymerases at the 3' end of exons since the nascent RNA associated with the elongation complex is longer protected (and thus, measurable). These observations agree with our kinetic measurements combining TT-seq and mNET-seq data of steady state human K562 cells (**Figure 34 e**, *unpublished*). Our approach allows to extract kinetic parameters at high resolution, and to compare the elongation velocity between segments of different genes which might have very diverse initiation frequencies,  $I_{local}$  which could not be interpreted by occupancy data alone.

Furthermore, it would be exciting to investigate if spacing between consecutive polymerases (spacing defined by the promoter-proximal pause duration) protects against polymerase traffic jams that could occur later due to pauses at splice sites. So far, it is unknown if promoter-proximal pause durations correlate with pause durations at exons *in vivo*. Thus, the analysis of transcription kinetics has the potential to reveal novel aspects of co-transcriptional splicing regulation.

*Elongation velocity and a dynamic histone code.* Thus far, we and others<sup>130,272,335,336,500</sup> have observed links between Pol II elongation velocity and histone modifications but much of this data is correlative in nature (**Figure 50**). To determine whether histone modifications are causative for fast or slow elongation velocity one would need to test mutants which cause more or less modifications and measure if elongation velocities would be changed. Another possibility is that modifiers traveling with Pol II have different time windows to act on the underlying chromatin during fast or slow elongation.



**Figure 34. Elongation velocity at high resolution across human genes.**

(a) Gene architecture. Colors as in (b) and (d). (b) Metagenome analysis of local elongation velocity [kbp min<sup>-1</sup>] along 9,329 genes in human K562 cells (steady state), depicted as scaled genomic position, aligned at first TSS and last pA site. (c) Boxplot of elongation velocity for different transcript classes annotated in human K562 cells: 886 eRNAs, 248 conRNAs, 235 uaRNAs, 877 sincRNAs, 1,281 asRNAs, 157 lincRNAs, and 4,582 mRNAs. Black bars represent median, boxes mark upper and lower quartiles, whiskers represent 1.5 times the inter-quartile range. (d) Boxplot of elongation velocity [kbp min<sup>-1</sup>] at different transcript segments: first exon ( $n = 8,674$ ) (dark blue), intermediate exons (middle) ( $n = 7,954$ ) (grey), last exons ( $n = 7,911$ ) (white), or introns ( $n = 8,788$ ) (emerald green). (e) Metagenome analysis of elongation velocity [kbp min<sup>-1</sup>] aligned at different transcript segments (reference point highlighted in green): aligned at TSS ( $n = 8,821$ ) (dark blue, left), exon start (3'SS) and end (5'SS) ( $n = 69,184$ ) (middle), or pA site ( $n = 8,072$ ) (right). List of splice sites was generated from RefSeq annotations.

## V. Supplementary Information

This chapter contains information in support of the Introduction, Materials and Methods, Results and Discussion which could not be integrated within the respective chapter due to space limitations.

### 1. Introduction

**Supplementary Note 1. The struggle to clarify the mechanistic basis of promoter-proximal events.**

For simplicity, ‘promoter-proximal pausing’ was introduced as an umbrella term for promoter-proximal events including Pol II stalling, pausing, abortive elongation, termination and attenuation. The observed Pol II peaks close to mammalian promoters could, however, be explained by rapidly initiating and then terminating polymerases (promoter-proximal attenuation). There has been a lot of effort in the field to measure and eventually understand the behavior of Pol II at promoter-proximal windows. But, to date, there is no method available to distinguish attenuation from long pause duration. None of the Pol II occupancy methods (see Introduction I.3.1.2), not even short capped RNA assays are able to distinguish the frequency of these events. Short capped RNA measurements are as well biased by the residence time of the polymerases since the nascent RNA associated with the elongation complex is longer protected (and thus, measurable) for loci with longer pause durations. In summary, it is technically challenging to assess the frequency of these events, and it will be a crucial task for future methods development to quantify the extent of each in order to define the contribution to transcription regulation in humans (see Discussion IV.3.3). At this point, I want to emphasize that the events upstream of the pause site are irrelevant to our model and its conclusions.

**Table 27. Genome-wide approaches used to study Pol II occupancy.**

Abbreviations: native elongating transcript sequencing (NET-seq), precision nuclear run-on sequencing (PRO-seq), immunoprecipitation (IP), global run-on sequencing (GRO-seq). ‘All’ Pol II states refers to initiating, elongation-competent, paused, backtracked, stalled, terminating Pol II states. Note that approaches listed here determine Pol II occupancies of potentially heterogenous populations of cells.

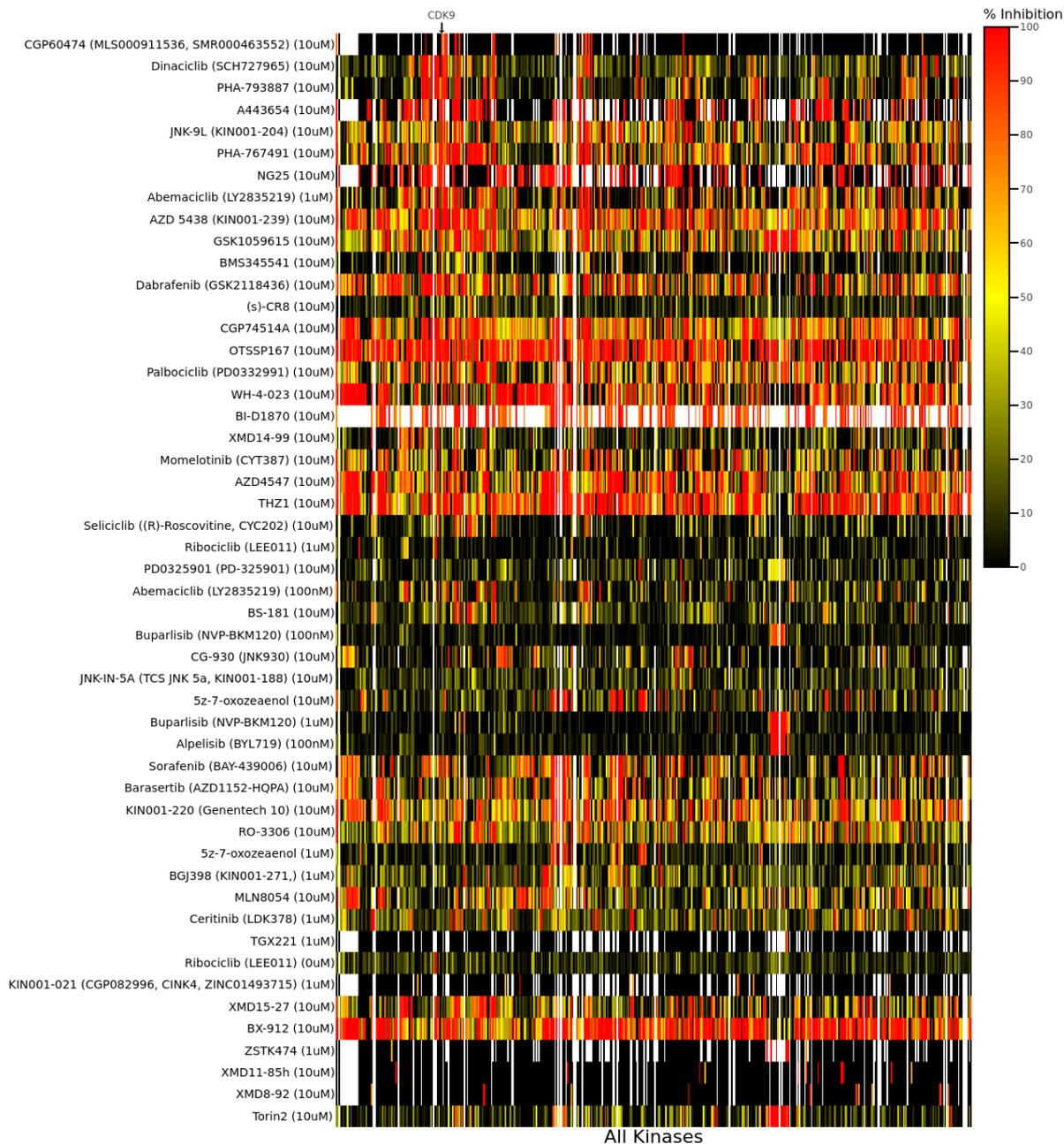
Method	Features						Refs
	Treatment	Chromatin solubilization	IP	Strand-specific?	Resolution	Mapping of Pol II states	
ChIP-seq	<i>in vivo</i> crosslinking	‘random’ sonication	Pol II	no	>200 bp	all	382-384,501
ChIP-exo, ChIP-nexus	<i>in vivo</i> crosslinking	$\lambda$ -exonuclease digest	Pol II	no	~50 bp	all	385,386
GRO-seq	<i>in vitro</i> run-on using sarkosyl	-	RNA	yes	~50 nt	only elongation-competent Pol II	346
PRO-seq	<i>in vitro</i> run-on using sarkosyl	-	RNA	yes	~1 nt	only elongation-competent Pol II	347
3’NT-seq, NET-seq	cellular fractionation	-	-	yes	~1 nt	all, except initiating and terminating Pol II	348,349
mNET-seq	cellular fractionation	MNase digest	Pol II	yes	~1 nt	all, except initiating Pol II	350,387

**Table 28. Comparison of Pol II elongation rate estimates derived from 1976 to 2019.**

‘Wash-out’ refers to the removal of a reversible drug (DRB), and subsequent tracking of the retrieving wave of Pol II molecules or its transcripts. Thus, a wash-out experiment resembles the outcome of gene activation by natural stimuli (e.g. TNF $\alpha$ , estradiol, heat shock). ‘Shut-down’ refers to disruption of transcription, and subsequent tracking of the declining wave of Pol II molecules or its transcripts. For abbreviations used below refer to **Table 38**.

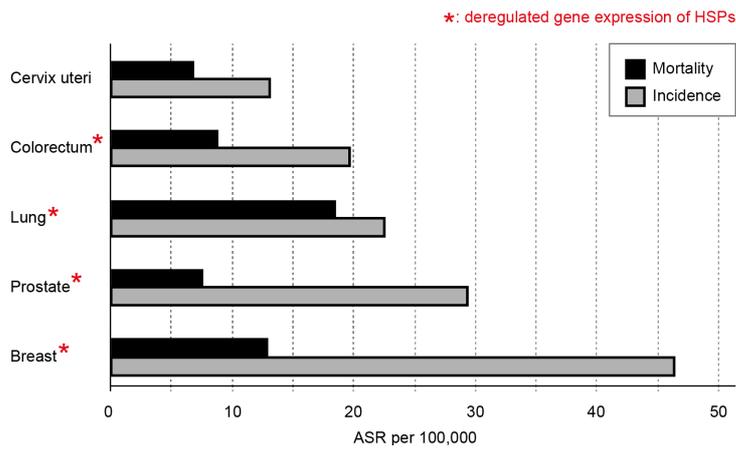
Year	Gene(s)	Organism or cell line	Elongation rate [kbp/min]	Pol II occupancy profiling	(nascent) RNA	Drug / stimulus	Ref
1976	hnRNA	Human HeLa cells	3–6	-	Radioisotope labeling	DRB (0-180 sec)	319
1984	MMTV (7.8 kbp)	Rat HTC cells	1.5	Nuclease protection assay	-	-	502
1990	E74 (60 kbp)	Drosophila	1.1	-	Northern blotting	-	503
1991	Ubx (78 kbp)	Drosophila	1.4	-	In situ hybridization	-	504
1993	Hsp70 (2.4 kbp)	Drosophila S2 cells	1.2-1.5	Pol II imaging, NRO assay	-	Heat shock	326
1995	DMD (2,300 kbp)	Human myogenic cells	2.4	-	RT-PCR	-	323
1998	ACTB (3 kbp)	Normal rat kidney cells	1.3	-	RNA FISH	-	329
2005	Hybrid gene (8 kbp)	Yeast	2	Pol II ChIP-qPCR	-	Galactose	322
2007	Hsp70 (2.4 kbp)	Drosophila	1.2-1.5	Pol II imaging (FRAP), NRO assay	-	-	327
2007	Engineered HIV reporter (3.8 kbp)	Human U2OS cells	1.9	Pol II imaging	RNA FRAP	-	330
2007	Engineered gene cassette (3.3 kbp)	Human U2OS cells	4.3	Pol II imaging	-	-	328
2009	10 native genes (100–580 kbp)	Human Tet-21 cells	3.8	-	RT-qPCR	DRB wash-out	324
2009	5 genes (>100 kbp)	Human HUVEC cells	3.1	Pol II ChIP-qPCR, tiling microarrays	-	TNF $\alpha$ (0-3 h)	325
2009	heat shock genes (Hsp70, Hsp83)	Drosophila S2 cells	1.1	Pol II imaging, Pol II ChIP-qPCR	-	Heat shock (0-20 min)	272
2010	Mouse ACTB (3.6 kbp)	Human U2OS Tet-On stable cells	3.3	-	RNA imaging	Doxycycline-induced	331
2011	4 genes (e.g. $\beta$ -globin mini-gene)	Human U2OS Tet-On stable cells	3.6	-	RNA FISH, FRAP	Actinomycin shut-down	332
2011	Yeast gene MDN1 (15 kbp)	Saccharomyces cerevisiae	1.2	-	RNA imaging	-	333
2013	140 genes (MCF-7) / 26 genes (AC16)	Human MCF-7 / AC16 cells	2-3	GRO-seq	-	E2 / TNF $\alpha$ (0-40 min)	334

2014	1,577 genes	Human HeLa cells	3.5	-	4sU-seq	DRB wash-out (0-8 min)	335
2014	8,529 genes	Human HEK293	3.13	GRO-seq	-	DRB wash-out (0-40 min)	129
2014	938 genes	mESCs		GRO-seq	-	TRP / FP shut-down (0-50 min)	130
2014	1,932-2,702 genes	Human fibroblasts (HF1, TM, CS-B), cancer-derived cells (K562, MCF-7)	1.5	-	BrU-seq	DRB wash-out (0-10 min)	336
2014	4,587 genes	Human HEK293 Flp-in cell	1.7	GRO-seq	-	DRB wash-out (0-20 min)	132
2016	88-2,786 genes	Mouse ESCs	1.8-2.4	PRO-seq		Heat shock (0-1 h),	231
2017	reporter genes	Drosophila embryos	2.4-3.0	-	RNA imaging	-	505
2017	2,443 genes	Human Raji B cells	2.3	mNET-seq	TT-seq	1-NA-PP1 (0-15 min)	381
2018		Human HEK293, KBM-7 cells		TV-PRO-seq	-	-	506
2019	600-1,200 genes	Human HEK293 cells	2-3	Pol II ChIP-seq	BrU-seq	DRB wash-out (0-20 min)	305



**Figure 35. Commercially available CDK9 kinase inhibitors compared by KInhibition.**

Heatmap shows the inhibition profile of the top 50 CDK9 kinase inhibitors out of 133 compounds compared. Rows represent the tested compounds, columns represent the kinases screened. CDK9 is highlighted on the top with an arrow. White color highlights missing data. Data for ‘CDK9’ were compared using KInhibition, a kinase inhibitor selection portal (<https://kininhibition.fredhutch.org>)<sup>408</sup>. Resources are listed in <sup>408</sup>. Note that CDK9 kinase inhibitors available to date are unspecific. Full inhibition of the on-target CDK9 comes at the price of severe off-target inhibition (other kinases).



**Figure 36. Heat shock proteins are deregulated in human cancer cells.**

Bar plot shows the most common cancers worldwide (top five: breast, prostate, lung, colorectum, cervix uteri), estimated in 2018 of both sexes and all ages (estimated numbers from <sup>507</sup>). The age-standardised rate (ASR) per 100,000 estimates is depicted. Cancers with deregulated gene expression of heat shock proteins (HSPs) are marked with red asterisk. Transcriptome data are published for breast <sup>207,508</sup>, prostate <sup>509</sup>, lung <sup>207</sup>, and colon <sup>207</sup> cancer. For cervix uteri cancers, overexpression of Hsp27 was shown by immunohistochemistry <sup>510</sup>.

## 2. Materials and Methods

### 2.1 Sequencing data pre-processing

*Data pre-processing was performed by Dr. Björn Schwalb (MPI-bpc, Dept. of Molecular Biology).*

The methods presented in this section 2.1 have been published as:

**S. Gressel \***, B. Schwalb \* *et al.* CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* 6, doi:10.7554/eLife.29736 (2017).

(\* joint first authorship)

#### 2.1.1 TT-seq and RNA-seq

**TT-seq and RNA-seq data pre-processing.** Paired-end 50 or 75 base reads with additional 6 base reads of barcodes were obtained for each of the samples. Reads were demultiplexed and mapped with STAR<sup>403</sup> to the hg20/hg38 (GRCh38) genome assembly (Human Genome Reference Consortium). Samtools<sup>404</sup> 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<sup>406</sup> environment. We used a spike-in (RNAs) normalization strategy essentially as described<sup>168</sup> to allow observation of global shifts  $\sigma_j$  (sequencing depth), cross-contamination rate  $\epsilon_j$  (proportion of unlabeled reads purified in the TT-seq samples) and antisense bias ratio  $c_j$  (ratio of spurious reads originating from the opposite strand introduced by the reverse transcription reaction). Read counts  $k_{ij}$  for spike-ins were calculated using HTSeq<sup>405</sup>. Calculations for each parameter are described in the following in more detail.

**Antisense bias ratio  $c_j$ .** Antisense bias ratios were calculated for each sample  $j$  according to

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

for all available spike-ins  $i$ .

**Sequencing depth  $\sigma_j$  and cross-contamination rate  $\epsilon_j$ .** Sequencing depths were calculated for each sample  $j$  according to

$$\sigma_j = \operatorname{median}_i \left( \frac{k_{ij}}{l_i} \right)$$

using all available spike-ins  $i$  for the RNA-seq samples and the labeled spike-ins  $i$  for the TT-seq samples. The cross-contamination rate  $\epsilon_j$  was calculated for each sample  $j$  as

$$\epsilon_j = \operatorname{median}_i \left( \frac{k_{ij}}{l_i} \right) / \sigma_j$$

using the unlabeled spike-ins  $i$  for TT-seq samples. Note that  $\epsilon_j$  is set to 1 for the RNA-seq samples.

**Definition of transcription units (TUs).** For each annotated gene, transcription units (TUs) were defined as the union of all existing inherent transcript isoforms (UCSC RefSeq GRCh38). Read counts for all features were calculated using HTSeq<sup>405</sup> and corrected for antisense bias using antisense bias ratios  $c_j$  calculated as described above. The real number of read counts  $s_{ij}$  for transcribed unit  $i$  in sample  $j$  was calculated as

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

where  $S_{ij}$  and  $A_{ij}$  are the observed number of read counts on the sense and antisense strand. Read counts per kilobase (RPK) were calculated upon bias corrected read counts falling into the region of a transcribed unit divided by its length in kilobases. Based on the antisense bias corrected RPKs a subgroup of expressed TUs was defined to comprise all TUs with an RPK of 100 or higher in two summarized replicates of TT-seq without treatment. An RPK of 100 corresponds to approximately a coverage of 10 per sample due to an average fragment size of 200. This subset was used throughout the analysis unless stated otherwise.

**Calculation of the number of transcribed bases.** Aligned duplicated fragments were discarded for each sample. Of the resulting unique fragment isoforms only those were kept that exhibited a positive inner mate distance. The number of transcribed bases ( $tb_j$ ) for all samples was 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 (UCSC RefSeq GRCh38, ~ 96% of all fragments) in addition to the sum of the coverage of non-evident fragment parts (entire fragment).

### 2.1.2 mNET-seq data

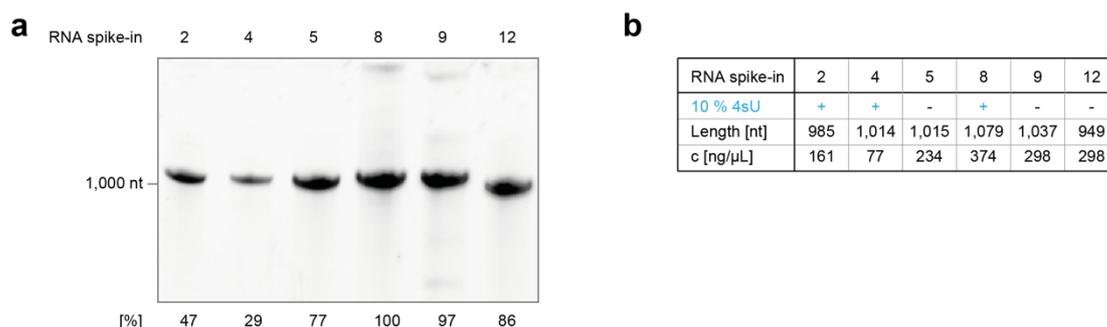
**mNET-seq data pre-processing.** Paired-end 50 or 75 base reads with additional 6 base reads of barcodes were obtained for each of the samples. Reads were demultiplexed and mapped with STAR<sup>403</sup> to the hg20/hg38 (GRCh38) genome assembly (Human Genome Reference Consortium). Samtools<sup>404</sup> 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<sup>406</sup> environment. Antisense bias (ratio of spurious reads originating from the opposite strand introduced by the RT reactions) was determined using positions in regions without antisense annotation with a coverage of at least 100 according to Refseq annotated genes (UCSC RefSeq GRCh38).

## 2.2 Additional tables and figures

**Table 29. Instruments and laboratory equipment ordered by application.**

Abbreviations for suppliers are listed in Materials II.1.

Instrument	Supplier	Application
Eppendorf ThermoMixer® C basic device, add thermoblock (1.5 mL, 2 mL)	[L]	Basic equipment
Biometra ProfessionalTRIO Thermocycler	Analytik Jena	Basic equipment
Refrigerated bench top centrifuge (rotor and adapters 0.5-2 mL tubes)	[L], 5424 R	Basic equipment
Pipetboy (Thermo Scientific™ S1 Pipettierhilfe)	[B], 9501	Basic equipment
Vortex mixer (Vortex-Genie 2)	Scientific Industries	Basic equipment
Wizard Mini-Fuge	Scientific Industries	Basic equipment
Reinstwassersystem Milli-Q® Advantage A10	VWR / Merck	Basic equipment
Cell culture water bath	LAUDA, Aqualine AL12	Cell culture
Bottle weights (for 500 mL)	[B], A1098801	Cell culture
VWR® Lead Ring Flask Weights	VWR, 13196-076	Cell culture
Cell culture fume hood, Type A2 Biological Safety Cabinets (incl. UV light)	Herasafe™ KS (NSF) Class II	Cell culture
Refrigerated bench top centrifuge (rotor and adapters for 50 mL and 15 mL tubes)	[L], 5702 RH	Cell culture
Fluid aspiration system, Biochem VacuuCenter, BVC professional	VWR, 181-0295	Cell culture
Incubator (37 °C, 5 % CO <sub>2</sub> )	HERAcell 150 i	Cell culture
Nalgene Cryo 1 °C Freezing Container (1-2 mL tubes)	[B], 5100-0001	Cell culture
Light microscope to check confluency and cell morphology	Leica	Cell culture
Hemocytometer, counting chamber (Neubauer)	VWR, 631-0696	Cell culture
Scepter™ 2.0 Cell Counter	Millipore Merck	Cell culture
Thermometer (water bath)		Cell culture
DynaMag™ - 15 Magnet (15 mL)	[B], 12301D	mNET-seq
DynaMag™ - 2 Magnet (1.5 mL, 2 mL)	[B], 12321D	mNET-seq
Typhoon™ FLA 9500	GE Healthcare	mNET-seq
Swing-out rotor (up to 20,000 x g), refrigerated bench top centrifuge (adapters for 15 mL tubes)	Sigma 2-16 KL	RNA isolation
NanoDrop 2000 (dsDNA, RNA quantification >100 ng/μL)	[B]	RNA/DNA quantification
Qubit Fluorometric Quantification	[B]	RNA/DNA
Fragment Analyzer Automated CE System	Advanced Analytical, [N]	RNA/DNA size
TapeStation	[N]	RNA/DNA size
qTOWER 2.0/2.2 instrument	Analytik Jena AG	RT-qPCR
Magnetic stand and separator	[M], 130-042-303	TT-seq
Magnetic Stand-96 for U-bottom microplates	[B], AM10027	TT-seq
Covaris S220 Series Focused- ultrasonicators (S220) and chiller	Covaris	TT-seq
Rotating wheel, rotator SB3	stuart	TT-seq, mNET-seq
Xcell II™ Blot Module (Western blot)	[B], EI9051	Western blot
Intas Science imaging, Advanced fluorescence Imager	Intas	Western blot



**Figure 37. RNA spike-in pool preparation for TT-seq.**

(a) Purified RNA spike-ins (left to right: 2, 4, 5, 8, 9, 12) were analyzed on a 5 % denaturing PAGE. Bands were visualized by SYBRgold staining and quantified using Fiji/Image J. The uncropped image can be found in **Figure 40**. (b) Table depicting information regarding 4sU content (in blue), RNA length [nt], and concentration [ng/ $\mu$ L] measured by NanoDrop spectrometer.

**Table 30. Normalized Ct values (RT-qPCR, Figure 7).**

Normalized Ct values,  $\Delta$ Ct, were calculated by multiplication of the primer efficiency (E) with the mean of the technical replicates. E was 0.99 for both primers. Two biological replicates were prepared for each condition.

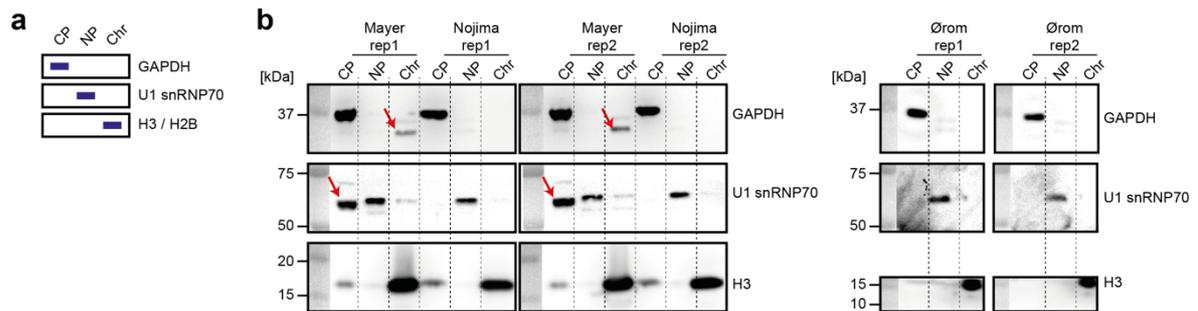
No.	RNA	Cell line	Replicate no.	Treatment	$\Delta$ Ct <sub>spike in 4</sub>	$\Delta$ Ct <sub>spike in 12</sub>	$\Delta$ $\Delta$ Ct	S.D.
1	total	Raji B CDK9 <sup>as</sup>	1	0.05 % v/v DMSO, 15 min.	14.77	14.26	-0.51	0.41
2	total	Raji B CDK9 <sup>as</sup>	2	0.05 % v/v DMSO, 15 min.	14.53	14.45	-0.08	0.32
3	total	Raji B CDK9 <sup>as</sup>	1	5 $\mu$ M 1-NA-PP1, 15 min.	14.12	14.17	0.06	0.17
4	total	Raji B CDK9 <sup>as</sup>	2	5 $\mu$ M 1-NA-PP1, 15 min.	14.02	13.92	-0.09	0.08
5	4sU-labeled	Raji B CDK9 <sup>as</sup>	1	0.05 % v/v DMSO, 15 min.	13.65	19.78	6.13	0.50
6	4sU-labeled	Raji B CDK9 <sup>as</sup>	2	0.05 % v/v DMSO, 15 min.	14.12	20.49	6.36	0.40
7	4sU-labeled	Raji B CDK9 <sup>as</sup>	1	5 $\mu$ M 1-NA-PP1, 15 min.	12.69	19.12	6.43	0.34
8	4sU-labeled	Raji B CDK9 <sup>as</sup>	2	5 $\mu$ M 1-NA-PP1, 15 min.	13.35	19.33	5.98	0.67
9	total	K562	1	-	14.29	14.75	0.46	0.01
10	total	K562	2	-	14.44	14.64	0.20	0.02
11	total	K562	1	42 °C, 15 min.	14.40	14.45	0.05	0.02
12	total	K562	2	42 °C, 15 min.	14.28	14.51	0.23	0.01
13	total	K562	1	42 °C, 30 min.	14.36	14.28	-0.09	0.07
14	total	K562	2	42 °C, 30 min.	14.51	14.31	-0.20	0.03
15	4sU-labeled	K562	1	-	12.87	18.14	5.27	0.09

<b>16</b>	4sU-labeled	K562	2	-	12.93	18.53	5.60	0.26
<b>17</b>	4sU-labeled	K562	1	42 °C, 15 min.	12.91	18.50	5.59	0.05
<b>18</b>	4sU-labeled	K562	2	42 °C, 15 min.	12.81	18.32	5.51	0.15
<b>19</b>	4sU-labeled	K562	1	42 °C, 30 min.	12.94	18.59	5.65	0.37
<b>20</b>	4sU-labeled	K562	2	42 °C, 30 min.	12.86	18.45	5.60	0.22
<b>21</b>	total	Raji B CDK9 <sup>as</sup>	1	0.05 % v/v DMSO, 35 min.	13.19	13.37	0.19	0.08
<b>22</b>	total	Raji B CDK9 <sup>as</sup>	2	0.05 % v/v DMSO, 35 min.	13.23	13.37	0.14	0.06
<b>23</b>	total	Raji B CDK9 <sup>as</sup>	1	5 µM 1-NA-PP1, 35 min.	17.26	16.63	-0.63	0.08
<b>24</b>	total	Raji B CDK9 <sup>as</sup>	2	5 µM 1-NA-PP1, 35 min.	17.43	16.83	-0.59	0.25
<b>25</b>	total	Raji B CDK9 <sup>as</sup>	1	0.05 % v/v DMSO, 35 min; after 5 min: 42 °C, 30 min.	15.31	15.86	0.55	0.47
<b>26</b>	total	Raji B CDK9 <sup>as</sup>	2	0.05 % v/v DMSO, 35 min; after 5 min: 42 °C, 30 min.	14.31	14.40	0.09	0.24
<b>27</b>	total	Raji B CDK9 <sup>as</sup>	1	0.05 % v/v DMSO, 35 min; after 5 min: 42 °C, 30 min.	16.72	16.13	-0.59	0.08
<b>28</b>	total	Raji B CDK9 <sup>as</sup>	2	0.05 % v/v DMSO, 35 min; after 5 min: 42 °C, 30 min.	16.34	16.26	-0.09	0.07
<b>29</b>	4sU-labeled	Raji B CDK9 <sup>as</sup>	1	0.05 % v/v DMSO, 35 min.	13.59	19.61	6.02	0.20
<b>30</b>	4sU-labeled	Raji B CDK9 <sup>as</sup>	2	0.05 % v/v DMSO, 35 min.	13.54	19.51	5.97	0.09
<b>31</b>	4sU-labeled	Raji B CDK9 <sup>as</sup>	1	5 µM 1-NA-PP1, 35 min.	15.86	22.77	6.91	0.06
<b>32</b>	4sU-labeled	Raji B CDK9 <sup>as</sup>	2	5 µM 1-NA-PP1, 35 min.	16.13	23.47	7.35	0.38
<b>33</b>	4sU-labeled	Raji B CDK9 <sup>as</sup>	1	0.05 % v/v DMSO, 35 min; after 5 min: 42 °C, 30 min.	14.14	19.75	5.61	0.07
<b>34</b>	4sU-labeled	Raji B CDK9 <sup>as</sup>	2	0.05 % v/v DMSO, 35 min; after 5 min: 42 °C, 30 min.	14.82	20.49	5.67	0.27
<b>35</b>	4sU-labeled	Raji B CDK9 <sup>as</sup>	1	0.05 % v/v DMSO, 35 min; after 5 min: 42 °C, 30 min.	16.69	24.66	7.97	0.12
<b>36</b>	4sU-labeled	Raji B CDK9 <sup>as</sup>	2	0.05 % v/v DMSO, 35 min; after 5 min: 42 °C, 30 min.	16.72	24.56	7.85	0.09

### 3. Results

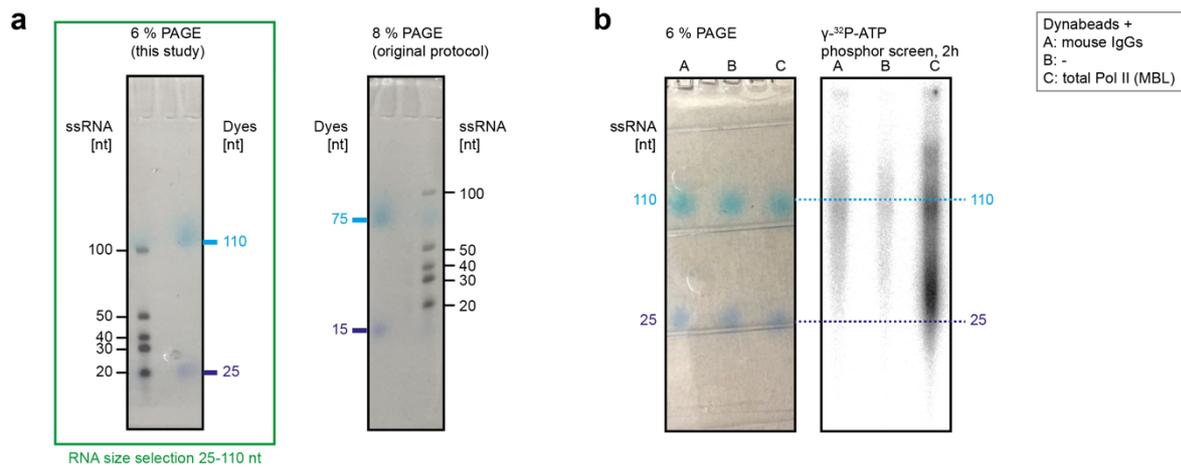
#### 3.1 mNET-seq optimization in suspension cells

This section contains supplementary information (figures) regarding the mNET-seq optimization.



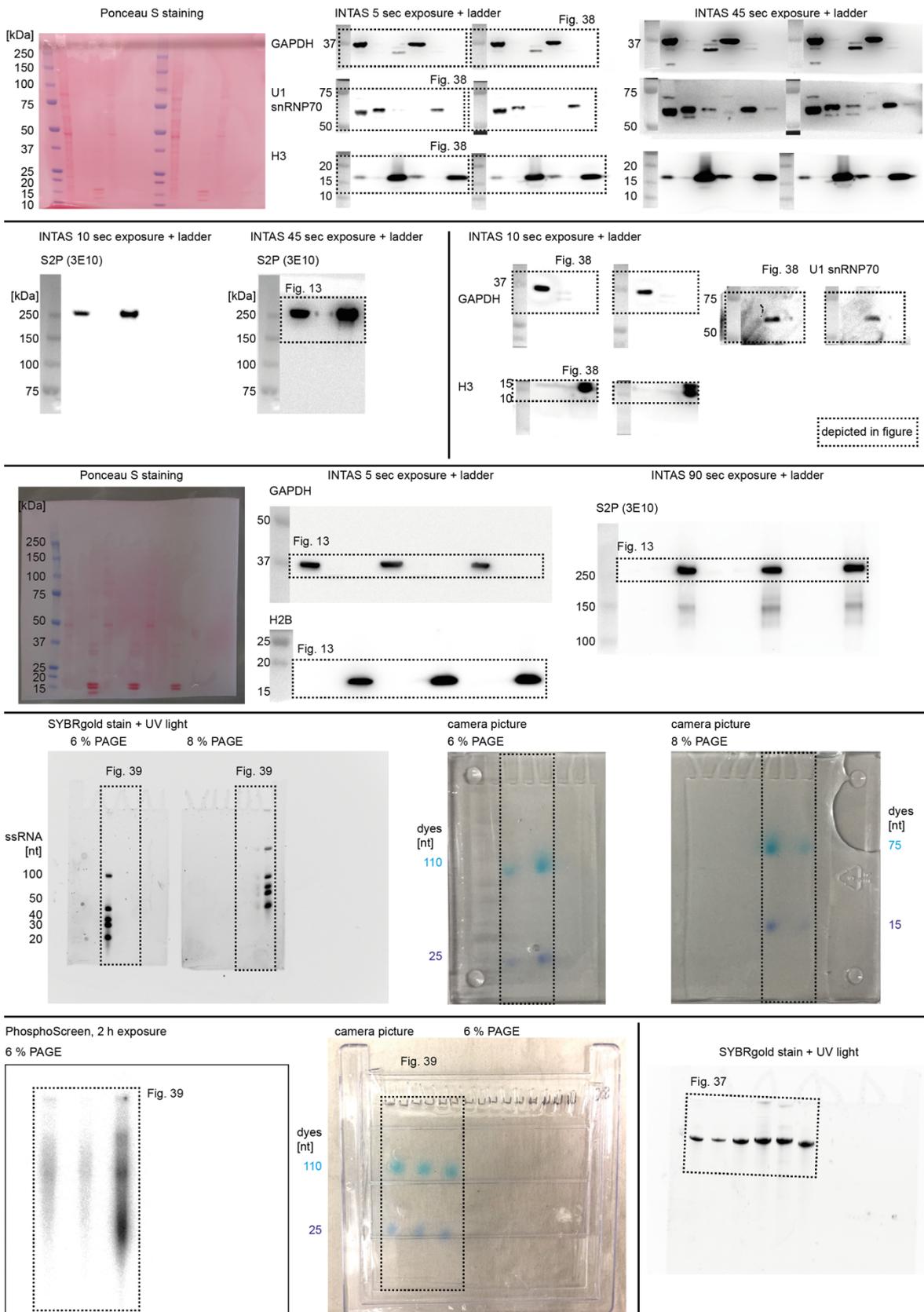
**Figure 38. Comparison of cell fractionation protocols by marker proteins.**

(a) Schematic depicting the outcome of a high-quality fractionation: GAPDH marks cytoplasm (CP), U1snRNP70 marks nucleoplasm (NP), histone 3 (H3) marks chromatin (Chr). (b) Western blot analysis of cytoplasm (CP), nucleoplasm (NP) and solubilized chromatin (Chr) fractions from K562 cells using indicated antibodies. Different fractionation protocols were compared in two biological replicates: Mayer<sup>431</sup>, Nojima<sup>388</sup>, and Ørom<sup>418</sup>. Chemiluminescence images were obtained on an INTAS ChemoCam Imager. Uncropped images of the PVDF membrane can be found in **Figure 40**.



**Figure 39. Optimizing nascent RNA size selection by PAGE (6 % versus 8 %).**

(a) The protocol from the Proudfoot lab used an 8 % PAGE for nascent RNA size selection in combination with radioactive monitoring by  $\gamma$ - $^{32}\text{P}$ -ATP labeling <sup>388</sup>. I tested different polyacrylamide concentrations (5-8 %) to define the concentration where the two blue tracking dyes frame the RNA target size (25-110 nt). Shown is an overlap of a camera picture of the PAGE (left: 6 %, right: 8 %) prior to SYBRgold staining (tracking dyes) and post SYBRgold staining (RNA marker), visualized on an INTAS gel imager. Small single stranded RNA (ssRNA) marker was used (Abnova). Uncropped images can be found in **Figure 40**. (b) Nascent RNA of 25-110 nt can be selected on a 6 % PAGE by cutting between the two blue dyes. K562 cell were fractionated by the Ørom protocol, digested by MNase. For 1 h of immunoprecipitation (IP), Dynabeads coupled to mouse IgG (A), Dynabeads only (B), or Dynabeads coupled to MBL antibody (total POLR2A) were used. See **Table 11** for antibodies. Left: camera picture of 6 % PAGE. Right: corresponding phosphor screen image of the 6 % PAGE after 2 h exposure. The screen was scanned on a Typhoon<sup>TM</sup> FLA 9500 unit. Uncropped images can be found in **Figure 40**. Experiments were performed at least two times, no replication attempt failed.

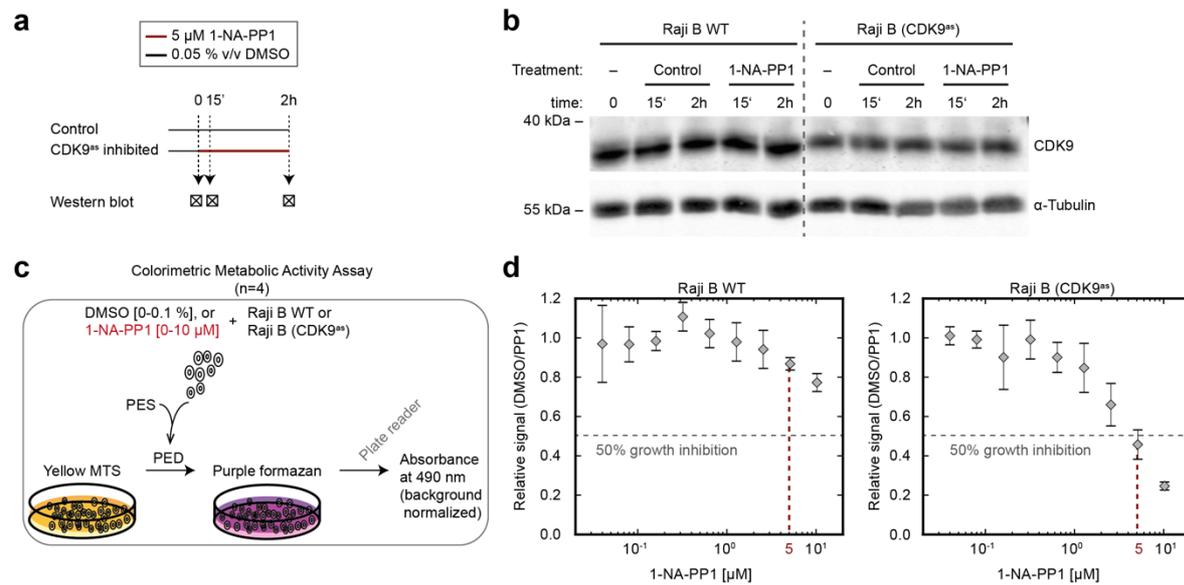


**Figure 40. Uncropped gel images and membrane scans for figures.**

Dotted boxes indicate region used for figures. Top rows: protein standard was used as size marker on NuPAGE™ 4-12 % Bis-Tris Protein Gels (see Methods). If available, Ponceau S staining and longer exposure of the PVDF membranes are also shown (INTAS ChemoCam imager). Bottom rows: UV image of 6 or 8 % PAGE after SYBRgold staining (INTAS gel imager); camera picture of 6 or 8 % PAGE prior to SYBRgold staining. Phosphor screen after 2 h exposure (Typhoon™ FLA 9500); camera picture of 6 % PAGE prior to phosphorimaging.

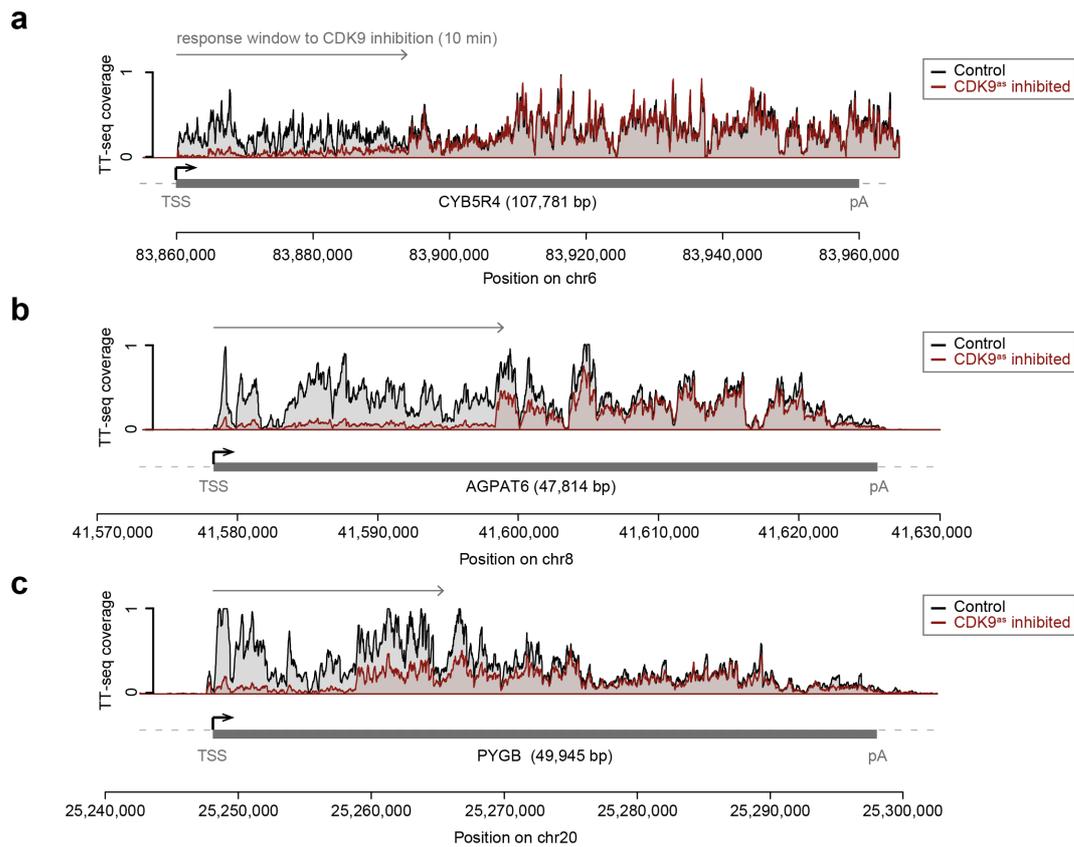
### 3.2 CDK9-dependent Pol II pausing controls transcription initiation

This section contains supplementary information (figures and tables) regarding the study ‘CDK9-dependent Pol II pausing controls transcription initiation’.



**Figure 41. CDK9 (WT or analog-sensitive) protein levels and cell viability.**

(a) Experimental schematic for Western blot. (b) Raji B wildtype (WT) and CDK9<sup>as</sup> cells were treated with 10  $\mu\text{M}$  of 1-NA-PP1 for 0 min, 15 min or 2 h. DMSO was used as solvent control. Stable CDK9 protein levels were detected by Western blotting.  $\alpha$ -Tubulin was used as loading control. (c) Experimental schematic for metabolic activity assay (MTS assay). (d) MTS assay to compare cell viability upon 1-NA-PP1 treatment (0-10  $\mu\text{M}$ ) for 72 h in wildtype and CDK9<sup>as</sup> cells. (b, d) Western blot and MTS assay were performed by Dr. Tim M. Decker (Helmholtz Center Munich, Dept. of Molecular Epigenetics, group of Prof. Dr. Dirk Eick) (methods are described in <sup>381</sup>).



**Figure 42. TT-seq tracks transcription after 10 min of CDK9<sup>as</sup> inhibition.**

Genome browser views of TT-seq signals of control (grey) and CDK9<sup>as</sup> inhibited samples at the CYB5R4 gene locus (107,781 [bp]) on chromosome 6 (a), at the AGPAT6 gene locus (47,814 [bp]) on chromosome 8 (b), and at the PYGB gene locus (49,945 [bp]) on chromosome 20 (c). Grey box depicts transcript body from TSS (black arrow) to polyA site (pA).

**Table 31. Published data sets used in this study (eLife 2017).**

Data sets and references used in the correlation study against genomic features and measures of transcriptional context. Abbreviations: chromatin immunoprecipitation (ChIP), dimethyl sulfate (DMS), deoxyribonuclease I (DNase I), formaldehyde-assisted isolation of regulatory elements-sequencing (FAIRE), chromosome conformation capture (Hi-C), minimum free energy (MFE), micrococcal nuclease (MNase).

Experiment	Cell type	GEO ID	Source	Read-out	Used in
AT content	<i>in silico</i>	This study			Figure 16e
Bisulfite-seq	K562	GSE27584	Myers <sup>144</sup>	Sites of DNA hypermethylated CpG islands	Figure 16e
CG content	<i>in silico</i>	This study			Figure 16e
ChIP-seq Brd4	HEK293T	GSE51633	Rosenfeld <sup>511</sup>	Protein-DNA interactions	Figure 21e
ChIP-seq Brd4	HeLa	GSE51633	Rosenfeld <sup>511</sup>	Protein-DNA interactions	Figure 21e
ChIP-seq CDK9	HEK293T	GSE51633	Rosenfeld <sup>511</sup>	Protein-DNA interactions	Figure 21e
ChIP-seq CDK9	HCT116	GSE70408	Shilatifard <sup>440</sup>	Protein-DNA interactions	Figure 21e
ChIP-seq NELF-E	K562	GSE31477	Struhl <sup>144</sup>	Protein-DNA interactions	Figure 16e; Figure 21e
ChIP-seq total Pol II (N-20)	HCT116	GSE70408	Shilatifard <sup>440</sup>	Protein-DNA interactions	Figure 16e
ChIP-seq TOP1	HCT116	GSE57628	Levens <sup>512</sup>	Protein-DNA interactions	Figure 16e
CpG islands		This study			Figure 16e
DMS-seq (sample '300 $\mu$ L <i>in vivo</i> ')	K562	GSE45803	Weißmann <sup>425</sup>	RNA secondary structure	Figure 21b
DNA-RNA basepairing energy	<i>in silico</i>	This study	Wilhelm <sup>513</sup>		Figure 16e
DNA-RNA melting temperature	<i>in silico</i>	This study	Wilhelm <sup>513</sup>		Figure 16e
DNase-seq	K562	GSE32970	Stamatoyannopoulos <sup>144</sup>	Active regulatory chromatin (nucleosome-depleted)	Figure 16e
DNaseI-seq	K562	GSE29692	Sandstrom <sup>144</sup>	Active regulatory chromatin (nucleosome-depleted)	Figure 16e
FAIRE-seq	K562	GSE35239	Lieb <sup>144</sup>	Active regulatory chromatin (nucleosome-depleted)	Figure 16e
Hi-C	K562	GSE56869	Crawford <sup>437</sup>	3D genome structure	Figure 21f
MFE	<i>in silico</i>		Stadler <sup>426</sup>	RNA secondary structure	Figure 21c
MNase-seq	K562	GSE35586	Snyder <sup>144</sup>	Active regulatory chromatin (nucleosome-depleted)	Figure 16e
MNase-seq	Raji B	GSE52914	Andrau <sup>514</sup>	Active regulatory chromatin (nucleosome-depleted)	Figure 16e
TOP1-seq	HCT116	GSE57628	Levens <sup>512</sup>		Figure 16e

---

### 3.3 The pause-initiation limit restricts transcription activation

This section contains supplementary information (figures, tables, and notes) regarding the study ‘The pause-initiation limit restricts transcription activation in human cells’.

---

The results presented in section 3.3 are currently under peer review for publication:

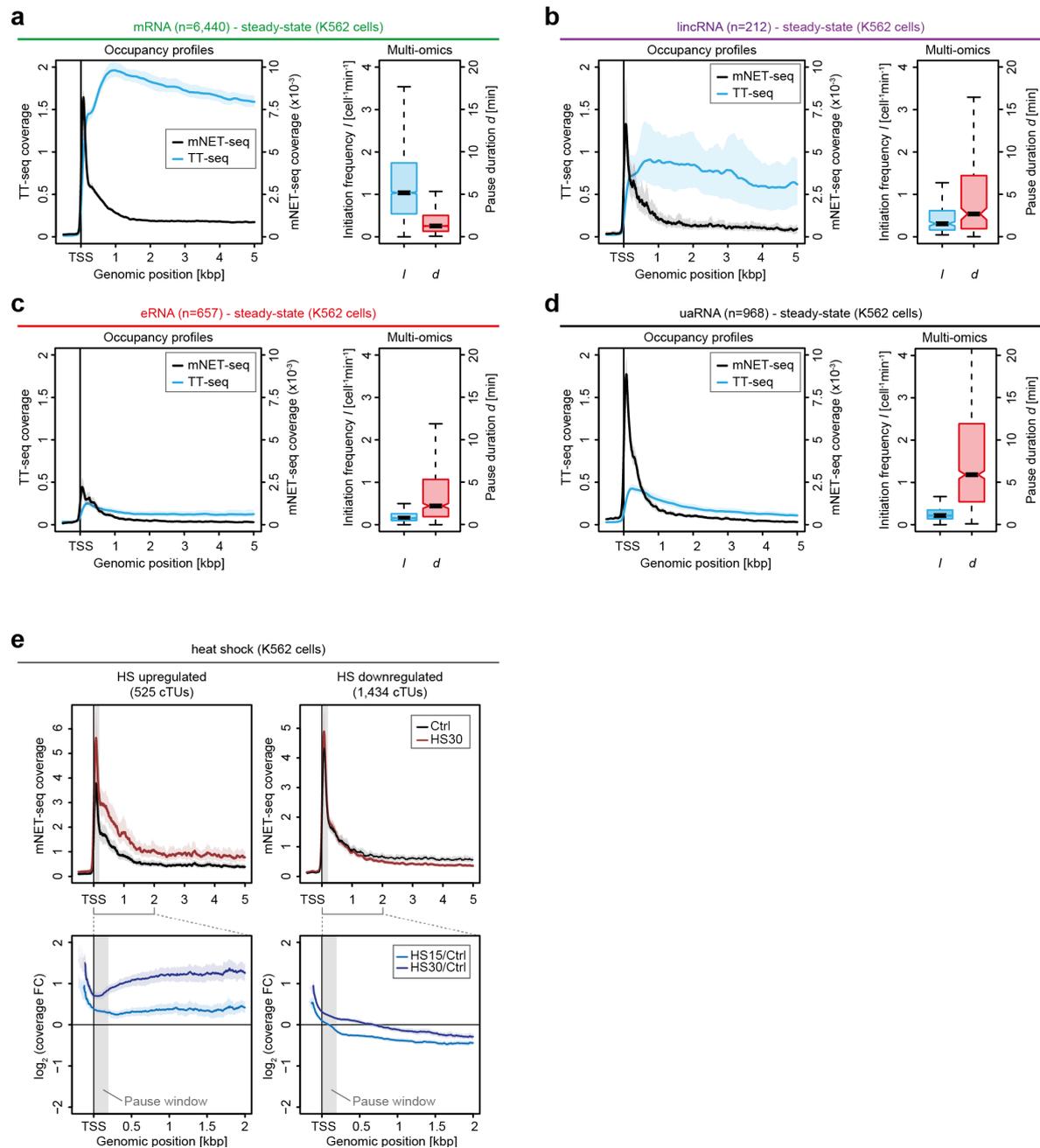
**S. Gressel** \*, B. Schwalb \* *et al.* The pause-initiation limit restricts transcription activation in human cells. *Nature Communications*, in revision.

(\* joint first authorship)

---

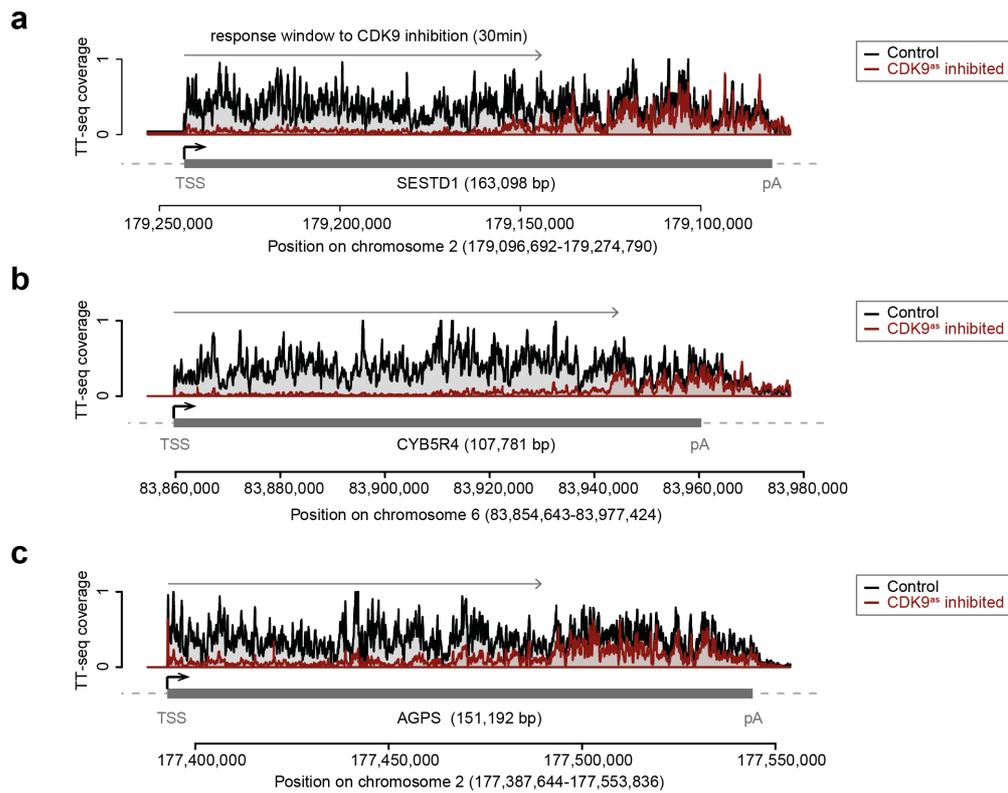
#### Supplementary Note 2. Occupancy profiling alone does not reveal kinetics.

To test whether Pol II pausing behavior can be inferred from Pol II occupancy changes alone, we investigated our mNET-seq data after 15 and 30 min heat shock response (**Figure 43**). After 15 min of heat shock, the average mNET-seq signal for upregulated genes increased in the pause window and in the gene body, and this does not allow for conclusions. After 30 min of heat shock, the signal had increased further in the gene body but not in the pause window, an observation that could have been correctly interpreted as a decrease in  $d$  and an increase in  $I$  (**Figure 43**, left). For downregulated genes after 15 min of heat shock, the signal in the pause window remained unchanged, whereas the signal in the gene body decreased, again not allowing for definitive conclusions with respect to the duration of pausing, because the same profile change is expected by a decrease in productive initiation frequency. After 30 min of heat shock, the signal in the pause window increased, whereas the signal in the gene body decreased further, and this could be interpreted as an increase in  $d$  and a decrease in  $I$  (**Figure 43**, right). Assuming unchanged elongation velocity in the gene body upon heat shock<sup>231</sup>, these Pol II occupancy changes are generally consistent with reduced and increased pause durations for up- and down-regulated genes, respectively, although the interpretation of mNET-seq data at single time points alone would not always have led to the correct conclusions. Taken together, the observed changes in Pol II pausing behavior may in favorable cases be inferred from an analysis of occupancy changes alone, but this is not always conclusive and cannot provide kinetics.



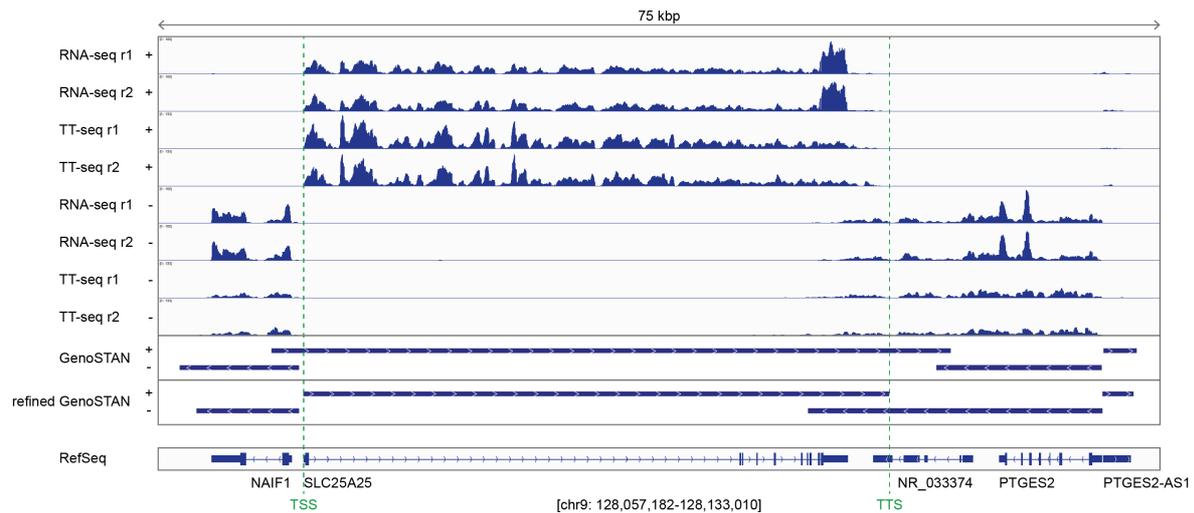
**Figure 43. Occupancy profiling by mNET-seq alone does not reveal kinetics.**

Multi-omics, but not occupancy profiling alone, can reveal transcription kinetics. (a-d) For comparison among different transcript classes, TT-seq coverage (left y-axis) and mNET-seq coverage (right y-axis) show the same range for all metagene profiles. Coverage per cell is shown for two biological replicates of steady-state K562 cells. (a-b) mRNAs (6,440, green) and lincRNAs (212, purple) show similar total Pol II peak heights in mNET-seq (with Empigen BB). However, the TT-seq signal is significantly lower for lincRNAs indicating less productive initiation events compared to mRNAs. (c-d) eRNAs (657, red) and uaRNAs (968, black) show a similar height of the TT-seq signal. However, Pol II occupancy strongly varies these transcript classes. As a consequence, our multi-omics approach shows that pause durations are longer for uaRNAs than for eRNAs. (e) Top: mNET-seq coverage aligned at the TSS for 525 significantly upregulated (left), and 1,434 downregulated genes (right) with a minimum length of 5 kbp upon 30 min of heat shock (HS). Bottom: close-up showing coverage fold change upon 15 min (light blue) and 30 min (dark blue) of heat shock for genes as in top panel. Shaded areas show confidence intervals. The pause window is highlighted as grey shaded area



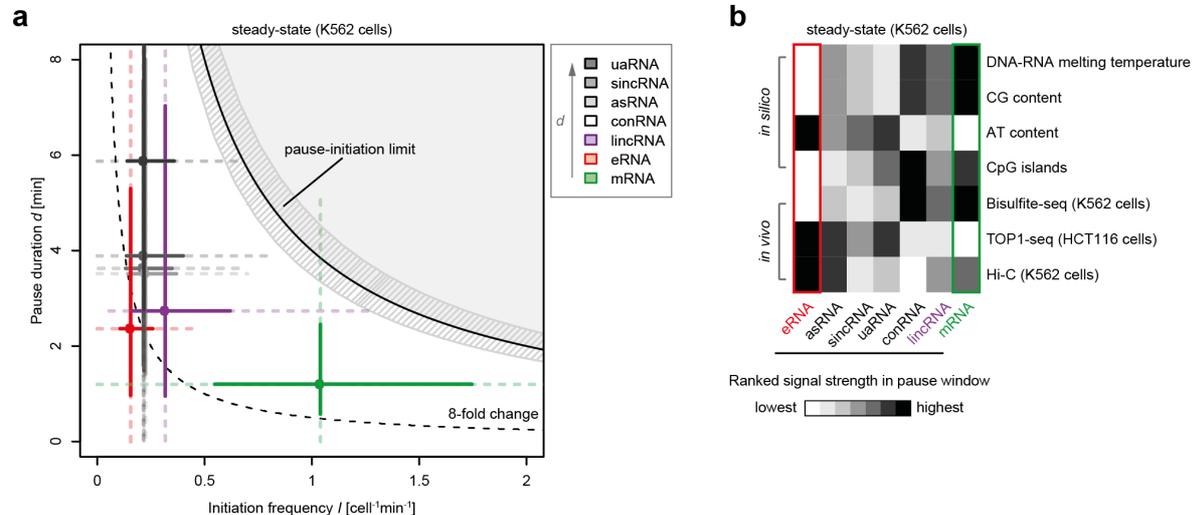
**Figure 44. T'T-seq tracks transcription after 30 min of CDK9<sup>as</sup> inhibition.**

Based on the assumption that the inhibitor is evenly distributed across cells and within cells, the portion of CDK9<sup>as</sup> molecules that has been fully inhibited by 35 min treatment with 1-NA-PP1 is very high. Note that inhibitory effects on pause release by unspecific inhibition of other kinases upon heat shock can be excluded with the use of the chemical biology (Shokat) approach, thus, inhibition of CDK9 is highly specific. Genome browser views of T'T-seq signals of control (grey) and CDK9<sup>as</sup> inhibited samples at the SESTD1 gene locus (163,098 [bp]) on chromosome 2 (a), at the CYB5R4 gene locus (107,781 [bp]) on chromosome 6 (b), and at the AGPS gene locus (151,192 [bp]) on chromosome 2 (c). Grey box depicts transcript body from TSS (black arrow) to polyA site (pA).



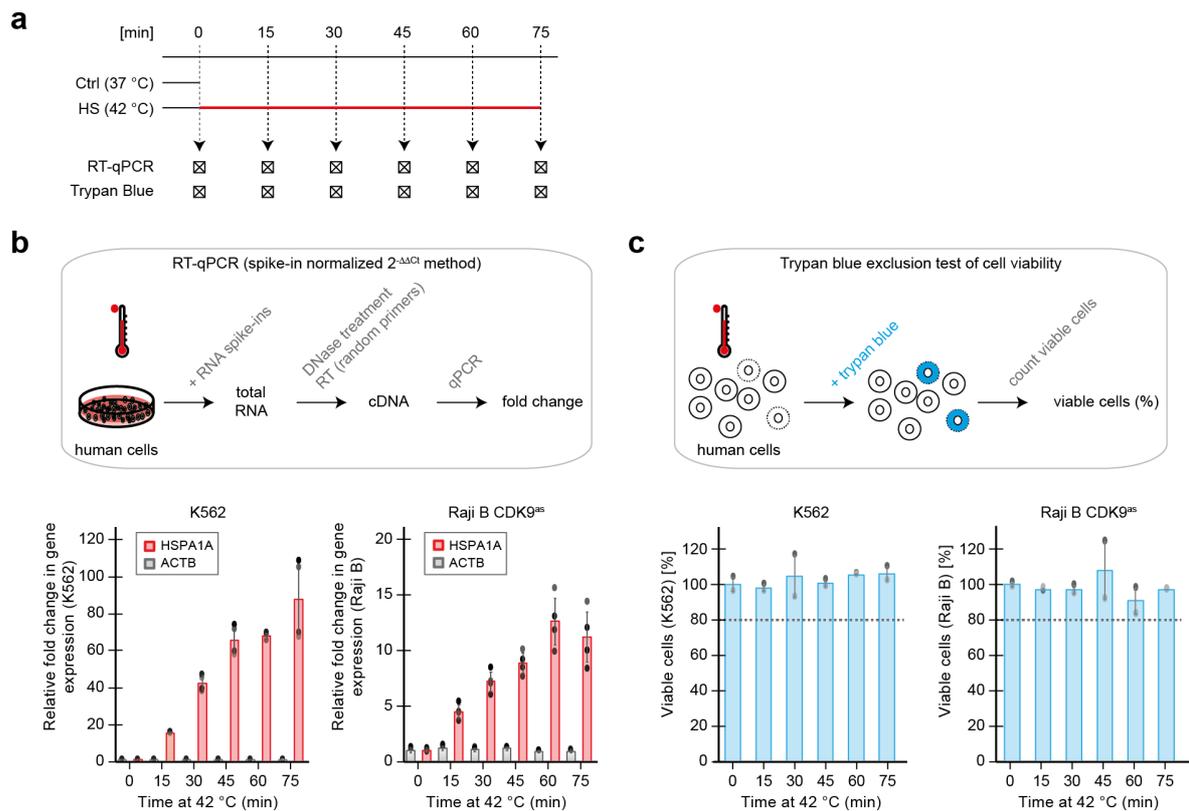
**Figure 45. TSS and TTS are better refined in new annotation of human K562 cells.**

Comparison of published GenoSTAN annotation<sup>168</sup>, and new refined annotation (this study) which includes new TT-seq data (this study) and published GRO-cap<sup>50</sup> data in K562 cells for the annotation of TSSs. Note that TT-seq libraries in this study were generated using only random priming while the previous TT-seq library in K562 was generated using a mixture of random and poly(dT) primers<sup>168</sup>. Abbreviations: transcription start site (TSS), transcription termination site (TTS).



**Figure 46. Transcription kinetics of different transcript classes in steady state.**

(a) Plot shows the median productive initiation frequency  $I$  [ $\text{cell}^{-1}\text{min}^{-1}$ ] depicted against the median pause duration  $d$  [min] for all transcript classes (circles). The two solid perpendicular lines define the inter-quartile range, the dotted whiskers represent 1.5 times the inter-quartile range of the respective estimate for the entire transcript class. The grey shaded area depicts impossible combinations of  $I$  and  $d$ . Striped area shows confidence intervals of the pause-initiation limit. The dotted line defines an 8-fold possible fold change until a gene would be restricted by the pause-initiation limit. (b) Chromatin features in the pause window (Methods) of different transcript classes. Data are ranked by each row across different transcript types to better highlight the contrast of individual features. Published data sets are listed in **Table 34**.



**Figure 47. Optimal time points for heat shock time-course of human hematopoietic cells.**

(a) Human K562 or Raji B (CDK9<sup>as</sup>) cells were subjected to a heat shock (42 °C) time-course of 0 to 75 min. Every 15 min, total RNA was isolated and analyzed by RT-qPCR. Cell viability was tested by trypan blue exclusion test. (b) Top: experimental set-up of RT-qPCR of heat shock time-course. Bottom: relative fold change in gene expression of b-actin (ACTB, in grey), or hsp70 (HSPA1A, in red) in K562 (left bar plot) or Raji B (CDK9<sup>as</sup>) (right bar plot) cells (spike-ins normalized). Error bars represent the standard deviation. Differential expression observed for HSPA1A in K562 and Raji B (CDK9<sup>as</sup>) cells agrees very well with other studies<sup>195</sup>. Corresponding normalized Ct values ( $\Delta Ct$ ) and relative fold changes per target gene ( $2^{\Delta\Delta Ct}$ ) can be found in **Table 32**. (c) Top: experimental set-up of trypan blue exclusion assay. Bottom: viable cells [%] for K562 (left bar plot) or Raji B (CDK9<sup>as</sup>) (right bar plot) upon heat shock time-course. Corresponding cell counts can be found in **Table 33**.

**Table 32. Normalized Ct values (RT-qPCR, Figure 47 b).**

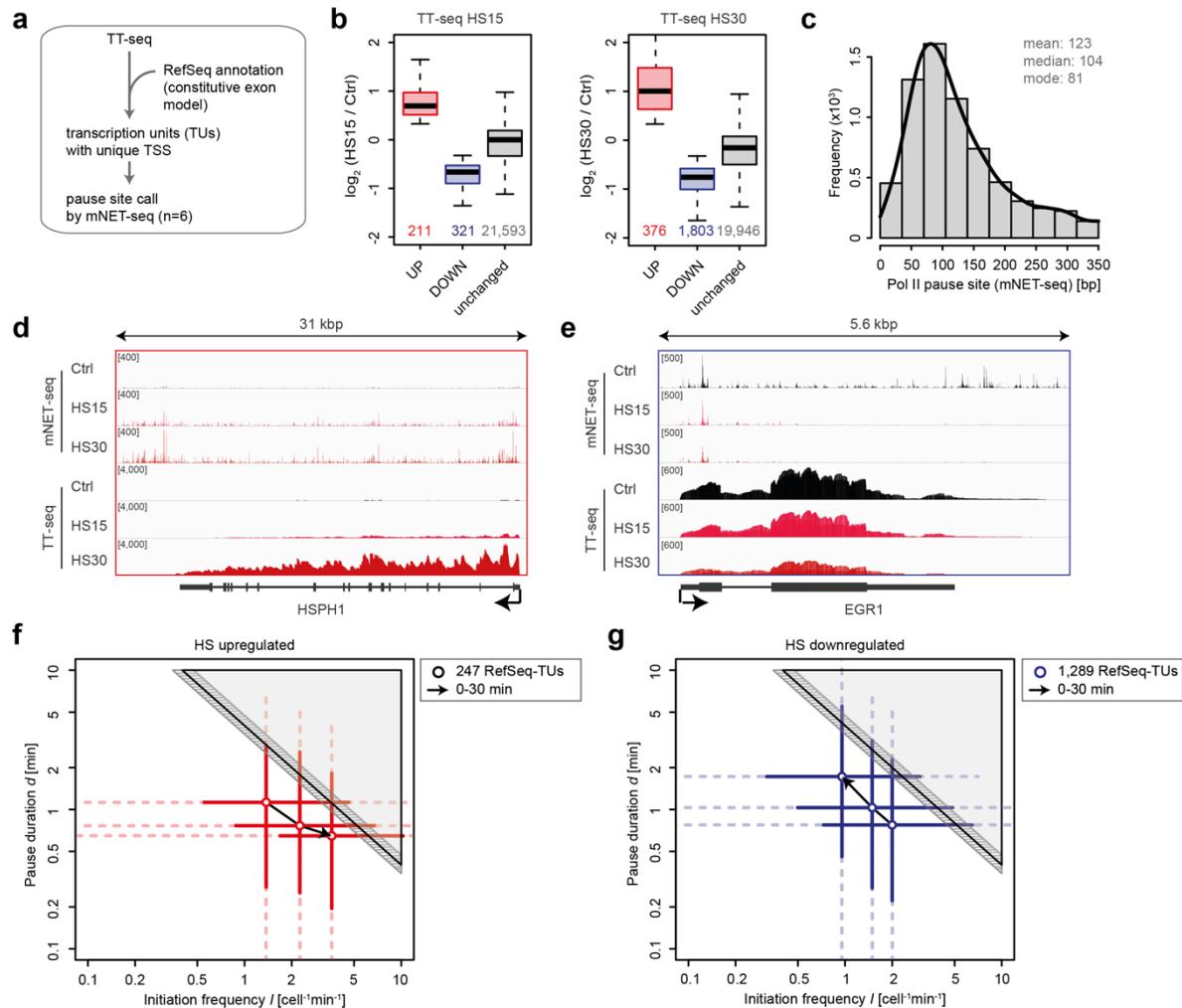
Normalized Ct values,  $\Delta Ct$ , were calculated by multiplication of the primer efficiency (E) with the mean of the technical replicates. E was 0.99 for spike-in primers (4, 5, 12), 1.04 for ACTB primers, and 1.16 for HSPA1A primers. Two biological replicates were prepared for each condition.

No.	Cell line	Time at 42°C (min)	$\Delta Ct$					$2^{\Delta\Delta Ct}$		S.D.	
			Spike-in 4	Spike-in 5	Spike-in 12	ACTB	HSPA1A	ACTB	HSPA1A	ACTB	HSPA1A
1	K562	0	16.08	16.08	15.26	15.87	24.94	1.00	1.00	0.01	0.02
2	K562	0	15.78	15.79	15.38	15.74	24.84				
3	K562	15	15.37	15.70	15.09	15.50	20.57	0.94	15.92	0.04	0.30
4	K562	15	15.48	15.47	15.13	15.59	20.52				
5	K562	30	15.40	15.74	14.93	15.84	19.21	0.87	43.66	0.11	4.05
6	K562	30	15.38	15.63	14.91	15.41	18.90				
7	K562	45	15.52	15.58	14.88	15.60	18.26	0.88	67.86	0.01	7.40
8	K562	45	15.64	15.73	14.72	15.61	18.61				
9	K562	60	15.50	15.61	15.07	15.83	18.44	0.86	70.02	0.08	1.49
10	K562	60	15.39	15.75	15.18	15.60	18.46				
11	K562	75	15.67	15.66	16.01	15.97	18.15	0.81	90.70	0.11	19.66
12	K562	75	15.49	15.58	15.23	16.04	18.44				
13	Raji B CDK9 <sup>as</sup>	0	14.25	14.35	13.40	13.45	24.62	1.02	1.01	0.20	0.13
14	Raji B CDK9 <sup>as</sup>	0	14.36	14.22	13.44	13.06	24.36				
15	Raji B CDK9 <sup>as</sup>	15	14.20	14.24	13.69	13.17	22.53	1.19	4.49	0.18	0.62
16	Raji B CDK9 <sup>as</sup>	15	14.27	14.28	14.08	13.11	22.39				
17	Raji B CDK9 <sup>as</sup>	30	14.34	14.33	14.03	13.36	21.99	1.06	7.22	0.15	0.89
18	Raji B CDK9 <sup>as</sup>	30	14.30	14.39	13.96	13.46	21.74				
19	Raji B CDK9 <sup>as</sup>	45	14.19	14.13	13.71	13.11	21.27	1.15	8.91	0.16	0.92
20	Raji B CDK9 <sup>as</sup>	45	14.32	14.21	13.34	12.99	21.37				
21	Raji B CDK9 <sup>as</sup>	60	14.40	14.41	13.99	13.68	20.90	0.87	12.70	0.12	2.09
22	Raji B CDK9 <sup>as</sup>	60	14.24	14.25	13.60	13.56	21.07				
23	Raji B CDK9 <sup>as</sup>	75	14.28	14.27	14.06	13.54	20.96	0.93	11.25	0.13	2.27
24	Raji B CDK9 <sup>as</sup>	75	14.34	14.38	13.74	13.56	21.44				

**Table 33. Cell counts (trypan blue exclusion assay, Figure 47 c).**

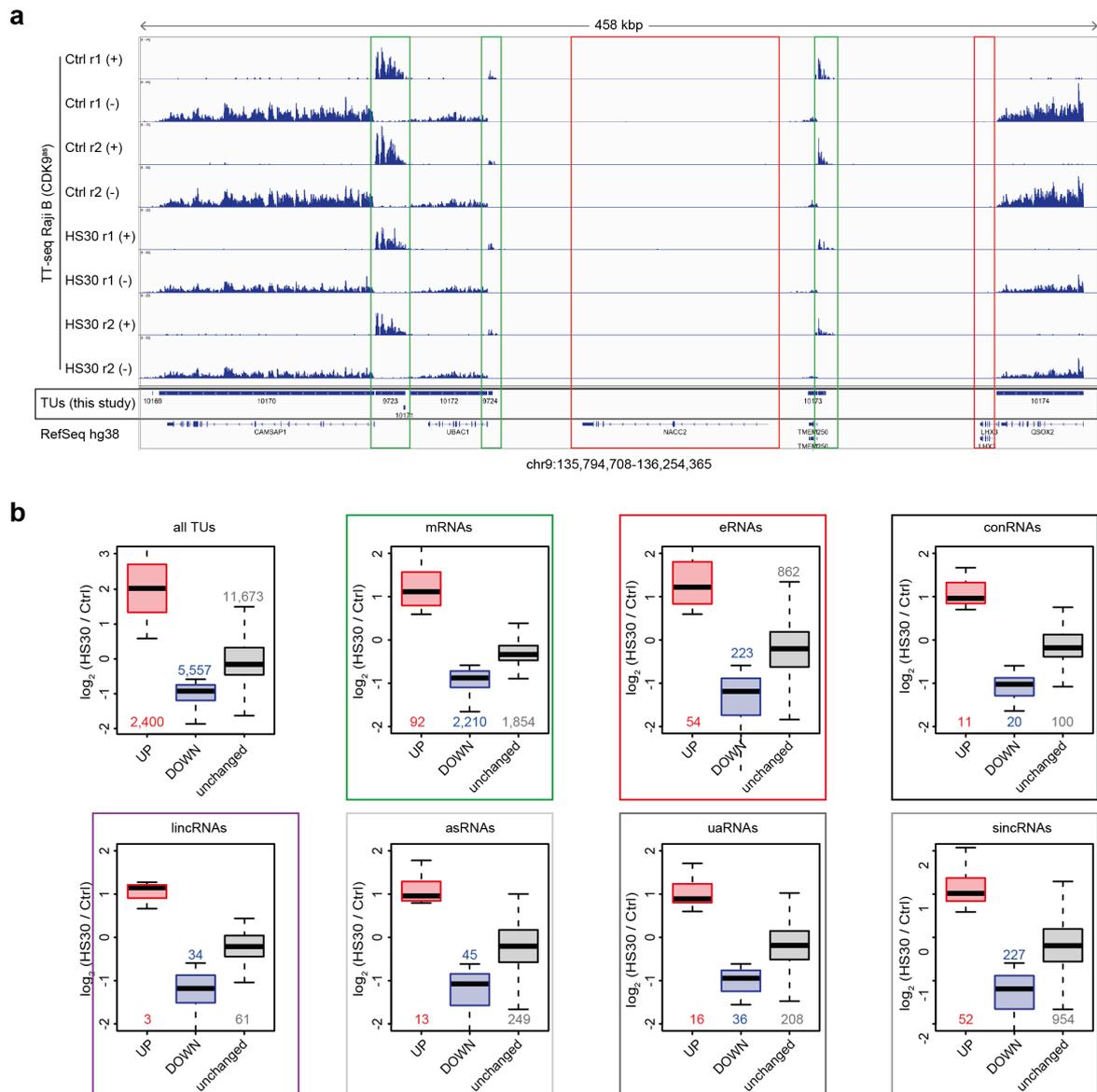
For experimental details see **Methods II.2.4.1**. Expect for 60 min of heat shock (HS) in K562, two biological replicates of all time-points were collected.

No.	Cell line	Replicate no.	Time at 42 °C (min)	Cell count		Mean cell number/mL	Viable cells (%)	S.D.
				1	2			
1	K562	1	0	77	87			
2	K562	2	0	76	76	158	100	3
3	K562	1	15	83	75			
4	K562	2	15	74	78	155	102	2
5	K562	1	30	108	76			
6	K562	2	30	73	73	165	96	9
7	K562	1	45	63	99			
8	K562	2	45	86	70	159	99	9
9	K562	1	60	NA	NA			
10	K562	2	60	75	91	166	95	5
11	K562	1	75	80	93			
12	K562	2	75	77	84	167	95	4
13	Raji B CDK9 <sup>as</sup>	1	0	87	81			
14	Raji B CDK9 <sup>as</sup>	2	0	76	88	166	100	3
15	Raji B CDK9 <sup>as</sup>	1	15	75	85			
16	Raji B CDK9 <sup>as</sup>	2	15	97	66	162	103	7
17	Raji B CDK9 <sup>as</sup>	1	30	89	76			
18	Raji B CDK9 <sup>as</sup>	2	30	74	83	165	101	4
19	Raji B CDK9 <sup>as</sup>	1	45	91	115			
20	Raji B CDK9 <sup>as</sup>	2	45	74	78	179	93	9
21	Raji B CDK9 <sup>as</sup>	1	60	78	85			
22	Raji B CDK9 <sup>as</sup>	2	60	78	61	151	110	6
23	Raji B CDK9 <sup>as</sup>	1	75	82	79			
24	Raji B CDK9 <sup>as</sup>	2	75	78	83	161	103	1



**Figure 48. Constitutive exons analysis agrees with analysis on new annotation.**

(a) Diagram illustrating the main selection steps of constitutive exons in all RefSeq GRCh38 annotated isoforms. (b) Differential expression (DE) analysis of upregulated (light red), downregulated (dark blue), and unchanged TUs (grey) in human K562 cells upon 15 min (left box plot) or 30 min (right box plot). Heat shock samples (HS15, or HS30) were compared to respective control (Ctrl). Spike-in normalized TT-seq data were analyzed for DE of constitutive exons in all RefSeq GRCh38 annotated isoforms. A minimum fold change of 1.25 as cutoff and a maximum adjusted p-value of 0.001 was set for calling a significant expression change. Black bars represent medians, boxes mark upper and lower quartiles, whiskers represent 1.5 times the inter-quartile range. (c) Distribution of pause site distance from the TSS for 7,406 investigated genes measured by mNET-seq depicted as a histogram with respect to TSS of RefSeq annotated genes (mean 123 [bp], median 104 [bp], mode 81 [bp]). Note that this in contrast to GRO-cap refined TSSs of our new K562 annotation for which called pause sites were distributed around a maximum located ~50 bp downstream of the TSS (**Figure 22 c**). (d-e) Representative genome browser view of mNET-seq and TT-seq data at the HSPH1 gene locus on chromosome 13 (minus strand) (d), or at the EGR1 gene locus on chromosome 5 (plus strand) (e). Visualized with the Integrative Genomics Viewer (IGV, version 2.4.10; human hg38)<sup>400</sup>. Shown is the strand-specific Pol II occupancy with single nucleotide resolution (mNET-seq, top panels), and the number of transcribed bases (TT-seq, bottom panels). Two biological replicates are merged. (f-g) Median of pause-initiation trajectories upon heat shock time-course of 247 significantly upregulated (red circles, f) or 1,289 downregulated genes (dark blue circles, g) in log scale. The two solid perpendicular lines define the inter-quartile range, the dotted whiskers represent 1.5 times the inter-quartile range of the respective estimate for the entire transcript class. The grey shaded area depicts impossible combinations of  $I$  and  $d$ . Striped areas show confidence intervals of the pause-initiation limit.



**Figure 49. GenoSTAN annotation and DE analysis of TUs in Raji B cells.**

(a) Genome browser track of 458 kbp of chromosome 9 (chr9:135,794,708-136,254,365) visualized with the Integrative Genomics Viewer (IGV, version 2.4.10)<sup>400</sup>. From top to bottom tracks represent: TT-seq coverage in Raji B (CDK9<sup>as</sup>) cells (n=4), new TU annotation (Methods), RefSeq GRCh38 annotation. TUs missing in the RefSeq annotation are highlighted in a green. RefSeq transcripts not present in Raji B cells are highlighted in red.

(b) TUs and transcript classes as annotated in this study: all TUs, mRNAs (green frame), eRNAs (light red frame), conRNAs (black frame), lincRNAs (purple frame), asRNAs, uaRNAs, sincRNAs (different shades of grey). Boxplots show differential expression (DE) analysis of upregulated (light red), downregulated (dark blue), and unchanged TUs (grey) in human Raji B cells upon 30 min of heat shock. Heat shock samples (HS30) were compared to respective controls (Ctrl) of spike-in normalized TT-seq data. A minimum fold change of 1.5 in all conditions and a maximum adjusted P-value of 0.1 was set for calling a significant expression change.

**Table 34. Published data sets used in this study.**

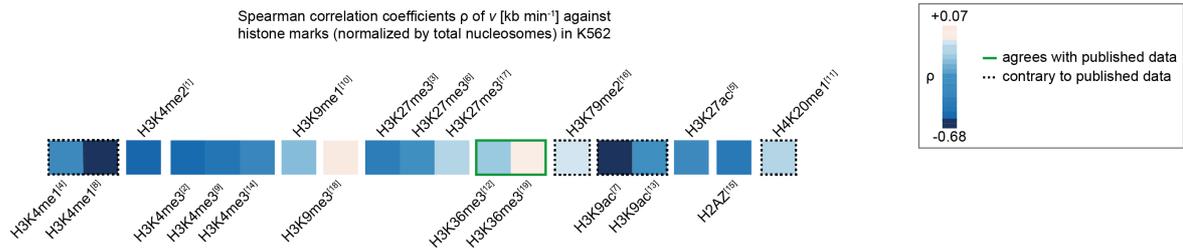
For chromatin state annotation of K562 cells (no. 1), Benedikt Zacher *et al.* used 18 chromatin states leading to a definition of promoter and enhancer states. Specifically, the segmentation algorithm GenoSTAN considers K562-specific DNase-seq and CHIP-seq data of 8 histone modifications (H3K27me3, H3K36me3, H4K20me1, H3K4me1-3, H3K27ac, H3K9ac) and of the histone acetyltransferase P300 <sup>168,177</sup>.

No.	Data	Cell line	Used for	Reference	Available at
1	Chromatin state annotation K562	K562	New annotation (cTU) of K562	Gagneur <sup>177</sup>	<a href="https://i12g-gagneurweb.in.tum.de/public/paper/GenoSTAN/">https://i12g-gagneurweb.in.tum.de/public/paper/GenoSTAN/</a>
2	GRO-cap	K562	New annotation (cTU) of K562	Lis <sup>50</sup>	NCBI Gene Expression Omnibus (GSE60456)
3	H3K4me1	Raji B	New annotation (TU) of Raji B	Andrau <sup>514</sup>	NCBI Gene Expression Omnibus (GSE52914)
4	H3K4me3	Raji B	New annotation (TU) of Raji B	Andrau <sup>514</sup>	NCBI Gene Expression Omnibus (GSE52914)
5	mNET-seq	Raji B (CDK9 <sup>as</sup> )	Pause duration <i>d</i> of Raji B	Cramer <sup>381</sup>	NCBI Gene Expression Omnibus (GSE96056)
6	TT-seq	Raji B (CDK9 <sup>as</sup> )	Pause duration <i>d</i> of Raji B	Cramer <sup>381</sup>	NCBI Gene Expression Omnibus (GSE96056)
7	In situ Hi-C	K562	Chromatin features of the PS (Figure 46)	Rao et al. 2014	NCBI Gene Expression Omnibus (GSM1551620)
8	MNase-seq	K562	Chromatin features of the PS (Figure 46)	Kundaje et al. 2012	NCBI Gene Expression Omnibus (GSE35586)
9	AT content	<i>in silico</i>	Chromatin features of the PS (Figure 46)	this study	NA
10	TOP1-seq	HCT116	Chromatin features of the PS (Figure 46)	Baranello et al. 2016	NCBI Gene Expression Omnibus (GSE57628)
11	Bisulfite-seq (RRBS)	K562	Chromatin features of the PS (Figure 46)	Myers et al. 2011	NCBI Gene Expression Omnibus (GSE27584)
12	CG content	<i>in silico</i>	Chromatin features of the PS (Figure 46)	this study	NA
13	CpG island	<i>in silico</i>	Chromatin features of the PS (Figure 46)	this study	NA
14	DNA-RNA melting temperature	<i>in silico</i>	Chromatin features of the PS (Figure 46)	this study	NA
15	DNase I hypersensitive sites (DHSs)	K562	Chromatin features of the PS (Figure 46)	Crawford, Ohler et al. 2012	NCBI Gene Expression Omnibus (GSE32970)

## 4. Discussion

**Table 35. Top five players for TPD in combination with multi-omics.**

No.	Candidates	Reasoning for targeted protein degradation (TPD)
1	NELF complex	Establishes pause <i>in vitro</i> and depletion of NELF leads to reduced levels of Pol II pausing in <i>Drosophila</i> <sup>241,515</sup> . However, so far observed changes in pausing are surprisingly small <sup>352</sup> .
2	CTCF	Covered in the main text.
3	PAF1 complex	TPD will clarify on-going discussion about PAF1 acting as pausing factor <i>in vivo</i> <sup>516</sup> - which is unlikely due to recent structural insights showing that NELF and PAF1 binding to Pol II is mutually exclusive <i>in vitro</i> <sup>110</sup> .
4	Cap-binding complex (CBC)	Facilitates P-TEFb recruitment by interaction with P-TEFb and Pol II <sup>517</sup> . It was further reported that CBC interacts with DSIF and NELF <sup>518,519</sup> .
5	CCNT1	Cyclin subunit of P-TEFb, might be involved in establishing different transcription condensates (see LLPS, Introduction I.1.1) <sup>121</sup> .



**Figure 50. Elongation velocity and a dynamic histone code.**

Spearman correlation coefficients  $\rho$  of elongation velocity  $v$  [kbp min<sup>-1</sup>] against histone marks. Marks were normalized by total nucleosome occupancy. Published datasets of ChIP-seq in K562 cells are listed below in **Table 36**. Agreement with published data is highlighted in green, while correlations contrary to the literature are highlighted with dotted lines (see **Table 37** for references). Note, that none of the cited papers is normalizing to total nucleosome occupancy.

**Table 36. Published data sets used in Figure 50.**

Data sets and references used in the correlation study of elongation velocity against histone features. All ChIP-seq data of histone marks were normalized to total nucleosome occupancy (estimated by MNase-seq).

Experiment	Cell type	GEO ID	Source	Read-out
MNase-seq	K562	GSE35586	Snyder <sup>144</sup>	Active regulatory chromatin (nucleosome-depleted)
ChIP-seq H3K4me2	K562	GSE26320	Bernstein <sup>520</sup>	Protein-DNA interactions
ChIP-seq H3K4me3	K562	GSE32970	Stamatoyannopoulos <sup>144</sup>	Protein-DNA interactions
ChIP-seq H3K27me3	K562	GSM788088	Farnham	Protein-DNA interactions
ChIP-seq H3K4me1	K562	GSE26320	Bernstein <sup>520</sup>	Protein-DNA interactions
ChIP-seq H3K27ac	K562	GSE26320	Bernstein <sup>520</sup>	Protein-DNA interactions
ChIP-seq H3K27me3	K562	GSE32970	Stamatoyannopoulos <sup>144</sup>	Protein-DNA interactions
ChIP-seq H3K9ac	K562	GSM788082	Farnham	Protein-DNA interactions
ChIP-seq H3K4me1	K562	GSM788085	Farnham	Protein-DNA interactions
ChIP-seq H3K4me3	K562	GSM788087	Farnham	Protein-DNA interactions
ChIP-seq H3K9me1	K562	GSE26320	Bernstein <sup>520</sup>	Protein-DNA interactions
ChIP-seq H4K20me1	K562	GSE26320	Bernstein <sup>520</sup>	Protein-DNA interactions
ChIP-seq H3K36me3	K562	GSE32970	Stamatoyannopoulos <sup>144</sup>	Protein-DNA interactions
ChIP-seq H3K9ac	K562	GSE26320	Bernstein <sup>520</sup>	Protein-DNA interactions
ChIP-seq H3K4me3	K562	GSE26320	Bernstein <sup>520</sup>	Protein-DNA interactions
ChIP-seq H2AZ	K562	GSM733786	Bernstein	Protein-DNA interactions
ChIP-seq H3K79me2	K562	GSM733653	Bernstein	Protein-DNA interactions
ChIP-seq H3K27me3	K562	GSE26320	Bernstein <sup>520</sup>	Protein-DNA interactions
ChIP-seq H3K9me3	K562	GSE26320	Bernstein <sup>520</sup>	Protein-DNA interactions
ChIP-seq H3K36me3	K562	GSE26320	Bernstein	Protein-DNA interactions

**Table 37. Comparison of correlation in this study with the literature.**

A positive correlation indicates that marks occur at genes transcribed at fast elongation rates. Neutral highlights marks with no influence on elongation rates. Abbreviation: not assigned (NA).

Histone mark	Reported correlation	Refs (if available)	Correlation (this study)
H2AZ			negative
H2Bub1	positive	500,521	NA
H3ac, H4ac	positive	522	NA (strong negative, see H3K9ac)
H3K27ac			negative
H3K27me3			negative
H3K36me3	neutral	272,335,336	neutral
H3K4me1	positive	130	strong negative
H3K4me2			negative
H3K4me3			negative
H3K79me2	positive	130,335,336	neutral
H3K9ac			strong negative
H3K9me1			neutral
H3K9me3			neutral
H4K20me1	positive / negative	336 / 19	neutral

---

## References

- 1 Dahm, R. Friedrich Miescher and the discovery of DNA. *Dev Biol* **278**, 274-288, doi:10.1016/j.ydbio.2004.11.028 (2005).
- 2 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).
- 3 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351, doi:10.1126/science.1058040 (2001).
- 4 Comings, D. E. The rationale for an ordered arrangement of chromatin in the interphase nucleus. *Am J Hum Genet* **20**, 440-460 (1968).
- 5 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).
- 6 Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**, 737-738 (1953).
- 7 Weiss, S. B. & Gladstone, L. A MAMMALIAN SYSTEM FOR THE INCORPORATION OF CYTIDINE TRIPHOSPHATE INTO RIBONUCLEIC ACID1. *Journal of the American Chemical Society* **81**, 4118-4119, doi:10.1021/ja01524a087 (1959).
- 8 Chambon, P., Karon, H., Ramuz, M. & Mandel, P. The influence of ionic strength and a polyanion on transcription in vitro. II. Effects on the template efficiency of rat liver chromatin for a purified bacterial RNA polymerase. *Biochim Biophys Acta* **157**, 520-531 (1968).
- 9 Roeder, R. G. & Rutter, W. J. Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature* **224**, 234-237 (1969).
- 10 Krakow, J. S. & Fronk, E. Azotobacter vinelandii ribonucleic acid polymerase. 8. Pyrophosphate exchange. *J Biol Chem* **244**, 5988-5993 (1969).
- 11 Steitz, T. A. A mechanism for all polymerases. *Nature* **391**, 231-232, doi:10.1038/34542 (1998).
- 12 Crick, F. H. On protein synthesis. *Symp Soc Exp Biol* **12**, 138-163 (1958).
- 13 Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).
- 14 Cobb, M. 60 years ago, Francis Crick changed the logic of biology. *PLoS Biol* **15**, e2003243, doi:10.1371/journal.pbio.2003243 (2017).
- 15 Cech, T. R. & Steitz, J. A. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157**, 77-94, doi:10.1016/j.cell.2014.03.008 (2014).
- 16 Morris, K. V. & Mattick, J. S. The rise of regulatory RNA. *Nat Rev Genet* **15**, 423-437, doi:10.1038/nrg3722 (2014).
- 17 Shandilya, J. & Roberts, S. G. The transcription cycle in eukaryotes: from productive initiation to RNA polymerase II recycling. *Biochim Biophys Acta* **1819**, 391-400, doi:10.1016/j.bbagr.2012.01.010 (2012).
- 18 Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**, 720-731, doi:10.1038/nrg3293 (2012).
- 19 Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16**, 167-177, doi:10.1038/nrm3953 (2015).

- 20 Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol* **17**, 743-755, doi:10.1038/nrm.2016.104 (2016).
- 21 Uhler, C. & Shivashankar, G. V. Regulation of genome organization and gene expression by nuclear mechanotransduction. *Nat Rev Mol Cell Biol* **18**, 717-727, doi:10.1038/nrm.2017.101 (2017).
- 22 Dekker, J. *et al.* The 4D nucleome project. *Nature* **549**, 219-226, doi:10.1038/nature23884 (2017).
- 23 Gibcus, J. H. & Dekker, J. The hierarchy of the 3D genome. *Mol Cell* **49**, 773-782, doi:10.1016/j.molcel.2013.02.011 (2013).
- 24 Cutter, A. R. & Hayes, J. J. A brief review of nucleosome structure. *FEBS Lett* **589**, 2914-2922, doi:10.1016/j.febslet.2015.05.016 (2015).
- 25 Venkatesh, S. & Workman, J. L. Histone exchange, chromatin structure and the regulation of transcription. *Nat Rev Mol Cell Biol* **16**, 178-189, doi:10.1038/nrm3941 (2015).
- 26 Talbert, P. B. & Henikoff, S. Histone variants on the move: substrates for chromatin dynamics. *Nat Rev Mol Cell Biol* **18**, 115-126, doi:10.1038/nrm.2016.148 (2017).
- 27 Richmond, T. J., Finch, J. T., Rushton, B., Rhodes, D. & Klug, A. Structure of the nucleosome core particle at 7 Å resolution. *Nature* **311**, 532-537 (1984).
- 28 Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251-260, doi:10.1038/38444 (1997).
- 29 Zhou, K., Gaullier, G. & Luger, K. Nucleosome structure and dynamics are coming of age. *Nat Struct Mol Biol* **26**, 3-13, doi:10.1038/s41594-018-0166-x (2019).
- 30 Gomez-Diaz, E. & Corces, V. G. Architectural proteins: regulators of 3D genome organization in cell fate. *Trends Cell Biol* **24**, 703-711, doi:10.1016/j.tcb.2014.08.003 (2014).
- 31 Larson, A. G. & Narlikar, G. J. The Role of Phase Separation in Heterochromatin Formation, Function, and Regulation. *Biochemistry* **57**, 2540-2548, doi:10.1021/acs.biochem.8b00401 (2018).
- 32 Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**, 292-301, doi:10.1038/35066075 (2001).
- 33 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).
- 34 Furlong, E. E. M. & Levine, M. Developmental enhancers and chromosome topology. *Science* **361**, 1341-1345, doi:10.1126/science.aau0320 (2018).
- 35 Pickersgill, H. *et al.* Characterization of the *Drosophila melanogaster* genome at the nuclear lamina. *Nat Genet* **38**, 1005-1014, doi:10.1038/ng1852 (2006).
- 36 Guelen, L. *et al.* Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-951, doi:10.1038/nature06947 (2008).
- 37 van Steensel, B. & Belmont, A. S. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **169**, 780-791, doi:10.1016/j.cell.2017.04.022 (2017).
- 38 van Koningsbruggen, S. *et al.* High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. *Mol Biol Cell* **21**, 3735-3748, doi:10.1091/mbc.E10-06-0508 (2010).

- 
- 39 Pederson, T. The nucleolus. *Cold Spring Harb Perspect Biol* **3**, doi:10.1101/cshperspect.a000638 (2011).
- 40 Clapier, C. R., Iwasa, J., Cairns, B. R. & Peterson, C. L. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nat Rev Mol Cell Biol* **18**, 407-422, doi:10.1038/nrm.2017.26 (2017).
- 41 Lai, W. K. M. & Pugh, B. F. Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nat Rev Mol Cell Biol* **18**, 548-562, doi:10.1038/nrm.2017.47 (2017).
- 42 Lorch, Y., LaPointe, J. W. & Kornberg, R. D. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell* **49**, 203-210 (1987).
- 43 Kujirai, T. *et al.* Structural basis of the nucleosome transition during RNA polymerase II passage. *Science* **362**, 595-598, doi:10.1126/science.aau9904 (2018).
- 44 Farnung, L., Vos, S. M. & Cramer, P. Structure of transcribing RNA polymerase II-nucleosome complex. *Nat Commun* **9**, 5432, doi:10.1038/s41467-018-07870-y (2018).
- 45 Ehara, H. *et al.* Structural insight into nucleosome transcription by RNA polymerase II with elongation factors. *Science*, doi:10.1126/science.aav8912 (2019).
- 46 Landick, R. The regulatory roles and mechanism of transcriptional pausing. *Biochem Soc Trans* **34**, 1062-1066, doi:10.1042/BST0341062 (2006).
- 47 Krishnamurthy, S. & Hampsey, M. Eukaryotic transcription initiation. *Curr Biol* **19**, R153-156, doi:10.1016/j.cub.2008.11.052 (2009).
- 48 Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol* **19**, 621-637, doi:10.1038/s41580-018-0028-8 (2018).
- 49 Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L. & Myers, R. M. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome Res* **16**, 1-10, doi:10.1101/gr.4222606 (2006).
- 50 Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**, 1311-1320, doi:10.1038/ng.3142 (2014).
- 51 Murakami, K. *et al.* Architecture of an RNA polymerase II transcription pre-initiation complex. *Science* **342**, 1238724, doi:10.1126/science.1238724 (2013).
- 52 Schilbach, S. *et al.* Structures of transcription pre-initiation complex with TFIID and Mediator. *Nature* **551**, 204-209, doi:10.1038/nature24282 (2017).
- 53 Dienemann, C., Schwalb, B., Schilbach, S. & Cramer, P. Promoter Distortion and Opening in the RNA Polymerase II Cleft. *Mol Cell* **73**, 97-106 e104, doi:10.1016/j.molcel.2018.10.014 (2019).
- 54 Sainsbury, S., Bernecky, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16**, 129-143, doi:10.1038/nrm3952 (2015).
- 55 Bram, R. J. & Kornberg, R. D. Specific protein binding to far upstream activating sequences in polymerase II promoters. *Proc Natl Acad Sci U S A* **82**, 43-47 (1985).
- 56 Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299-308 (1981).
- 57 Gruss, P., Dhar, R. & Khoury, G. Simian virus 40 tandem repeated sequences as an element of the early promoter. *Proc Natl Acad Sci U S A* **78**, 943-947 (1981).

- 
- 58 Benoist, C. & Chambon, P. In vivo sequence requirements of the SV40 early promoter region. *Nature* **290**, 304-310 (1981).
- 59 Trinklein, N. D. *et al.* An abundance of bidirectional promoters in the human genome. *Genome Res* **14**, 62-66, doi:10.1101/gr.1982804 (2004).
- 60 Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484-1488, doi:10.1126/science.1138341 (2007).
- 61 Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849-1851, doi:10.1126/science.1162253 (2008).
- 62 Ibrahim, M. M. *et al.* Determinants of promoter and enhancer transcription directionality in metazoans. *Nat Commun* **9**, 4472, doi:10.1038/s41467-018-06962-z (2018).
- 63 Alpert, T., Herzelt, L. & Neugebauer, K. M. Perfect timing: splicing and transcription rates in living cells. *Wiley Interdiscip Rev RNA* **8**, doi:10.1002/wrna.1401 (2017).
- 64 Porrua, O. & Libri, D. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat Rev Mol Cell Biol* **16**, 190-202, doi:10.1038/nrm3943 (2015).
- 65 Proudfoot, N. J. Ending the message: poly(A) signals then and now. *Genes Dev* **25**, 1770-1782, doi:10.1101/gad.17268411 (2011).
- 66 Andersen, P. K., Jensen, T. H. & Lykke-Andersen, S. Making ends meet: coordination between RNA 3'-end processing and transcription initiation. *Wiley Interdiscip Rev RNA* **4**, 233-246, doi:10.1002/wrna.1156 (2013).
- 67 Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nat Rev Genet* **15**, 163-175, doi:10.1038/nrg3662 (2014).
- 68 Martinez-Rucobo, F. W. *et al.* Molecular Basis of Transcription-Coupled Pre-mRNA Capping. *Mol Cell* **58**, 1079-1089, doi:10.1016/j.molcel.2015.04.004 (2015).
- 69 Shuman, S. RNA capping: progress and prospects. *RNA* **21**, 735-737, doi:10.1261/rna.049973.115 (2015).
- 70 Wahl, M. C., Will, C. L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701-718, doi:10.1016/j.cell.2009.02.009 (2009).
- 71 Herzelt, L., Ottoz, D. S. M., Alpert, T. & Neugebauer, K. M. Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nat Rev Mol Cell Biol* **18**, 637-650, doi:10.1038/nrm.2017.63 (2017).
- 72 Bazak, L. *et al.* A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res* **24**, 365-376, doi:10.1101/gr.164749.113 (2014).
- 73 Nishikura, K. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem* **79**, 321-349, doi:10.1146/annurev-biochem-060208-105251 (2010).
- 74 Huang, H. *et al.* Histone H3 trimethylation at lysine 36 guides m(6)A RNA modification co-transcriptionally. *Nature* **567**, 414-419, doi:10.1038/s41586-019-1016-7 (2019).
- 75 Battaglia, S. *et al.* RNA-dependent chromatin association of transcription elongation factors and Pol II CTD kinases. *Elife* **6**, doi:10.7554/eLife.25637 (2017).
- 76 Van Nostrand, E. L. *et al.* A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *bioRxiv*, 179648, doi:10.1101/179648 (2018).
- 77 Bjorkegren, C. & Baranello, L. DNA Supercoiling, Topoisomerases, and Cohesin: Partners in Regulating Chromatin Architecture? *Int J Mol Sci* **19**, doi:10.3390/ijms19030884 (2018).

- 
- 78 Lawrence, M., Daujat, S. & Schneider, R. Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends Genet* **32**, 42-56, doi:10.1016/j.tig.2015.10.007 (2016).
- 79 Roeder, R. G. Transcriptional regulation and the role of diverse coactivators in animal cells. *FEBS Lett* **579**, 909-915, doi:10.1016/j.febslet.2004.12.007 (2005).
- 80 Lewis, B. A. Understanding large multiprotein complexes: applying a multiple allosteric networks model to explain the function of the Mediator transcription complex. *J Cell Sci* **123**, 159-163, doi:10.1242/jcs.057216 (2010).
- 81 Sanso, M. & Fisher, R. P. Pause, play, repeat: CDKs push RNAP II's buttons. *Transcription* **4**, 146-152 (2013).
- 82 Eick, D. & Geyer, M. The RNA polymerase II carboxy-terminal domain (CTD) code. *Chem Rev* **113**, 8456-8490, doi:10.1021/cr400071f (2013).
- 83 Gniazdowski, M., Mandel, J. L., Jr., Gissinger, F., Kedinger, C. & Chambon, P. Calf thymus RNA polymerases exhibit template specificity. *Biochem Biophys Res Commun* **38**, 1033-1040 (1970).
- 84 Kedinger, C., Gniazdowski, M., Mandel, J. L., Jr., Gissinger, F. & Chambon, P. Alpha-amanitin: a specific inhibitor of one of two DNA-pendent RNA polymerase activities from calf thymus. *Biochem Biophys Res Commun* **38**, 165-171 (1970).
- 85 Weinmann, R., Raskas, H. J. & Roeder, R. G. Role of DNA-dependent RNA polymerases II and III in transcription of the adenovirus genome late in productive infection. *Proc Natl Acad Sci U S A* **71**, 3426-3439 (1974).
- 86 Kelly, J. L. & Lehman, I. R. Yeast mitochondrial RNA polymerase. Purification and properties of the catalytic subunit. *J Biol Chem* **261**, 10340-10347 (1986).
- 87 Hantsche, M. & Cramer, P. The Structural Basis of Transcription: 10 Years After the Nobel Prize in Chemistry. *Angew Chem Int Ed Engl* **55**, 15972-15981, doi:10.1002/anie.201608066 (2016).
- 88 Burley, S. K. & Roeder, R. G. TATA box mimicry by TFIID: autoinhibition of pol II transcription. *Cell* **94**, 551-553 (1998).
- 89 Boehning, M. *et al.* RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nat Struct Mol Biol* **25**, 833-840, doi:10.1038/s41594-018-0112-y (2018).
- 90 Cramer, P. A tale of chromatin and transcription in 100 structures. *Cell* **159**, 985-994, doi:10.1016/j.cell.2014.10.047 (2014).
- 91 Gottesfeld, J. M. Milestones in transcription and chromatin published in the Journal of Biological Chemistry. *J Biol Chem* **294**, 1652-1660, doi:10.1074/jbc.TM118.004162 (2019).
- 92 Sims, R. J., 3rd, Belotserkovskaya, R. & Reinberg, D. Elongation by RNA polymerase II: the short and long of it. *Genes Dev* **18**, 2437-2468, doi:10.1101/gad.1235904 (2004).
- 93 Liu, Z. & Tjian, R. Visualizing transcription factor dynamics in living cells. *J Cell Biol* **217**, 1181-1191, doi:10.1083/jcb.201710038 (2018).
- 94 Mugridge, J. S., Collier, J. & Gross, J. D. Structural and molecular mechanisms for the control of eukaryotic 5'-3' mRNA decay. *Nat Struct Mol Biol* **25**, 1077-1085, doi:10.1038/s41594-018-0164-z (2018).
- 95 Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13-23, doi:10.1016/j.cell.2017.02.007 (2017).
- 96 Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* **25**, 2227-2241, doi:10.1101/gad.176826.111 (2011).

- 
- 97 Archer, T. K., Cordingley, M. G., Wolford, R. G. & Hager, G. L. Transcription factor access is mediated by accurately positioned nucleosomes on the mouse mammary tumor virus promoter. *Mol Cell Biol* **11**, 688-698 (1991).
- 98 Segal, E. & Widom, J. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet* **10**, 443-456, doi:10.1038/nrg2591 (2009).
- 99 Kouzine, F., Levens, D. & Baranello, L. DNA topology and transcription. *Nucleus* **5**, 195-202, doi:10.4161/nucl.28909 (2014).
- 100 Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010-1022, doi:10.1101/gad.2037511 (2011).
- 101 Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705, doi:10.1016/j.cell.2007.02.005 (2007).
- 102 Kebede, A. F., Schneider, R. & Daujat, S. Novel types and sites of histone modifications emerge as players in the transcriptional regulation contest. *FEBS J* **282**, 1658-1674, doi:10.1111/febs.13047 (2015).
- 103 Farrelly, L. A. *et al.* Histone serotonylation is a permissive modification that enhances TFIID binding to H3K4me3. *Nature*, doi:10.1038/s41586-019-1024-7 (2019).
- 104 Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292**, 1863-1876, doi:10.1126/science.1059493 (2001).
- 105 Bernecky, C., Herzog, F., Baumeister, W., Plitzko, J. M. & Cramer, P. Structure of transcribing mammalian RNA polymerase II. *Nature* **529**, 551-554, doi:10.1038/nature16482 (2016).
- 106 Corden, J. L., Cadena, D. L., Ahearn, J. M., Jr. & Dahmus, M. E. A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. *Proc Natl Acad Sci U S A* **82**, 7934-7938 (1985).
- 107 Harlen, K. M. & Churchman, L. S. The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nat Rev Mol Cell Biol* **18**, 263-273, doi:10.1038/nrm.2017.10 (2017).
- 108 Svejstrup, J. Q. *et al.* Evidence for a mediator cycle at the initiation of transcription. *Proc Natl Acad Sci U S A* **94**, 6075-6078 (1997).
- 109 Fuda, N. J. *et al.* Fcp1 dephosphorylation of the RNA polymerase II C-terminal domain is required for efficient transcription of heat shock genes. *Mol Cell Biol* **32**, 3428-3437, doi:10.1128/MCB.00247-12 (2012).
- 110 Vos, S. M. *et al.* Structure of activated transcription complex Pol II-DSIF-PAF-SPT6. *Nature* **560**, 607-612, doi:10.1038/s41586-018-0440-4 (2018).
- 111 Chun, Y. *et al.* Selective kinase inhibition shows that Bur1 (Cdk9) phosphorylates the Rpb1 linker in vivo. *bioRxiv*, 507251, doi:10.1101/507251 (2018).
- 112 Dieci, G. & Sentenac, A. Detours and shortcuts to transcription reinitiation. *Trends Biochem Sci* **28**, 202-209, doi:10.1016/S0968-0004(03)00054-9 (2003).
- 113 Kuriyan, J. & Eisenberg, D. The origin of protein interactions and allostery in colocalization. *Nature* **450**, 983-990, doi:10.1038/nature06524 (2007).
- 114 Konig, J., Zarnack, K., Luscombe, N. M. & Ule, J. Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet* **13**, 77-83, doi:10.1038/nrg3141 (2012).
- 115 Hyman, A. A., Weber, C. A. & Julicher, F. Liquid-liquid phase separation in biology. *Annu Rev Cell Dev Biol* **30**, 39-58, doi:10.1146/annurev-cellbio-100913-013325 (2014).

- 
- 116 Keating, C. D. Aqueous phase separation as a possible route to compartmentalization of biological molecules. *Acc Chem Res* **45**, 2114-2124, doi:10.1021/ar200294y (2012).
- 117 Strzyz, P. Concentrating on intrinsic disorder. *Nat Rev Mol Cell Biol* **19**, 544, doi:10.1038/s41580-018-0039-5 (2018).
- 118 Shakya, A. & King, J. T. DNA Local-Flexibility-Dependent Assembly of Phase-Separated Liquid Droplets. *Biophys J* **115**, 1840-1847, doi:10.1016/j.bpj.2018.09.022 (2018).
- 119 Trcek, T. & Lehmann, R. All about the RNA after all. *Elife* **6**, doi:10.7554/eLife.24106 (2017).
- 120 Kwon, I. *et al.* Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. *Cell* **155**, 1049-1060, doi:10.1016/j.cell.2013.10.033 (2013).
- 121 Lu, H. *et al.* Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature* **558**, 318-323, doi:10.1038/s41586-018-0174-3 (2018).
- 122 Reynolds, N., O'Shaughnessy, A. & Hendrich, B. Transcriptional repressors: multifaceted regulators of gene expression. *Development* **140**, 505-512, doi:10.1242/dev.083105 (2013).
- 123 Nevado, J., Gaudreau, L., Adam, M. & Ptashne, M. Transcriptional activation by artificial recruitment in mammalian cells. *Proc Natl Acad Sci U S A* **96**, 2674-2677 (1999).
- 124 Allen, B. L. & Taatjes, D. J. The Mediator complex: a central integrator of transcription. *Nat Rev Mol Cell Biol* **16**, 155-166, doi:10.1038/nrm3951 (2015).
- 125 Brent, R. & Ptashne, M. A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. *Cell* **43**, 729-736 (1985).
- 126 Ma, J. & Ptashne, M. The carboxy-terminal 30 amino acids of GAL4 are recognized by GAL80. *Cell* **50**, 137-142 (1987).
- 127 Spiegelman, B. M. & Heinrich, R. Biological control through regulated transcriptional coactivators. *Cell* **119**, 157-167, doi:10.1016/j.cell.2004.09.037 (2004).
- 128 Breyer, W. A. & Matthews, B. W. A structural basis for processivity. *Protein Sci* **10**, 1699-1711, doi:10.1110/ps.10301 (2001).
- 129 Saponaro, M. *et al.* RECQL5 controls transcript elongation and suppresses genome instability associated with transcription stress. *Cell* **157**, 1037-1049, doi:10.1016/j.cell.2014.03.048 (2014).
- 130 Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* **3**, e02407, doi:10.7554/eLife.02407 (2014).
- 131 Moore, M. J. & Proudfoot, N. J. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* **136**, 688-700, doi:10.1016/j.cell.2009.02.001 (2009).
- 132 Fong, N. *et al.* Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev* **28**, 2663-2676, doi:10.1101/gad.252106.114 (2014).
- 133 Tanny, J. C. Chromatin modification by the RNA Polymerase II elongation complex. *Transcription* **5**, e988093, doi:10.4161/21541264.2014.988093 (2014).
- 134 Al-Hashimi, H. M. & Walter, N. G. RNA dynamics: it is about time. *Curr Opin Struct Biol* **18**, 321-329, doi:10.1016/j.sbi.2008.04.004 (2008).
- 135 Lunde, B. M. *et al.* Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain. *Nat Struct Mol Biol* **17**, 1195-1201, doi:10.1038/nsmb.1893 (2010).
- 136 Ansari, A. & Hampsey, M. A role for the CPF 3'-end processing machinery in RNAP II-dependent gene looping. *Genes Dev* **19**, 2969-2978, doi:10.1101/gad.1362305 (2005).

- 137 Mapendano, C. K., Lykke-Andersen, S., Kjems, J., Bertrand, E. & Jensen, T. H. Crosstalk between mRNA 3' end processing and transcription initiation. *Mol Cell* **40**, 410-422, doi:10.1016/j.molcel.2010.10.012 (2010).
- 138 Anamika, K., Gyenis, A. & Tora, L. How to stop: the mysterious links among RNA polymerase II occupancy 3' of genes, mRNA 3' processing and termination. *Transcription* **4**, 7-12, doi:10.4161/trns.22300 (2013).
- 139 Zobeck, K. L., Buckley, M. S., Zipfel, W. R. & Lis, J. T. Recruitment timing and dynamics of transcription factors at the Hsp70 loci in living cells. *Mol Cell* **40**, 965-975, doi:10.1016/j.molcel.2010.11.022 (2010).
- 140 Strauss, B. S. Biochemical Genetics and Molecular Biology: The Contributions of George Beadle and Edward Tatum. *Genetics* **203**, 13-20, doi:10.1534/genetics.116.188995 (2016).
- 141 Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc Natl Acad Sci U S A* **74**, 3171-3175 (1977).
- 142 Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**, 1-8 (1977).
- 143 Witkowski, J. A. The discovery of 'split' genes: a scientific revolution. *Trends Biochem Sci* **13**, 110-113 (1988).
- 144 Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).
- 145 Gerstein, M. B. *et al.* What is a gene, post-ENCODE? History and updated definition. *Genome Res* **17**, 669-681, doi:10.1101/gr.6339607 (2007).
- 146 Martens, J. A., Wu, P. Y. & Winston, F. Regulation of an intergenic transcript controls adjacent gene transcription in *Saccharomyces cerevisiae*. *Genes Dev* **19**, 2695-2704, doi:10.1101/gad.1367605 (2005).
- 147 Ebisuya, M., Yamamoto, T., Nakajima, M. & Nishida, E. Ripples from neighbouring transcription. *Nat Cell Biol* **10**, 1106-1113, doi:10.1038/ncb1771 (2008).
- 148 Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nat Rev Genet* **19**, 789-800, doi:10.1038/s41576-018-0060-8 (2018).
- 149 Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *Elife* **3**, e03523, doi:10.7554/eLife.03523 (2014).
- 150 Salzberg, S. L. Open questions: How many genes do we have? *BMC Biol* **16**, 94, doi:10.1186/s12915-018-0564-x (2018).
- 151 Zhu, L. *et al.* Patterns of exon-intron architecture variation of genes in eukaryotic genomes. *BMC Genomics* **10**, 47, doi:10.1186/1471-2164-10-47 (2009).
- 152 Sakharkar, M. K., Chow, V. T. & Kanguane, P. Distributions of exons and introns in the human genome. *In Silico Biol* **4**, 387-393 (2004).
- 153 Gigli, M. *et al.* A Review of the Giant Protein Titin in Clinical Molecular Diagnostics of Cardiomyopathies. *Front Cardiovasc Med* **3**, 21, doi:10.3389/fcvm.2016.00021 (2016).
- 154 Sakharkar, M. K., Perumal, B. S., Sakharkar, K. R. & Kanguane, P. An analysis on gene architecture in human and mouse genomes. *In Silico Biol* **5**, 347-365 (2005).
- 155 Zhang, M. Q. Statistical features of human exons and their flanking regions. *Hum Mol Genet* **7**, 919-932 (1998).
- 156 Mudge, J. M. & Harrow, J. The state of play in higher eukaryote gene annotation. *Nat Rev Genet* **17**, 758-772, doi:10.1038/nrg.2016.119 (2016).
- 157 Kung, J. T., Colognori, D. & Lee, J. T. Long noncoding RNAs: past, present, and future. *Genetics* **193**, 651-669, doi:10.1534/genetics.112.146704 (2013).

- 
- 158 Project, T. G. Statistics about the current GENCODE Release (version 29). *EMBL-EBI* (2018).
- 159 Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* **47**, D766-D773, doi:10.1093/nar/gky955 (2019).
- 160 Hon, C. C. *et al.* An atlas of human long non-coding RNAs with accurate 5' ends. *Nature* **543**, 199-204, doi:10.1038/nature21374 (2017).
- 161 Nitsche, A., Rose, D., Fasold, M., Reiche, K. & Stadler, P. F. Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved. *RNA* **21**, 801-812, doi:10.1261/rna.046342.114 (2015).
- 162 Beaulieu, Y. B., Kleinman, C. L., Landry-Voyer, A. M., Majewski, J. & Bachand, F. Polyadenylation-dependent control of long noncoding RNA expression by the poly(A)-binding protein nuclear 1. *PLoS Genet* **8**, e1003078, doi:10.1371/journal.pgen.1003078 (2012).
- 163 Wilusz, J. E. *et al.* A triple helix stabilizes the 3' ends of long noncoding RNAs that lack poly(A) tails. *Genes Dev* **26**, 2392-2407, doi:10.1101/gad.204438.112 (2012).
- 164 Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* **4**, e08890, doi:10.7554/eLife.08890 (2015).
- 165 Mele, M. & Rinn, J. L. "Cat's Cradling" the 3D Genome by the Act of lncRNA Transcription. *Mol Cell* **62**, 657-664, doi:10.1016/j.molcel.2016.05.011 (2016).
- 166 Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223-227, doi:10.1038/nature07672 (2009).
- 167 Ransohoff, J. D., Wei, Y. & Khavari, P. A. The functions and unique features of long intergenic non-coding RNA. *Nat Rev Mol Cell Biol* **19**, 143-157, doi:10.1038/nrm.2017.104 (2018).
- 168 Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352**, 1225-1228, doi:10.1126/science.aad9841 (2016).
- 169 Brown, C. J. & Willard, H. F. The human X-inactivation centre is not required for maintenance of X-chromosome inactivation. *Nature* **368**, 154-156, doi:10.1038/368154a0 (1994).
- 170 Galupa, R. & Heard, E. X-Chromosome Inactivation: A Crossroads Between Chromosome Architecture and Gene Regulation. *Annu Rev Genet* **52**, 535-566, doi:10.1146/annurev-genet-120116-024611 (2018).
- 171 Tripathi, V. *et al.* The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol Cell* **39**, 925-938, doi:10.1016/j.molcel.2010.08.011 (2010).
- 172 Amodio, N. *et al.* MALAT1: a druggable long non-coding RNA for targeted anti-cancer approaches. *J Hematol Oncol* **11**, 63, doi:10.1186/s13045-018-0606-4 (2018).
- 173 Barman, P., Reddy, D. & Bhaumik, S. R. Mechanisms of Antisense Transcription Initiation with Implications in Gene Expression, Genomic Integrity and Disease Pathogenesis. *Noncoding RNA* **5**, doi:10.3390/ncrna5010011 (2019).
- 174 Pelechano, V. & Steinmetz, L. M. Gene regulation by antisense transcription. *Nat Rev Genet* **14**, 880-893, doi:10.1038/nrg3594 (2013).
- 175 Wang, G. Q. *et al.* Sirt1 AS lncRNA interacts with its mRNA to inhibit muscle formation by attenuating function of miR-34a. *Sci Rep* **6**, 21865, doi:10.1038/srep21865 (2016).

- 176 Munroe, S. H. & Lazar, M. A. Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA. *J Biol Chem* **266**, 22083-22086 (1991).
- 177 Zacher, B. *et al.* Accurate Promoter and Enhancer Identification in 127 ENCODE and Roadmap Epigenomics Cell Types and Tissues by GenoSTAN. *PLoS One* **12**, e0169249, doi:10.1371/journal.pone.0169249 (2017).
- 178 Hobson, D. J., Wei, W., Steinmetz, L. M. & Svejstrup, J. Q. RNA polymerase II collision interrupts convergent transcription. *Mol Cell* **48**, 365-374, doi:10.1016/j.molcel.2012.08.027 (2012).
- 179 Young, R. S., Kumar, Y., Bickmore, W. A. & Taylor, M. S. Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers. *Genome Biol* **18**, 242, doi:10.1186/s13059-017-1379-8 (2017).
- 180 Halfon, M. S. Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases. *Trends Genet* **35**, 93-103, doi:10.1016/j.tig.2018.11.004 (2019).
- 181 Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol Cell* **49**, 825-837, doi:10.1016/j.molcel.2013.01.038 (2013).
- 182 Henriques, T. *et al.* Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev* **32**, 26-41, doi:10.1101/gad.309351.117 (2018).
- 183 Dao, L. T. M. *et al.* Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat Genet* **49**, 1073-1081, doi:10.1038/ng.3884 (2017).
- 184 Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat Methods* **14**, 629-635, doi:10.1038/nmeth.4264 (2017).
- 185 Buffry, A. D., Mendes, C. C. & McGregor, A. P. The Functionality and Evolution of Eukaryotic Transcriptional Enhancers. *Adv Genet* **96**, 143-206, doi:10.1016/bs.adgen.2016.08.004 (2016).
- 186 Catarino, R. R. & Stark, A. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev* **32**, 202-223, doi:10.1101/gad.310367.117 (2018).
- 187 Li, W. *et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516-520, doi:10.1038/nature12210 (2013).
- 188 Lai, F. *et al.* Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* **494**, 497-501, doi:10.1038/nature11884 (2013).
- 189 Guenther, M. G. & Young, R. A. Transcription. Repressive transcription. *Science* **329**, 150-151, doi:10.1126/science.1193995 (2010).
- 190 Wray, G. A. *et al.* The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* **20**, 1377-1419, doi:10.1093/molbev/msg140 (2003).
- 191 Clos, J., Rabindran, S., Wisniewski, J. & Wu, C. Induction temperature of human heat shock factor is reprogrammed in a Drosophila cell environment. *Nature* **364**, 252-255, doi:10.1038/364252a0 (1993).
- 192 Rougvie, A. E. & Lis, J. T. The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of *D. melanogaster* is transcriptionally engaged. *Cell* **54**, 795-804 (1988).
- 193 Guertin, M. J., Petesch, S. J., Zobeck, K. L., Min, I. M. & Lis, J. T. Drosophila heat shock system as a general model to investigate transcriptional regulation. *Cold Spring Harb Symp Quant Biol* **75**, 1-9, doi:10.1101/sqb.2010.75.039 (2010).

- 
- 194 Vihervaara, A., Duarte, F. M. & Lis, J. T. Molecular mechanisms driving transcriptional stress responses. *Nat Rev Genet* **19**, 385-397, doi:10.1038/s41576-018-0001-6 (2018).
- 195 Leppa, S., Kajanne, R., Arminen, L. & Sistonen, L. Differential induction of Hsp70-encoding genes in human hematopoietic cells. *J Biol Chem* **276**, 31713-31719, doi:10.1074/jbc.M104375200 (2001).
- 196 Morimoto, R. I. Regulation of the heat shock transcriptional response: cross talk between a family of heat shock factors, molecular chaperones, and negative regulators. *Genes Dev* **12**, 3788-3796 (1998).
- 197 Lindquist, S. The heat-shock response. *Annu Rev Biochem* **55**, 1151-1191, doi:10.1146/annurev.bi.55.070186.005443 (1986).
- 198 Gonzalez Plaza, J. J., Hulak, N., Zhumadilov, Z. & Akilzhanova, A. Fever as an important resource for infectious diseases research. *Intractable Rare Dis Res* **5**, 97-102, doi:10.5582/irdr.2016.01009 (2016).
- 199 Roti Roti, J. L. Cellular responses to hyperthermia (40-46 degrees C): cell killing and molecular events. *Int J Hyperthermia* **24**, 3-15, doi:10.1080/02656730701769841 (2008).
- 200 Lindquist, S. & Kim, G. Heat-shock protein 104 expression is sufficient for thermotolerance in yeast. *Proc Natl Acad Sci U S A* **93**, 5301-5306 (1996).
- 201 Verghese, J., Abrams, J., Wang, Y. & Morano, K. A. Biology of the heat shock response and protein chaperones: budding yeast (*Saccharomyces cerevisiae*) as a model system. *Microbiol Mol Biol Rev* **76**, 115-158, doi:10.1128/MMBR.05018-11 (2012).
- 202 Lis, J. T. Imaging *Drosophila* gene activation and polymerase pausing in vivo. *Nature* **450**, 198-202, doi:10.1038/nature06324 (2007).
- 203 Kregel, K. C. Heat shock proteins: modifying factors in physiological stress responses and acquired thermotolerance. *J Appl Physiol (1985)* **92**, 2177-2186, doi:10.1152/jappphysiol.01267.2001 (2002).
- 204 Akerfelt, M., Morimoto, R. I. & Sistonen, L. Heat shock factors: integrators of cell stress, development and lifespan. *Nat Rev Mol Cell Biol* **11**, 545-555, doi:10.1038/nrm2938 (2010).
- 205 Cramer, M. N. & Jay, O. Biophysical aspects of human thermoregulation during heat stress. *Auton Neurosci* **196**, 3-13, doi:10.1016/j.autneu.2016.03.001 (2016).
- 206 Gomez-Pastor, R., Burchfiel, E. T. & Thiele, D. J. Regulation of heat shock transcription factors and their roles in physiology and disease. *Nat Rev Mol Cell Biol* **19**, 4-19, doi:10.1038/nrm.2017.73 (2018).
- 207 Mendillo, M. L. *et al.* HSF1 drives a transcriptional program distinct from heat shock to support highly malignant human cancers. *Cell* **150**, 549-562, doi:10.1016/j.cell.2012.06.031 (2012).
- 208 Ciocca, D. R., Arrigo, A. P. & Calderwood, S. K. Heat shock proteins and heat shock factor 1 in carcinogenesis and tumor development: an update. *Arch Toxicol* **87**, 19-48, doi:10.1007/s00204-012-0918-z (2013).
- 209 Shalgi, R., Hurt, J. A., Lindquist, S. & Burge, C. B. Widespread inhibition of posttranscriptional splicing shapes the cellular transcriptome following heat shock. *Cell Rep* **7**, 1362-1370, doi:10.1016/j.celrep.2014.04.044 (2014).
- 210 Meyer, K. D. *et al.* 5' UTR m(6)A Promotes Cap-Independent Translation. *Cell* **163**, 999-1010, doi:10.1016/j.cell.2015.10.012 (2015).
- 211 Zhou, J. *et al.* Dynamic m(6)A mRNA methylation directs translational control of heat shock response. *Nature* **526**, 591-594, doi:10.1038/nature15377 (2015).

- 212 Shalgi, R. *et al.* Widespread regulation of translation by elongation pausing in heat shock. *Mol Cell* **49**, 439-452, doi:10.1016/j.molcel.2012.11.028 (2013).
- 213 Raychaudhuri, S. *et al.* Interplay of acetyltransferase EP300 and the proteasome system in regulating heat shock transcription factor 1. *Cell* **156**, 975-985, doi:10.1016/j.cell.2014.01.055 (2014).
- 214 Lindquist, S. & Craig, E. A. The heat-shock proteins. *Annu Rev Genet* **22**, 631-677, doi:10.1146/annurev.ge.22.120188.003215 (1988).
- 215 Sati, S. & Cavalli, G. Chromosome conformation capture technologies and their impact in understanding genome function. *Chromosoma* **126**, 33-44, doi:10.1007/s00412-016-0593-6 (2017).
- 216 Ray, J. *et al.* Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. *bioRxiv*, 527838, doi:10.1101/527838 (2019).
- 217 Teves, S. S. & Henikoff, S. Heat shock reduces stalled RNA polymerase II and nucleosome turnover genome-wide. *Genes Dev* **25**, 2387-2397, doi:10.1101/gad.178079.111 (2011).
- 218 Brown, S. A. & Kingston, R. E. Disruption of downstream chromatin directed by a transcriptional activator. *Genes Dev* **11**, 3116-3121 (1997).
- 219 Vihervaara, A. *et al.* Transcriptional response to stress is pre-wired by promoter and enhancer architecture. *Nat Commun* **8**, 255, doi:10.1038/s41467-017-00151-0 (2017).
- 220 Scharf, K. D., Berberich, T., Ebersberger, I. & Nover, L. The plant heat stress transcription factor (Hsf) family: structure, function and evolution. *Biochim Biophys Acta* **1819**, 104-119, doi:10.1016/j.bbagr.2011.10.002 (2012).
- 221 Vihervaara, A. *et al.* Transcriptional response to stress in the dynamic chromatin environment of cycling and mitotic cells. *Proc Natl Acad Sci U S A* **110**, E3388-3397, doi:10.1073/pnas.1305275110 (2013).
- 222 Niskanen, E. A. *et al.* Global SUMOylation on active chromatin is an acute heat stress response restricting transcription. *Genome Biol* **16**, 153, doi:10.1186/s13059-015-0717-y (2015).
- 223 Hentze, N., Le Breton, L., Wiesner, J., Kempf, G. & Mayer, M. P. Molecular mechanism of thermosensory function of human heat shock transcription factor Hsf1. *Elife* **5**, doi:10.7554/eLife.11576 (2016).
- 224 Trinklein, N. D., Murray, J. I., Hartman, S. J., Botstein, D. & Myers, R. M. The role of heat shock transcription factor 1 in the genome-wide regulation of the mammalian heat shock response. *Mol Biol Cell* **15**, 1254-1261, doi:10.1091/mbc.e03-10-0738 (2004).
- 225 Inouye, S. *et al.* Heat shock transcription factor 1 opens chromatin structure of interleukin-6 promoter to facilitate binding of an activator or a repressor. *J Biol Chem* **282**, 33210-33217, doi:10.1074/jbc.M704471200 (2007).
- 226 Poirier, G. G., de Murcia, G., Jongstra-Bilen, J., Niedergang, C. & Mandel, P. Poly(ADP-ribosyl)ation of polynucleosomes causes relaxation of chromatin structure. *Proc Natl Acad Sci U S A* **79**, 3423-3427 (1982).
- 227 Fujimoto, M., Takii, R., Katiyar, A., Srivastava, P. & Nakai, A. Poly(ADP-Ribose) Polymerase 1 Promotes the Human Heat Shock Response by Facilitating Heat Shock Transcription Factor 1 Binding to DNA. *Mol Cell Biol* **38**, doi:10.1128/MCB.00051-18 (2018).
- 228 Zanton, S. J. & Pugh, B. F. Full and partial genome-wide assembly and disassembly of the yeast transcription machinery in response to heat shock. *Genes Dev* **20**, 2250-2265, doi:10.1101/gad.1437506 (2006).

- 
- 229 Murray, J. I. *et al.* Diverse and specific gene expression responses to stresses in cultured human cells. *Mol Biol Cell* **15**, 2361-2374, doi:10.1091/mbc.e03-11-0799 (2004).
- 230 Bouchama, A. *et al.* A Model of Exposure to Extreme Environmental Heat Uncovers the Human Transcriptome to Heat Stress. *Sci Rep* **7**, 9429, doi:10.1038/s41598-017-09819-5 (2017).
- 231 Mahat, D. B., Salamanca, H. H., Duarte, F. M., Danko, C. G. & Lis, J. T. Mammalian Heat Shock Response and Mechanisms Underlying Its Genome-wide Transcriptional Regulation. *Mol Cell* **62**, 63-78, doi:10.1016/j.molcel.2016.02.025 (2016).
- 232 Aprile-Garcia, F., Tomar, P., Hummel, B., Khavaran, A. & Sawarkar, R. Nascent-protein ubiquitination is required for heat shock-induced gene downregulation in human cells. *Nat Struct Mol Biol* **26**, 137-146, doi:10.1038/s41594-018-0182-x (2019).
- 233 Gariglio, P., Bellard, M. & Chambon, P. Clustering of RNA polymerase B molecules in the 5' moiety of the adult beta-globin gene of hen erythrocytes. *Nucleic Acids Res* **9**, 2589-2598, doi:10.1093/nar/9.11.2589 (1981).
- 234 Gilmour, D. S. & Lis, J. T. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in *Drosophila melanogaster* cells. *Mol Cell Biol* **6**, 3984-3989 (1986).
- 235 Strobl, L. J. & Eick, D. Hold back of RNA polymerase II at the transcription start site mediates down-regulation of c-myc in vivo. *EMBO J* **11**, 3307-3314 (1992).
- 236 Krumm, A., Meulia, T., Brunvand, M. & Groudine, M. The block to transcriptional elongation within the human c-myc gene is determined in the promoter-proximal region. *Genes Dev* **6**, 2201-2213 (1992).
- 237 Plet, A., Eick, D. & Blanchard, J. M. Elongation and premature termination of transcripts initiated from c-fos and c-myc promoters show dissimilar patterns. *Oncogene* **10**, 319-328 (1995).
- 238 Bentley, D. L. & Groudine, M. A block to elongation is largely responsible for decreased transcription of c-myc in differentiated HL60 cells. *Nature* **321**, 702-706, doi:10.1038/321702a0 (1986).
- 239 Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. *Science* **327**, 335-338, doi:10.1126/science.1181421 (2010).
- 240 Rahl, P. B. *et al.* c-Myc regulates transcriptional pause release. *Cell* **141**, 432-445, doi:10.1016/j.cell.2010.03.030 (2010).
- 241 Core, L. J. *et al.* Defining the status of RNA polymerase at promoters. *Cell Rep* **2**, 1025-1035, doi:10.1016/j.celrep.2012.08.034 (2012).
- 242 Sigova, A. A. *et al.* Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proc Natl Acad Sci U S A* **110**, 2876-2881, doi:10.1073/pnas.1221904110 (2013).
- 243 Bunch, H. *et al.* RNA polymerase II promoter-proximal pausing in mammalian long non-coding genes. *Genomics* **108**, 64-77, doi:10.1016/j.ygeno.2016.07.003 (2016).
- 244 Flynn, R. A., Almada, A. E., Zamudio, J. R. & Sharp, P. A. Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *Proc Natl Acad Sci U S A* **108**, 10460-10465, doi:10.1073/pnas.1106630108 (2011).
- 245 Glover-Cutter, K., Kim, S., Espinosa, J. & Bentley, D. L. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat Struct Mol Biol* **15**, 71-78, doi:10.1038/nsmb1352 (2008).

- 246 Moteki, S. & Price, D. Functional coupling of capping and transcription of mRNA. *Mol Cell* **10**, 599-609 (2002).
- 247 Tome, J. M., Tippens, N. D. & Lis, J. T. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nat Genet* **50**, 1533-1541, doi:10.1038/s41588-018-0234-5 (2018).
- 248 Nechaev, S. & Adelman, K. Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. *Biochim Biophys Acta* **1809**, 34-45, doi:10.1016/j.bbagr.2010.11.001 (2011).
- 249 Elagib, K. E. *et al.* Cross-talk of GATA-1 and P-TEFb in megakaryocyte differentiation. *Blood* **112**, 4884-4894, doi:10.1182/blood-2008-03-145722 (2008).
- 250 Liu, X., Kraus, W. L. & Bai, X. Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways. *Trends Biochem Sci* **40**, 516-525, doi:10.1016/j.tibs.2015.07.003 (2015).
- 251 Adelman, K. *et al.* Immediate mediators of the inflammatory response are poised for gene activation through RNA polymerase II stalling. *Proc Natl Acad Sci U S A* **106**, 18207-18212, doi:10.1073/pnas.0910177106 (2009).
- 252 Galbraith, M. D. *et al.* HIF1A employs CDK8-mediator to stimulate RNAPII elongation in response to hypoxia. *Cell* **153**, 1327-1339, doi:10.1016/j.cell.2013.04.048 (2013).
- 253 Boettiger, A. N. & Levine, M. Synchronous and stochastic patterns of gene activation in the Drosophila embryo. *Science* **325**, 471-473, doi:10.1126/science.1173976 (2009).
- 254 Hamperl, S. & Cimprich, K. A. Conflict Resolution in the Genome: How Transcription and Replication Make It Work. *Cell* **167**, 1455-1467, doi:10.1016/j.cell.2016.09.053 (2016).
- 255 Palmer, A. C., Egan, J. B. & Shearwin, K. E. Transcriptional interference by RNA polymerase pausing and dislodgement of transcription factors. *Transcription* **2**, 9-14, doi:10.4161/trns.2.1.13511 (2011).
- 256 Hendrix, D. A., Hong, J. W., Zeitlinger, J., Rokhsar, D. S. & Levine, M. S. Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. *Proc Natl Acad Sci U S A* **105**, 7762-7767, doi:10.1073/pnas.0802406105 (2008).
- 257 Larson, M. H. *et al.* A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science* **344**, 1042-1047, doi:10.1126/science.1251871 (2014).
- 258 Wada, T. *et al.* DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev* **12**, 343-356 (1998).
- 259 Yamaguchi, Y. *et al.* NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97**, 41-51 (1999).
- 260 Gilchrist, D. A. *et al.* Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell* **143**, 540-551, doi:10.1016/j.cell.2010.10.004 (2010).
- 261 Jimeno-Gonzalez, S., Ceballos-Chavez, M. & Reyes, J. C. A positioned +1 nucleosome enhances promoter-proximal pausing. *Nucleic Acids Res* **43**, 3068-3078, doi:10.1093/nar/gkv149 (2015).
- 262 Cheng, B. *et al.* Functional association of Gdown1 with RNA polymerase II poised on human genes. *Mol Cell* **45**, 38-50, doi:10.1016/j.molcel.2011.10.022 (2012).
- 263 Marshall, N. F. & Price, D. H. Control of formation of two distinct classes of RNA polymerase II elongation complexes. *Mol Cell Biol* **12**, 2078-2090 (1992).

- 
- 264 Marshall, N. F. & Price, D. H. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J Biol Chem* **270**, 12335-12338 (1995).
- 265 Price, D. H. P-TEFb, a cyclin-dependent kinase controlling elongation by RNA polymerase II. *Mol Cell Biol* **20**, 2629-2634 (2000).
- 266 Shi, X. *et al.* Paf1p, an RNA polymerase II-associated factor in *Saccharomyces cerevisiae*, may have both positive and negative roles in transcription. *Mol Cell Biol* **16**, 669-676 (1996).
- 267 Zhu, B. *et al.* The human PAF complex coordinates transcription with events downstream of RNA synthesis. *Genes Dev* **19**, 1668-1673, doi:10.1101/gad.1292105 (2005).
- 268 Yu, M. *et al.* RNA polymerase II-associated factor 1 regulates the release and phosphorylation of paused RNA polymerase II. *Science* **350**, 1383-1386, doi:10.1126/science.aad2338 (2015).
- 269 Swanson, M. S., Carlson, M. & Winston, F. SPT6, an essential gene that affects transcription in *Saccharomyces cerevisiae*, encodes a nuclear protein with an extremely acidic amino terminus. *Mol Cell Biol* **10**, 4935-4941 (1990).
- 270 Kaplan, C. D., Laprade, L. & Winston, F. Transcription elongation factors repress transcription initiation from cryptic sites. *Science* **301**, 1096-1099, doi:10.1126/science.1087374 (2003).
- 271 Endoh, M. *et al.* Human Spt6 stimulates transcription elongation by RNA polymerase II in vitro. *Mol Cell Biol* **24**, 3324-3336 (2004).
- 272 Ardehali, M. B. *et al.* Spt6 enhances the elongation rate of RNA polymerase II in vivo. *EMBO J* **28**, 1067-1077, doi:10.1038/emboj.2009.56 (2009).
- 273 Adelman, K. *et al.* Efficient release from promoter-proximal stall sites requires transcript cleavage factor TFIIIS. *Mol Cell* **17**, 103-112, doi:10.1016/j.molcel.2004.11.028 (2005).
- 274 Garriga, J. *et al.* CDK9 is constitutively expressed throughout the cell cycle, and its steady-state expression is independent of SKP2. *Mol Cell Biol* **23**, 5165-5173 (2003).
- 275 Papparidis, N. F., Durvale, M. C. & Canduri, F. The emerging picture of CDK9/P-TEFb: more than 20 years of advances since PITALRE. *Mol Biosyst* **13**, 246-276, doi:10.1039/c6mb00387g (2017).
- 276 Bacon, C. W. & D'Orso, I. CDK9: a signaling hub for transcriptional control. *Transcription*, 1-19, doi:10.1080/21541264.2018.1523668 (2018).
- 277 Grana, X. *et al.* PITALRE, a nuclear CDC2-related protein kinase that phosphorylates the retinoblastoma protein in vitro. *Proc Natl Acad Sci U S A* **91**, 3834-3838 (1994).
- 278 Peng, J., Zhu, Y., Milton, J. T. & Price, D. H. Identification of multiple cyclin subunits of human P-TEFb. *Genes Dev* **12**, 755-762 (1998).
- 279 Wei, P., Garber, M. E., Fang, S. M., Fischer, W. H. & Jones, K. A. A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell* **92**, 451-462 (1998).
- 280 Kohoutek, J. & Blazek, D. Cyclin K goes with Cdk12 and Cdk13. *Cell Div* **7**, 12, doi:10.1186/1747-1028-7-12 (2012).
- 281 Nguyen, V. T., Kiss, T., Michels, A. A. & Bensaude, O. 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature* **414**, 322-325, doi:10.1038/35104581 (2001).

- 282 Yang, Z., Zhu, Q., Luo, K. & Zhou, Q. The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature* **414**, 317-322, doi:10.1038/35104575 (2001).
- 283 Li, Q. *et al.* Analysis of the large inactive P-TEFb complex indicates that it contains one 7SK molecule, a dimer of HEXIM1 or HEXIM2, and two P-TEFb molecules containing Cdk9 phosphorylated at threonine 186. *J Biol Chem* **280**, 28819-28826, doi:10.1074/jbc.M502712200 (2005).
- 284 AJ, C. Q., Bugai, A. & Barboric, M. Cracking the control of RNA polymerase II elongation by 7SK snRNP and P-TEFb. *Nucleic Acids Res* **44**, 7527-7539, doi:10.1093/nar/gkw585 (2016).
- 285 Peterlin, B. M., Brogie, J. E. & Price, D. H. 7SK snRNA: a noncoding RNA that plays a major role in regulating eukaryotic transcription. *Wiley Interdiscip Rev RNA* **3**, 92-103, doi:10.1002/wrna.106 (2012).
- 286 Eberhardy, S. R. & Farnham, P. J. Myc recruits P-TEFb to mediate the final step in the transcriptional activation of the cad promoter. *J Biol Chem* **277**, 40156-40162, doi:10.1074/jbc.M207441200 (2002).
- 287 Barboric, M., Nissen, R. M., Kanazawa, S., Jabrane-Ferrat, N. & Peterlin, B. M. NF-kappaB binds P-TEFb to stimulate transcriptional elongation by RNA polymerase II. *Mol Cell* **8**, 327-337 (2001).
- 288 McNamara, R. P., McCann, J. L., Gudipaty, S. A. & D'Orso, I. Transcription factors mediate the enzymatic disassembly of promoter-bound 7SK snRNP to locally recruit P-TEFb for transcription elongation. *Cell Rep* **5**, 1256-1268, doi:10.1016/j.celrep.2013.11.003 (2013).
- 289 Takahashi, H. *et al.* Human mediator subunit MED26 functions as a docking site for transcription elongation factors. *Cell* **146**, 92-104, doi:10.1016/j.cell.2011.06.005 (2011).
- 290 Jang, M. K. *et al.* The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Mol Cell* **19**, 523-534, doi:10.1016/j.molcel.2005.06.027 (2005).
- 291 Yang, Z. *et al.* Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. *Mol Cell* **19**, 535-545, doi:10.1016/j.molcel.2005.06.029 (2005).
- 292 Shelton, S. B. *et al.* Crosstalk between the RNA Methylation and Histone-Binding Activities of MePCE Regulates P-TEFb Activation on Chromatin. *Cell Rep* **22**, 1374-1383, doi:10.1016/j.celrep.2018.01.028 (2018).
- 293 McNamara, R. P. *et al.* KAP1 Recruitment of the 7SK snRNP Complex to Promoters Enables Transcription Elongation by RNA Polymerase II. *Mol Cell* **61**, 39-53, doi:10.1016/j.molcel.2015.11.004 (2016).
- 294 Larochelle, S. *et al.* Cyclin-dependent kinase control of the initiation-to-elongation switch of RNA polymerase II. *Nat Struct Mol Biol* **19**, 1108-1115, doi:10.1038/nsmb.2399 (2012).
- 295 Ebmeier, C. C. *et al.* Human TFIIH Kinase CDK7 Regulates Transcription-Associated Chromatin Modifications. *Cell Rep* **20**, 1173-1186, doi:10.1016/j.celrep.2017.07.021 (2017).
- 296 Ivanov, D., Kwak, Y. T., Guo, J. & Gaynor, R. B. Domains in the SPT5 protein that modulate its transcriptional regulatory properties. *Mol Cell Biol* **20**, 2970-2983, doi:10.1128/mcb.20.9.2970-2983.2000 (2000).

- 
- 297 Fujinaga, K. *et al.* Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. *Mol Cell Biol* **24**, 787-795, doi:10.1128/mcb.24.2.787-795.2004 (2004).
- 298 Marshall, N. F., Peng, J., Xie, Z. & Price, D. H. Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase. *J Biol Chem* **271**, 27176-27183 (1996).
- 299 Ahn, S. H., Kim, M. & Buratowski, S. Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing. *Mol Cell* **13**, 67-76 (2004).
- 300 Ni, Z., Schwartz, B. E., Werner, J., Suarez, J. R. & Lis, J. T. Coordination of transcription, RNA processing, and surveillance by P-TEFb kinase on heat shock genes. *Mol Cell* **13**, 55-65 (2004).
- 301 Garber, M. E. *et al.* CDK9 autophosphorylation regulates high-affinity binding of the human immunodeficiency virus type 1 tat-P-TEFb complex to TAR RNA. *Mol Cell Biol* **20**, 6958-6969 (2000).
- 302 Laitem, C. *et al.* CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II-transcribed genes. *Nat Struct Mol Biol* **22**, 396-403, doi:10.1038/nsmb.3000 (2015).
- 303 Sanso, M. *et al.* P-TEFb regulation of transcription termination factor Xrn2 revealed by a chemical genetic screen for Cdk9 substrates. *Genes Dev* **30**, 117-131, doi:10.1101/gad.269589.115 (2016).
- 304 Kwak, H. & Lis, J. T. Control of transcriptional elongation. *Annu Rev Genet* **47**, 483-508, doi:10.1146/annurev-genet-110711-155440 (2013).
- 305 Sheridan, R. M., Fong, N., D'Alessandro, A. & Bentley, D. L. Widespread Backtracking by RNA Pol II Is a Major Effector of Gene Activation, 5' Pause Release, Termination, and Transcription Elongation Rate. *Mol Cell* **73**, 107-118 e104, doi:10.1016/j.molcel.2018.10.031 (2019).
- 306 Michel, M. & Cramer, P. Transitions for regulating early transcription. *Cell* **153**, 943-944, doi:10.1016/j.cell.2013.04.050 (2013).
- 307 Vos, S. M., Farnung, L., Urlaub, H. & Cramer, P. Structure of paused transcription complex Pol II-DSIF-NELF. *Nature* **560**, 601-606, doi:10.1038/s41586-018-0442-2 (2018).
- 308 Saba, J. *et al.* The elemental mechanism of transcriptional pausing. *Elife* **8**, doi:10.7554/eLife.40981 (2019).
- 309 Armache, K. J., Kettenberger, H. & Cramer, P. The dynamic machinery of mRNA elongation. *Curr Opin Struct Biol* **15**, 197-203, doi:10.1016/j.sbi.2005.03.002 (2005).
- 310 Robinson, P. J. *et al.* Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. *Cell* **166**, 1411-1422 e1416, doi:10.1016/j.cell.2016.08.050 (2016).
- 311 Louder, R. K. *et al.* Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. *Nature* **531**, 604-609, doi:10.1038/nature17394 (2016).
- 312 He, Y. *et al.* Near-atomic resolution visualization of human transcription promoter opening. *Nature* **533**, 359-365, doi:10.1038/nature17970 (2016).
- 313 Plaschka, C. *et al.* Transcription initiation complex structures elucidate DNA opening. *Nature* **533**, 353-358, doi:10.1038/nature17990 (2016).
- 314 Kettenberger, H., Armache, K. J. & Cramer, P. Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS. *Mol Cell* **16**, 955-965, doi:10.1016/j.molcel.2004.11.040 (2004).

- 315 Ehrensberger, A. H., Kelly, G. P. & Svejstrup, J. Q. Mechanistic interpretation of promoter-proximal peaks and RNAPII density maps. *Cell* **154**, 713-715, doi:10.1016/j.cell.2013.07.032 (2013).
- 316 Herbert, K. M., Greenleaf, W. J. & Block, S. M. Single-molecule studies of RNA polymerase: motoring along. *Annu Rev Biochem* **77**, 149-176, doi:10.1146/annurev.biochem.77.073106.100741 (2008).
- 317 Larson, M. H. *et al.* Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. *Proc Natl Acad Sci U S A* **109**, 6555-6560, doi:10.1073/pnas.1200939109 (2012).
- 318 Tamm, I., Folkers, K., Shunk, C. H. & Horsfall, F. L., Jr. Inhibition of influenza virus multiplication by N-glycosides of benzimidazoles-N. *J Exp Med* **99**, 227-250 (1954).
- 319 Sehgal, P. B., Derman, E., Molloy, G. R., Tamm, I. & Darnell, J. E. 5,6-Dichloro-1-Beta-D-ribofuranosylbenzimidazole inhibits initiation of nuclear heterogeneous RNA chains in HeLa cells. *Science* **194**, 431-433 (1976).
- 320 Chodosh, L. A., Fire, A., Samuels, M. & Sharp, P. A. 5,6-Dichloro-1-beta-D-ribofuranosylbenzimidazole inhibits transcription elongation by RNA polymerase II in vitro. *J Biol Chem* **264**, 2250-2257 (1989).
- 321 Kephart, D. D., Marshall, N. F. & Price, D. H. Stability of Drosophila RNA polymerase II elongation complexes in vitro. *Mol Cell Biol* **12**, 2067-2077 (1992).
- 322 Mason, P. B. & Struhl, K. Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. *Mol Cell* **17**, 831-840, doi:10.1016/j.molcel.2005.02.017 (2005).
- 323 Tennyson, C. N., Klamut, H. J. & Worton, R. G. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nat Genet* **9**, 184-190, doi:10.1038/ng0295-184 (1995).
- 324 Singh, J. & Padgett, R. A. Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol* **16**, 1128-1133, doi:10.1038/nsmb.1666 (2009).
- 325 Wada, Y. *et al.* A wave of nascent transcription on activated human genes. *Proc Natl Acad Sci U S A* **106**, 18357-18361, doi:10.1073/pnas.0902573106 (2009).
- 326 O'Brien, T. & Lis, J. T. Rapid changes in Drosophila transcription after an instantaneous heat shock. *Mol Cell Biol* **13**, 3456-3463 (1993).
- 327 Yao, J., Ardehali, M. B., Fecko, C. J., Webb, W. W. & Lis, J. T. Intranuclear distribution and local dynamics of RNA polymerase II during transcription activation. *Mol Cell* **28**, 978-990, doi:10.1016/j.molcel.2007.10.017 (2007).
- 328 Darzacq, X. *et al.* In vivo dynamics of RNA polymerase II transcription. *Nat Struct Mol Biol* **14**, 796-806, doi:10.1038/nsmb1280 (2007).
- 329 Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H. Visualization of single RNA transcripts in situ. *Science* **280**, 585-590 (1998).
- 330 Boireau, S. *et al.* The transcriptional cycle of HIV-1 in real-time and live cells. *J Cell Biol* **179**, 291-304, doi:10.1083/jcb.200706018 (2007).
- 331 Ben-Ari, Y. *et al.* The life of an mRNA in space and time. *J Cell Sci* **123**, 1761-1774, doi:10.1242/jcs.062638 (2010).
- 332 Brody, Y. *et al.* The in vivo kinetics of RNA polymerase II elongation during co-transcriptional splicing. *PLoS Biol* **9**, e1000573, doi:10.1371/journal.pbio.1000573 (2011).

- 333 Larson, D. R., Zenklusen, D., Wu, B., Chao, J. A. & Singer, R. H. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science* **332**, 475-478, doi:10.1126/science.1202142 (2011).
- 334 Danko, C. G. *et al.* Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol Cell* **50**, 212-222, doi:10.1016/j.molcel.2013.02.015 (2013).
- 335 Fuchs, G. *et al.* 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biol* **15**, R69, doi:10.1186/gb-2014-15-5-r69 (2014).
- 336 Veloso, A. *et al.* Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res* **24**, 896-905, doi:10.1101/gr.171405.113 (2014).
- 337 Morales, F. & Giordano, A. Overview of CDK9 as a target in cancer research. *Cell Cycle* **15**, 519-527, doi:10.1080/15384101.2016.1138186 (2016).
- 338 Bishop, A. C. *et al.* A chemical switch for inhibitor-sensitive alleles of any protein kinase. *Nature* **407**, 395-401, doi:10.1038/35030148 (2000).
- 339 Lopez, M. S., Kliegman, J. I. & Shokat, K. M. The logic and design of analog-sensitive kinases and their small molecule inhibitors. *Methods Enzymol* **548**, 189-213, doi:10.1016/B978-0-12-397918-6.00008-2 (2014).
- 340 Bishop, A. C., Buzko, O. & Shokat, K. M. Magic bullets for protein kinases. *Trends Cell Biol* **11**, 167-172 (2001).
- 341 Larochelle, S. *et al.* Dichotomous but stringent substrate selection by the dual-function Cdk7 complex revealed by chemical genetics. *Nat Struct Mol Biol* **13**, 55-62, doi:10.1038/nsmb1028 (2006).
- 342 Larochelle, S. *et al.* Requirements for Cdk7 in the assembly of Cdk1/cyclin B and activation of Cdk2 revealed by chemical genetics in human cells. *Mol Cell* **25**, 839-850, doi:10.1016/j.molcel.2007.02.003 (2007).
- 343 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821, doi:10.1126/science.1225829 (2012).
- 344 Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096, doi:10.1126/science.1258096 (2014).
- 345 Hsu, P. D., Lander, E. S. & Zhang, F. Development and applications of CRISPR-Cas9 for genome engineering. *Cell* **157**, 1262-1278, doi:10.1016/j.cell.2014.05.010 (2014).
- 346 Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845-1848, doi:10.1126/science.1162228 (2008).
- 347 Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950-953, doi:10.1126/science.1229386 (2013).
- 348 Weber, C. M., Ramachandran, S. & Henikoff, S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol Cell* **53**, 819-830, doi:10.1016/j.molcel.2014.02.014 (2014).
- 349 Mayer, A. *et al.* Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161**, 541-554, doi:10.1016/j.cell.2015.03.010 (2015).
- 350 Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161**, 526-540, doi:10.1016/j.cell.2015.03.027 (2015).

- 351 Zentner, G. E. & Henikoff, S. High-resolution digital profiling of the epigenome. *Nat Rev Genet* **15**, 814-827, doi:10.1038/nrg3798 (2014).
- 352 Muse, G. W. *et al.* RNA polymerase is poised for activation across the genome. *Nat Genet* **39**, 1507-1511, doi:10.1038/ng.2007.21 (2007).
- 353 Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the *Drosophila melanogaster* embryo. *Nat Genet* **39**, 1512-1516, doi:10.1038/ng.2007.26 (2007).
- 354 Reppas, N. B., Wade, J. T., Church, G. M. & Struhl, K. The transition between transcriptional initiation and elongation in *E. coli* is highly variable and often rate limiting. *Mol Cell* **24**, 747-757, doi:10.1016/j.molcel.2006.10.030 (2006).
- 355 Henriques, T. *et al.* Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. *Mol Cell* **52**, 517-528, doi:10.1016/j.molcel.2013.10.001 (2013).
- 356 Li, B., Weber, J. A., Chen, Y., Greenleaf, A. L. & Gilmour, D. S. Analyses of promoter-proximal pausing by RNA polymerase II on the hsp70 heat shock gene promoter in a *Drosophila* nuclear extract. *Mol Cell Biol* **16**, 5433-5443 (1996).
- 357 Buckley, M. S., Kwak, H., Zipfel, W. R. & Lis, J. T. Kinetics of promoter Pol II on Hsp70 reveal stable pausing and key insights into its regulation. *Genes Dev* **28**, 14-19, doi:10.1101/gad.231886.113 (2014).
- 358 Vispe, S. *et al.* Triptolide is an inhibitor of RNA polymerase I and II-dependent transcription leading predominantly to down-regulation of short-lived mRNA. *Mol Cancer Ther* **8**, 2780-2790, doi:10.1158/1535-7163.MCT-09-0549 (2009).
- 359 Titov, D. V. *et al.* XPB, a subunit of TFIIH, is a target of the natural product triptolide. *Nat Chem Biol* **7**, 182-188, doi:10.1038/nchembio.522 (2011).
- 360 Krebs, A. R. *et al.* Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol Cell* **67**, 411-422 e414, doi:10.1016/j.molcel.2017.06.027 (2017).
- 361 Chen, F., Gao, X. & Shilatifard, A. Stably paused genes revealed through inhibition of transcription initiation by the TFIIH inhibitor triptolide. *Genes Dev* **29**, 39-47, doi:10.1101/gad.246173.114 (2015).
- 362 Melvin, W. T., Milne, H. B., Slater, A. A., Allen, H. J. & Keir, H. M. Incorporation of 6-thioguanosine and 4-thiouridine into RNA. Application to isolation of newly synthesised RNA by affinity chromatography. *Eur J Biochem* **92**, 373-379 (1978).
- 363 Cleary, M. D., Meiring, C. D., Jan, E., Guymon, R. & Boothroyd, J. C. Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nat Biotechnol* **23**, 232-237, doi:10.1038/nbt1061 (2005).
- 364 Kenzelmann, M. *et al.* Microarray analysis of newly synthesized RNA in cells and animals. *Proc Natl Acad Sci U S A* **104**, 6164-6169, doi:10.1073/pnas.0610439104 (2007).
- 365 Dölken, L. *et al.* High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA* **14**, 1959-1972, doi:10.1261/rna.1136108 (2008).
- 366 Miller, C. *et al.* Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol Syst Biol* **7**, 458, doi:10.1038/msb.2010.112 (2011).

- 367 Sun, M. *et al.* Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Res* **22**, 1350-1359, doi:10.1101/gr.130161.111 (2012).
- 368 Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat Biotechnol* **29**, 436-442, doi:10.1038/nbt.1861 (2011).
- 369 Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337-342, doi:10.1038/nature10098 (2011).
- 370 Schulz, D. *et al.* Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell* **155**, 1075-1087, doi:10.1016/j.cell.2013.10.024 (2013).
- 371 Tani, H. & Akimitsu, N. Genome-wide technology for determining RNA stability in mammalian cells: historical perspective and recent advantages based on modified nucleotide labeling. *RNA Biol* **9**, 1233-1238, doi:10.4161/rna.22036 (2012).
- 372 Weintz, G. *et al.* The phosphoproteome of toll-like receptor-activated macrophages. *Mol Syst Biol* **6**, 371, doi:10.1038/msb.2010.29 (2010).
- 373 Wachutka, L. & Gagneur, J. Measures of RNA metabolism rates: Toward a definition at the level of single bonds. *Transcription* **8**, 75-80, doi:10.1080/21541264.2016.1257972 (2017).
- 374 Michel, M. *et al.* TT-seq captures enhancer landscapes immediately after T-cell stimulation. *Mol Syst Biol* **13**, 920, doi:10.15252/msb.20167507 (2017).
- 375 Wouters, J., Kalender Atak, Z. & Aerts, S. Decoding transcriptional states in cancer. *Curr Opin Genet Dev* **43**, 82-92, doi:10.1016/j.gde.2017.01.003 (2017).
- 376 Zylicz, J. J. *et al.* The Implication of Early Chromatin Changes in X Chromosome Inactivation. *Cell*, doi:10.1016/j.cell.2018.11.041 (2018).
- 377 Riml, C. *et al.* Osmium-Mediated Transformation of 4-Thiouridine to Cytidine as Key To Study RNA Dynamics by Sequencing. *Angew Chem Int Ed Engl* **56**, 13479-13483, doi:10.1002/anie.201707465 (2017).
- 378 Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C. & Simon, M. D. TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat Methods* **15**, 221-225, doi:10.1038/nmeth.4582 (2018).
- 379 Herzog, V. A. *et al.* Thiol-linked alkylation of RNA to assess expression dynamics. *Nat Methods* **14**, 1198-1204, doi:10.1038/nmeth.4435 (2017).
- 380 Uvarovskii, A., Naarmann-de Vries, I. & Dieterich, C. On the optimal design of metabolic RNA labeling experiments. *bioRxiv*, 428862, doi:10.1101/428862 (2018).
- 381 Gressel, S. *et al.* CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* **6**, doi:10.7554/eLife.29736 (2017).
- 382 Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129**, 823-837, doi:10.1016/j.cell.2007.05.009 (2007).
- 383 Mikkelsen, T. S. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560, doi:10.1038/nature06008 (2007).
- 384 Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497-1502, doi:10.1126/science.1141319 (2007).
- 385 Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408-1419, doi:10.1016/j.cell.2011.11.013 (2011).

- 386 He, Q., Johnston, J. & Zeitlinger, J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nat Biotechnol* **33**, 395-401, doi:10.1038/nbt.3121 (2015).
- 387 Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368-373, doi:10.1038/nature09652 (2011).
- 388 Nojima, T., Gomes, T., Carmo-Fonseca, M. & Proudfoot, N. J. Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. *Nat Protoc* **11**, 413-428, doi:10.1038/nprot.2016.012 (2016).
- 389 Cai, H. & Luse, D. S. Transcription initiation by RNA polymerase II in vitro. Properties of preinitiation, initiation, and elongation complexes. *J Biol Chem* **262**, 298-304 (1987).
- 390 Wuarin, J. & Schibler, U. Physical isolation of nascent RNA chains transcribed by RNA polymerase II: evidence for cotranscriptional splicing. *Mol Cell Biol* **14**, 7219-7225 (1994).
- 391 Schlackow, M. *et al.* Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs. *Mol Cell* **65**, 25-38, doi:10.1016/j.molcel.2016.11.029 (2017).
- 392 Nojima, T. *et al.* RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing. *Mol Cell* **72**, 369-379 e364, doi:10.1016/j.molcel.2018.09.004 (2018).
- 393 Lowthert, L. A., Ku, N. O., Liao, J., Coulombe, P. A. & Omary, M. B. Empigen BB: a useful detergent for solubilization and biochemical analysis of keratins. *Biochem Biophys Res Commun* **206**, 370-379, doi:10.1006/bbrc.1995.1051 (1995).
- 394 Crawford, C. R. *et al.* Use of zwitterionic detergent for the preparation of an influenza virus vaccine. 1: Preparation and characterization of disrupted virions. *Vaccine* **2**, 193-198 (1984).
- 395 Shendure, J. & Lieberman Aiden, E. The expanding scope of DNA sequencing. *Nat Biotechnol* **30**, 1084-1094, doi:10.1038/nbt.2421 (2012).
- 396 Norrander, J., Kempe, T. & Messing, J. Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. *Gene* **26**, 101-106 (1983).
- 397 Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281-2308, doi:10.1038/nprot.2013.143 (2013).
- 398 Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9**, 671-675 (2012).
- 399 Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-386 (2000).
- 400 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).
- 401 Andrews, S. FastQC: a quality control tool for high throughput sequence data. (2010).
- 402 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *2011* **17**, 3, doi:10.14806/ej.17.1.200 (2011).
- 403 Dobin, A. & Gingeras, T. R. Mapping RNA-seq Reads with STAR. *Curr Protoc Bioinformatics* **51**, 11 14 11-19, doi:10.1002/0471250953.bi1114s51 (2015).
- 404 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
- 405 Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2015).

- 406 Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80, doi:10.1186/gb-2004-5-10-r80 (2004).
- 407 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539, doi:10.1038/msb.2011.75 (2011).
- 408 Bello, T. & Gujral, T. S. Kinhibition: A Kinase Inhibitor Selection Portal. *iScience* **8**, 49-53, doi:10.1016/j.isci.2018.09.009 (2018).
- 409 American Type Culture Collection Standards Development Organization Workgroup, A. S. N. Cell line misidentification: the beginning of the end. *Nat Rev Cancer* **10**, 441-448, doi:10.1038/nrc2852 (2010).
- 410 Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. in *Statistical Analysis of Next Generation Sequencing Data* (eds Somnath Datta & Dan Nettleton) 169-190 (Springer International Publishing, 2014).
- 411 Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-408, doi:10.1006/meth.2001.1262 (2001).
- 412 Thomas, J. M. & Hodes, M. E. A new discontinuous buffer system for the electrophoresis of cationic proteins at near-neutral pH. *Anal Biochem* **118**, 194-196 (1981).
- 413 Goldman, A., Harper, S. & Speicher, D. W. Detection of Proteins on Blot Membranes. *Curr Protoc Protein Sci* **86**, 10.18.11-10.18.11, doi:10.1002/cpps.15 (2016).
- 414 Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129-141, doi:10.1016/j.cell.2010.03.009 (2010).
- 415 Lebedeva, S. *et al.* Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol Cell* **43**, 340-352, doi:10.1016/j.molcel.2011.06.008 (2011).
- 416 Farazi, T. A. *et al.* Identification of distinct miRNA target regulation between breast cancer molecular subtypes using AGO2-PAR-CLIP and patient datasets. *Genome Biol* **15**, R9, doi:10.1186/gb-2014-15-1-r9 (2014).
- 417 Chomczynski, P. & Mackey, K. Short technical reports. Modification of the TRI reagent procedure for isolation of RNA from polysaccharide- and proteoglycan-rich sources. *Biotechniques* **19**, 942-945 (1995).
- 418 Conrad, T. & Ørom, U. A. Cellular Fractionation and Isolation of Chromatin-Associated RNA. *Methods Mol Biol* **1468**, 1-9, doi:10.1007/978-1-4939-4035-6\_1 (2017).
- 419 Albright, L. M. & Slatko, B. E. Denaturing polyacrylamide gel electrophoresis. *Curr Protoc Nucleic Acid Chem* **Appendix 3**, Appendix 3B, doi:10.1002/0471142700.nca03bs00 (2001).
- 420 Yost, H. J. & Lindquist, S. RNA splicing is interrupted by heat shock and is rescued by heat shock protein synthesis. *Cell* **45**, 185-193 (1986).
- 421 Mulholland, C. B. *et al.* A modular open platform for systematic functional studies under physiological conditions. *Nucleic Acids Res* **43**, e112, doi:10.1093/nar/gkv550 (2015).
- 422 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106, doi:10.1186/gb-2010-11-10-r106 (2010).
- 423 Sugimoto, N. *et al.* Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry* **34**, 11211-11216 (1995).
- 424 SantaLucia, J., Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci U S A* **95**, 1460-1465 (1998).

- 425 Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701-705, doi:10.1038/nature12894 (2014).
- 426 Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26, doi:10.1186/1748-7188-6-26 (2011).
- 427 Mahat, D. B. & Lis, J. T. Use of conditioned media is critical for studies of regulation in response to rapid heat shock. *Cell Stress Chaperones* **22**, 155-162, doi:10.1007/s12192-016-0737-x (2017).
- 428 Strober, W. Trypan blue exclusion test of cell viability. *Curr Protoc Immunol Appendix* **3**, Appendix 3B, doi:10.1002/0471142735.ima03bs21 (2001).
- 429 Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94, doi:10.1186/1471-2105-11-94 (2010).
- 430 Huber, W., Toedling, J. & Steinmetz, L. M. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* **22**, 1963-1970, doi:10.1093/bioinformatics/btl289 (2006).
- 431 Mayer, A. & Churchman, L. S. Genome-wide profiling of RNA polymerase transcription at nucleotide resolution in human cells with native elongating transcript sequencing. *Nat Protoc* **11**, 813-833, doi:10.1038/nprot.2016.047 (2016).
- 432 Axel, R. Cleavage of DNA in nuclei and chromatin with staphylococcal nuclease. *Biochemistry* **14**, 2921-2925 (1975).
- 433 Allan, J., Fraser, R. M., Owen-Hughes, T. & Keszenman-Pereyra, D. Micrococcal nuclease does not substantially bias nucleosome mapping. *J Mol Biol* **417**, 152-164, doi:10.1016/j.jmb.2012.01.043 (2012).
- 434 Chung, H. R. *et al.* The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS One* **5**, e15754, doi:10.1371/journal.pone.0015754 (2010).
- 435 Nojima, T. *et al.* Deregulated Expression of Mammalian lncRNA through Loss of SPT6 Induces R-Loop Formation, Replication Stress, and Cellular Senescence. *Mol Cell* **72**, 970-984 e977, doi:10.1016/j.molcel.2018.10.011 (2018).
- 436 Kireeva, M. L., Komissarova, N., Waugh, D. S. & Kashlev, M. The 8-nucleotide-long RNA:DNA hybrid is a primary stability determinant of the RNA polymerase II elongation complex. *J Biol Chem* **275**, 6530-6536 (2000).
- 437 Ma, W. *et al.* Fine-scale chromatin interaction maps reveal the cis-regulatory landscape of human lincRNA genes. *Nat Methods* **12**, 71-78, doi:10.1038/nmeth.3205 (2015).
- 438 Ghavi-Helm, Y. *et al.* Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **512**, 96-100, doi:10.1038/nature13417 (2014).
- 439 Liu, W. *et al.* Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell* **155**, 1581-1595, doi:10.1016/j.cell.2013.10.056 (2013).
- 440 Chen, F. X. *et al.* PAF1, a Molecular Regulator of Promoter-Proximal Pausing by RNA Polymerase II. *Cell* **162**, 1003-1015, doi:10.1016/j.cell.2015.07.042 (2015).
- 441 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).
- 442 Schwalb, B. *et al.* Measurement of genome-wide RNA synthesis and decay rates with Dynamic Transcriptome Analysis (DTA). *Bioinformatics* **28**, 884-885, doi:10.1093/bioinformatics/bts052 (2012).

- 443 Wolfe, M. B., Goldstrohm, A. C. & Freddolino, P. L. Global analysis of RNA metabolism using bio-orthogonal labeling coupled with next-generation RNA sequencing. *Methods* **155**, 88-103, doi:10.1016/j.ymeth.2018.12.001 (2019).
- 444 Shao, W. & Zeitlinger, J. Paused RNA polymerase II inhibits new transcriptional initiation. *Nat Genet* **49**, 1045-1051, doi:10.1038/ng.3867 (2017).
- 445 Coulon, A., Chow, C. C., Singer, R. H. & Larson, D. R. Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nat Rev Genet* **14**, 572-584, doi:10.1038/nrg3484 (2013).
- 446 Pope, S. D. & Medzhitov, R. Emerging Principles of Gene Expression Programs and Their Regulation. *Mol Cell* **71**, 389-397, doi:10.1016/j.molcel.2018.07.017 (2018).
- 447 Vvedenskaya, I. O. *et al.* Interactions between RNA polymerase and the "core recognition element" counteract pausing. *Science* **344**, 1285-1289, doi:10.1126/science.1253458 (2014).
- 448 Traut, T. W. Physiological concentrations of purines and pyrimidines. *Mol Cell Biochem* **140**, 1-22 (1994).
- 449 Kireeva, M. L. & Kashlev, M. Mechanism of sequence-specific pausing of bacterial RNA polymerase. *Proc Natl Acad Sci U S A* **106**, 8900-8905, doi:10.1073/pnas.0900407106 (2009).
- 450 Homann, O. R. & Johnson, A. D. MochiView: versatile software for genome browsing and DNA motif analysis. *BMC Biol* **8**, 49, doi:10.1186/1741-7007-8-49 (2010).
- 451 Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res* **14**, 1188-1190, doi:10.1101/gr.849004 (2004).
- 452 Dubois, M. F. *et al.* Heat-shock inactivation of the TFIIF-associated kinase and change in the phosphorylation sites on the C-terminal domain of RNA polymerase II. *Nucleic Acids Res* **25**, 694-700 (1997).
- 453 Egyhazi, E. *et al.* Heat-shock-specific phosphorylation and transcriptional activity of RNA polymerase II. *Exp Cell Res* **242**, 211-221, doi:10.1006/excr.1998.4112 (1998).
- 454 Schwartz, B. E., Larochelle, S., Suter, B. & Lis, J. T. Cdk7 is required for full activation of Drosophila heat shock genes and RNA polymerase II phosphorylation in vivo. *Mol Cell Biol* **23**, 6876-6886 (2003).
- 455 Bunch, H. *et al.* TRIM28 regulates RNA polymerase II promoter-proximal pausing and pause release. *Nat Struct Mol Biol* **21**, 876-883, doi:10.1038/nsmb.2878 (2014).
- 456 Dubois, M. F. *et al.* Heat shock of HeLa cells inactivates a nuclear protein phosphatase specific for dephosphorylation of the C-terminal domain of RNA polymerase II. *Nucleic Acids Res* **27**, 1338-1344 (1999).
- 457 Zyllicz, J. J. *et al.* The Implication of Early Chromatin Changes in X Chromosome Inactivation. *Cell* **176**, 182-197 e123, doi:10.1016/j.cell.2018.11.041 (2019).
- 458 Duffy, E. E. *et al.* Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry. *Mol Cell* **59**, 858-866, doi:10.1016/j.molcel.2015.07.023 (2015).
- 459 Marzi, M. J. & Nicassio, F. Uncovering the Stability of Mature miRNAs by 4-Thio-Uridine Metabolic Labeling. *Methods Mol Biol* **1823**, 141-152, doi:10.1007/978-1-4939-8624-8\_11 (2018).
- 460 Duffy, E. E., Canzio, D., Maniatis, T. & Simon, M. D. Solid phase chemistry to covalently and reversibly capture thiolated RNA. *Nucleic Acids Res* **46**, 6996-7005, doi:10.1093/nar/gky556 (2018).
- 461 Stasevich, T. J. *et al.* Regulation of RNA polymerase II activation by histone acetylation in single living cells. *Nature* **516**, 272-275, doi:10.1038/nature13714 (2014).

- 462 The enzyme-linked immunosorbent assay (ELISA). *Bull World Health Organ* **54**, 129-139 (1976).
- 463 Nguyen-Duc, T., Peeters, E., Muyldermans, S., Charlier, D. & Hassanzadeh-Ghassabeh, G. Nanobody(R)-based chromatin immunoprecipitation/micro-array analysis for genome-wide identification of transcription factor DNA binding sites. *Nucleic Acids Res* **41**, e59, doi:10.1093/nar/gks1342 (2013).
- 464 Hamers-Casterman, C. *et al.* Naturally occurring antibodies devoid of light chains. *Nature* **363**, 446-448, doi:10.1038/363446a0 (1993).
- 465 Muyldermans, S. Nanobodies: natural single-domain antibodies. *Annu Rev Biochem* **82**, 775-797, doi:10.1146/annurev-biochem-063011-092449 (2013).
- 466 MPI-bpc, P. R. Animal Welfare Prize for Max Planck researchers Dirk Görlich and Tino Pleiner. (2018).
- 467 Lopes, R., Agami, R. & Korkmaz, G. GRO-seq, A Tool for Identification of Transcripts Regulating Gene Expression. *Methods Mol Biol* **1543**, 45-55, doi:10.1007/978-1-4939-6716-2\_3 (2017).
- 468 Gardini, A. Global Run-On Sequencing (GRO-Seq). *Methods Mol Biol* **1468**, 111-120, doi:10.1007/978-1-4939-4035-6\_9 (2017).
- 469 Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods* **9**, 72-74, doi:10.1038/nmeth.1778 (2011).
- 470 Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* **11**, 163-166, doi:10.1038/nmeth.2772 (2014).
- 471 Fu, Y., Wu, P. H., Beane, T., Zamore, P. D. & Weng, Z. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics* **19**, 531, doi:10.1186/s12864-018-4933-1 (2018).
- 472 Bonhoure, N. *et al.* Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. *Genome Res* **24**, 1157-1168, doi:10.1101/gr.168260.113 (2014).
- 473 Orlando, D. A. *et al.* Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Rep* **9**, 1163-1170, doi:10.1016/j.celrep.2014.10.018 (2014).
- 474 Portz, B. *et al.* Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. *Nat Commun* **8**, 15231, doi:10.1038/ncomms15231 (2017).
- 475 Lu, H. *et al.* AFF1 is a ubiquitous P-TEFb partner to enable Tat extraction of P-TEFb from 7SK snRNP and formation of SECs for HIV transactivation. *Proc Natl Acad Sci U S A* **111**, E15-24, doi:10.1073/pnas.1318503111 (2014).
- 476 Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* **2**, 343-372, doi:10.1146/annurev.genom.2.1.343 (2001).
- 477 Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional Addiction in Cancer. *Cell* **168**, 629-643, doi:10.1016/j.cell.2016.12.013 (2017).
- 478 Fujita, T. & Fujii, H. Efficient isolation of specific genomic regions and identification of associated proteins by engineered DNA-binding molecule-mediated chromatin immunoprecipitation (enChIP) using CRISPR. *Biochem Biophys Res Commun* **439**, 132-136, doi:10.1016/j.bbrc.2013.08.013 (2013).
- 479 Hamidian, A. *et al.* Promoter-associated proteins of EPAS1 identified by enChIP-MS - A putative role of HDX as a negative regulator. *Biochem Biophys Res Commun* **499**, 291-298, doi:10.1016/j.bbrc.2018.03.150 (2018).

- 480 Groehler, A. t., Degner, A. & Tretyakova, N. Y. Mass Spectrometry-Based Tools to Characterize DNA-Protein Cross-Linking by Bis-Electrophiles. *Basic Clin Pharmacol Toxicol* **121 Suppl 3**, 63-77, doi:10.1111/bcpt.12751 (2017).
- 481 Lin, S., Staahl, B. T., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *Elife* **3**, e04766, doi:10.7554/eLife.04766 (2014).
- 482 Adli, M. The CRISPR tool kit for genome editing and beyond. *Nat Commun* **9**, 1911, doi:10.1038/s41467-018-04252-2 (2018).
- 483 Cheng, J. C., Moore, T. B. & Sakamoto, K. M. RNA interference and human disease. *Mol Genet Metab* **80**, 121-128 (2003).
- 484 Sakamoto, K. M. *et al.* Protacs: chimeric molecules that target proteins to the Skp1-Cullin-F box complex for ubiquitination and degradation. *Proc Natl Acad Sci U S A* **98**, 8554-8559, doi:10.1073/pnas.141230798 (2001).
- 485 Kronke, J. *et al.* Lenalidomide causes selective degradation of IKZF1 and IKZF3 in multiple myeloma cells. *Science* **343**, 301-305, doi:10.1126/science.1244851 (2014).
- 486 Nabet, B. *et al.* The dTAG system for immediate and target-specific protein degradation. *Nat Chem Biol* **14**, 431-441, doi:10.1038/s41589-018-0021-8 (2018).
- 487 Natsume, T., Kiyomitsu, T., Saga, Y. & Kanemaki, M. T. Rapid Protein Depletion in Human Cells by Auxin-Inducible Degron Tagging with Short Homology Donors. *Cell Rep* **15**, 210-218, doi:10.1016/j.celrep.2016.03.001 (2016).
- 488 Cristina Mayor-Ruiz, G. E. W. Identification and characterization of cancer vulnerabilities via targeted protein degradation. *Drug Discovery Today: Technologies*, doi:https://doi.org/10.1016/j.ddtec.2018.12.003 (2019).
- 489 Parelho, V. *et al.* Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422-433, doi:10.1016/j.cell.2008.01.011 (2008).
- 490 Laitem, C. *et al.* CTCF regulates NELF, DSIF and P-TEFb recruitment during transcription. *Transcription* **6**, 79-90, doi:10.1080/21541264.2015.1095269 (2015).
- 491 Kang, H. & Lieberman, P. M. Mechanism of glycyrrhizic acid inhibition of Kaposi's sarcoma-associated herpesvirus: disruption of CTCF-cohesin-mediated RNA polymerase II pausing and sister chromatid cohesion. *J Virol* **85**, 11159-11169, doi:10.1128/JVI.00720-11 (2011).
- 492 Oesterreich, F. C. *et al.* Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell* **165**, 372-381, doi:10.1016/j.cell.2016.02.045 (2016).
- 493 Wallace, E. W. J. & Beggs, J. D. Extremely fast and incredibly close: cotranscriptional splicing in budding yeast. *RNA* **23**, 601-610, doi:10.1261/rna.060830.117 (2017).
- 494 Close, P. *et al.* DBIRD complex integrates alternative mRNA splicing with RNA polymerase II transcript elongation. *Nature* **484**, 386-389, doi:10.1038/nature10925 (2012).
- 495 de la Mata, M. *et al.* A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell* **12**, 525-532 (2003).
- 496 Munoz, M. J. *et al.* DNA damage regulates alternative splicing through inhibition of RNA polymerase II elongation. *Cell* **137**, 708-720, doi:10.1016/j.cell.2009.03.010 (2009).
- 497 Dujardin, G. *et al.* How slow RNA polymerase II elongation favors alternative exon skipping. *Mol Cell* **54**, 683-690, doi:10.1016/j.molcel.2014.03.044 (2014).

- 498 Godoy Herz, M. A. *et al.* Light Regulates Plant Alternative Splicing through the Control of Transcriptional Elongation. *Mol Cell* **73**, 1066-1074 e1063, doi:10.1016/j.molcel.2018.12.005 (2019).
- 499 Taft, R. J. *et al.* Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nat Struct Mol Biol* **17**, 1030-1034, doi:10.1038/nsmb.1841 (2010).
- 500 Pavri, R. *et al.* Histone H2B monoubiquitination functions cooperatively with FACT to regulate elongation by RNA polymerase II. *Cell* **125**, 703-717, doi:10.1016/j.cell.2006.04.029 (2006).
- 501 Solomon, M. J. & Varshavsky, A. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proc Natl Acad Sci U S A* **82**, 6470-6474 (1985).
- 502 Ucker, D. S. & Yamamoto, K. R. Early events in the stimulation of mammary tumor virus RNA synthesis by glucocorticoids. Novel assays of transcription rates. *J Biol Chem* **259**, 7416-7420 (1984).
- 503 Thummel, C. S., Burtis, K. C. & Hogness, D. S. Spatial and temporal patterns of E74 transcription during *Drosophila* development. *Cell* **61**, 101-111 (1990).
- 504 Shermoen, A. W. & O'Farrell, P. H. Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. *Cell* **67**, 303-310 (1991).
- 505 Fukaya, T., Lim, B. & Levine, M. Rapid Rates of Pol II Elongation in the *Drosophila* Embryo. *Curr Biol* **27**, 1387-1391, doi:10.1016/j.cub.2017.03.069 (2017).
- 506 Zhang, J., Cavallaro, M. & Hebenstreit, D. Timing Polymerase Pausing with TV-PRO-seq. *bioRxiv*, 461442, doi:10.1101/461442 (2018).
- 507 Ferlay J, E. M., Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. (2018).
- 508 Zoppino, F. C. M., Guerrero-Gimenez, M. E., Castro, G. N. & Ciocca, D. R. Comprehensive transcriptomic analysis of heat shock proteins in the molecular subtypes of human breast cancer. *BMC Cancer* **18**, 700, doi:10.1186/s12885-018-4621-1 (2018).
- 509 Tang, D. *et al.* Expression of heat shock proteins and heat shock protein messenger ribonucleic acid in human prostate carcinoma in vitro and in tumors in vivo. *Cell Stress Chaperones* **10**, 46-58 (2005).
- 510 Ono, A. *et al.* Overexpression of heat shock protein 27 in squamous cell carcinoma of the uterine cervix: a proteomic analysis using archival formalin-fixed, paraffin-embedded tissues. *Hum Pathol* **40**, 41-49, doi:10.1016/j.humpath.2008.06.010 (2009).
- 511 Liu, W. *et al.* Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell* **155**, 1581-1595, doi:10.1016/j.cell.2013.10.056 (2013).
- 512 Baranello, L. *et al.* RNA Polymerase II Regulates Topoisomerase 1 Activity to Favor Efficient Transcription. *Cell* **165**, 357-371, doi:10.1016/j.cell.2016.02.036 (2016).
- 513 Friedel, M., Nikolajewa, S., Suhnel, J. & Wilhelm, T. DiProDB: a database for dinucleotide properties. *Nucleic Acids Res* **37**, D37-40, doi:10.1093/nar/gkn597 (2009).
- 514 Descostes, N. *et al.* Tyrosine phosphorylation of RNA polymerase II CTD is associated with antisense promoter transcription and active enhancers in mammalian cells. *Elife* **3**, e02105, doi:10.7554/eLife.02105 (2014).
- 515 Wu, C. H. *et al.* NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in *Drosophila*. *Genes Dev* **17**, 1402-1414, doi:10.1101/gad.1091403 (2003).

- 
- 516 Van Oss, S. B., Cucinotta, C. E. & Arndt, K. M. Emerging Insights into the Roles of the Paf1 Complex in Gene Regulation. *Trends Biochem Sci* **42**, 788-798, doi:10.1016/j.tibs.2017.08.003 (2017).
- 517 Lenasi, T., Peterlin, B. M. & Barboric, M. Cap-binding protein complex links pre-mRNA capping to transcription elongation and alternative splicing through positive transcription elongation factor b (P-TEFb). *J Biol Chem* **286**, 22758-22768, doi:10.1074/jbc.M111.235077 (2011).
- 518 Mandal, S. S. *et al.* Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II. *Proc Natl Acad Sci U S A* **101**, 7572-7577, doi:10.1073/pnas.0401493101 (2004).
- 519 Wen, Y. & Shatkin, A. J. Transcription elongation factor hSPT5 stimulates mRNA capping. *Genes Dev* **13**, 1774-1779 (1999).
- 520 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:10.1038/nature09906 (2011).
- 521 Fuchs, G. *et al.* RNF20 and USP44 regulate stem cell differentiation by modulating H2B monoubiquitylation. *Mol Cell* **46**, 662-673, doi:10.1016/j.molcel.2012.05.023 (2012).
- 522 Zhou, H. L. *et al.* Hu proteins regulate alternative splicing by inducing localized histone hyperacetylation in an RNA-dependent manner. *Proc Natl Acad Sci U S A* **108**, E627-635, doi:10.1073/pnas.1103344108 (2011).

# Appendix

## Listed Contents

### List of figures

#### I. Introduction

Figure 1. Key regulatory concepts of chromatin transcription in human cells.....	3
Figure 2. Structural modeling of promoter-proximal pausing and initiation.....	11
Figure 3. Studying transcription kinetics genome-wide.....	14
Figure 4. Multi-omics analysis of transcription kinetics in human cells.....	18

#### II. Materials and Methods

Figure 5. STR electropherograms of human hematopoietic cell lines used in this study. ....	32
Figure 6. Detailed experimental workflow of TT-seq and mNET-seq experiments. ....	36
Figure 7. Labeled spike-ins are enriched in TT-seq compared to RNA-seq samples.....	39
Figure 8. Size distribution of cDNA libraries generated in this study. ....	40
Figure 9. CRISPR/Cas9 engineering of analog-sensitive CDK9 in human Raji B cells.....	48
Figure 10. Correlations and fragment sizes of cDNA libraries (Raji B cells). ....	53
Figure 11. Correlations and fragment sizes of cDNA libraries (K562 cells). ....	60
Figure 12. Correlations and fragment sizes of TT-seq libraries (Raji B). ....	61

#### III. Results

Figure 13. Establishing mNET-seq of total Pol II in human hematopoietic cell lines. ....	64
Figure 14. Inhibition of the human CDK9as kinase activity specific.....	66
Figure 15. Decrease of RNA synthesis in the 5'-region of genes upon CDK9 inhibition. ....	67
Figure 16. Pol II elongation velocity is gene-specific. ....	68
Figure 17. Characteristics of the promoter-proximal pause site.....	69
Figure 18. Pol II pausing generally limits transcription initiation ('pause-initiation limit').	71
Figure 19. Increasing Pol II pause duration decreases the frequency of initiation.....	72
Figure 20. CDK9 inhibition leads to increased pause duration. ....	73
Figure 21. Determinants of long promoter-proximal pause duration d.....	74
Figure 22. Multi-omics analysis and annotation of different gene classes in K562 cells.....	76
Figure 23. Transcription kinetics of different gene classes in steady-state.....	78
Figure 24. Regulation of transcription upon heat shock time-course in human K562 cells...	80
Figure 25. DE analysis of transcript classes and kinetics of HSPA1A upon heat shock. ....	81
Figure 26. GO analysis of differentially expressed RefSeq-TUs upon heat shock.....	82
Figure 27. Enhancer transcription is generally not pause limited.....	83
Figure 28. Gene classes vary in CDK9-induced pause duration changes upon heat shock....	85

## IV. Discussion and Outlook

Figure 29. Summary of the novel multi-omics analysis. ....	87
Figure 30. Pause site specification by sequence in flies, bacteria and humans. ....	90
Figure 31. Target regions of human POLR2A antibodies. ....	94
Figure 32. Future developments and applications of our multi-omics approach. ....	98
Figure 33. The challenge to quantify promoter-proximal attenuation. ....	99
Figure 34. Elongation velocity at high resolution across human genes. ....	101

## V. Supplementary Information

Figure 35. Commercially available CDK9 kinase inhibitors compared by Kinhibition. ....	106
Figure 36. Heat shock proteins are deregulated in human cancer cells. ....	107
Figure 37. RNA spike-in pool preparation for TT-seq. ....	111
Figure 38. Comparison of cell fractionation protocols by marker proteins. ....	113
Figure 39. Optimizing nascent RNA size selection by PAGE (6 % versus 8 %). ....	114
Figure 40. Uncropped gel images and membrane scans for figures. ....	115
Figure 41. CDK9 (WT or analog-sensitive) protein levels and cell viability. ....	116
Figure 42. TT-seq tracks transcription after 10 min of CDK9as inhibition. ....	117
Figure 43. Occupancy profiling by mNET-seq alone does not reveal kinetics. ....	120
Figure 44. TT-seq tracks transcription after 30 min of CDK9as inhibition. ....	121
Figure 45. TSS and TTS are better refined in new annotation of human K562 cells. ....	122
Figure 46. Transcription kinetics of different transcript classes in steady-state. ....	122
Figure 47. Optimal time points for heat shock time-course of human hematopoietic cells. ....	123
Figure 48. Constitutive exons analysis agrees with analysis on new annotation. ....	126
Figure 49. GenoSTAN annotation and DE analysis of TUs in Raji B cells. ....	127
Figure 50. Elongation velocity and a dynamic histone code. ....	129

### List of tables

## II. Materials and Methods

Table 1. Chemicals and reagents used in this study in alphabetical order. ....	19
Table 2. Enzymes used in this study in alphabetical order. ....	21
Table 3. Kits used in this study in alphabetical order. ....	21
Table 4. Consumables used in this study. ....	22
Table 5. Protein, DNA and RNA ladder used in this study. ....	22
Table 6. Plasmids used as IVT template of RNA spike-ins in this study. ....	23
Table 7. Plasmid used for CRISPR/Cas9 engineering in this study. ....	23
Table 8. RT-qPCR primers used in this study. ....	23
Table 9. CDK9as screening: PCR and sequencing primers. ....	24
Table 10. Other oligonucleotides used in this study. ....	25
Table 11. Primary antibodies used in this study. ....	25
Table 12. HRP-coupled secondary antibodies used for Western blotting in this study. ....	26
Table 13. Human cell lines used in this study. ....	26

Table 14. Growth and freeze medium composition used in this study.....	26
Table 15. Cell culture consumables used in this study.....	27
Table 16. Western blot: buffers used in this study.....	27
Table 17. TT-seq: solutions and buffers used in this study.....	28
Table 18. mNET-seq: solutions and buffers used in this study.....	28
Table 19. Overview of software and databases used in this thesis.....	30
Table 20. Characteristics of 4sU-labeled and unlabeled RNA spike-ins.....	35
Table 21. Information on experimental conditions used in this study.....	49
Table 22. Sequencing statistics of 12 libraries generated in this study.....	52
Table 23. Information on experimental conditions used in this study.....	55
Table 24. Annotation of RNA classes in human hematopoietic cell lines (K562, Raji B).....	58
Table 25. Sequencing statistics of 26 libraries generated in this study.....	59

### III. Results

Table 26. Comparison of three cellular fractionation protocols.....	62
---	----

### V. Supplementary Information

Table 27. Genome-wide approaches used to study Pol II occupancy.....	103
Table 28. Comparison of Pol II elongation rate estimates derived from 1976 to 2019.....	104
Table 29. Instruments and laboratory equipment ordered by application.....	110
Table 30. Normalized Ct values (RT-qPCR, Figure 7).....	111
Table 31. Published data sets used in this study (eLife 2017).....	118
Table 32. Normalized Ct values (RT-qPCR, Figure 47 b).....	124
Table 33. Cell counts (tryphan blue exclusion assay, Figure 47 c).....	125
Table 34. Published data sets used in this study.....	128
Table 35. Top five players for TPD in combination with multi-omics.....	129
Table 36. Published data sets used in Figure 50.....	130
Table 37. Comparison of correlation in this study with the literature.....	130
Table 38. Abbreviations.....	166

---

## List of items from publications

A summary of items and chapters which are excerpts from my first author publications can be found below. The corresponding pages numbers of the items can be found in Listed Contents (figures, tables), or in the Table of Contents (chapters). A detailed list of author contributions can be found on page VII.

Below: ‘x of this thesis’ corresponds to ‘y of the above listed publication’.

(\* joint first authorship)

## Figures

Minor parts of the figures were modified for visual purposes, i.e. to unify the journal specific style of published figures.

- S. Gressel \*, B. Schwalb \* *et al.* CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* 6, doi:10.7554/eLife.29736 (2017):
  - ‘Figure 2 b’ corresponds to ‘Figure 1 – figure supplement 1’
  - ‘Figure 14 b-c’ corresponds to ‘Figure 1-figure supplement 2 c’
  - ‘Figure 15’ corresponds to ‘Figure 1’
  - ‘Figure 16 a-d’ corresponds to ‘Figure 2’
  - ‘Figure 16 e’ corresponds to ‘Figure 2-figure supplement 1 c’
  - ‘Figure 17 a-b’ corresponds to ‘Figure 3 a-b’
  - ‘Figure 18’ corresponds to ‘Figure 4’
  - ‘Figure 19 a-c’ correspond to ‘Figure 5’
  - ‘Figure 19 d-f’ correspond to ‘Figure 4-figure supplement 1 d-f’
  - ‘Figure 20’ corresponds to ‘Figure 6 a-c’
  - ‘Figure 21 b, d, f’ corresponds to ‘Figure 7’
  - ‘Figure 21 c, e’ corresponds to ‘Figure 7 – figure supplement 1 a, c’
  - ‘Figure 30 d, left’ corresponds to ‘Figure 3 b’
  - ‘Figure 41 b, d’ corresponds to ‘Figure 1 – figure supplement 2 d-e’
  - ‘Figure 42’ corresponds to ‘Author response image 1’
- S. Gressel \*, K. Lidschreiber \* *et al.* Transient transcriptome sequencing: experimental protocol to monitor genome-wide RNA synthesis including enhancer transcription. *Methods Mol Biol*, *in revision*:
  - ‘Figure 6 a’ corresponds to ‘Figure 1’
- S. Gressel \*, B. Schwalb \* *et al.* The pause-initiation limit restricts transcription activation in human cells. *Nature Communications*, *in revision*:
  - ‘Figure 11 a, c’ corresponds to ‘Supplementary Figure 1’
  - ‘Figure 12 a’ corresponds to ‘Supplementary Figure 8 a’
  - ‘Figure 22’ corresponds to ‘Figure 1’
  - ‘Figure 23’ corresponds to ‘Figure 2’
  - ‘Figure 24’ corresponds to ‘Figure 3’
  - ‘Figure 25’ corresponds to ‘Supplementary Figure 4’

- 'Figure 26' corresponds to 'Supplementary Figure 7'
- 'Figure 27' corresponds to 'Figure 4'
- 'Figure 28' corresponds to 'Supplementary Figure 10'
- 'Figure 43' corresponds to 'Supplementary Figure 6'
- 'Figure 44' corresponds to 'Supplementary Figure 8 b-d'
- 'Figure 46' corresponds to 'Supplementary Figure 2'
- 'Figure 47' corresponds to 'Supplementary Figure 3'
- 'Figure 48' corresponds to 'Supplementary Figure 5'
- 'Figure 49' corresponds to 'Supplementary Figure 9'

## Tables

- S. Gressel \*, B. Schwalb \* *et al.* CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* 6, doi:10.7554/eLife.29736 (2017):
  - 'Table 31' corresponds to 'Supplementary file 1'
- S. Gressel \*, K. Lidschreiber \* *et al.* Transient transcriptome sequencing: experimental protocol to monitor genome-wide RNA synthesis including enhancer transcription. *Methods Mol Biol*, in revision:
  - 'Table 20' corresponds to 'Table 1'
- S. Gressel \*, B. Schwalb \* *et al.* The pause-initiation limit restricts transcription activation in human cells. *Nature Communications*, in revision:
  - 'Table 8' corresponds to 'Supplementary Table 3'
  - 'Table 23' corresponds to 'Supplementary Table 1'
  - 'Table 25' corresponds to 'Supplementary Table 2'
  - 'Table 34' corresponds to 'Supplementary Table 5'

## Chapters

- The following chapters are excerpts from S. Gressel \*, B. Schwalb \* *et al.* CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* 6, doi:10.7554/eLife.29736 (2017):
  - Abstract
  - Acknowledgements
  - Methods II.2.2.4, II.2.3.1 (first paragraph), II.2.3.4, II.2.3.5
  - Results III.2
  - Discussion IV.1.2 (two paragraphs)
  - Supplementary Information V.2.1
- The following chapters are excerpts from S. Gressel \*, K. Lidschreiber \* *et al.* Transient transcriptome sequencing: experimental protocol to monitor genome-wide RNA synthesis including enhancer transcription.

*Methods Mol Biol, in revision:*

- Introduction I.3.2.1 (two paragraphs)
- Methods II.2.2.2

- The following chapters are excerpts from S. Gressel \*, B. Schwab \* *et al.* The pause-initiation limit restricts transcription activation in human cells. *Nature Communications, in revision:*
  - Abstract
  - Acknowledgements
  - Methods II.2.2.4, II.2.4
  - Results III.3
  - Discussion IV.1.4, IV.1.5
  - Supplementary Information V.3.3

## Abbreviations

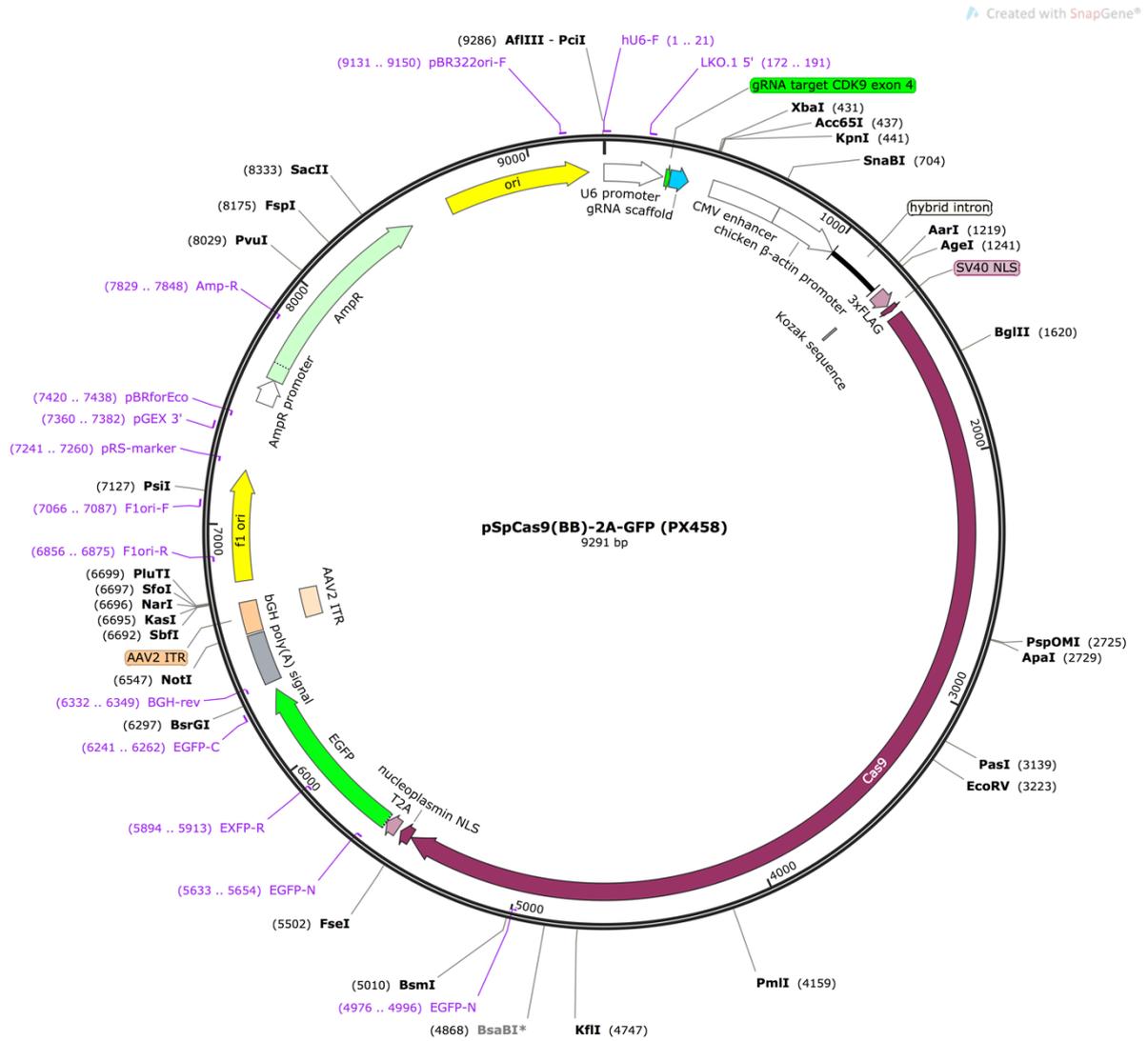
**Table 38. Abbreviations.**

-	minus strand
$\rho$ (rho)	Spearman correlation rho
[Å]	Ångstrom
[U]	Units
[v/v]	volume per volume
[w/v]	weight per volume
+	plus strand
1-NA-PP1	1-Naphthyl-PP1; 1-(1,1-dimethylethyl)-3-(1-naphthalenyl)-1H-pyrazolo[3,4-d]pyrimidin-4-amine
4-thio-UTP	4-thio-uridine triphosphate
4sU	4-thiouracil
A	alanine
ACTB	$\beta$ -actin
AID	auxin-inducible degron system
APS	ammonium persulfate
asRNA	antisense RNA
ATM	Ataxia Telangiectasia Mutated, cell cycle checkpoint kinase
ATP	adenosine triphosphate
BRD4	acetylated histone binding protein bromodomain containing 4
BrU	5-bromouridine 5'-triphosphate
BSA	bovine serum albumin
CBC	cap-binding complex
CDK7	cyclin-dependent kinase 7
CDK9 <sup>as</sup>	analog-sensitive cyclin-dependent kinase 9
cDNA	complementary DNA
ChIP-seq	chromatin immunoprecipitation sequencing
Chr	chromosome
conRNA	convergent RNA
CRBN	Cereblon (E3 protein ligase)
Ct	mean threshold cycle (RT-qPCR)
CTCF	CCCTC-binding factor (zinc finger protein)
CTD	RNA polymerase II C-terminal domain
Ctrl	control
cTU	capped transcription unit
$d$ [min]	pause duration
DMD	Dystrophin
DMF	N,N-Dimethylformamide
DMSO	dimethylsulfoxid
DNA	deoxyribonucleic acid
DNA-PK	DNA-dependent protein kinase
DNase	deoxyribonuclease
DSIF	DRB sensitivity inducing factor
dTAG	degradation tag system

DTT	DL-Dithiothreitol
E	amplification efficiency (RT-qPCR)
EC	elongation complex
ECL	enhanced luminol-based chemiluminescent
EDTA	ethylenediaminetetraacetic acid
EGTA	ethylene-glycol-bis( $\beta$ -aminoethyl ether)-N,N,N',N'-tetraacetic acid
enChIP	engineered DNA-binding molecule mediated chromatin immunoprecipitation
eRNA	enhancer RNA
F	phenylalanine
FBS	fetal bovine serum
FCP1	CTD phosphatase
GAPDH	glyceraldehyde 3-phosphate dehydrogenase
GenoSTAN	genomic state annotation segmentation algorithm
GRINL1A, GDOWN1	RNA polymerase subunit M
GRO-seq	global nuclear run-on and sequencing
hnRNA	heterogeneous nuclear RNA, intron-containing pre-mRNA
HS	heat shock
HSPA1A, HSP70	heat shock protein 70
$I$ [cell <sup>-1</sup> min <sup>-1</sup> ]	productive initiation frequency
IgG	Immunoglobulin G
IGV	Integrative Genomics Viewer
IP	immunoprecipitation
IVT	<i>in vitro</i> transcription
K562	ENCODE cell line, human
KD	knockdown
Length [bp]	base pair(s)
Length [kbp]	kilo base pair(s)
Length [nt]	nucleotide(s)
lincRNA	long intergenic noncoding RNA
MEF [kcal mol <sup>-1</sup> ]	minimum free energy
MMTV	murine mammary tumor virus
MNase	micrococcal nuclease
mNET-seq	mammalian native elongating transcript sequencing
MOPS	3-(N-morpholino)propanesulfonic acid
mRNA	protein-coding, messenger RNA
NaCl	sodium chloride
NaOAc	sodium acetate
NCBI	National Center for Biotechnology Information
ncRNA	long noncoding RNA
NELF	negative elongation factor
NET-seq	native elongating transcript sequencing
NF $\kappa$ B	nuclear factor $\kappa$ B
NRO	nuclear run-on assay
NTP	nucleotide triphosphate
o/n	overnight
$p$	number of polymerases

P-TEFb	positive transcription elongation factor b
pA	polyA site
PAF	polymerase associated factor
PAF1	Pol II associated factor 1
PAGE	polyacrylamide gel electrophoresis
Pausing	promoter-proximal pausing of Pol II
PB	pause button
PCR	polymerase chain reaction
PE	polyethylen
PE	pause element
PEC	paused elongation complex
PI	protease inhibitor cocktail
PIC	pre-initiation complex
PNK	polynucleotide kinase
Pol II	DNA-dependent RNA polymerase II
PP	polypropylene
PRO-seq	precision nuclear run-on and sequencing
PS	pause site
PVDF	polyvinylidene difluoride
PW	pause window
Raji B	Raji Burkitt's lymphoma
RNA	ribonucleic acid
RNase	ribonuclease
RPK	read counts per kilobase
RT	reverse transcription, reverse transcriptase
RT-qPCR	quantitative reverse transcription PCR
S	slope of standard curve (RT-qPCR)
sincRNA	short intergenic noncoding RNA
Sirt1	silent information regulator factor 2 related enzyme 1
SUPT6H, SPT6	transcription elongation factor 6
<i>tb</i>	transcribed bases
TEMED	tetramethylethylenediamine
Temperature [°C]	degree Celsius
TF	basal transcription factor
TFIIS	transcription elongation factor S
Time [h]	hour(s)
Time [min]	minute(s)
TIR1	transport Inhibitor Response 1 (E3 protein ligase)
TPD	targeted protein degradation
TRIM28, KAP-1	tripartite Motif Containing 28
TSS	transcription start site
TT-seq	transient transcriptome sequencing
TTS	transcription termination site
TU	transcription unit
uaRNA	upstream antisense RNA
$v$ [bp min <sup>-1</sup> ]	elongation velocity
XIST	X-inactive specific transcript

## Sequences and plasmid maps

*S. pyogenes* Cas9 plasmid

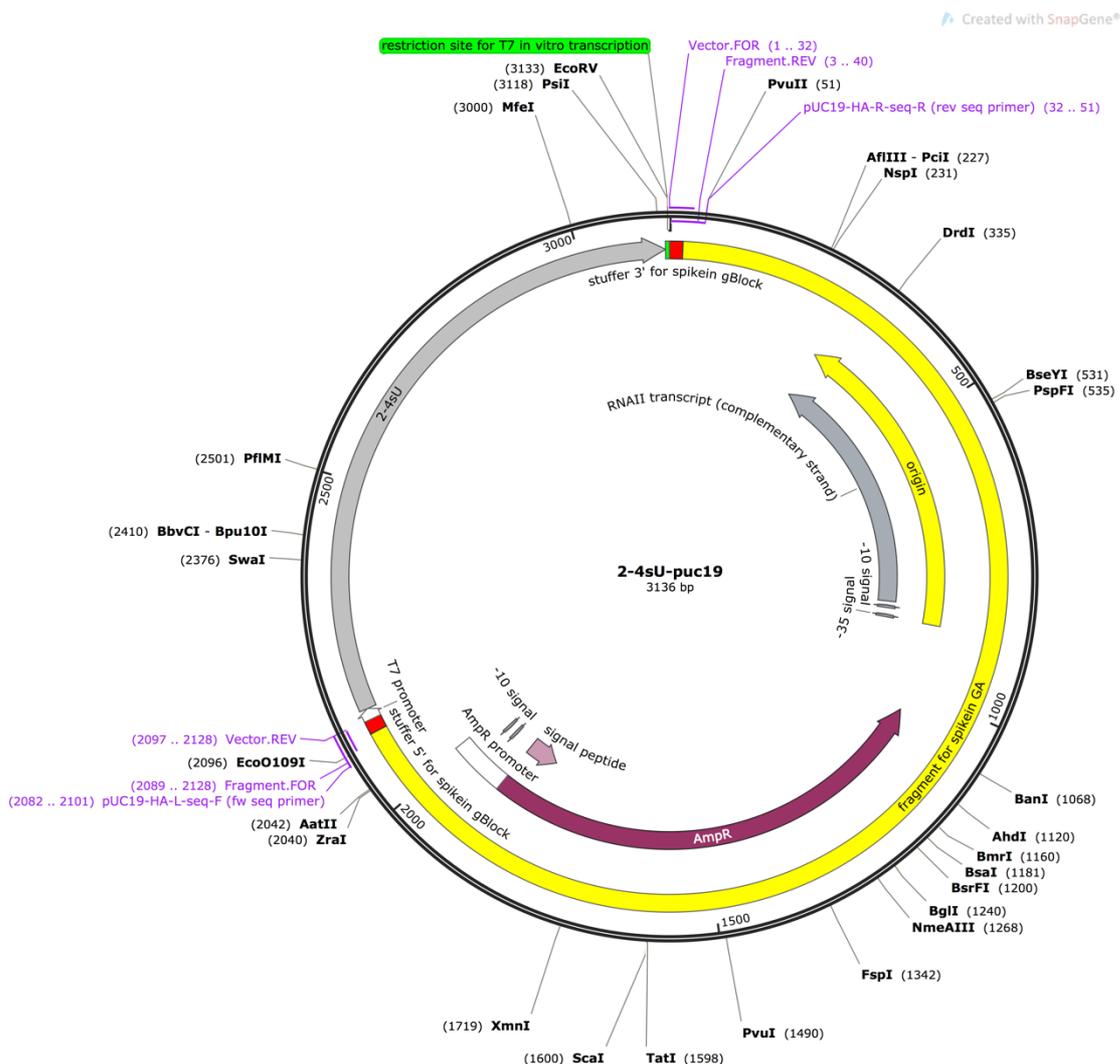
Plasmid map of pSpCas9(BB)-2A-GFP (PX458). Plasmid map was generated using SnapGene.

## Spike-ins (TT-seq)

General note: Spike-ins are derived from ERCC Spike-in Mix but are not polyadenylated (see II.2.2.1 and **Table 20** for details). At the 5'-end of the transcript 'GGG' (included in length) is added to facilitate in vitro transcription by T7 polymerase (highlighted by underline).

Sequence (5' to 3') of **spike-in 2** (length 985 nt) derived from ERCC-00043:

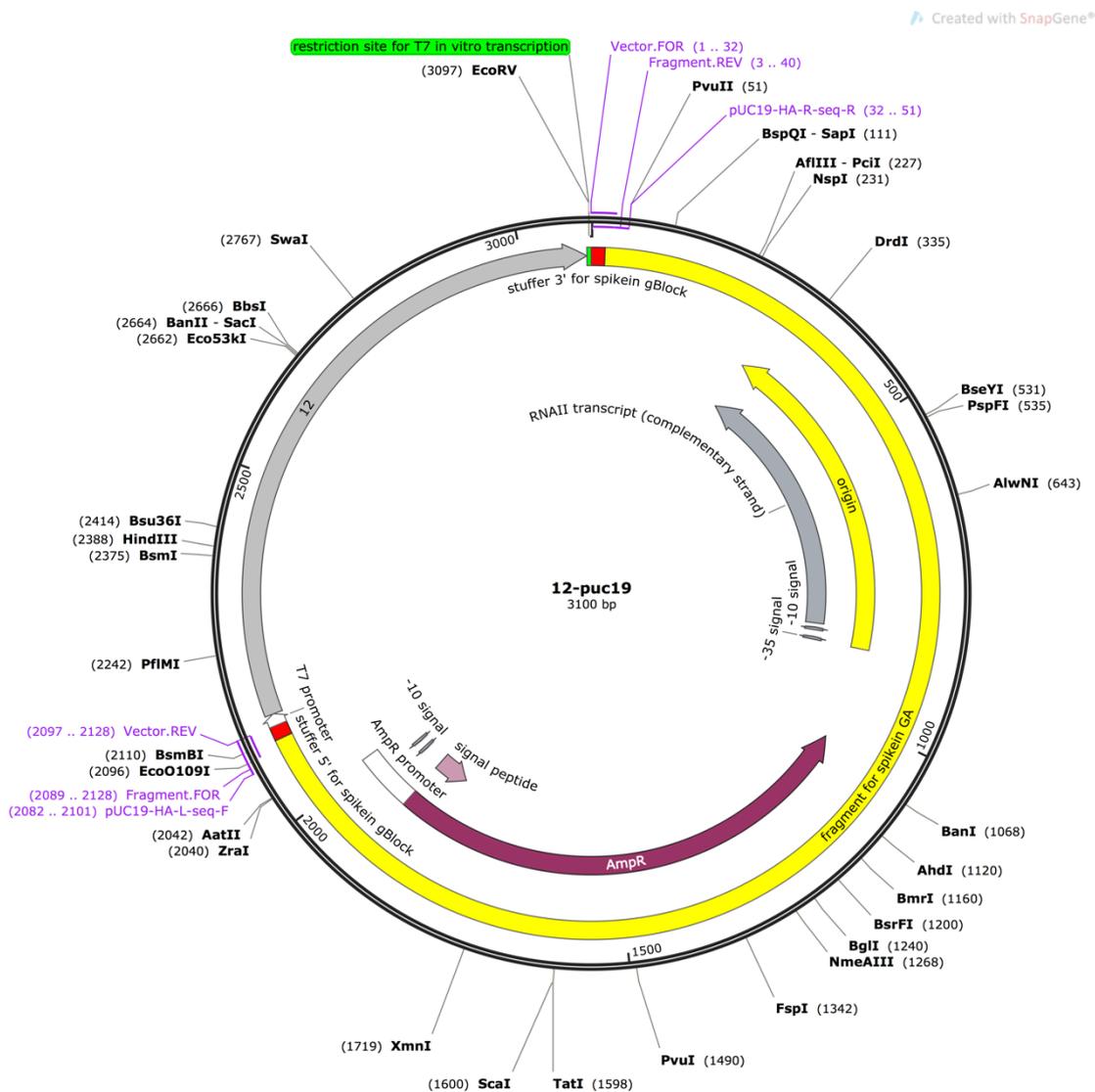
GGGTGCTTTAACAAGAGGAAATTGTGTTTTTGCCTTAATTAAGACCTAATTTAATAGTTAAACCATTA  
ACCTTAGTTGTTCCAAGGCATAATATAGAGAGTGAGATACAGGATGAGCTATTTTCAGGGAGTTATT  
CAGTATGCAGTTGCCAAGGCAGTTGCTGATTTAGATTTAGATGAAGATTTAAAGGTTGTTGTCTCT  
GTTAATGTCCCAGAGGTTCCAATAACCAATTTAAATAAAAAGAAAACCTCTTCCAATACTTCTATGCCT  
CAGCAAAGTTAGCTATAAACAGAGCTTTAAATGAATATCCTTCAAAAAGAGAAGGTAAGAAAAGAGA  
AATATAGAGCTTTGCATCCATTAGTTGGATTTAGGGATGTTAGATTGGAGTATCCTCCATATCTACA  
AATTGCTTTGGATGTCCCAACTATGGAGAATTTGGAATTTTTGTTACAAACAATTCCAAATAGCGAC  
CACATCATCTTAGAGGCTGGAACACCACTAATTA AAAAAGTTTGGTTTAGAGGTTATTGAAATAATG  
AGAGAATATTTTATGATGGCTTTATTGTTGCTGATTTAAAAACCTTAGACACTGGAAGGGTTGAGGTA  
AGATTGGCATTGGAAGCAACAGCTAATGCAGTGGCAATAAGTGGAGTAGCACCAAAATCAACAATA  
ATTAAGCTATCCACGAATGTCAAAAATGTGGTTTAATCAGCTATTTGGATATGATGAACGTCTCTG  
AACCTCAAAAATTATATGATTCATTA AAAAATTAAGCCAGATGTTGTTATCTTGCATAGAGGGATTGA  
TGAGGAGACATTTGGAATTA AAAAAGGAATGAAATTTAAGGAAAACCTGCTTATTAGCAATTGCTGG  
AGGAGTTGGTGTGGAGAATGTTGAAGAGCTTTTAAAAGAATATCAAATATTAATCGTTGGTAGAGC  
AATTACAAAATCAAAAAGACCCAGGAAGAGTAATTAGGATTTTATAAACAAGATGG



Plasmid map of spike-in 2 pUC19. Plasmid map was generated using SnapGene.

Sequence (5' to 3') of **spike-in 12** (length 949 nt) derived from ERCC-00170:

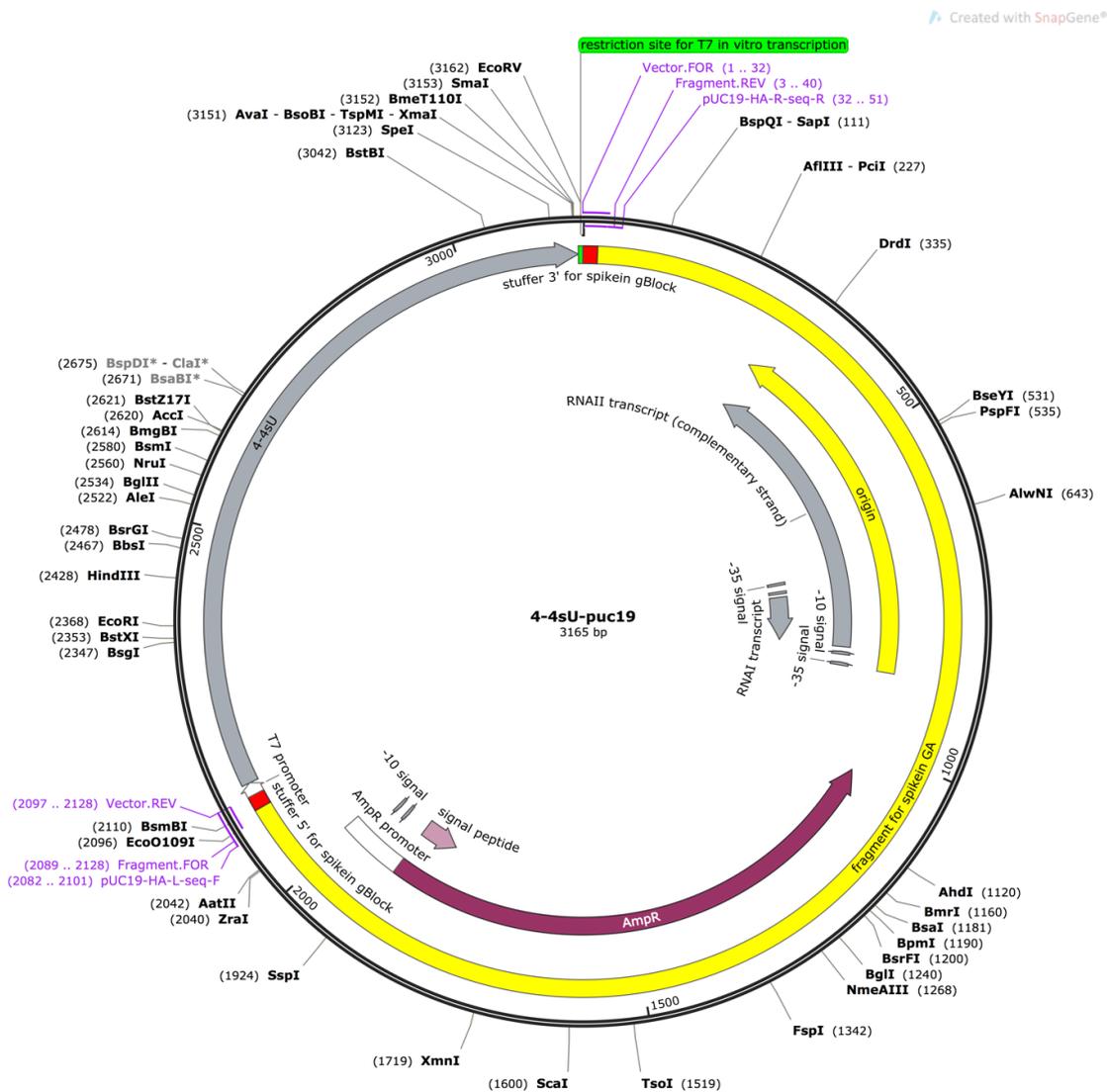
GGGGCACAAGTTGCTGAAGTTGCGAGAGGGGCGATAAGTGAGGCAGACAGGCATAATATAAGAGG  
 GGAGAGAATTAGCGTAGATACTCTTCCAATAGTTGGTGAAGAAAATTTATATGAGGCTGTTAAAGC  
 TGTAGCAACTCTTCCACGAGTAGGAATTTTGTAGTTTGTAGCTGGCTCTTTAATGGGAGGGAAGATAAC  
 TGAAGCAGTTAAAGAATTAAGGAAAAGACTGGCATTCCCCTGATAAGCTTAAAGATGTTTGGCTC  
 TGTTCCCTAAGGTTGCTGATTTGGTTGTTGGAGACCCATTGCAGGCAGGGGTTTGTAGCTGTTATGGC  
 TATTGCTGAAACAGCAAAATTTGATATAAATAAGGTTAAAGGTAGGGTGTCTATAAAGATAATTTAA  
 TAATTTTTGATGAAACCGAAGCGTTAGCTTTGGGTTATGAACTCCATGATTTTCATTTAATTTTTT  
 CCTATTAATTTTCTCCTAAAAAGTTTCTTTAACATAAATAAGGTTAAAGGGAGAGCTCTATGATTGT  
 CTTCAAAAATACAAAGATTATTGATGTATATACTGGAGAGGTTGTTAAAGGAAATGTTGCAGTTGA  
 GAGGGATAAAATATCCTTTGTGGATTTAAATGATGAAATTGATAAGATAATTGAAAAAATAAAGGA  
 GGATGTTAAAGTTATTGACTTAAAAGGAAAATATTTATCTCCAACATTTATAGATGGGCATATACAT  
 ATAGAATCTTCCCATCTCATCCCATCAGAGTTTGAGAAATTTGTATTA AAAAGCGGAGTTAGCAAAG  
 TAGTTATAGACCCGCATGAAATAGCAAATATTGCTGGAAAAGAAGGAATTTTGTTTATGTTGAATG  
 ATGCCAAAATTTTAGATGTCTATGTTATGCTTCCTTCTGTGTTCCAGCTACAAACTTAGAAACAAG  
 TGGAGCTGAGATTACAGCAGA



Plasmid map of 12-pUC19. Plasmid map was generated using SnapGene.

Sequence (5' to 3') of **spike-in 4** (length 1,014 nt) derived from ERCC-00136:

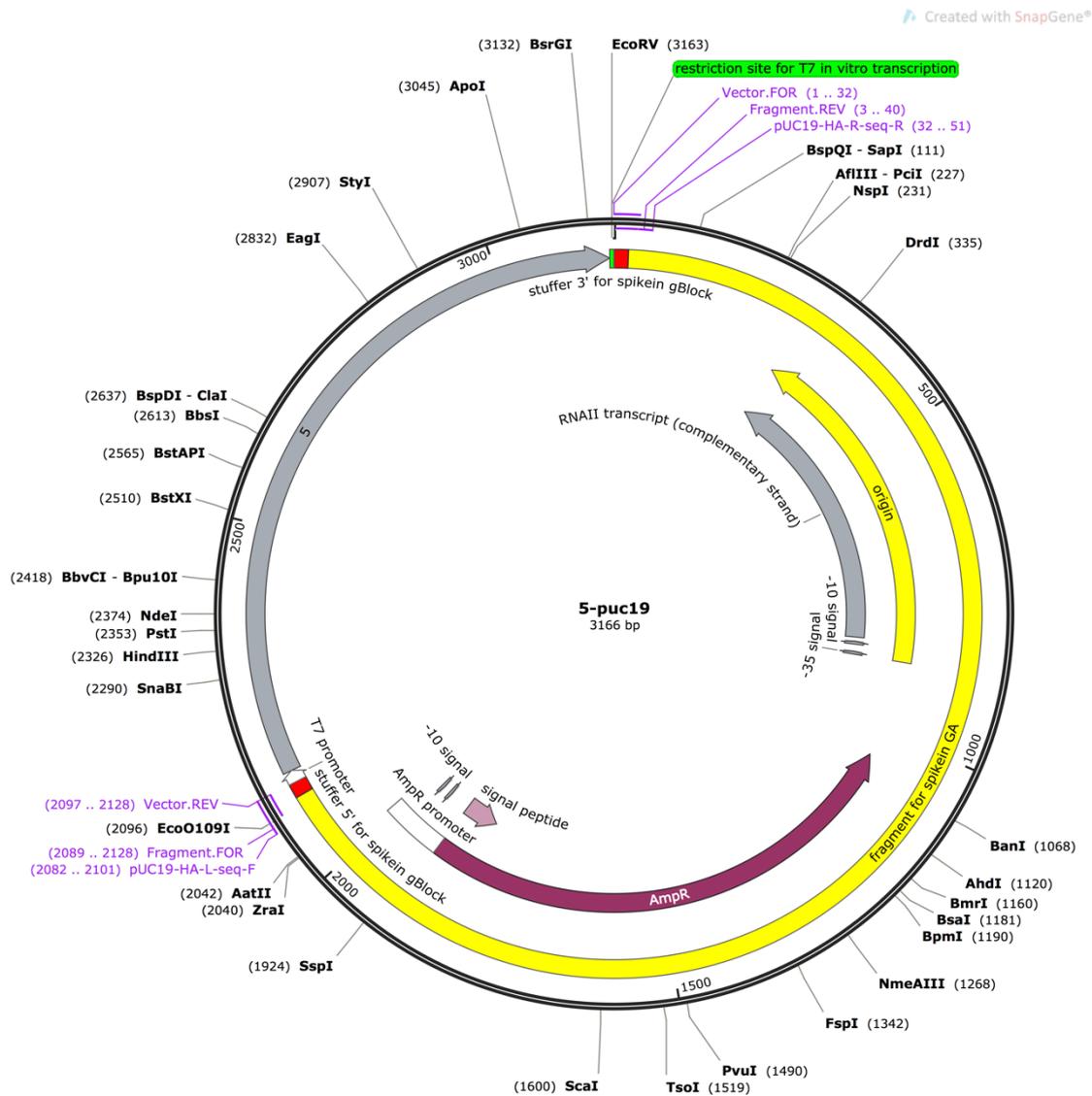
GGGTTTTCGACGTTTGAAGGAGGGTTTTAAGTAATGATCGAGATTGAAAAACCAAAAATCGAAACG  
 GTTGAAATCAGCGACGATGCCGAATTTGGTAAGTTTGTCTAGAGCCACTTGAGCGTGGATATGGT  
 ACAACTCTGGGTAACCTTACGTCGTATCCTCTTATCCTCACTCCCTGGTGCCGCTGTAACATCAA  
 TCCAGATAGATGGTGTACTGCACGAATTCTCGACAATTGAAGGCGTTGTGGAAGATGTTACAACGA  
 TTATCTTACACATTAAGCTTGCATTGAAAATCTACTCTGATGAAGAGAAGACGCTAGAAAATTGA  
 TGTACAGGGTGAAGGAAGTAAACGGCAGCTGATATTACACACGATAGTGATGTAGAGATCTTAAA  
 TCCTGATCTTCATATCGCGACTCTTGGTGAGAATGCGAGTTTCCGAGTTCGCCTTACTGCTCAAAGA  
 GGACGTGGGTATACGCCTGCTGACGCAACAAGAGAGGGGATCAGCCAATCGGCGTGATTCCGATC  
 GATTCTATCTATACGCCAGTTTCCCGTGTATCTTATCAGGTAGAGAACACTCGTGTAGGCCAAGTTG  
 CAAACTATGATAAACTTACACTTGATGTTTGGACTGATGGAAGCACTGGACCGAAAGAAGCAATTG  
 CGTTGGTTCAAAGATTTTAACTGAACACCTTAATATATTCGCTGGTTTAACTGACGAAGCTCAACA  
 TGCTGAAATCATGTTGAAGAAGAAGAAGATCAAAAAGAGAAAGTTCTTCAAATGACAATTGAAGA  
 ATTGGATCTTTCTGTTCTTACAACCTGCTTAAAGCGTGCGGGTATTAACACGGTTC AAGAGCTT  
 GCGAACAAGACGGAAGAAGATATGATGAAAGTTCGAAATCTAGGACGCAAATCACTTGAAGAAGTG  
 AAAGCGAGACTAGAAGAAGTGGACTCGGACTTCGCAAAGACGATTGACTAGTTTCCCTTGTGAAC  
 TAGGATTTTCCCGGTAC



Plasmid map of spike-in 4 pUC19. Plasmid map was generated using SnapGene.

Sequence (5' to 3') of **spike-in 5** (length 1,015 nt) derived from ERCC-00145:

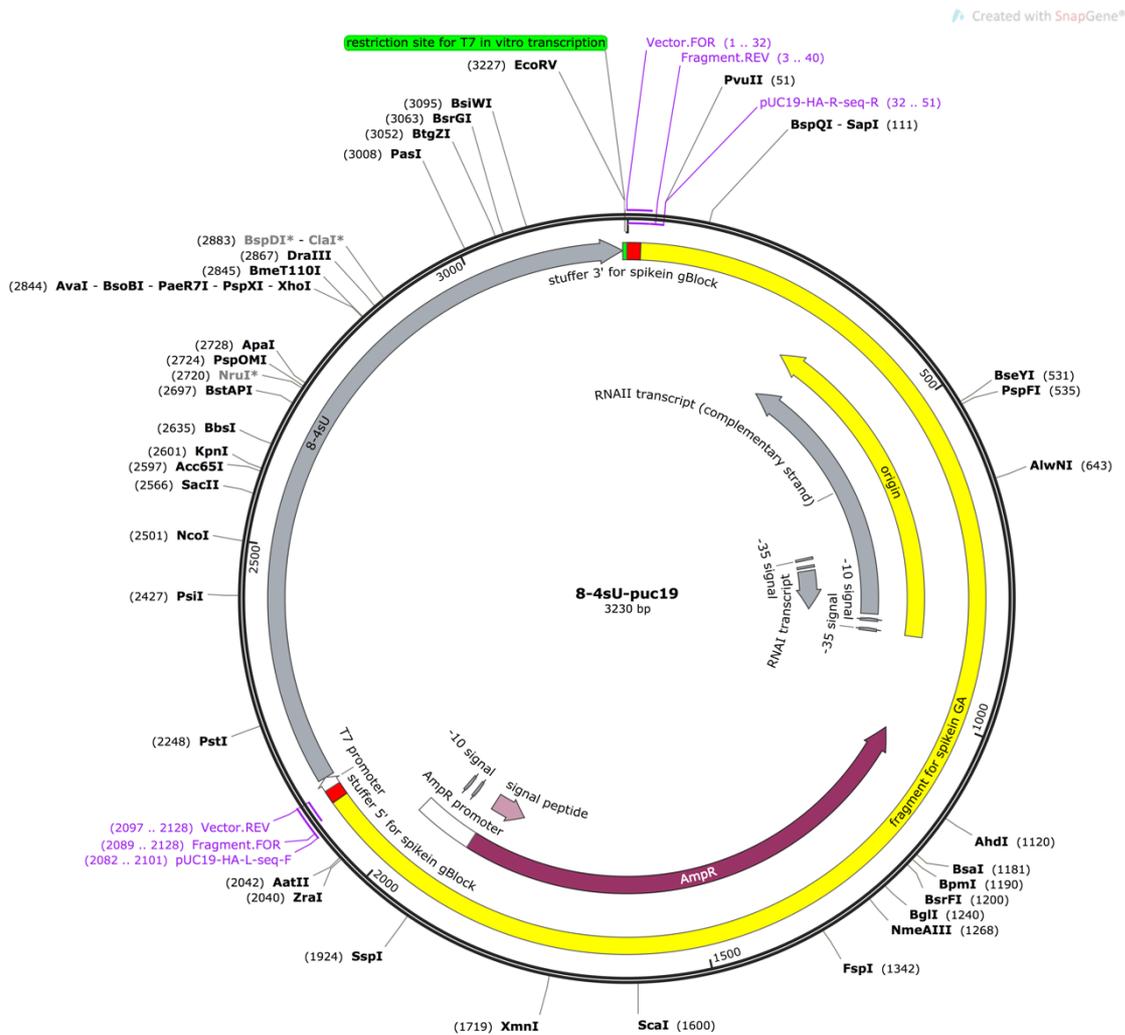
GGGACTGTCCTTTCATCCATAAGCGGAGAAAGAGGGAATGACATTGTTCTTACACGGCACAAGCAG  
 AAAAAATCAACATGGTCATTTAGAAATCGGAGGTGTGGATGCTCTCTATTTAGCGGAGAAATATGG  
 TACACCTCTTTACGTATATGATGTGGCTTTAATACGTGAGCGTGCTAAAAGCTTTAAGCAGCGCTTT  
 ATTTCTGCAGGGCTGAAAGCACAGGTGGCATATGCGAGCAAAGCATTCTCATCAGTCGCAATGATT  
 CAGCTCGCTGAGGAAGAGGGACTTTCTTTAGATGTGCGATCCGGAGGAGAGCTATATACGGCTGTT  
 GCAGCAGGCTTTCCGGCAGAACGCATCCACTTTCATGGAAACAATAAGAGCAGGGAAGAAGTGGC  
 ATGGCGCTTGAGCACCGCATCGGCTGCATTGTGGTGGATAATTTCTATGAAATCGCGCTTCTTGAA  
 GACCTATGTAAGAAACGGGTCACTCCATCGATGTTCTTCTTCGGATCACGCCCGGAGTAGAAGCG  
 CATA CGCATGACTACATTACAACGGGCCAGGAAGATTCAAAGTTTGGTTTCGATCTTCATAACGGA  
 CAAACTGAACGGGCCATTGAACAAGTATTACAATCGGAACACATTCAGCTGCTGGGTGTCCATTGC  
 CATATCGGCTCGAAATCTTTGATACGGCCGGTTTTGTGTTAGCAGCGGAAAAAATCTTCAAAAAA  
 CTAGACGAATGGAGAGATTCATATTCATTTGTATCCAAGGTGCTGAATCTTGGAGGAGGTTTCGGC  
 ATTCGTTATACGGAAGATGATGAACCGCTTCATGCCACTGAATACGTTGAAAAAATTATCGAAGCT  
 GTGAAAGAAAATGCTTCCCGTTACGGTTTTGACATTCGGGAAATTTGGATCGAACCGGGCCGTTCT  
 CTCGTGGGAGACGCAGGCACAACCTTTTATACGGTTGGCTCTCAAAAAGAAGTGGATAAGCTGTAC  
 AATCGTTTCATCATTCGGCGTGCG



Plasmid map of spike-in 5 pUC19. Plasmid map was generated using SnapGene.

Sequence (5' to 3') of **spike-in 8** (length 1,079 nt) derived from ERCC-00092:

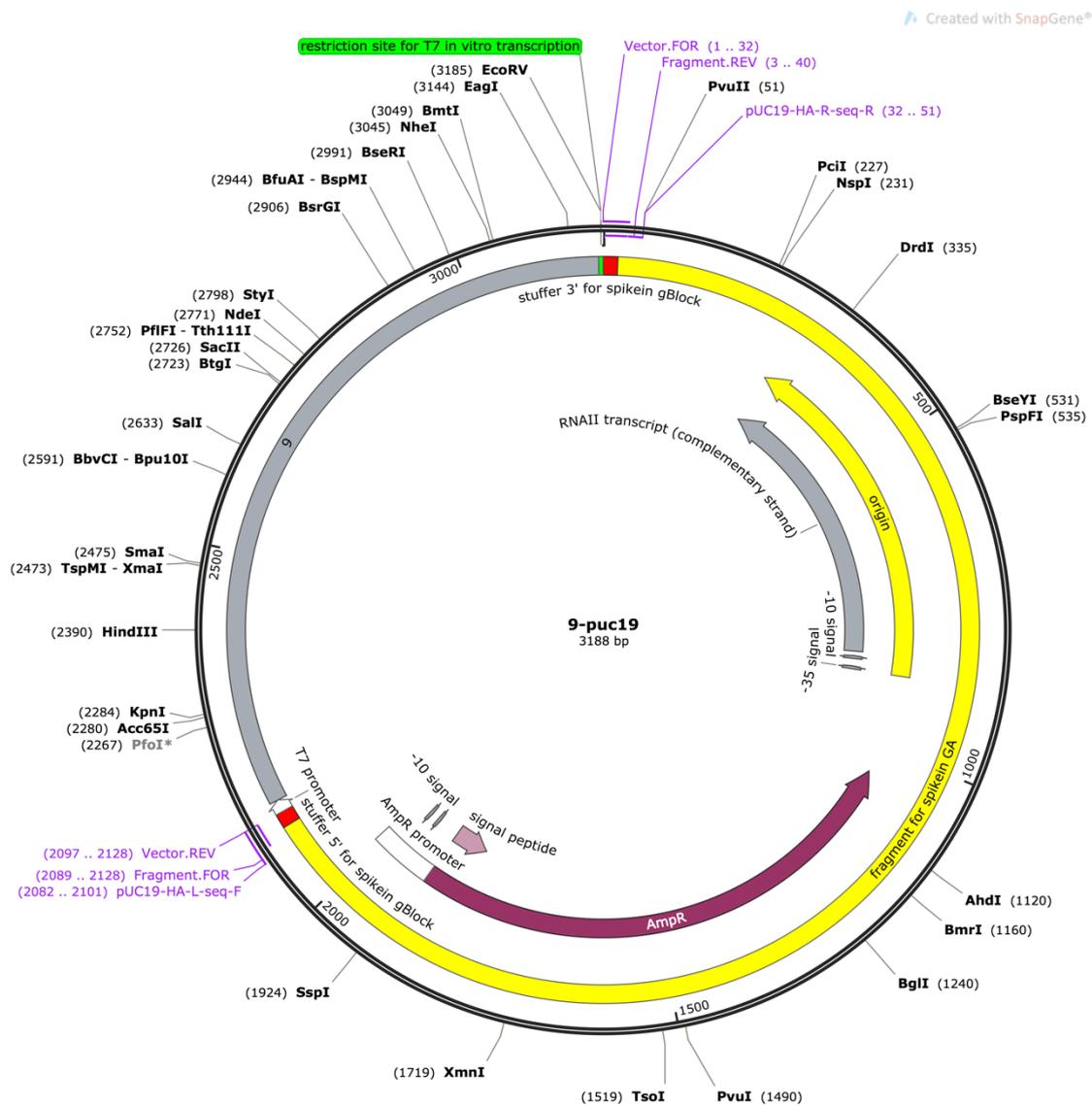
GGGGATGTCCTTGGACGGGGTGGCGCAGTATTACTGCAAGAGAGCGGACAGATTAGTGTGTTGGAG  
 CCGACACATCAAAGGTTTCGTCCGGGGACCGATCTGCAGCCTACGGGACATTTATCCGTA AAAAGCAT  
 GGCGCTGTTTCGTA CTTATCGGAGGCCAGGTATCGTCGCGGCGAGTCTCCCGACGACGGAGATGG  
 GCGTACTATCTGGGCCGTCTCGTACTCTGTTACTTGGCACAGATGCGAGCCCTCGTAATGTGCATC  
 AGCTAAGGGCGATATTATAATGCGACGTTTGTACGGATTTCGTTACTAACGTGTTGGACGCTAGTGG  
 AATATGTGTTCGTTGGTTAGCCTACCCATGGCTTTCGCGGCGACACATGCTTAGACTCTTTCAAACCT  
 TCGGTGAAGTTCACTCAAGCCGCGGAGCGCCGTCGTAATTCCTAGGGATGGCGGTACCCGTGCCC  
 GTCCGATTTCGTAGCAACCTGCATCACGATTTTGTCTTCGGGCGACTTATCAGATACGGTAATGTAAA  
 TACCTGGCATTGTTGGGCACTTCTTTCGCTTTAAGCGGGAAAGATCGCGAGGGCCCCGCTATTTGCGATA  
 CTTCCCATGTCGGTGCCGTGCGCTCTATGTACTCGGAGACGTTAATGCAGAGGCTAAGGACAATTT  
 ACCATGACTCGGTAATCCGTTTCGTCAAGCAGGTAGCTCGAGTCTCCCCACGGACACGTAAGTGGGTT  
 TGTAACGATCGATACCGAGTCTTTTGTCTAGTAGAACCAACCAACCATTAAGGAGTTCCTACTAGCAC  
 ATCTTTGCGACCCGATCGTCCGTGTGTCGCGTAATACTTTTGTATGACGAGACATACGCTCAAGCC  
 CTGGGTAGCTAGTCGCGGAGGCACGTTACCGCGCACAAACCCCTATTCGTTTACATGTACATCGCATC  
 TGAGGTAGTACACTTCCGGCGTACGTGAGTATTTGCGCGTAATAAGCGCGTGTTTAGCTGATCCCC  
 TCTCGTATCGAGGTTAAGGCAGATTAGTGCCAGTAATTGCGTTTTTTTGTCTGTTGTCGAGAACG  
 CGATTTGCTCCGAAAGC



Plasmid map of spike-in 8 pUC19. Plasmid map was generated using SnapGene.

Sequence (5' to 3') of **spike-in 9** (length 1,037 nt) derived from ERCC-00002:

GGGCCAGATTACTTCCATTTCCGCCCAAGCTGCTCACAGTATACGGGCGTCGGCATCCAGACCGTCTG  
 GCTGATCGTGGTTTTACTAGGCTAGACTAGCGTACGAGCACTATGGTCAGTAATTCCTGGAGGAAT  
 AGGTACCAAGAAAAAACGAACCTTTGGGTTCCAGAGCTGTACGGTCGCACTGAACTCGGATAGGT  
 CTCAGAAAAACGAAATATAGGCTTACGGTAGGTCCGAATGGCACAAAGCTTGTTCGGTTAGCTGGC  
 ATAAGATTCCATGCCTAGATGTGATACACGTTTCTGGAAACTGCCTCGTCATGCGACTGTTCCCCGG  
 GGTCAGGGCCGCTGGTATTTGCTGTAAGAGGGGGCGTTGAGTCCGTCGGACTTCACTGCCCCCTTT  
 CAGCCTTTTGGGTCTGTATCCCAATTCTCAGAGGTCCCGCCGTACGCTGAGGACCACCTGAAACGG  
 GCATCGTCGCTCTTCGTTGTTTCGTCGACTTCTAGTGTGGAGACGAATTGCCAGAATTATTAACCTGCG  
 CAGTTAGGGCAGCGTCTGAGGAAGTTTGCTGCGGTTTCGCCTTGACCGCGGGAAGGAGACATAACG  
 ATAGCGACTCTGTCTCAGGGGATCTGCATATGTTTGCAGCATACTTTAGGTGGGCCTTGCGTTCCTT  
 CCGCAGTCAAACCCGCGCAATTATCCCCGTCTGATTTACTGGACTCGCAACGTGGGTCCATCAGTT  
 GTCCGTATACCAAGACGTCTAAGGGCGGTGTACACCCTTTTGAGCAATGATTGCACAACCTGCGAT  
 CACCTTATACAGAATTATCAATCAAGCTCCCCGAGGAGCGGACTTGTAAGGACCGCCGCTTTCGCTC  
 GGGTCTGCGGGTTATAGCTTTTTCAGTCTCGACGGGCTAGCACACATCTGGTTGACTAGGCGCATAG  
 TCGCCATTACAGATTTGCTCGGCAATCAGTACTGGTAGGCGTTAGACCCCGTACTCGTGGCTGA  
 ACGGCCGTACAACCTCGACAGCCGGTGCTTGCCTTTTACC



Plasmid map of spike-in 9 pUC19. Plasmid map was generated using SnapGene.