Timing Polymerase Pausing with TV-PRO-seq

By

Jie Zhang

A thesis submitted for the degree of Doctor of Philosophy

University of Warwick, School of Life Sciences

Septermber, 2019

Table of Contents

List of Tables

List of Figures

Acknowledgments

Here I want to thank:

My parents, grandparents and other family members for funding my Ph.D. and supporting my life and work.

Dr. Daniel Hebenstreit for supervision and funding of my Ph.D.

Dr. Massimo Cavallaro for helping me with the mathematics and modelling work.

Dr. Jose Gutierrez-Marcos and Dr. Yin Chen for advising.

Dr. Michael Huemer, Dr. Mark Walsh, Dr. Nathan Archer and Steven Servín González for helping me resolve the problems about programming and experiment.

Other members in DH group and my friends.

Declarations

I hereby declare that this theses entitles ‘Timing Polymerase Pausing with TV-PRO-seq’ is an original work and has not been submitted for a degree or diploma or other qualification at any university.

The mathematics and modelling work of section 2.2.5 is finished under the help of Dr. Massimo Cavallaro.

The work have been preprint on bioRxiv:

Zhang, J., Cavallaro, M. & Hebenstreit, D. Timing Polymerase Pausing with TV-PRO-seq. bioRxiv, 461442 (2018).

Abstract

Transcription of many genes in metazoans is subject to polymerase pausing, which corresponds to the transient arrest of transcriptionally engaged polymerase. It occurs mainly at promoter proximal regions and is not well understood. In particular, a genome-wide measurement of pausing times at high resolution has been lacking.

I present in this thesis an extension of PRO-seq, time variant PRO-seq (TV-PRO-seq), that allowed researchers to estimate genome-wide pausing times at single base resolution. Its application to human cells reveals that promoter proximal pausing is surprisingly short compared to other regions and displays an intricate pattern. Furthermore, I found precisely conserved pausing profiles at tRNA and rRNA genes and identified DNA motifs associated with pausing time. I also found histone acetylation repressor H3K36me3 can cause long polymerase pausing. Finally, our result suggest that regulation of elongation is based on joint effect of multiple position rather than single position.

Abbreviation

APM	Accurate pausing motif
APS	Ammonium persulfate
bp	base pair
BrU	5-Bromouridine 5′-triphosphate
ChIP	Chromatin immunoprecipitation
ChIP-seq	Chromatin immunoprecipitation and DNA sequencing
chr	chromosome
chRO-seq	chromatin run-on and sequencing
coPRO-seq	coordinated precision run-on and sequencing
CTD	C-terminal domain
CV	Coefficient of Variation
DEPC	Diethyl pyrocarbonate
DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DSIF	DRB sensitivity inducing factor
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid

EGTA	ethylene glycol tetraacetic acid
FISH	Precise run-on sequencing
FP	Flavopiridol
FRAP	Fluorescence recovery after photobleaching
GRO-seq	Global run-on sequencing
ICR	Internal control region
IMDM	Iscove's Modified Dulbecco's Medium
LLPT	Liquid Liquid Phase Transition
lncRNA	Long non-coding RNAs
LOESS	locally estimated scatterplot smoothing
LPI	Local pausing index
mRNA	Messenger RNA
ncRNA	non-coding RNA
NELF	negative elongation factor
NET-seq	Native elongation transcript sequencing
nt	nucleotide
NTP	Nucleoside triphosphate
P-TEFb	positive transcription elongation factor
PAGE	Polyacrylamide
PCR	Polymerase chain reaction

PI	Pausing index
Pol I	RNA Polymerase I
Pol II	RNA Polymerase II
Pol III	RNA Polymerase III
POLRMT	RNA Polymerase Mitochondrial
PPM	position probability matrices
PPR	Promoter proximal region
PRO-seq	Precise run-on sequencing
RNA	Ribonucleic acid
RppH	RNA 5´ Pyrophosphohydrolase
rRNA	ribosomal RNA
RT	Reverse transcription
scRNA-seq	short capped RNA sequencing
SD	Standard Deviation
TEMED	Tetramethylethylenediamine
TES	Transcription end site
TF	Transcription factor
TFIIS	Transcription elongation factor IIS
TR	tandem repeat
tRNA	Transfer RNA

Trp	triptolide
TSS	Transcription start site
TV-PRO-seq	Time variant pecise run-on sequencing
XPB	xeroderma pigmentosum type B

Chapter 1 Introduction and background

1.1 Motivation for study

Enrichment of RNA polymerase II has been found in the promoter proximal region of highly regulated genes of metazoans[1, 2]. This enrichment has been suggested to be caused by the longer residence time of polymerases in this region[2-4]. However, several recent studies proposed that polymerases tend to abort transcription before entering productive elongation[5-7]. This phenomenon will also lead to polymerase enrichment in the promoter proximal region. To tackle this problem, I designed TV-PRO-seq, the first method which can estimate pausing times of polymerases at specific genome locations genome-widely. As a result, TV-PRO-seq is minimally influenced by the turnover rate of polymerase (See chapter 3.1); it can be used to test if the pausing time of polymerases in promoter proximal regions is indeed longer than other region of genes.

Because pausing time calculated by TV-PRO-seq is based on the growth rate of reads of pausing sites, it is not influenced by the gene expression level. This advantage allows analysis for pausing sites across meta-genes to, for example, analyse elemental pausing. It also makes TV-PRO-seq data well suited to integration with other sequencing data such as ChIP-seq datasets for genome-wide analysis.

As TV-PRO-seq is not limiterd to Pol II, it also can reveal pausing patterns of Pol I and Pol III transcribed genes.

1.2 Gene expression and transcription

DNA is the macromolecule that stores the genetic information of organisms. It is composed of monomeric units, nucleotides. Each nucleotide contains one of four kinds of nitrogen-containing nucleobases: A, T, C or G. DNA determines the phenotype of organisms indirectly. It is stable and identical in most cells during development and differentiation of organisms. While cells in the same organism share the genetic information, they perform different tasks, which is largely determined by the proteins in the cells. The process finally resulting in the generation protein is ultimately determined by the information stored in DNA, and is called gene expression.

Gene expression consists mainly of two parts, transcription and translation[8] (Figure 1.1). As an analogy, consider a cell as a computer. DNA will be the code stored in the hard drive, and proteins are the image we can see on the monitor. Code itself does not have a function, but it decides the reaction of software towards input. The process of the code running, and output images, is gene expression. In the same way that the image on screen corresponds to the input, the expression of genes exhibit spatial and temporal differences according to internal and external signals. The regulation of gene expression orchestrates functional specification in different cell types and is thus essential for development, differentiation, stress response, and adaptability in organisms.

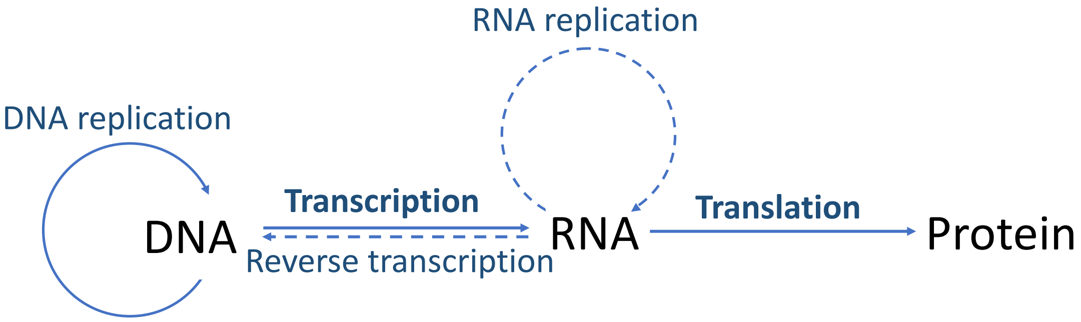
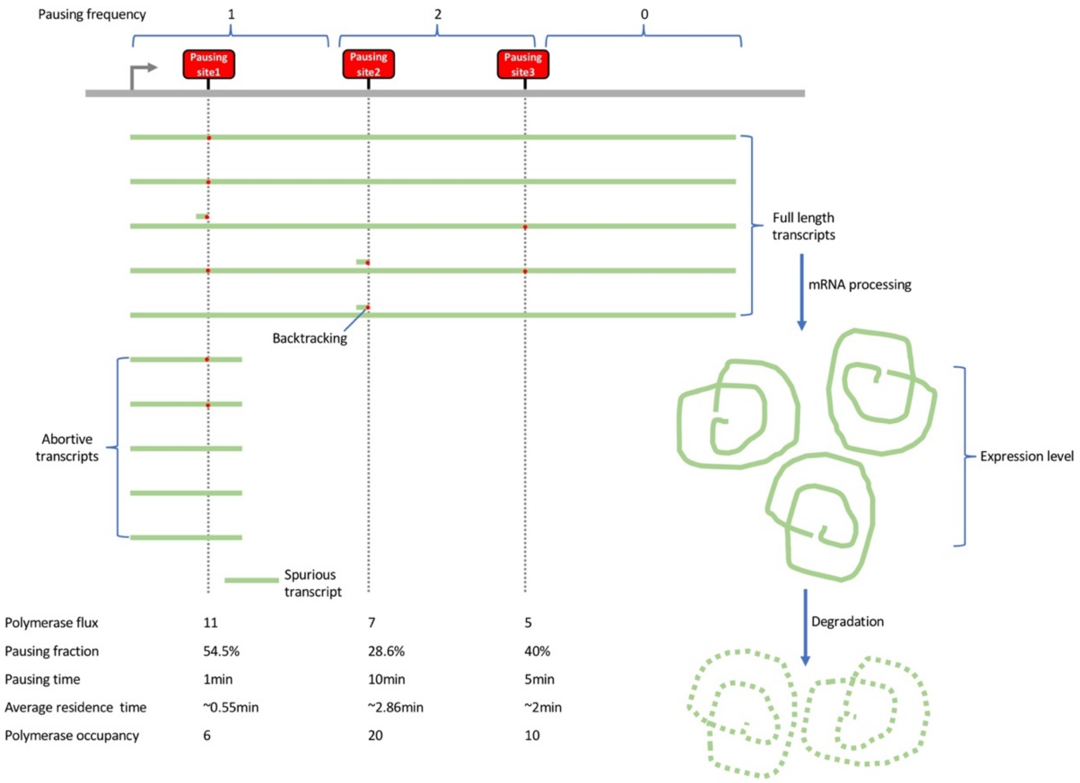


Figure 1.1 Central 'dogma' of molecular biology

1.3 Polymerase pausing

RNA polymerases are the key players of transcription. Three different types of RNA polymerases have roles in the nucleus of eukaryote cells: Pol I, Pol II and Pol III (RNA polymerase I, II and III). Pol I, Pol II and Pol III transcribe different classes of genes. Pol I transcribes 18S, 5.8S and 28S rRNA[9] (ribosomal RNA); Pol III mainly transcribes short structured RNAs, including 5S rRNA, tRNA[10] (transfer RNA). Pol II is highly researched, as it transcribes mRNA (message RNA), the template for proteins[11]. During transcription, RNA polymerase binds to template DNA; nascent RNA is generated according to the sequence of the DNA template as the RNA polymerase moves forward. The speed of RNA polymerase is not uniform. RNA polymerases have been found enriched in particular positions of genes and proposed to stay longer on these compared to other positions[2, 4, 12-15]. This phenomenon, the transient stop of polymerase in certain genome locations during elongation, has been termed ‘polymerase pausing’. As pausing is a controlled process[16-19], its dynamics are expected to be relatively complex. In this thesis, I introduce various terms for a better understanding of polymerase pausing (Figure 1.2, detailed explanation of terms see 1.2.1-1.2.4).



Polymerase flux	11	7	5
Pausing fraction	54.5%	28.6%	40%
Pausing time	1min	10min	5min
Average residence time	~0.55min	~2.86min	~2min
Polymerase occupancy	6	20	10

Figure 1.2 Dissecting polymerase pausing with different parameters

Three examples of pausing sites have been posited. The grey line represents template DNA, and green lines represent RNAs. The red points on the green lines mean that pausing has happened at the corresponding pausing sites during transcription. Polymerase flux refers to the number of polymerase that pass each position. Genes with a higher expression level have a higher polymerase flux. Abortive transcription will increase the polymerase flux of positions in the promoter proximal region without influencing the pausing time of each paused polymerase. Backtracking allows polymerases to pass the same positions twice, which also increases the polymerase flux of these positions. As shown, we propose that not all the polymerases necessarily pause at all pausing sites while they transcribe. The percentage of paused polymerases among the polymerase flux is termed the pausing fraction. The sum of polymerase *residence time of paused and non-paused polymerase divided by the polymerase flux yields the average residence time. Because non-paused polymerase contributes little to the sum of polymerase residence time, the average residence time can be approximated pausing time times pausing fraction.*

1.3.1 Polymerase occupancy

Polymerase occupancy corresponds to the enrichment level of polymerase on a specific genome position, which can be measured by ChIP-seq. The experimental procedure of ChIP-seq starts with fixation of the polymerases on to chromatin and then breaks the chromatin into small fragments. Antibodies against polymerase are then used for immunoprecipitation, which enriches DNA fragments bound to polymerase, which in turn allows sequencing of the DNA. Reads of ChIP-seq are aligned to a reference genome. The average number of aligned reads on specific genome positions is used as a measure of coverage[20, 21]. Various sequencing methods have been invented to discover the polymerase occupancy on specific DNA strands and/or higher resolution, such as GRO-seq (Global run-on sequencing)[22], scRNA-seq (short capped RNA sequencing)[23], nascent RNA sequencing[24], NET-seq (Native elongation transcript sequencing)[25] and PRO-seq (Precise run-on sequencing)[26].

1.3.2 Polymerase flux and average residence time

The number of polymerases that move past a given position in a unit of time is defined as ‘polymerase flux’. Polymerase flux is positively correlated to polymerase occupancy. Higher polymerase flux means more polymerases pass a genome position in a certain time period. As shown in Figure 1.3 A and B, genes with higher expression level will have higher polymerase flux, thus have higher polymerase occupancy.

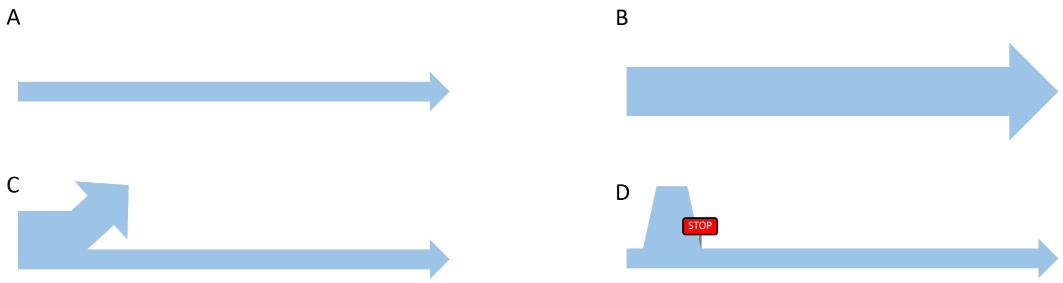


Figure 1.3 Diagrams of polymerase occupancy difference of different statement

A. *This diagram shows the polymerase occupancy of a mock gene.*

B. *The polymerase occupancy of a mock gene with higher expression level than the gene in (A).*

C. *The polymerase occupancy of a mock gene with the same expression level as in (A), but where a majority of polymerases will turn over in the PPR (promoter proximal region).*

D. *The polymerase occupancy of a mock gene with the same expression level as in (A), but with a pausing site in the PPR that stops polymerases.*

This is a simplified situation where polymerase flux stays constant within the same gene. However, the polymerase flux along a gene is not necessarily the same (Figure 1.2). For instance, not all the polymerases might generate full-length transcripts. As shown in Figure 1.3 C, a majority of polymerases might turn over before they enter productive transcription, and will generate abortive transcripts[5-7]. This will make the polymerase flux in the region *before* the early transcription termination position higher than *after* it. Beyond abortive transcription, there are other transcription events that can make the polymerase flux different between positions in the same gene. In 'backtracking', for example, some polymerases will be blocked at a certain position during transcription; the polymerases have to move backwards first, then become arrested before going forward again[23, 25, 27, 28]. It means polymerases will go through the backtracking region twice, thus have higher polymerase flux.

If we divide polymerase occupancy by polymerase flux, we can get the 'average residence time':

polymerase occupancy = polymerase flux * average residence time (1)

The average residence time represents the average period of time a polymerase spends at certain positions. Polymerases that stay longer at certain positions of a gene will also give rise to higher polymerase occupancy at those points (Figure 1.3 A and D).

1.3.3 Pausing fraction and pausing time

Polymerase pausing is subject to regulation[16-19]. This means that the profile of pausing at genes is potentially different upon responding to the environment. After a heat shock, for instance, polymerases are likely to pass unimpededly the pausing sites of response genes[29]. I defined the average fraction of polymerases that pause at a pausing site as the 'pausing fraction'. Pausing sites with a higher pausing fraction should have a higher polymerase occupancy.

In contrast to the average residence time, if we only consider the residence time of polymerase *really paused* at a certain position, we get the 'pausing time'. As the polymerase moves fast during elongation (it only spends ~0.01 to 0.06 seconds at each nucleotide[3], it contributes little to the polymerase occupancy of polymerases at the pausing site. Since the polymerase occupancy is the product of polymerase flux and average residence time (Eq. 1), we can deduce that the average residence time is approximately equal to the product of pausing fraction and pausing time (Eq. 5):

average residence time = (polymerase occupancy$_n$ + polymerase occupancy$_p$) / (polymerase flux$_n$ + polymerase flux$_p$), (2)

where *n* denotes non-paused polymerase and *p* denotes paused polymerase.

average residence time ≈ polymerase occupancy$_p$ / (polymerase flux$_n$ + polymerase flux$_p$) (3)

⇒

average residence time ≈ (polymerase occupancy$_p$/ polymerase flux$_p$) * [polymerase flux$_p$ / (polymerase flux$_n$ + polymerase flux$_p$)] (4)

$$\textbf{average residence time} \approx \textbf{pausing time * pausing fraction} \tag{5}$$

Base on Eq1 and Eq5, polymerase occupancy can also be calculated as:

$$\textbf{polymerase occupancy} \approx \textbf{polymerase flux * pausing time * pausing fraction} \tag{6}$$

1.3.4 Pausing frequency

As pausing typically occurs at multiple positions in a transcribed region[25, 30], I define the *density* of pausing sites, i.e., their number within a length of sequence, as 'pausing frequency'. Pausing frequency is an important parameter that influences transcriptional dynamics. The more pausing sites in a gene, the slower the speed of polymerase engaging in that region can be expected. Also, pausing frequency has been suggested to influence the dispersion of mRNAs in individual cells (transcriptional noise)[31].

1.4 Polymerase pausing and transcription regulation

As one of the first steps of gene expression, transcription is highly regulated. The process of transcription can be divided into three phases: initiation, RNA polymerase binds to chromatin; elongation, RNA polymerase moves to product nascent RNA; termination, RNA polymerase is released from the DNA template[32]. Studies about the mechanisms of gene regulation are mostly focused on the assembly of the pre-initiation complex (PIC)[33]. However, recent research emphasizes the importance of regulation downstream of transcription initiation, as polymerase pausing has been found to be widespread throughout the whole genome[20, 21, 25, 26]. Polymerase pausing influences all three phases of transcription and plays various roles in the regulation of transcription.

1.4.1 Pausing and initiation

Most attention in pausing related literature is focused on Pol II enrichment downstream of TSS. The enrichment has been interpreted as polymerase that pauses for a longer time in this region[2, 13, 34] (Figure 1.3 D). This promoter proximal pausing has been suggested to be a rate-limiting step for gene expression, as it has been found to dominate among genes with high expression level[35]. Polymerase enrichment in the PPR (promoter proximal region) has been suggested to inhibit the formation of nucleosome. Thus, the promoter can maintain an open chromatin state to permit higher expression[1, 36]. This phenomenon has mainly been found in genes that are high regulated, but not in housekeeping genes[1]. Beyond that, pausing in the PPR has also been suggested to occupy the region downstream of TSS in order to inhibit initiation of successive rounds of transcription[4].

However, recent studies suggest that the reason that Pol II enriches in the PPR may also be caused by a high turnover rate of Pol II[5-7] (Figure 1.3 C). More than 90% of initiated polymerase appears to drop off the DNA template and generate abortive transcripts before it enters productive transcription[6]. Studies that measure polymerase pausing times are desirable to distinguish among different reasons for polymerase enrichment in the PPR.

1.4.2 Pausing and elongation

Pausing of polymerase is in principle not restricted to the PPR, but has been found throughout the entire length of genes[25, 30]. Nucleosome loss and/or histone acetylation after heat shocks have been proposed to loosen chromatin, thereby facilitating elongation by reducing polymerase pausing[29]. Beyond that, RNA splicing, the process that removes the intron from pre-mRNA by spliceosomes, has been shown to correlate with polymerase pausing by a series of works.

Splicing occurs during transcription; more than half of splicing takes place only within 45nt downstream of intron/exon boundaries[37]. Higher polymerase occupancy has been found around splicing sites, which suggests that polymerase pauses for splicing[26, 30, 38]. Also, a slow-down of the polymerase can help the spliceosome bind to the alternative exon[39]. Furthermore, the transcription factor CTCF can induce polymerase pausing and lead to the retention of weak upstream splicing sites[40].

1.4.3 Pausing and termination

A region with high polymerase occupancy has also been found downstream of the TES (transcription end site)[38]. This suggests that termination of transcription requires polymerase pausing or slowing as well[41]. Mutants of Pol II with different elongation rates are consistent with this suggestion; Pol II with faster elongation terminates transcription further downstream while slowly moving Pol II terminates transcription upstream[42]. In addition, dominant-negative TFIIS (Transcription factor IIS), which inhibits the rescue of backtracked polymerase, also facilitates termination of transcription just downstream of TES[43].

1.5 Deeper understanding of pausing by measuring pausing time

During the last 12 years, about 10 different next generation sequencing methods have been developed or used for understanding pausing[4, 20-26, 44-46]. However, all of these methods in principle can only measure polymerase occupancy, not pausing. As illustrated in Figure 1.2, polymerase occupancy is influenced by polymerase flux, pausing time and pausing fraction. Individual cases of altered/elevated polymerase occupancy can have completely different biological explanations. For example, the polymerase enrichment in PPR can be caused by longer pausing time (Figure 1.3 D) or higher polymerase flux (Figure 1.3 C).

Various methods have been used to study pausing time. However, all of these have certain limitations. FRAP (fluorescence recovery after photobleaching) can reveal overall pausing *in vivo*[6, 47], but it cannot detect the genomic locations of polymerases. Nascent transcription RNA FISH (fluorescence *in situ* hybridization) can reveal pausing sites of individual cells, but only for a small number of genes with designed probes[48]. Trp (Triptolide), a covalent inhibitor of the TFIIH subunit XPB, has been employed to inhibit transcription initiation prior to sequencing. Fitting decay curves to the polymerase occupancy of the region downstream of TSS upon a Trp treatment time series allows estimation of the average pausing times at the PPRs of all genes[3, 4, 49]. However, this method cannot estimate pausing time in regions other than the PPR and the measurements have low positional resolution. Furthermore, recent research suggests that uptake of Trp is slow, which will lead to overestimates of pausing time by this method[50].

For these reasons, I developed TV-PRO-seq, a method that can estimate pausing times in genome-wide fashion at single-base resolution. TV-PRO-seq allowed me a meta-analysis for pausing times in different gene regions. In addition, it can be used for analysis of short genes, such as tRNAs and lncRNAs (long non-coding RNAs), which was hitherto impossible.

My results showed that promoter proximal pausing is actually shorter than pausing in other regions (this result can also be due to the effect of sarkosyl). The polymerase actually does not pause for longer time in this region, but shorter. This result is consistent with previous research showing that the majority of Pol II drops off from the DNA template before entering productive transcription[5-7]. My results also highlight the importance of pausing in the gene body for transcription regulation. As polymerase pauses about every 20nt to 100nt[13] in a typical gene, a widespread pausing mechanism should exist also for this gene region. Previous research has shown that nucleosomes can act as barriers for Pol II[51, 52]. My results extend this by demonstrating that polymerase is paused for a long time in front of nucleosomes with modification such as H3K9me3 and H3K36me3. Beyond that, I have defined various new sequence motifs that correlate with pausing. I am proposing that these motifs and nascent RNAs can form DNA-RNA hybrid helices which then leads to pausing. Finally, I analyse the relationship between pausing and transcriptional dynamics, which establishes the importance of pausing frequency in transcriptional regulation.

Chapter 2 Timing pausing with TV-PRO-seq

2.1 Introduction

Pausing has been known for decades. It was first found *in vitro* for RNA polymerase of *Escherichia coli* in the early 1970s[53, 54], and was finally confirmed by *in vivo* experiments in hen erythrocytes for the beta-globin gene in 1981[55]. Pausing in the promoter proximal region has been suggested to play an important role in gene expression by various mechanisms. These include the maintenance of an open chromatin state at the promoter region for activation of expression[1], the blocking of further initiation for successive rounds of transcription[4], along with enabling rapid responses to the environment[2, 29] and synchronous expression of genes[12]. Pausing in the gene body has been suggested to be functionally interdependent with co-transcriptional splicing[37, 56], and pausing after the TES faciliates termination of transcription[41-43].

In more recent years, ChIP (chromatin immunoprecipitation) of polymerase[57] and nuclear run-on assays[58] were introduced to the study of polymerase pausing. ChIP encompasses immunoprecipitation of a target protein, i.e., Pol II, by antibody, followed by isolation of the DNA/RNA bound to it. Nuclear run-on assays, on the other hand, are based on the addition of labelled NTP (Nucleoside triphosphate) into the cells suspension, followed by extraction of the labelled nascent RNAs. Using these types of methods, pausing, specifically in the region close to the TSS (transcription start site), has been proved to occur at several other genes in the following decade[59-61]. This phenomenon of Pol II enrichment within 20 nt to 100 nt downstream of TSS has been termed ‘promoter proximal pausing’. The promoter proximal pausing has been confirmed with genome-wide ChIP-chip (chromatin immunoprecipitation microarray) experiments[21].

Various sequencing methods have been developed/used for the research of transcriptional dynamics and similar topics. For ChIP-seq, DNA fragments bound to

Pol II are selected by ChIP, followed by sequencing of the fragments to reveal the genomic locations of Pol II[20, 21] (Figure 2.1A). As the fragments of DNAs are usually between 100 to 500 nt of size, ChIP-seq produces results at comparatively low positional resolution. To improve this aspect, ChIP-exo[62] (exonuclease) and its advanced version ChIP-nexus[63] (nucleotide resolution through exonuclease, unique barcode and single ligation) were devised. These two methods degrade overhanging DNA by exonuclease after ChIP, while the central, protein-bound part is protected. This narrows down the detected positions towards the 5' borders of the DNA actually bound by the protein[4, 64] (Figure 2.1B). NET-seq sequences the nascent RNA attached to Pol II after ChIP of the latter[25] (Figure 2.1C). It produces a strand-specific map of polymerases at single nucleotide resolution. scRNA-seq (short capped RNA-seq, not to be confused with single cell RNA-seq) is aimed at sequencing short RNAs with 5' caps[23]. The procedure of scRNA-seq is simple; uncapped RNAs such as rRNAs are removed by 5' monophosphate-dependent terminator exonuclease, followed by selection of RNAs between 25nt to 120nt of size by electrophoresis for sequencing (Figure 2.1D). Its yields high resolution results but is limited to the region right downstream of TSSs.

GRO-seq as the first run-on based nascent RNA sequencing method was developed in 2008[22]. It is based in the addition of BrU (5-Bromouridine 5′-triphosphate) to isolated nuclei. Active polymerases will then incorporate BrU into their nascent RNAs. This permits enrichment of the labelled RNAs by beads-bound antibodies and in turn their sequencing after reverse transcription and PCR (Polymerase chain reaction) (Figure 2.1E). PRO-seq is an advanced verions of GRO-seq that was developed in 2013[26]. Instead of BrU, the labelling step of PRO-seq is done with biotin-NTPs. These biotin-NTPs can block transcription and thus record the precise position of polymerase pausing (Figure 2.1F). Several assays based on PRO-seq extend its application. coPRO-seq allows the joint analysis of pausing, TSS and the 5' cap's state[44]. Finally, chRO-seq enables mapping of the polymerase distribution of input sample with degraded RNA[46].

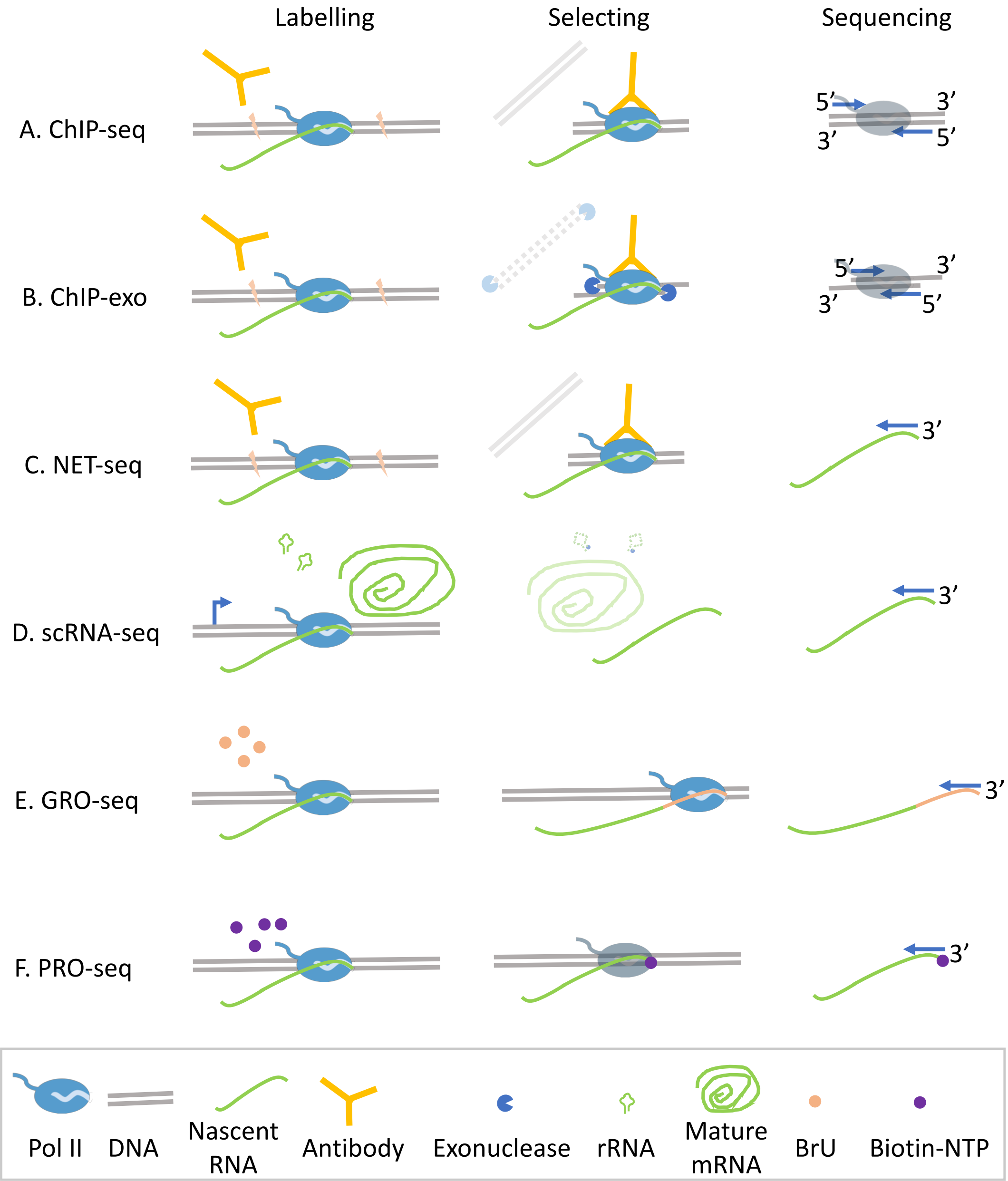


Figure 2.1 Principle of sequencing methods used to investigate polymerase pausing

A. *Chromatin is fragmented and crosslinked with Pol II, followed by immunoprecipitation of Pol II-bound chromatin fragments with an antibody directed against the polymerase. The DNA fragments are the processed and subjected to sequencing.*

B. *Based on A, but the DNA fragments bound to Pol II are degraded by exonuclease from the 5' end.*

C. *Similar to A, but nascent RNAs are processed to sequencing instead of the DNA.*

D. *Uncapped RNAs are removed by exonuclease, and long RNAs are removed by electrophoresis. Remaining short capped RNAs are processed to sequencing.*

E. *BrU is added to isolated nuclei or permeabilized cells and nascent RNAs transcribed by active polymerases become labelled. The labelled RNAs are processed to sequencing.*

F. *Similar to E, but using biotin-NTP instead of BrU.*

Despite the number of various sequencing methods that have been developed, they are restricted to reveal polymerase occupancy only. A method that can measure the pausing times of pausing sites in genome-wide fashion is critical for the in-depth study of the complex dynamics of transcription. Here, I developed time-variant PRO-seq (TV-PRO-seq), which is essentially a time series of individual PRO-seq[65] samples and which can be used to investigate the pausing time across the whole genome. For analysing TV-PRO-seq results, I devised a peak calling procedure that outputs results with single nucleotide resolution. Finally, I used a Bayesian framework that models saturation curves to infer estimated pausing times of each peak.

2.2 Methods

2.2.1 Reagents

Reagents	**Company**	**Part Number**
DEPC water	Fisher Scientific	10514065
NaCl	Sigma-Aldrich	S9888
KCl	Sigma-Aldrich	P9333
$CaCl_2$	Sigma-Aldrich	C1016
$MgCl_2 . 6H_2O$	Sigma-Aldrich	M2670
EDTA	Sigma-Aldrich	E9884
NaOAc	Sigma-Aldrich	S2889
NH_4Ac	Sigma-Aldrich	A1542
$MgAc_2$	Sigma-Aldrich	M5661
EGTA	Sigma-Aldrich	E3889
Sarkosyl	Sigma-Aldrich	L5125
Sucrose	Sigma-Aldrich	S0389
NaOH	Fisher chemical	10396240
DTT	Sigma-Aldrich	D0632
Glycerol	Sigma-Aldrich	G5516
DMSO	Sigma-Aldrich	41640
TWEEN-20	Sigma-Aldrich	P9416
Triton X-100	Sigma-Aldrich	X100
Tris-HCl pH 6.8 1M	VWR International Ltd	A4987
Tris-HCl pH 7.4 1M	Sigma-Aldrich	T2663
Tris-HCl pH 8.0 1M	Sigma-Aldrich	93283
1 X PBS pH 7.4	Fisher Scientific	10728775

Absolute Ethanol	Fisher Scientific	BP2818
Isopropanol	Fisher Scientific	BP2618
Chloroform	Fisher Scientific	10488400
Biotin-11-CTP	PerkinElmer	NEL542001EA
Biotin-11-UTP	PerkinElmer	NEL543001EA
Biotin-11-ATP	PerkinElmer	NEL544001EA
Biotin-11-GTP	PerkinElmer	NEL545001EA
ATP	New England Biolabs	P0756S
GTP	Fisher Scientific	10698085
P-30 column	Bio-Rad	732-6250
Streptavidin M280 beads	Fisher Scientific	10465723
Trizol	Fisher Scientific	15608948
Trizol LS	Fisher Scientific	15867521
GlycoBlue	Fisher Scientific	10301575
Phenol:chloroform	Sigma-Aldrich	77617
RNase inhibitor	Fisher Scientific	10773267
T4 RNA ligase I	New England Biolabs	M0204S
RppH	New England Biolabs	M0356S
T4 Polynucleotide Kinase	New England Biolabs	M0201L
Superscript III	Fisher Scientific	12087539
dNTP mix	New England Biolabs	N0447
Q5 master mix	New England Biolabs	M0544
TEMED	VWR International Ltd	443083G
APS	Sigma-Aldrich	A3678
Acrylamide	Sigma-Aldrich	A3449
Orange loading dye 6X	New England Biolabs	B7022S

SYBR Gold	Fisher Scientific	10358492
25-700bp DNA ladder	Fisher Scientific	10784881

All RNA/DNA oligos (RNA adaptors and DNA primers) synthesis was done by Sigma-Aldrich. The sequences are the same as described in the published PRO-seq protocol[65].

Sequencing was performed on an Illumina NextSeq 500 machine for 51bp single end by the Genomics Facility of the School of Life Sciences, University of Warwick.

2.2.2 Library building of TV-PRO-seq

2.2.2.1 Cell culture

HEK293 cells were cultured at 37°C and 5% CO_2 in DMEM containing 10% FBS in a 175 cm^2 flask. KBM-7 cells were cultured in the same way, using IMDM instead of DMEM. S2 cells were cultured at 28°C in Schneider's D. melanogaster Medium supplemented with 10% heat-inactivated FBS in a 25 cm^2 flask. When confluency reached 60%, the culture medium was replaced with fresh medium for one day. (For triptolide (Trp) and Flavopiridol (FP) treatments of HEK293 cells, Trp and FP were added at concentrations of 500 nM and 300 nM, respectively, and cells were incubated at 37°C for 10 min before cell permeabilization.)

2.2.2.2 Cell permeabilization

1. Cells from 2.2.2.1 were harvested and collected in a 50mL falcon tube, followed by 1000g, 4°C centrifugation for 5min (for TV-PRO-seq, at least 5×10^8 cells are required in this step).

2. Cells were washed with ice-cold PBS and centrifuged again.

3. Cells were resuspended in 20ml ice-cold permeabilization buffer (Table 2.1), then incubated for 5min on ice, followed by centrifugation.

Table 2.1 Permeabilization buffer

1M Sucrose	15mL
1% Tween-20	2.5mL
1M Tris-HCl pH 7.4	500μL
0.1M EGTA	500μL
10% NP40	500μL
2M KCl	250μL
1M $MgCl_2$	250μL
1M DTT	25μL
RNase inhibitor	5μL
Protease inhibitor	1 tablet
Total (by adding DEPC water)	50mL

4. Cells were washed by 15ml ice-cold permeabilization buffer following centrifugation.

5. Repeat Step 4.

6. Cells were resuspended in storage buffer (Table 2.2) to a concentration of about 10^7 cells in 100μL in 1.5ml tubes (TV-PRO-seq needs at least 20 tubes samples when considering 4 timepoints, duplicates, and 4 reserve samples). Tubes were flash-frozen in liquid nitrogen and stored at -80℃ (the permeabilized cells can be stored at -80℃ for up to 6 months).

Table 2.2 Storage buffer

0.5M EDTA	0.4μL
1M Tris-HCl pH 8.0	20μL
1M $MgCl_2$	10μL

1M DTT	10μL
Glycerol	500μL
DEPC water	1460μL

2.2.2.3 Buffer preparation before TV-PRO-seq

The solutions for TV-PRO-seq were prepared beforehand and can be stored at room temperature for 6 months:

5M NaCl: 58.4g NaCl was dissolved in 200mL DEPC water, then stored overnight after mixing. The solution was then autoclaved.

2M KCl: 2.982g KCl was dissolved in 20mL DEPC water, then stored overnight after mixing. The solution was then autoclaved.

1M $MgCl_2$: 4.066 $MgCl_2 \cdot 6H_2O$ was dissolved in 20mL DEPC water, then stored overnight after mixing. The solution was then autoclaved.

1M Sucrose: 34.23g Sucrose was dissolved in 100mL DEPC water, then stored overnight after mixing. The solution was then autoclaved.

5M $MgAc_2$: 4.289g $MgAc_2$ was dissolved in 20mL DEPC water, then stored overnight after mixing. The solution was then autoclaved.

1M NH_4Ac: 1.542g NH_4Ac was dissolved in 20mL DEPC water, then stored overnight after mixing. The solution was then autoclaved.

0.1M EGTA: 0.761g EGTA was dissolved in 20mL DEPC water, then stored overnight after mixing. The solution was then autoclaved.

1N NaOH: 2g NaOH was dissolved in 50mL DEPC water.

10% Triton-X100: 2mL Triton-X100 was added in 18mL DEPC water.

10% NP40: 2mL NP40 was added in 18mL DEPC water.

1% Tween-20: 0.2mL Tween-20 was added in 19.8mL DEPC water.

2% Sarkosyl: 0.4g Sarkosyl was dissolved in 10mL DEPC water, and then mixed. After dissolution, the solution was filtered by a 0.22 μm filter.

The buffers for washing streptavidin-coated magnetic beads were made before the library building. All buffers can be stored at 4℃ up to 1 week:

High-salt wash buffer: 1M Tris-HCl (pH 7.4) 2mL, 5M NaCl 16mL, 10% Triton X-100 2mL, DEPC water 20mL.

Binding buffer: 1M Tris-HCl (pH 7.4) 400μL, 5M NaCl 2.4mL, 10% Triton X-100 400μL, DEPC water 36.8mL.

Low-salt wash buffer: 1M Tris-HCl (pH 7.4) 200μL, 10% X-100 400μL, DEPC water 39.6mL.

Prewashed streptavidin-coated magnetic beads were prepared as:

1. For each sample, 90μL M280 beads were added into a 1.5mL tube and placed on a magnetic stand for 1min, followed by removal of the liquid by pipette.

2. The beads were washed once with buffer: 1N NaOH 100μL, 5M NaCl 10μl, DEPC water 890μL. The tubes were placed on the magnetic stand for 1min, followed by removal of the liquid by pipette.

3. The beads were washed twice with buffer: 5M NaCl 20μl, DEPC water 980μL. The tubes were placed on the magnetic stand for 1min, followed by removal of the liquid by pipette.

4. The beads were resuspended in binding buffer.

2.2.2.4 Nuclear run-on

1. Prepare run-on buffer (Table 2.3 for four biotin run-on and Table 2.4 for two biotin run-on). For TV-PRO-seq, we make 8.5X volume buffer; S2 cells and HEK293 cells used four biotin run-on, KBM7 used two biotin run-on:

Table 2.3 Four biotin run-on buffer

	1X	4.5X	8.5X
1M $MgCl_2$	0.5µL	2.25µL	4.25µL
1M Tris-HCl pH 8.0	1µL	4.5µL	8.5µL
0.1M DTT	1µL	4.5µL	8.5µL
RNase inhibitor	2µL	9µL	17µL
1mM biotin-ATP	5µL	22.5µL	42.5µL
1mM biotin-GTP	5µL	22.5µL	42.5µL
10mM biotin-UTP	0.5µL	2.25µL	4.25µL
10mM biotin-CTP	0.5µL	2.25µL	4.25µL
2M KCl	15µL	67.5µL	127.5µL
DEPC water	19.5µL	87.75µL	165.75µL
2% Sarkosyl	50µL	225µL	425µL

Table 2.4 Two biotin run-on buffer

	1X	4.5X	8.5X
1M $MgCl_2$	0.5µL	2.25µL	4.25µL
1M Tris-HCl pH8.0	1µL	4.5µL	8.5µL
0.1M DTT	1µL	4.5µL	8.5µL
RNase inhibitor	2µL	9µL	17µL
10mM ATP	2.5µL	11.25µL	21.25µL
10mM GTP	2.5µL	11.25µL	21.25µL
10mM biotin-UTP	0.5µL	2.25µL	4.25µL
10mM biotin-CTP	0.5µL	2.25µL	4.25µL
2M KCl	15µL	67.5µL	127.5µL

DEPC water	24.5µL	110.25µL	208.25µL
2% Sarkosyl	50µL	225µL	425µL

2. The permeabilized cells from 2.2.2.2 were preheated at 27℃ (D. melanogaster) or 37℃ [66] for 2min (For KBM7 Trp treatment, 1µL of 100µM Trp was added to 100µl permeabilized KBM-7 cells, followed by 10min incubation at 37℃).

3. 100µL run-on buffer was added into an 1.5mL tube with permeabilized cells for the designated run-on time (usually 4 timepoints are needed for a TV-PRO-seq series, prepared as duplicates; the four time points are 0.5min, 2min, 8min, and 32min). The tube was placed in an temperature block (human cells at 37℃ and D. melanogaster cells at 28℃), then mixed thoroughly by pipetting the liquid up and down about 15 times. The liquid was mixed every 3min by pipetting.

4. After run-on, 500µL Trizol LS was added to each sample, followed by vortexing. After this, the sample was placed on ice.

2.2.2.5 RNA extraction and fragmentation

1. After finishing the run-on of all samples, all tubes were thawed on the 37℃ temperature block for 2min, then placed at room temperature for 5min.

2. 130µL chloroform was added to each sample. The samples were vortexed vigorously for 15 s, followed by 2min incubation at room temperature.

3. The samples were centrifuged at 14,000g at 4 °C for 5 min. The aqueous phase of each tube was transferred to a new tube with 1µL of GlycoBlue.

4. 380µL of isopropanol was added into each tube, then vortexed for 10min, and followed with 10min incubation at room temperature.

5. The samples were centrifuged at 14,000g 4°C for 20 min, the RNA precipitate forms a gel-like pellet on the side and bottom of the tube.

6. The supernatant was removed, and the tubes with RNA pellets were opened for 5min to air-dry.

7. 20μL of DEPC water was added to each tube for re-dissolving the RNA pellet.

8. The tubes were placed on a 65°C heat block for 40s to heat-denature the RNA, and were then placed on ice.

9. 5 μL of ice-cold 1 N NaOH was added to each tube. The mixture was placed on ice for 10min.

10. P-30 columns were inverted to remove the bubble, then their tips were snapped off. The columns were then placed in 2.0mL tubes for 2min, and the flow-through discarded. The tubes with columns were then centrifuged at 1000g for 2min.

10. 25μL of 1 M Tris-HCl (pH 6.8) was added to each tube, followed by transfer of the mixture of each tube to a P-30 column prepared in step 10. All the P-30 columns were placed on 1.5mL tubes and centrifuged at 1000g for 4min.

11. The columns were discarded and 1μL of RNase inhibitor was added to each 1.5mL tube.

2.2.2.6 Biotin RNA enrichment

1. Each RNA sample from 2.2.2.5 was mixed with 50μL of prewashed streptavidin beads made from 2.2.2.3. The tubes with the mixtures were placed on a rotator for 20min incubation.

2. The tubes were placed on a magnetic stand for 1 min, followed by removal of the liquid.

3. 500μL ice-cold high-salt wash buffer (see 2.2.2.3) was added to each tube for washing the beads. The tubes were then placed on a magnetic stand for 1 min, followed by removal of the liquid.

4. 500μL ice-cold binding buffer (see 2.2.2.3) was added to each tube for washing the beads. The tubes were then placed on a magnetic stand for 1 min, followed by removal of the liquid.

5. Step 4 was repeated.

6. 500µL ice-cold low-salt buffer (see 2.2.2.3) was added to each tube for washing the beads. The tubes were then placed on a magnetic stand for 1 min, followed by removal of the liquid.

7. Step 6 was repeated.

8. The beads were resuspended in 300µL Trizol and vortexed vigorously, then incubated for 3min at room temperature.

9. 60µL of chloroform was add to each tube, vortexed vigorously, then incubated for 3min at room temperature.

10. Beads were centrifuged at 14,000g, 4°C for 5 min and the aqueous layer in each tube was transferred into a new tube.

11. The organic phase was removed and step 6-8 repeated for the beads. The collected aqueous layers were combined.

12. 360µL of isopropanol and 1µL GlycoBlue were added to each tube, then vortexed for 10s. The samples were then incubated at room temperature for 10min.

13. All samples were centrifuged at 14,000g at 4°C for 20min. The RNA precipitate formed a gel-like pellet on the side and bottom of each tube.

14. All supernatants were removed, followed by air-drying of the RNA pellets were for 5min.

2.2.2.7 Adaptor ligation

1. The RNA pellet of each tube was re-dissolved in 4µL 12.5µM 3' RNA adaptor.

2. The mixture was placed on a 65°C heat block for 20s for denaturing, then placed on ice.

3. 6µL adaptor ligation reagent (Table 2.5) was added to each tube. The mixture was then incubated at 20°C for 4h.

Table 2.5 Adaptor ligation reagent

	1X	4.5X	8.5X

T4 RNA ligase buffer (10X)	1μL	4.5μL	8.5μL
1mM ATP	1μL	4.5μL	8.5μL
RNase inhibitor	1μL	4.5μL	8.5μL
T4 RNA ligase I	1μL	4.5μL	8.5μL
50% PEG	2μL	9μL	17μL

4. 40μL DEPC water was added to each sample. Biotin enrichment was then performed on the samples according to 2.2.2.6.

5. The RNA pellet of each tube was re-dissolved in 7.5μL DEPC water, then incubated at 65°C for 20s for denaturing and finally placed on ice.

6. 2.5μL 5’ cap repair enzyme mix (Table 2.6) was added to each tube and placed in a 37°C incubator for 1h.

Table 2.6 5’ cap repair enzyme mix

	1X	4.5X	8.5X
Thermpol Reaction Buffer (10X)	1μL	4.5μL	8.5μL
RppH	1μL	4.5μL	8.5μL
RNase inhibitor	0.5μL	2.25μL	4.25μL

7. 90μL PNK mix (Table 2.7) was added to each tube, then placed in a 37°C incubator for 1h.

Table 2.7 PNK mix

	1X	4.5X	8.5X
DEPC water	65μL	292.5μL	552.5μL
1mM ATP	10μL	45μL	85μL

PNK buffer (10X)	10μL	45μL	85μL
RNA inhibitor	2.5μL	11.25μL	21.5μL
PNK	2.5μL	11.25μL	21.5μL

8. 300μL Trizol was added to each sample, following by vortexing. The samples were then placed at room temperature for 1min.

9. 60μL chloroform was added to each sample. The samples were vortexed vigorously for 15 s, followed by 2min incubation at room temperature.

10. The samples were centrifuged at 14,000g, 4 °C for 5 min. The aqueous phase of each tube was transferred to a new tube containing 1μL of GlycoBlue.

11. 280μL of isopropanol was added to each tube, and vortexed vigorously for 10min, followed by 10min incubation at room temperature.

12. The samples were centrifuged at 14,000g, 4°C for 20 min, with the RNA precipitate forming a gel-like pellet on the side and bottom of the tube.

13. The supernatant was removed and the tubes with the RNA pellets opened for 5min to air-dry.

14. The RNA pellet of each tube was re-dissolved in 4μL 12.5μM 5' RNA adaptor.

15. The mixture was incubated at 65°C for 20s for denaturing, then placed on ice.

16. 6μL adaptor ligation reagent (Table 2.5) was added to each tube. The mixture was then incubated at 20°C for 4h.

17. 40μL DEPC water was added to each sample. Biotin enrichment was performed on the samples according to 2.2.2.6.

2.2.2.8 Reverse transcription (RT) and PCR amplification

1. The RNA pellet of each tube was re-dissolved in 12.5μL RT primer mix (Table 2.8).

Table 2.8 RT primer mix

	1X	4.5X	8.5X
DEPC water	10.5μL	47.25μL	89.25μL
12.5mM dNTP mix	1μL	4.5μL	8.5μL
25μM RP1 primer	1μL	4.5μL	8.5μL

2. The mixture was placed on a 70°C heat block for 2min for denaturing, then placed on ice.

3. 7.5μL RT enzyme mix was added (Table 2.9) to each tube. The tubes were centrifuged for 10s.

Table 2.9 RT enzyme mix

	1X	4.5X	8.5X
First-stand buffer (5X)	4μL	18μL	34μL
RNase inhibitor	1μL	4.5μL	8.5μL
DTT (0.1M)	1μL	4.5μL	8.5μL
Superscript III RT enzyme	1.5μL	6.75μL	12.75μL

4. The mixture was transferred to 200μL PCR tubes, and subjected to a temperature ramp of the following scheme in a PCR machine: 37°C 5min, 45°C 15min, 50°C 40min, 55°C 10min, 70°C 15min, then 4°C forever.

5. 4μL DEPC water, 25μL Q5 PCR master and 1μL 25μM RPI-n primer were added to each tube

6. The samples were centrifuged for 10s, and then subjected to PCR using the following conditions: 95°C 2min, (95°C 30s, 56°C 30s, 72°C 30s) for 5 cycles, (95°C 30s, 65°C 30s, 72°C 30s) for 12 cycles, 72°C 10min, then 4°C forever.

2.2.2.9 Library recycling

1. The PCR product was transferred to 1.5mL tubes. 950μL purify mix (Table 2.10) was then added to each tube. The tubes were then centrifuged at 4°C for 30 min, with the DNA precipitate forming a gel-like pellet on the side and bottom of the tube.

Table 2.10 Purify mix

	1X	4.5X	8.5X
Ethanol	750μL	3375μL	6375μL
DEPC water	231μL	1039.5μL	1963.5μL
5M NaCl	18μL	81μL	153μL
GlycoBlue	1μL	4.5μL	8.5μL

2. 10μL water and 2μL 6X Orange G loading dye were added to each tube for re-dissolving the DNA pellet.

3. The samples were subjected to gel electrophoresis on a 10% native PAGE gel (the recipe is shown in Table 2.11, which suffices for 2 making two gels for one mini tank; the mixture needs approximately 1h for solidification), and were run along a 25bp DNA ladder on one side of the gel. The gels were first run at 15mA for 20min, then changed to 25mA until the Orange G dye run off the gel (about 45min).

Table 2.11 8% native PAGE gel

DEPC water	15.56mL
Acrylamide (30%)	6.67mL
TBE (10X)	1.25mL
APS (10%)	250μL
TEMED	25μL

4. After the electrophoresis, part of the gel covering DNA sized from about 140bp to 700bp was excised (the region below adaptor dimer band and above the last band of the ladder was targeted). The gel fragments of each sample were then placed into a 0.5mL microtube with a hole at the bottom (a heated 21 gauge needle was used to make the hole). The 0.5ml microtubes were then placed in 2mL tubes and centrifuged at 8000g for 2min at room temperature to ensure that the gel fragments were shredded into small pieces by extrusion.

5. 400μL gel elution buffer (Table 2.12) was added to each tube, and incubated in a temperature block with shaking function at 37°C, 500rpm for 2h.

Table 2.12 Gel elution buffer

1M NH_4Ac	25μL
0.5M EDTA	100μL
1M Tris-HCl pH8.0	500μL
1M $MgAc_2$	500μL
10% SDS	500μL
DEPC water	up to 50mL

6. The samples were centrifuged at 14000g and room temperature for 2min, and the liquid was transferred into new tubes. 400μL of gel elution buffer was then added to each tube with the gel pieces. These tubes were returned to incubation in the shaking temperature block at 37°C, 500rpm for an additional1h.

7. The gels were centrifuged at 14000g and room temperature for 2min, and the supernatants were combined with those from step 6. The liquids were then transferred into Spin-X filters and centrifuged at 6000g at room temperature for 2min.

8. After filtering, the liquid was transferred into 2mL tubes and the volume adjusted to 800μL. 800μL of buffered phenol:chloroform was then added into each tube. The tubes were vortexed vigorously and centrifuged at 14000g for 5min at 4°C. The equal

volume of aqueous layer from each tube was collected and transferred into two new 1.5mL tubes.

9. 2.5X volume of ethanol and 1μL of GlycoBlue were added into each 1.5 tube. The tubes were vortexed vigorously and incubated at -80°C overnight.

10. The tubes were centrifuged at 14000g for 20min at 4°C followed by removal of the liquid. The DNA pellets were then allowed to air-dry for 10min.

11. The DNA of each tube was re-dissolved in 20μL of H_2O, and 2μL of each sample was subjected to quantification with a Qubit device.

12. For TV-PRO-seq, the 8 samples were mixed at equal ratios of DNA mass (each sample should have at least 5ng DNA). The PAGE purification procedure from step 2 to step 11 was then repeated, but restricting the size selection in step 4 to 140bp to 500bp for the secondary purification. 10ng of the combined library sample were then sequenced for 51bp, single end reads, on a NextSeq 500 (Illumina) sequencer.

2.2.3 Processing of sequencing data

Raw data were converted into FASTQ format by bcl2fastq with 0 index mismatches allowed.

Reads were trimmed with Cutadapt version 1.14 [67], to remove sequences starting with the adaptor sequence 'TGGAATTCTCGGGTGCCAAGG' from the 3' end of reads, and reads shorter than 20bp after trimming were discarded:

```
cutadapt -a TGGAATTCTCGGGTGCCAAGG -m 20 -e 0.05
```

Trimmed reads were aligned to the best matched position of hg38 genome with Hisat2 version 2.1.0 [68], resulting in alignment rates above 80%:

```
hisat2 -p 4 -k 1 --no-unal -x -U -S
```

Because the ends of sequencing reads have lower sequencing quality, Hisat2 uses soft clipping for the reads, which moves the detected pausing site upstream of the actual pausing site. A custom script `Sam_enlong.pl` was used on the SAM files to extend the soft clipped reads to their original lengths.

Because sequencing depth also influences the process of peak calling of TV-PRO-seq, another script `Sam_cutter.pl` was used to reduce the 8 TV-PRO-seq SAM files for HEK293 cells to the same sizes by randomly selecting a subset of reads for each.

The processed SAM files were further converted to BAM files and were sorted with samtools version 0.1.19 using `samtools view -S -b` and `samtools sort` [69].

The sorted bam files were then converted to BEDGRAPH files [70]. The 5' end of a read corresponds to the position of the paused polymerase release site on the opposite strand:

Pausing on plus strand: `genomeCoverageBed -strand - -5 -bga -ibam`

Pausing on minus strand: `genomeCoverageBed -strand + -5 -bga -ibam`

I then combined the BEDGRAPH files for the various replicates and time points into two files, one for each strand, with the custom script `TV_bedGraph_merger.pl`. These files corresponded to tables with rows for each position and columns containing the read numbers across the samples and were used for the further analysis.

2.2.4 Single nucleotide resolution peak calling

I developed a custom procedure for peak calling from single-base resolution strand-specific sequencing experiments such as TV-PRO-seq. Rather generically, I require that the transcription level μ at a peak exceeds a threshold value Q_{bio} which depends on local fluctuations:

$$\mu \geq Q_{bio}. \qquad (1)$$

The actual procedure is based on the aggregated reads from all the experiments at different run-on times and for a specific position (hereafter, such total reads per nt will be simply referred to as the "total reads") and is detailed steps below:

Step1: A threshold *t* for the minimum number of reads on each single genomic position was set. More precisely, genomic positions with total reads higher than *t* were selected as 'candidate peaks' for further analysis. The basic threshold *t* has been heuristically set to 13 and will vary with sequencing depth (Type 1 peaks in Figure 2.2A have been excluded). In addition to this, I discard the candidate peaks if the number of reads is zero for all the replicates corresponding to a single one run-on time, at least (Type 2 peaks in Figure 2.2A have been excluded).

Step2: I address the fact that some polymerase pausing regions are wider than one nt [26]. An example of such a dispersed pausing region is illustrated in Figure 2.2A, within a 50nt fragment of plus strand of chromosome 1. In Figure 2.2A, we consider the position with most reads in the dispersed pausing region. To deal with this, I exclude a 'candidate peak' if another 'candidate peak' has more reads in its +/- three-nt neighbourhood (Type 3 peaks have been excluded). This ensures that only a single position is selected from a dispersed peak.

.For highly expressed genomic regions, it is likely that some positions have a large number of reads (viz., higher than the threshold *t*) and pass selection step 1, even if they correspond to regions with constant elongation rate and do not have significant pausing. Similarly, along the same non-pausing regions, step 2 returns the genomic positions that have the highest amount of reads, even if this is just due to random fluctuations. As an example, the genomic position 632561 in the fragment illustrated in Figure 2.2A corresponds to such a case. Therefore, a third step is necessary to filter the candidate peaks that are likely to be located in a region of constant elongation rate but cannot be discarded during steps 1 and 2. I perform a two-step procedure as explained below.

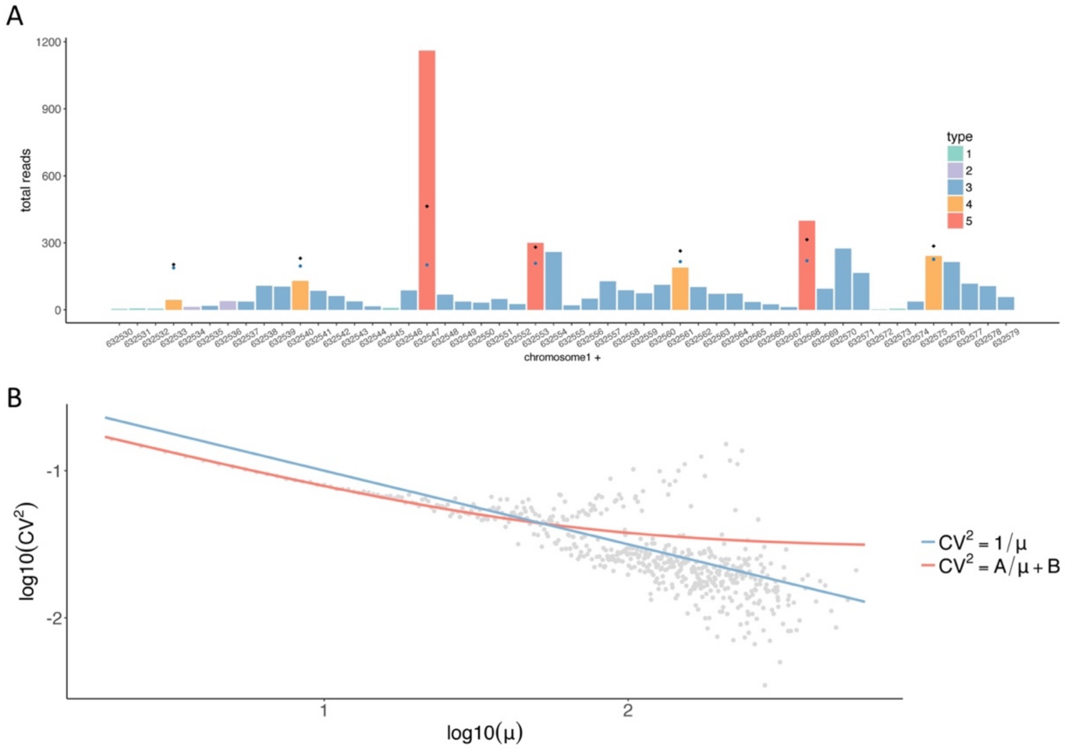


Figure 2.2 Peak calling

A. *A 50nt fragment of chromosome 1's plus strand. Positions that are excluded according to the various criteria explained in the main text are colour-coded into types 1-4. Red peaks (type 5) are identified as pausing sites for further analysis.*

B. *Scatterplot for sequencing noise. The red line represents a weighted nonlinear least-square fit $CV^2=A/\mu+B$, with parameters (A, B) = (0.53, 0.009), estimated by means of the random-search algorithm of the nls2 R package*[71]*. The blue line is the Poisson-predicted noise curve $CV^2=1/\mu$.*

Step3: The first sub-step consists of assessing the local biological fluctuations in the polymerase occupancy and deriving the threshold Q of condition (1). I assume that the polymerase occupancy in a constant elongation-rate region follows the Poisson distribution with parameter b. As the average elongation rate across the mammalian genome is about 33.3nt/sec [3], I expect that, in such non-pausing regions, all the polymerases are released by the time of the first run-on experiment/time-point (i.e., 30 seconds); therefore, for these regions, the differences observed between experiments at different run-on times are presumably due to statistical fluctuations, suggesting that we can actually ignore the dependence on run-on time and aggregate the reads across all experiments. I then focus on the reads across the +/-100nt neighbourhood around each candidate peak. Their mean read number, averaged over both the replicates and the 201nts, yields the expected number of reads b per nt[1] (in the neighbourhood). Based on a null local Poissonian assumption, as if reads were Poisson distributed with rate b, I associate an upper quantile Q_{bio} to each neighbourhood, where the value of q is heuristically chosen to control the number of (false positives) bases whose read number exceeds Q_{bio} purely due to statistical fluctuations. My (rather conservative) choice would be to allow only one false positive in the whole 'active genome'. I define the latter as all positions with at least one read. Since from my experiment there are 111868728 such bases, I heuristically set q=1/111868728.

Step4: We need to assess the sequencing noise as a function of the transcription level. To this end, I sequenced one of the replicates (specifically, the second 32-minute run-on replicate) twice, and trimmed the technical replicate with the highest total alignment reads to the same level as the other one. This trick gave me two replicates of identical total aligned reads, from which we computed the average reads for each nt. Further, by gathering the positions whose average read equals a certain number μ and computing their CV^2 I obtain the scatter plot of Figure 2.2B, which appears to closely follow the fitted standard noise model CV^2 = A/μ + B, and which can be expressed as

$$\varepsilon_\mu \sim \mathcal{N}\left(0, \sigma^2(\mu)\right),$$

[1] b is ideally estimated from the sample mean of read numbers at each of the 201 positions; however, many peaks are close to the TSS, which has many more reads downstream than upstream. To take account of this asymmetry, I assume that all the reads are downstream and average over the half-interval. This overestimates the background noise and is thus a conservative estimate.

where

$$\sigma^2(\mu) = A\,\mu + B\mu^2 \qquad (2)$$

(As an example, see Figure 2.2B for the empirical distribution of the reads centred at μ=20 alongside its Poisson and normal fit). Based on this model, the (observed) peak read is randomly drawn from

$$X = \mu + \varepsilon_\mu \qquad (3)$$

from which it follows that selecting the candidate peaks with more reads than the 0.99th quantile Q_{seq} of the normal distribution centred at Q_{bio} with variance $\sigma^2(\mu)$ satisfies condition (1) with probability 0.99,

$$Q_{seq}=\{x\text{:Prob}(x>Q_{bio}+\varepsilon_\mu)=0.99\}$$

Since we don't know the value of μ to insert into equation (2), we replace it with either Q_{bio} or the peak read number itself; the first choice underestimates Q_{seq} as $Q_{bio} < \mu$ (for all the non-trivial cases) and hence $\sigma^2(Q_{bio}) < \sigma^2(\mu)$, while the second choice has not such a bias as X is centred at μ. It is worth noting that there is an alternative but equivalent choice: one can compute the lower quantile of the distribution centred at the peak read x, $Q'_{seq}=\{q\text{: Prob}(q < x+\varepsilon)\}$, and require that $Q'_{seq} > Q_{bio}$.

In conclusion, we incorporate the polymerase noise model of point 3.1 and the sequencing noise model of point 3.2 into condition (1) by choosing the candidate peaks such that $x \geq Q_{seq}$, where Q_{seq} depends on Q_{bio} (Type 4 peaks have been peak excluded).

2.2.5 Calculation of pausing time

In this section, we derive a simple Bayesian model for TV-PRO-seq data and a procedure for their analysis on server CyVerse[72]. The mathematics and modelling parts in this section were carried out jointly with Massimo Cavallaro.

We are interested in the stochastic dynamics of biotin-NTP incorporation into a nascent mRNA which can be represented as the following simple reaction:

$$\text{nascent mRNA} + \text{biotin- NTP} \rightarrow \text{biotin} - \text{labelled mRNA}.$$

Such a reaction corresponds to one transcription step and is specific to the genomic position of the incorporation of the 3'-end nucleotide of the nascent mRNA. Assuming that the biotin-NTP population is large and remains constant during the reaction progress, we obtain

$$\text{nascent mRNA} \overset{\beta_i}{\rightarrow} \text{biotin- labeled mRNA},$$

which occurs at constant single-nucleotide transcription rate β_i. The average time that the Pol II spends on the base i is the reciprocal $1/\beta_i$, which we refer to the pausing time.

Let $y_i(t)$ and $x_i(t)$ denote the average populations of nascent-mRNA and biotin-labelled mRNA (specific to the genomic position i), respectively. The following rate equation is satisfied:

$$\frac{\mathrm{d}}{\mathrm{d}t} x_i(t) = \beta_i y_i(t).$$

As the presence of the biotin prevents further elongation and no new transcription is initiated, $y_i(t)$ naturally decays according to

$$\frac{\mathrm{d}}{\mathrm{d}t} y_i(t) = -\beta_i y_i(t).$$

Solving this simple system of ODEs with initial conditions

$$x_i(0) = 0,$$

$$y_i(0) = A_i,$$

yields

$$x_i(t) = A_i(1 - \mathrm{e}^{-t\beta_i}),$$

$$y_i(t) = A_i \mathrm{e}^{-t\beta_i},$$

predicting that the average population of the biotin-labelled mRNA increases up to the saturation point A_i while the unlabelled nascent mRNA is depleted according to exponential law.

Our analysis focuses on a subset of genomic positions $i \in S$, which we refer to as *peak* positions, where transcription level saturates to A_i at rate β_i. We speculate that a large number of genomic positions displays negligible pausing with Pol IIs stepping forwards shortly after biotin-NTP treatment and with transcription level concentrating around A_{bck}. We refer to such positions as *background*. Therefore, the expression level of the whole genome $x_{\mathrm{tot}}(t) = \sum_{i\in S} x_i\,(t) + x_{\mathrm{bck}}(t)$ grows according to

$$x_{\mathrm{tot}}(t) = \sum_{i\in S} A_i\ (1 - \mathrm{e}^{-\beta_i t}) + A_{\mathrm{bck}}(1 - \mathrm{e}^{-\beta_{\mathrm{bck}} t}).$$

While we have a model for the average transcription level $x_i(t)$ at genomic position $i \in S$ and run-on time t, the average number of reads $N_i(t)$ depends on the sequencing depth $\kappa(t)$, which is different for each sequencing experiment and therefore depends on the run-on time t, i.e.,

$$N_i(t) = \kappa(t) A_i (1 - \mathrm{e}^{-\beta_i t}).$$

It is convenient to study the ratio $\mathsf{x}_i = N_i(t)/N_{\mathrm{tot}}(t)$, where $N_{\mathrm{tot}}(t) = \kappa(t) x_{\mathrm{tot}}(t)$, as the dependence on $\kappa(t)$ cancels out. This represents the expected number of reads from the region of interest (e.g., from a peak position) normalised to the average total-genome reads at the same run-on time t.

We obtain the normalised model

$$\mathsf{x}_i(t) = \frac{x_i(t)}{x_{\mathrm{tot}}(t)} = \frac{(1 - \mathrm{e}^{-\beta_i t})}{\sum_{j\in S} \rho_{ij}\ (1 - \mathrm{e}^{-\beta_j t}) + \rho_{i,\mathrm{bck}}(1 - \mathrm{e}^{-\beta_{\mathrm{bck}} t})}, \qquad i \in S,$$

where $\rho_{ij} = A_j/A_i$ and $\rho_{i,\mathrm{bck}} = A_{\mathrm{bck}}/A_i$. We will later consider an approximated choice where the growth curve $x_{\mathrm{tot}}(t)$ is described by a single effective rate β_{tot}.

The quantities $\mathsf{x}_i(t)$, $i \in S$, can be organised into an $|S| \times T$ matrix X where T is the number of predictor observation run-on times. This allows us to use the compact notation

$$\mathsf{X} = (1 - \mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{t}}) \circ \left[\varrho\,(1 - \mathrm{e}^{-\beta^{\mathrm{T}}\mathbf{t}}) + \varrho_{\mathrm{bck}}^{\mathrm{T}}\,(1 - \mathrm{e}^{-\beta_{\mathrm{bck}}\mathbf{t}})\right]^{\circ -1}, \qquad (4)$$

where $\mathbf{t} = (t_1, t_2, \dots, t_T)$ is the vector of predictor observation run-on times, $\beta = (\beta_1, \beta_2, \dots, \beta_{|S|})$ is the vector of rates, $\varrho = \{\rho_{ij}\}$, $i, j \in S$, and $\varrho_{\mathrm{bck}} = (\rho_{1,\mathrm{bck}}, \rho_{2,\mathrm{bck}}, \dots \rho_{|S|,\mathrm{bck}})$ incorporates the relative saturation points. The notation $A \circ B$ is the Hadamard (element-wise) product of A and B while $A^{\circ -1}$ is the Hadamard inverse of A.

To simplify this model, we use the naïve form

$$N_{\mathrm{tot}}(t) = \kappa(t)x_{\mathrm{tot}}(t) = \kappa(t)A_{\mathrm{tot}}\left(1 - \mathrm{e}^{-\beta_{\mathrm{tot}}t}\right)$$

to approximate the growth of the average of total reads. As in the previous section, the mitochondrial chromosome can be thought of as being constant to $x_{\mathrm{chrM}} = \kappa(t)A_{\mathrm{chrM}}$ to a first approximation. We use them as a reference level. We divide the total reads by the chromosome-M reads and fit the model

$$\frac{x_{\mathrm{tot}}(t)}{x_{\mathrm{chrM}}} = \rho_{\mathrm{chrM,tot}}\left(1 - \mathrm{e}^{-\beta_{\mathrm{tot}}t}\right),$$

where $\rho_{\mathrm{chrM,tot}} = A_{\mathrm{tot}}/A_{\mathrm{chrM}}$, to such data using the random-search algorithm of the nls2 R package [71], which returned a significant fit with estimated parameters reported in the table below, see also Figure 2.3.

	Estimate	Std.err.	t value	Pr(>|t|)
$\rho_{\mathrm{chrM,tot}}$	46.621	2.769	16.839	0.000
β_{tot}	0.760	0.176	4.322	0.005

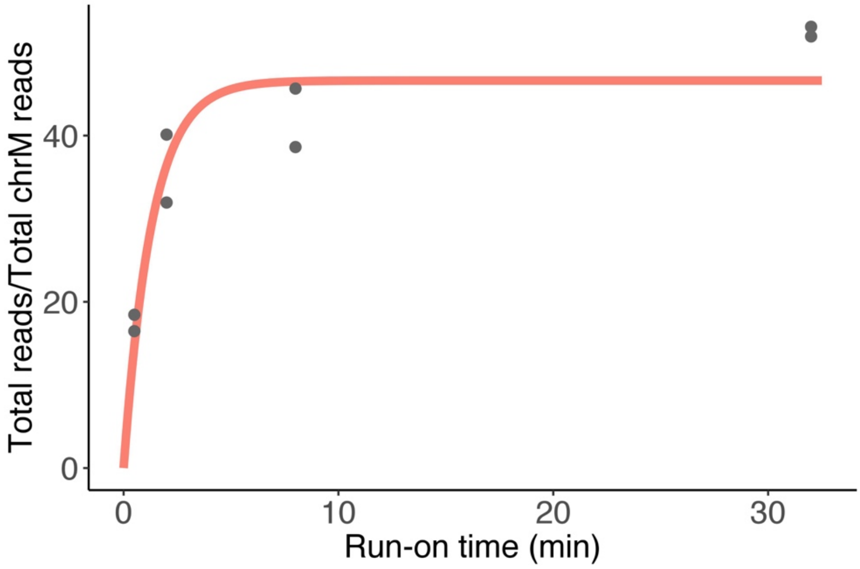


Figure 2.3 Saturation plot of the total-genomic reads normalized to the total-chrM reads.

Based on this consideration, our choice is to use the exponential model to approximate the growth of the average total-genome reads $N_{\text{tot}}(t)$, and study

$$\mathrm{x}_i(t) = \frac{1}{\rho_{i,\text{tot}}} \frac{(1-\mathrm{e}^{-\beta_i t})}{(1-\mathrm{e}^{-\beta_{\text{tot}} t})}, \tag{5}$$

where $i \in S$ and $\rho_{i,\text{tot}}$ are parameters fixed by data. In matrix form, we get

$$\mathrm{X} = (1 - \mathrm{e}^{-\beta^{\mathrm{T}} \mathbf{t}}) \circ \left[\varrho_{\text{tot}}^{\mathrm{T}} (1 - \mathrm{e}^{-\beta_{\text{tot}} \mathbf{t}})\right]^{\circ -1}, \tag{6}$$

where

$$\varrho_{\text{tot}} = (\rho_{1,\text{tot}}, \rho_{2,\text{tot}}, \dots \rho_{|S|,\text{tot}}).$$

We then chose the informative prior

$$\beta_{\text{tot}} \sim \text{Gamma}(1.1,1.1),$$

where $\text{Gamma}(\alpha, \beta)$ represents the Gamma distribution with mean α/β and variance α/β^2, which places substantial mass around 1 and little mass around 0^+. The peaks must have an average rate of the same order as the total growth rate, although the rates corresponding to pausing elements can be significantly smaller. Based on such a heuristic consideration we choose the informative priors

$$\beta_1, \beta_2, \dots, \beta_{|S|} \overset{i.i.d.}{\sim} \mathrm{Gamma}(0.1,0.1),$$

which have mean and variance equal to 1 and 10, respectively, and place a lot of mass at 0^+.

The next steps consist of incorporating noise and thus defining a Bayesian model to be fitted. We incorporate the noise in the model as follows. The sequencing reads are obtained after several amplification steps and are restricted to be positive. Hence we assume that the observables Y are subjected to multiplicative errors with lognormal distribution, i.e.,

$$\mathrm{Y} = \mathrm{X} \cdot \epsilon,$$

where

$$\log\epsilon \sim \mathcal{N}(0, \sigma^2).$$

As $\epsilon = \mathrm{e}^{\sigma Z}$ with $Z \sim \mathcal{N}(0,1)$, we get

$$\log\mathrm{Y} \sim \mathcal{N}(\log\mathrm{X}, \sigma^2).$$

To empirically guess a prior distribution for σ given the coefficient of variation of Y, we use the error-propagation formula

$$\mathrm{CV}^2\mathrm{Y} \approx \mathrm{CV}^2\epsilon,$$

where $\mathrm{CV}^2\mathrm{Y}$ is estimated from aggregated data. As ϵ is lognormal, we have

$$\mathrm{CV}^2\epsilon = \mathrm{e}^{\sigma^2} - 1,$$

and

$$\sigma^2 \approx \log[\mathrm{CV}^2\mathrm{Y} + 1],$$

which suggests the prior

$$\sigma \sim \mathrm{Gamma}(1.6,0.4).$$

An MCMC sampler to fit the model was implemented using the PyMC3 Library for Bayesian Statistical Modeling and Probabilistic Machine Learning [73]. PyMC3 relies on the **Theano** framework [74], which allows fast evaluation of matrix expressions, such as those in equations (4) and (6), and offers the powerful NUTS sampling algorithm to fit models with thousands of parameters. Nevertheless, we aim to infer the growth rate of up to ~ 170000 peaks. To ease the computational burden, we divide the peak list into chunks of ~ 3000 randomly chosen peaks. Further, we averaged the reads over the replicates, and the averages at 32 minutes of run-on time are used as saturation levels.

In addition to the estimates of the peak rates, the method returns estimates of β_{tot} from each chunk. These are very close to the rate of 0.1 min^{-1} obtained from the half-life measured in Jonkers, Kwak, and Lis [3]. Aggregating the individual-chunk estimates using the laws of total mean and variance yields:

$$\beta_{\mathrm{tot}} = 0.147 \pm 0.007 \text{ min}^{-1}.$$

To assess the sensitivity with respect to the prior distribution, we also ran the inference procedure using the vague prior distributions:

$$\beta_1, \beta_2, \dots, \beta_{|S|}, \beta_{\mathrm{tot}} \overset{i.i.d.}{\sim} \mathrm{Gamma}(0.001{,}0.001),$$

which results in a wider range of inferred β_i, whilst maintaining the same rank order.

2.3 Results

2.3.1 Principle of TV-PRO-seq

The procedure of TV-PRO-seq is based on PRO-seq[65]. As shown in Figure 2.1F, biotin labelled NTPs will replace the native NTPs to become incorporated into the 3' ends of nascent RNAs. The biotin will block transcription, thus the position +1nt with regards to the polymerase's position will be marked. If the polymerase remains at the pausing site during run-on period without moving, the biotin-NTP will not be added on the nascent RNA transcript by this particular polymerase. Increasing the run-on time allows more paused polymerase to become released from the pausing site, until eventually all paused polymerases will have become released (Figure 2.4A). The quicker polymerase releasing / shorter polymerase pausing, the faster nascent RNA will be labelled (Figure 2.4B). Thus, we can estimate pausing time by fitting saturation curves to the build-up of sequencing reads over the run-on time series at a particular position.

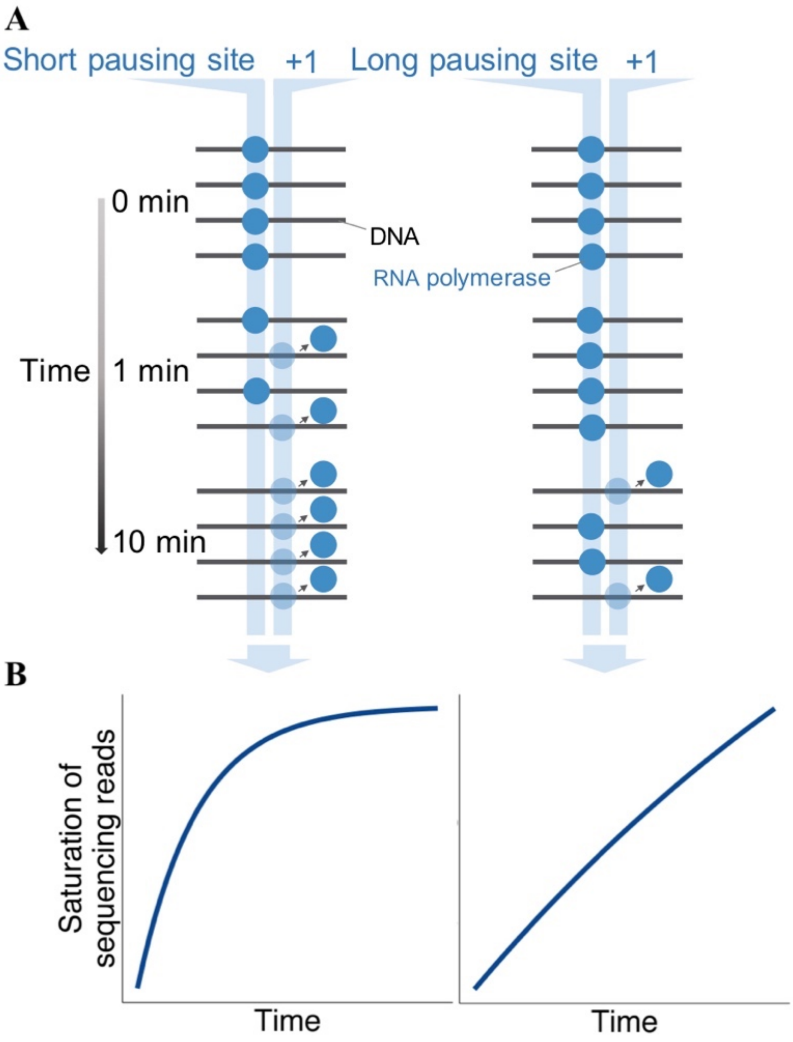


Figure 2.4 Principle of TV-PRO-seq

A. *Black lines represent template DNA and blue dots symbolize RNA polymerase. All polymerases are paused on pausing sites at the start (0 min) of the run-on period. Polymerases released from pausing site will be blocked by biotin-NTPs at the position one base downstream (+1) and drop off the DNA templates.*

B. Saturation curves of the example cases shown in (A).

2.3.2 Unique design features of TV-PRO-seq

In principle, TV-PRO-seq consists of 8 parallel PRO-seq reactions with four different run-on times as duplicates. The first run-on times I used where 3min, 6min, 12min and 24min in KBM7 cells. Because of the high level of biological and technical noise, I increased the ratio between neighbouring time points to 4 times. The final time points set is 0.5min, 2min, 8min and 32min. The main procedure of TV-PRO-seq is the same as PRO-seq's[65], although several modifications were made for TV-PRO-seq. PRO-seq entails three biotin enrichment and two RNA extraction steps before PCR amplification and nascent RNAs are only a small proportion of total RNAs. As a result, the library preparation is always struggling with low yields. TV-PRO-seq is based on preparation of 8 samples which requires a more robust procedure. I therefore removed the based the precipitation of RNA from Trizol extractions on 1X isopropanol instead of 2.5X ethanol and removed the washing step with 75% ethanol. Even though the purification steps were removed, TV-PRO-seq still yielded high quality result; more than the 80% of the trimmed reads could be aligned to the hg38 genome.

TV-PRO-seq is based on the assumption that the polymerase release rate of a certain position is fixed. Ideally, all the cells should be in the same condition before run-on. To reduce the variability between each sample of TV-PRO-seq, all cells for 8 samples were derived from the same tissue culture flask and were permeabilized together. PCR amplification was set to 17 cycles. For PAGE-purification, the primer and library DNAs above 700bp were removed first (Figure 2.5, the region outside of the dotted frames were discarded). After recovering the DNA from the PAGE gels, I mixed equal amounts of DNA from each sample together according to Qubit results. The pooled samples were PAGE purified again, this time discarding DNAs sized between 500bp to 700bp (Figure 2.5, DNAs above the solid were discarded), and the purified pooled library was sent for sequencing. The double-stepped purification prevents that the gel excision introduces a size bias of library DNAs. To obtain the best quality of my analysis results, I also trimmed the aligned reads of each sample to the same numbers. For HEK293 data (4-biotin run-on), each sample was trimmed to ~50million reads. For S2 cells (4-biotin run-on), the number was ~13million. KBM7 data (2-biotin run-on) was not trimmed; the total aligned reads of 8 samples are: 68.0million, 32.9million, 69.4million, 68.5million, 71.7million, 66.1million, 98.9million and 76.8million (Ordered by time and replicates as 0.5min R1, 0.5min R2 … 32min R2).

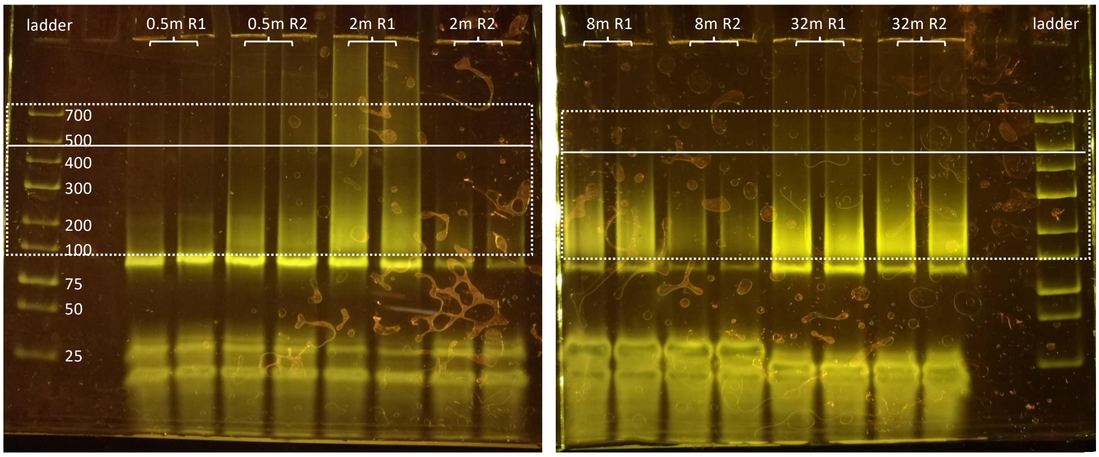


Figure 2.5 PAGE-gel for library purification of TV-PRO-seq

Native PAGE-gel for TV-PRO-seq. Each sample had been separated into two tubes for PCR. Gel pieces corresponding to a single tube of PCR product were purified and then loaded on to a single lane of a second PAGE-gel. 25bp DNA ladders were loaded at the side of each gel.

Since fluorescence signals of Illumina sequencing are generated un-synchronously at the start and end of the sequencing cycle, the quality at the (both) ends of reads sometimes is lower than the central part[75]. To remove this effect, read alignment software such as hisat2 will typically clip the end of reads prior to alignment[76]. This soft clipping will improve the aligning rate and accuracy but will unfortunately also lead to falsely reported locations of read ends. As the example analysis of published data[44] in Figure 2.6 shows, the 5' end of reads corresponds to the biotin labelled 3' ends of nascent RNAs. One nucleotide soft clipping will move the location of the 5' end of reads 1nt further 3'. Thus, I designed a custom script to extend the soft clipping reads back to the real pausing sites.

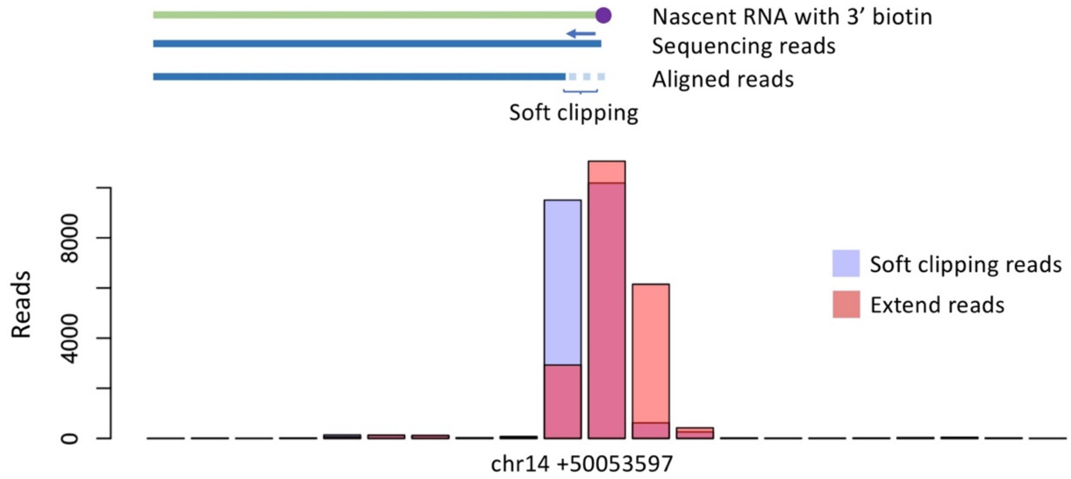


Figure 2.6 Soft clipping leading to false locating of pausing site

The soft clipped reads aligned by hisat2 are shown as the purple bar, the reads extended by the custom script were shown in the red bar. An example of a read with a 3' end at chr14 + 50053597 with one nucleotide soft clipping is shown. The soft clipping moves the 3' end of reads one nucleotide upstream of the real pausing release site, which the script reverts.

2.3.3 Evaluating pausing time by TV-PRO-seq result

After completing a TV-PRO-seq assay, each genome position will have 2 PRO-seq read numbers for 4 time points each. These numbers cannot be used for curve fitting directly, since the number of PRO-seq reads is not only related to polymerase occupancy but also influenced by the sequencing depth. As more polymerase will become released with increasing run-on time, the amount of labelled RNAs of each cell of early time points will be lower than later ones. Thus, directly normalizing peaks by total genome reads will be biased as well.

POLRMT (RNA Polymerase Mitochondrial) is a highly processive single subunit polymerase[77, 78]; I therefore assumed that pausing on mitochondrial DNA is shorter and reads will saturate quicker, resulting in approximately constant numbers of labelled nascent RNAs of chrM. I therefore used the total reads of chrM for normalization of read numbers. As shown in Figure 2.7, the total genome reads normalized by chrM reads display a saturation curve in accordance with theory. The polymerases at peaks (higher than Q_{bio} defined in 2.2.4) of chrM are released slower than the backgrounds (lower than Q_{bio}). Thus, the denominator of ‘Peaks’ (Fig. 2.7) is comparatively small and results in higher counts in early time points.

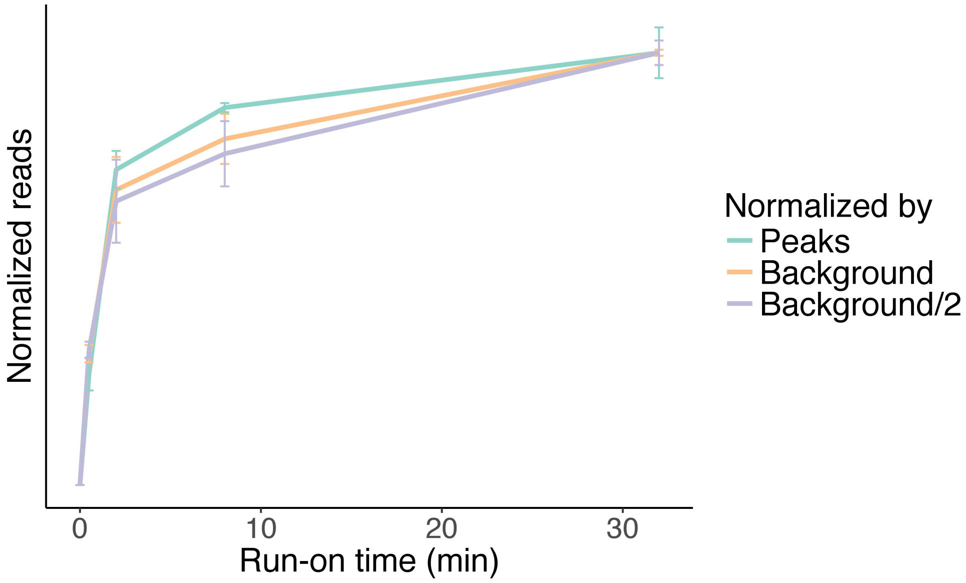


Figure 2.7 Total-genome reads of TV-PRO-seq samples normalized by chrM reads and selected at different heuristic thresholds (Background/2, Background, Peaks).

Peaks refers to the total reads number of positions with reads bigger than its Q_{seq} (See 2.2.4). Background refers to positions with reads lower than its Q_{seq}, and Background/2 means lower than *$Q_{seq}/2$. As polymerases release slower on the peak than background, the ratio of total reads of earlier time points comparing with last time points of peaks is lower than the background. Thus total genome reads/Peaks is bigger than total genome reads/Background in the earlier time points.*

M. Cavallaro developed a curve fitting script based on a Bayesian framework. Two example peaks of curve fitting result are shown in Figure 2.8. As the saturation curves of the red peak (chr21 + 8402177) grow slower than the blue one (chr21 + 8402194), the pausing time of the red peak is proposed to be longer.

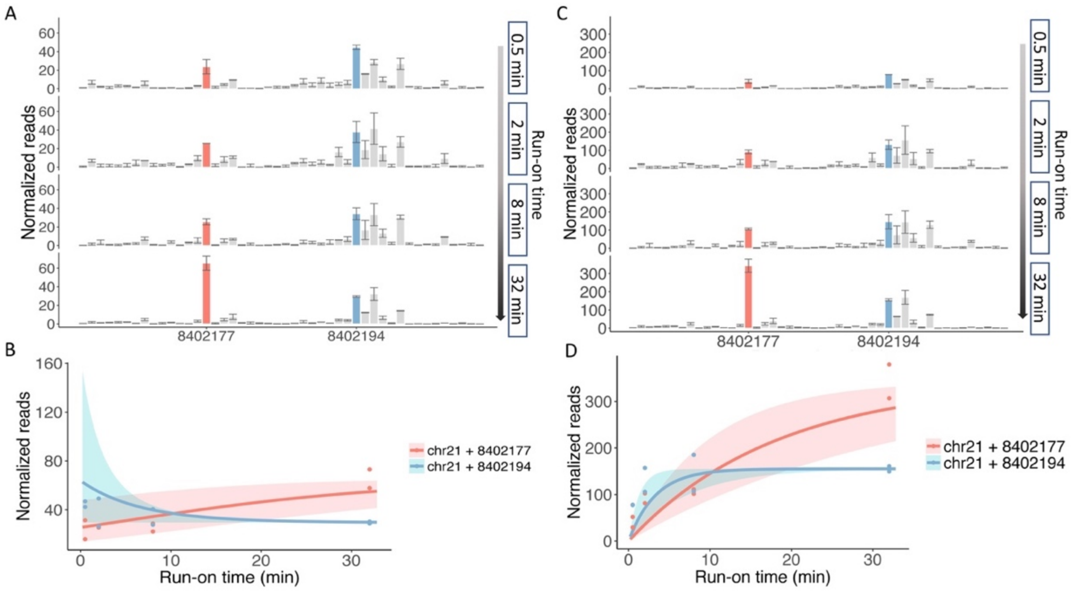


Figure 2.8 Example of pausing time estimation by TV-PRO-seq

A. *Reads are normalised by total-genome reads and rescaled by 10^7. The height of the bar is the mean of two replicates, the error bars correspond to the data range of two replicates.*

B. *Curve fitting result of (A); the shaded regions indicate lower and upper quartiles.*

C. *Reads are normalised by total chrM reads and rescaled by 10^6. The height of the bar is the mean of two replicates, the error bars correspond to the data range of two replicates.*

D. *Curve fitting result of (C), the shaded regions indicate lower and upper quartiles.*

2.4 Discussion

TV-PRO-seq is the first method that can estimate genome wide polymerase pausing times at single nucleotide resolution. FRAP[6, 47] cannot identify the genomic location of polymerases; nascent transcription RNA FISH[48] only is feasible for a small number of genes at low positional resolution; Trp treatment following sequencing[3, 4, 49] produces results at low resolution as well and are affected by slow Trp uptake[50].

The biggest challenge in the interpretation of sequencing data to investigate pausing is the removal of the influence of polymerase flux towards polymerase occupancy to obtain the average residence time of polymerase (See 1.2.2). The ‘pausing index’ was devised to achieve this. As polymerase flux has been suggested to be approximately constant throughout a gene, the polymerase density in ‘non-pausing regions' has been assumed to be positively correlated with the gene's expression level. This should permit using the polymerase density of the ‘non-pausing regions' to normalize the polymerase density of the ‘pausing regions' to correct for the polymerase flux's influence. The pausing index is based on these notions and is typically targeted at the high-occupancy PPR, whose signal is normalized by polymerase density downstream of it to calculate the index. The resulting values are thought to reflect the average residence time of polymerase and thus make pausing at different genes comparable. However, polymerase flux within the same gene is not always constant. Polymerase can pause, backtrack and even drop off the DNA template during transcription[6, 14, 51]; and the TSS and TES are also variable at many genes[44]. The estimation of pausing times by TV-PRO-seq is independent of polymerase occupancy, which allows us to ignore the influence of the complex confounding factors associated with the transcription process, such as alternative TSSs etc, and focus on pausing.

Some types of genes such as those coding for tRNAs and lncRNAs are too short to contain a ‘non-pausing region’ for calculation of the pausing index (Typically, this region is defined as the sequence from 500nt downstream of the TSS to the TES). As TV-PRO-seq estimates pausing times from data of a single genome position, the ‘non pausing region’ is not necessary for TV-PRO-seq. TV-PRO-seq thus provides a tool to research pausing in short genes.

As polymerase occupancy is composed of polymerase flux, pausing time and pausing fraction (Chapter 1.2 and Figure 1.2), TV-PRO-seq provides the second piece of the puzzle for dissection of polymerase pausing.

TV-PRO-seq is not free of limitations. TV-PRO-seq is based on PRO-seq which is performed *in vitro*, thus not an optimal reflection of the *in vivo* situation. As the biotin-NTP uptake also takes some time, the run-on time is not strictly the same as the polymerase pausing release time. Furthermore, biological variability may still introduce differences between samples, even if the permeabilized cells were prepared together. Technical noise that accumulates during the long and laborious experimental preparation and sequencing noise influence the data strongly. Although meta-analysis of TV-PRO-seq data across the genome gives statistically significant results, individual pausing time estimates, especially for pausing sites with lower reads, have to be treated with caution.

Chapter 3 Pausing in promoter proximal region

3.1 Introduction

The most conspicuous phenomenon which has been extensively researched in the polymerase pausing area is polymerase enrichment in the PPR (promoter proximal region) of metazoans. It was suggested to be caused by long duration of polymerase occupancy in this region[21, 22] (Figure 1.3D). The molecular principle of pausing in PPR and its biological function have been well demonstrated.

After transcription of the 5' end of the RNA and its capping, promoter proximal pausing of Pol II is found within 20nt to 60nt downstream of TSS[2] and was shown to involve several transcription factors; DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor) and P-TEFb (Positive transcription elongation factor b) are the key players. NELF and DSIF establish pausing in the PPR, and depletion of either will significantly reduce the polymerase occupancy in the PPR[1, 79]. P-TEFb promotes the release from pausing by phosphorylating NELF, DSIF and Ser2 of the Pol II CTD (Carboxy-Terminal Domain)[2, 79, 80] (Figure 3.1). Upon phosphorylation, NELF will be released from Pol II while DSIF will remain bound to Pol II but will have the opposite function[81, 82]. Inhibition of P-TEFb by FP (Flavopiridol) will prevent polymerase from becoming released into productive transcription, and this inhibition can be observed in nearly all active genes[3, 30, 83]. The widespread effect caused by FP treatment suggests the general importance of promoter proximal pausing in gene expression regulation.

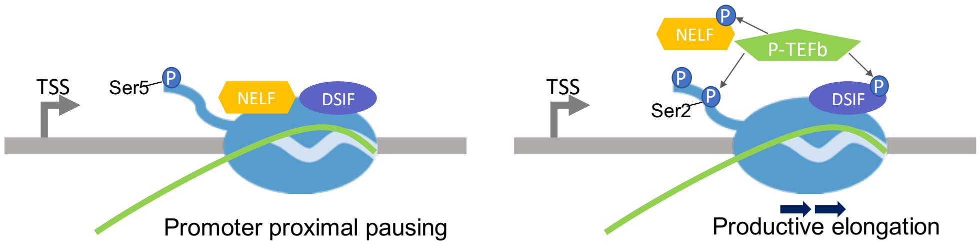


Figure 3.1 Mechanism of promoter proximal pausing

The grey line represents the template DNA, the green line shows the nascent RNA, and the blue ellipse with tail represents Pol II. 'P' in circles refers to phosphorylation. The left plot displays NELF- and DSIF-established promoter proximal pausing, and the right one shows how P-TEFb releases paused polymerase by phosphorylation of NELF, DSIF and Ser2 of the CTD.

However, longer residence time of polymerase can lead to higher polymerase occupancy, but higher polymerase occupancy is not necessarily caused by polymerase pausing. Two example peaks are shown in Figure 3.2, including their expected theoretical polymerase occupancy and pausing times and how these would appear in NET-seq and TV-PRO-seq data. In Figure 3.2A, two peaks are set to have the same pausing time, pausing fraction and polymerase flux (for definitions of terms, see 1.2). Figure 3.2B shows that if the pausing time of peak 1 were 5 times that of peak 2, the polymerase occupancy also became 5 fold different. However, differences of polymerase flux and pausing fraction between peak 1 and 2 will result in similar effects of polymerase occupancy (Figure 3.2C, D). Additional evidence is needed for establishing long pausing of polymerase in the PPR.

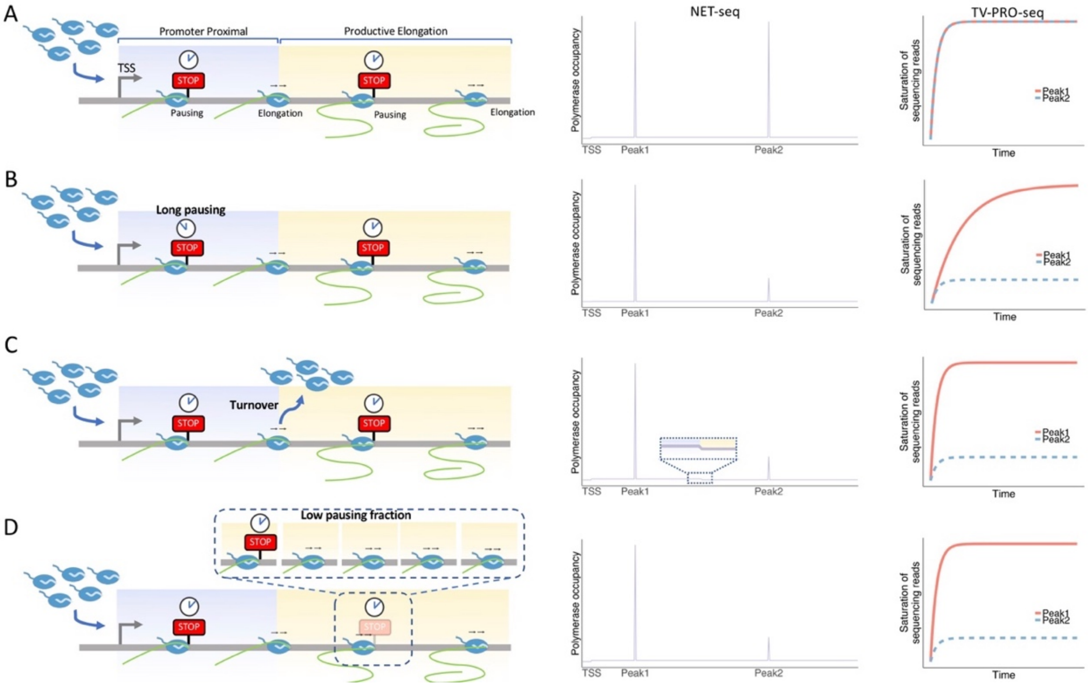


Figure 3.2 Polymerase occupancy and pausing time are influenced by different features

A. A schematic example is shown in the left panel; the PPR is shown with blue shading and productive elongation with beige shading; a single peak with identical pausing time, pausing fraction and polymerase flux has been set in each of the two regions. The polymerase occupancies measured by NET-seq (middle panel) and pausing times measured by TV-PRO-seq (right panel) are the same for both peaks.

B. As (A), but pausing time of peak 1 is set 5 times longer. Polymerase occupancy and pausing time of peak 1 will be measured to be 5 fold higher than for peak 2.

C. As (A), but 80% of polymerases are set to drop off at the end of PPR. Polymerase occupancy of peak 1 will be measured to be 5 fold higher than for peak 2. However, pausing times of the two peaks will be the same.

D. As (A), but only 20% of polymerases pause (= pausing fraction) at peak 2. Polymerase occupancy of peak 1 will be measured to be 5 fold higher than for peak 2. However, pausing times of the two peaks will be nearly the same.

The notion of long pausing of polymerase in the PPR has been supported by Trp treatment followed by sequencing-based assays. Various studies show that after blocking transcription initiation by Trp, the reduction of polymerase occupancy in the PPR is slow (about 10min)[3, 4, 84]. However, the slow uptake of Trp[50] requires a reconsideration of these results. Recently, several researches have suggested that the polymerase flux at the PPR is higher than at downstream regions due to polymerase turnover and generation of abortive transcripts[5-7]. Median of promoter proximal pausing was suggested to last only about 42 seconds, while productive elongation takes 1370 seconds. Comparing the latter figure to the surprisingly short residence time of polymerases in PPR implies that more than 90% of initiated transcription will terminate in the PPR. This will lead to a huge polymerase flux bias[6].

TV-PRO-seq can measure pausing time independently from polymerase flux and pausing fraction (Figure 3.2C, D). This allows re-examination of promoter proximal pausing. My result suggests that pausing in the PPR is actually shorter than in other regions in human cells (Or sarkosyl specifically reduces the pausing time of polymerase in the promoter proximal region). FP treatment followed by NET-seq or (conventional) PRO-seq assays produces results consistent with TV-PRO-seq data. My findings further suggest that Pol II bound to NELF and DSIF actually pauses shorter.

3.2 Methods

3.2.1 Peak annotation to relative genes

Annotation of peaks identified in 2.2.4 was done in two ways for different analyses:

1. Annotation to 3’ and 5’ end of exons

This annotation gave the absolute distance of peaks towards a certain annotation site. It can be used for getting the distance of peaks towards TSS, TES or splicing sites.

Two different reference databases were used. For annotation of mRNAs, the reference list was downloaded from UCSC table browser and the parameters had been set as: assembly - hg38, group - mRNA and EST, table - UCSC RefSeq, output format - all fields from selected table[85]. A custom script *Unique_annotation_maker.pl* was created for transforming the reference list for further analysis. The output of the script is: Column 1 – chromosome of gene; Column. 2 – strand of gene; Column 3 – position of gene; Column 4 – name of gene; Column 5 – type of annotated sites; if equal to ‘start’, the annotation site is the 5’ end of the annotated exon, otherwise it is the 3’ end; Column 6 and Column 7 are the min and max number of exons in the genes in different variant, respectively; TES are specifically marked as -1; Column 8 – the number of variants of a transcript having this annotation site; Column 9 is the total number of transcript variants the gene has.

For ncRNAs (non-coding RNAs) like rRNAs and tRNAs, tables were downloaded from RNAcentral (https://rnacentral.org/). The RNA gene classification information and corresponding genomic locations were store in different files as *rfam_annotations.tsv* and *Homo_sapiens.GRCh38*.bed[86], respectively. A custom script *rFAM_annotation_merger.pl* was used for merging the two tables for further analysis.

The transformed reference lists were used for annotating peaks, which was carried out with another custom script, *Peak_annotater.pl*. It can annotate peaks within a specific distance to annotation sites. For instance, the command for annotating peaks in *beta_summary* to the annotation sites of *All_mRNA* in a +/-4500nt region is:

```
perl Peak_annotater.pl All_mRNA Beta_summary 4500
```

The peaks with specific annotations can be extracted from the output file. For example, the peaks annotated to TSSs of genes with a unique TSS is: type = start & number_max = 1 & hit = variant. 'Type = start' means the annotation is the start site of an exon; 'number_max = 1' means the annotation is in exon1 of all aligned variants; 'hit = variant' means all variants of a gene have this annotation site. If 'number_min = 1' and 'number_max > 1' it means at least one alternative TSS is located upstream of this TSS. If 'number_max =1 & hit < variant', it means an alternative TSS is located downstream of this TSS.

2. Annotation within genic regions

This annotation gives the absolute and relative positions of peaks with regards to the annotated region they are located in. For RNA transcribed by different types of RNA polymerase, different pipelines have been used: Pol I transcribed rRNAs except 5S were extracted from merged lists from RNAcentral generated as described above; the custom script *rFAM_region.pl* was used to transform the list into a. BED-like format. For mRNAs, the script UCSC2bed.pl was used. Pol III annotation came from published data[87]; the table 'Potential Pol3 targets' was converted to hg38 from hg19 with the *UCSC liftOver*[85] tool. The output BED-like files contain 6 columns: chromosome, TSS, TES, gene name, gene type / transcript ID and DNA strand. The custom script *Annotation_region.pl* was generated for annotating peaks with these annotation files in BED format.

For the exons of mRNAs, another pipeline was used. The UCSC annotation list was transformed with the custom script whole_gene_annotation_list_maker.pl. The output contains 9 columns as: gene name; chromosome of gene; strand of gene; TSS of gene; TES of gene; variant of gene; start site of exon; end site of exon; number of exons that appear in different variants of the gene. Another custom script *whole_gene_annotater.pl* was used to extract the location of peaks within the regions. Its output file contains the 9 columns of the annotation file plus the information relating to the peaks. Peaks in introns were recorded as 'hit = 0'. With this output file, we can obtain the absolute and relative distances of peaks to the boundaries of the regions they are in.

3.2.2 Analysis of rDNA repeats

As rDNA are highly repeated, a special strategy was used for analysis of these. A special Hisat2 index built from repeat-masked hg38 genome and a standard rDNA sequence[9] was used as reference for alignment. The pipeline of peak calling and pausing time estimation was the same as in 2.2.4 and 2.2.5. 5.8S rRNA data is absent from standard rDNA[9] because UCSC does not mask it in the hg38 genome, resulting in its multiple occurrence in my analysis pipeline.

3.2.3 Meta gene analysis of pausing peaks

6562 genes which have unique TSSs and TESs and are longer than 3000ntwere used for meta gene analysis. I classified the peaks into 7 regions: 1. Promoter, 2. TSS related region, 3. earlier intron, 4. exon, 5. later intron, 6. region before TES and 7. pA related region.

I obtained regions 1, 2, 6 and 7 from the annotations of 3' and 5' ends of exons from the list generated with `Peak_annotater.pl`.

Promoter: 1000bp region upstream of TSS

TSS related region: 1000bp region downstream of TSS

region before TES: 500bp region upstream of TES

pA related region: 4500bp region downstream of TES

The peaks in the introns and exons were annotated with `whole_gene_annotater.pl`, using the annotation list generated with `whole_gene_annotation_list_maker.pl`. Only exons and introns not overlapping with the first 1000bp or last 500bp of transcripts were selected. If the intron's centre position was in the first half of the gene, I considered an intron to be an early intron. Otherwise I regarded it as a later intron.

Because most exons or introns have different lengths, I normalized the peak densities before plotting. First, the peaks in introns and exons were annotated with the relative location, that is the distance between the peak and the 5' end of the annotated region, divided by the length of the annotated region. Then I calculated the average length for each region, and multiplied it with the relative location.

To show the pausing times of the 7 regions defined above, a smoothed conditional mean plot with loess fitting was generated by the ggplot2 *R* package with parameter span=0.1 (Figure 2B). I also separately plotted the smoothed conditional mean plot for the promoter and TSS related region only (Figure S5). Peaks around TSS and TES of tRNA genes were plotted in the same way (Figure 3A, Figure S8).

3.2.4 Analysis of Trp treatment PRO-seq data

Trp treatment can inhibit the initiation of transcription and perturb the dynamical balance of polymerase occupancy downstream of TSSs. The polymerase occupancy of pausing sites will reduce after Trp treatment. The quicker polymerase is released from pausing, the faster polymerase occupancy will reduce. Even though sequencing reads are also influenced by sequencing depth, the ratio of reads of peaks before and after Trp treatment can still reflect the relative pausing length.

3.2.5 PPR definition by FP treatment data

I defined the promoter proximal region [88] as the region downstream of TSS, whose polymerase occupancy increases after FP treatment. The PPR starts from TSS; I further define the 'fold change' := reads of FP treatment / reads of DMSO or NO treatment; the script `PPR_definer.pl` calculates the 3' boundary of PPR for each gene based on the following steps (Figure 3.3):

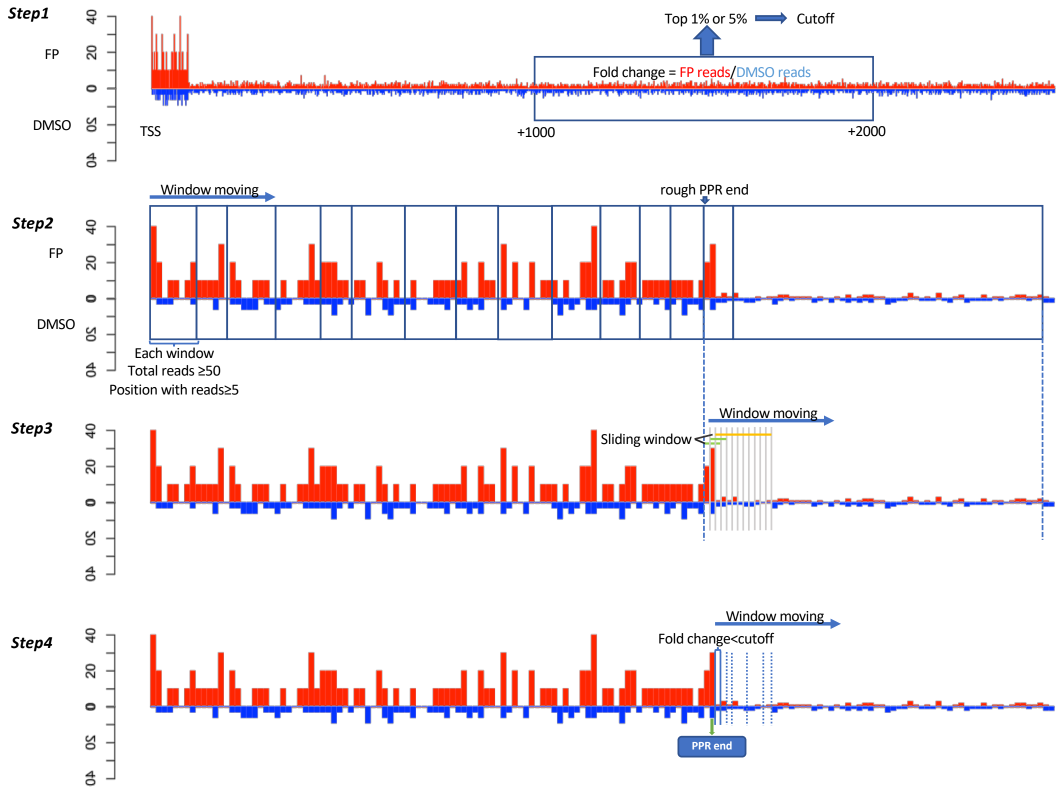


Figure 3.3 Defining the PPR.

Step 1. Get the ‘cutoff ’ value:

I assume that the region from TSS+1000 to TSS+2000 is sufficiently distant to the PPR and can thus use it as a negative control. I define the top 1% (5% for NET-seq data) of fold changes of all genes in the negative control region as *cutoff*.

Step 2. Get a ‘rough PPR end’:

In this step, the script finds a ‘rough PPR end’ based on the cutoff; this region’s boundary is downstream of the real PPR. The procedure is to set successive sequence windows in the 3’ direction from the TSS onwards. Because NET-seq/PRO-seq reads are sparse, each window needs to have at least 50 total reads and at least one read after FP treatment at its start and end positions and three other positions. The window size is flexible; its start and end positions are fixed once the criteria are met. Gaps inbetween two windows will be split in half and assigned to the adjacent windows. This ensures that the reads are not due to noise. Windows are being set as long as their calculated fold changes are above the cutoff. The beginning of the second last window is recorded as the rough PPR end.

Step 3. Zoom in:

The last window will have fold change < cutoff and the second last window fold change > cutoff. Both windows have the possibility to contain the precise end of the PPR. Therefore, I zoom into the last two windows; the script uses a sliding window in 3’ direction within this region, setting the two criteria as before to 10 and 3, respectively. It then compares the fold change for the total reads of the window and moves it to the next position with at least one FP read if the foldchange is still bigger than the cutoff. Once the fold change becomes smaller than the cutoff, the window stops.

Step 4. Get the PPR end:

The script then calculates the fold change of each position that has FP reads in the last sliding window. Once the fold change of a position becomes smaller than the cutoff, the PPR end is recorded as the position just 5’ of it.

3.3 Results

3.3.1 Profile of Pol III transcription

The peaks annotated to genes as described in section 3.2.1 were taken for further analysis. The peak density around TSSs is consistent with previous works (Figure 3.4). Sense peaks are enriched in the PPR region, and divergent transcription also results in enrichment of pausing sites. The distance of antisense peaks towards the TSS is greater than for sense peaks.

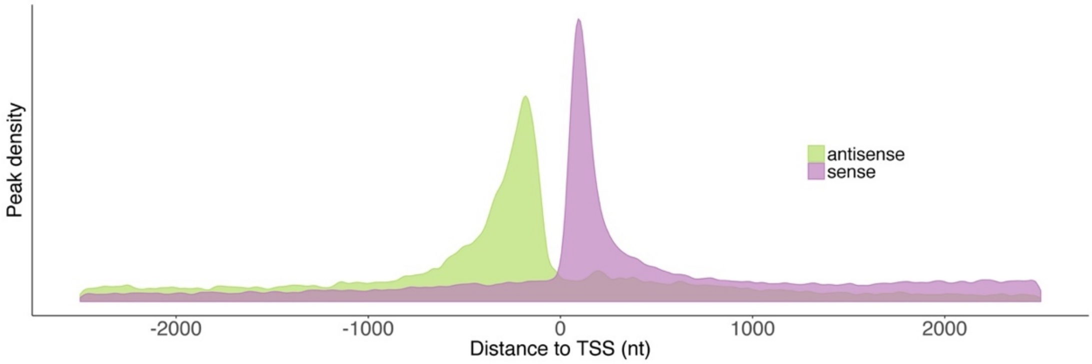


Figure 3.4 **Peak density around TSS**

The density plot shows the pausing sites of divergent transcription around TSS. Enrichment of pausing sites in the PPR are seen in both direction of transcription.

PRO-seq in principle detects nascent RNAs transcribed by all types of RNA polymerases[65]. By selecting regions according to the different polymerases they are transcribed by, we can identify the source of reads. As pausing time estimation by TV-PRO-seq is only based on the data of individual nucleotides, comparation of pausing lengths between different polymerase types is possible. The peaks annotated to chrM are transcribed by POLRMT (described in section 2.3.3). As expected, this highly processive single-subunit polymerase[77, 78] has shorter pausing times than Pol II (Figure 3.5A). Surprisingly, Pol III is the polymerase that pauses shortest, not POLRMT (Figure 3.5A). Pol III is responsible for about 20% of the nucleotide consumption in the nuclei[89]. It transcribes tRNAs, RNase P, RNase MRP and 5S rRNAs. Remarkably, these transcripts are mostly short noncoding RNAs[87, 89, 90]. Unlike mRNAs transcribed by Pol II, which have an average length longer than 60,000nt[85] and take more than 23mins to be transcribed [6], the short transcripts generated by Pol III seem not to have space for regulation of transcription. Also some of these RNAs, for example tRNA, are very stable. This suggests that the expression of these genes requires less regulation.

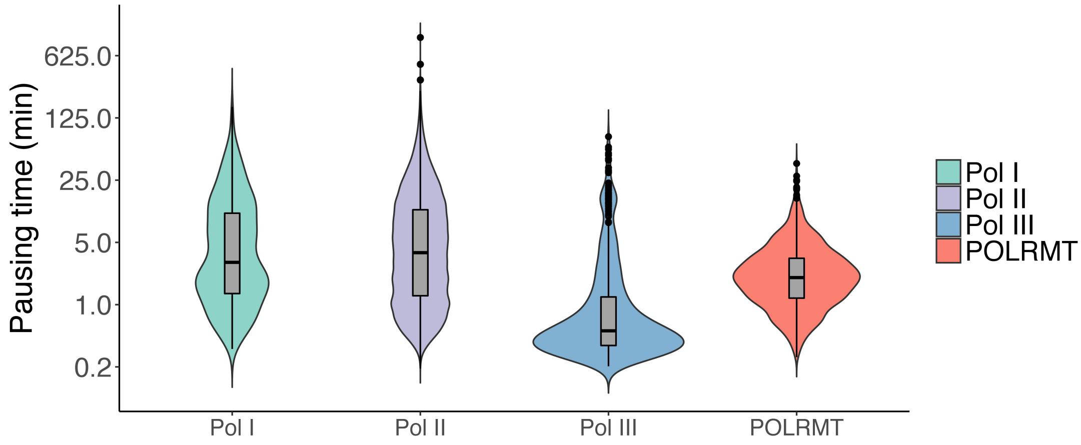


Figure 3.5 Pausing times of different types of polymerases

The violin plot shows the distribution of pausing times of pausing site within regions transcribed by Pol I, Pol II, Pol III and POLRMT. P-values from Bonferroni-corrected Mann-Whitney U test for all pairwise comparison except Pol I vs Pol II are smaller than 10^{-17}.

As polymerase pausing is short in Pol III transcribed genes, initiation and termination become the most likely rate-limit steps for Pol III transcription. The promoters of Pol III have been classified into three different types[91-93]. Type I specifically refers to 5S rRNAs. Type I promoters have a special region called the ICR (Internal control region). It is located 50nt to 90nt downstream of the TSS and is composed of three elements: A box, intermediate element and C box. Type II has mainly been found in tRNAs and contains A box and B box, which are located from +8 to +19 and +52 to +62, respectively. Type III were first found in mammalian U6 spliceosomal genes and later in 7SK genes, H1RNA, RNase P and RNase MRP. They contain a TATA box and a motif called proximal promoter element upstream of the former. By mutation of the TATA box, the Type III promoter can be converted into a Pol II transcribed U2 promoter, and vice versa [94].

I examined the pausing profiles of 5S rRNA, tRNA and U6 snRNA representing Type I, II, III promoters, respectively. I found polymerase to concentrate short pausing in 3 sites in the gene bodies of tRNAs (Figure 3.6A&B), while a much longer pausing site is located downstream of the TES of tRNAs (Figure 3.6A). This suggests that termination of transcription of tRNAs may play an important role in their expression control. Both 5S rRNAs and U6 snRNAs have a pausing site directly on the TES, but the associated pausing times appear to be very short.

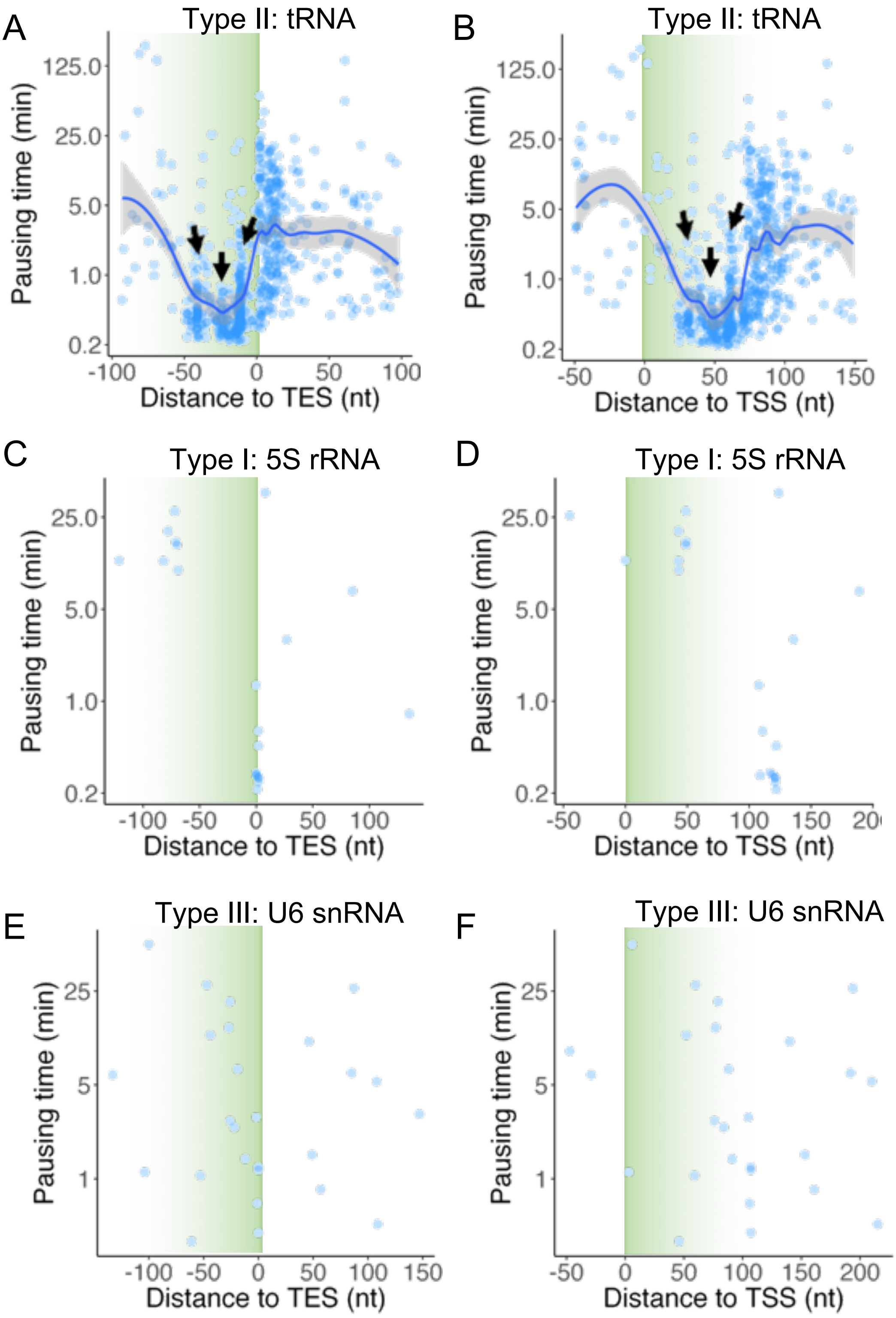


Figure 3.6 Pausing time of different Pol III transcribed genes

A. Pausing times and positions at tRNA genes. Each dot corresponds to a pausing peak. The blue line corresponds to the moving average with the gray shading indicating the 0.95 confidence interval (LOESS fit). The metagene is aligned at the

TES, where the pausing times interestingly increase. Three common pausing sites are marked with arrows.

B. Similar as (A), but aligned at TSSs instead of TESs.

C. Similar as (A), but for5S rRNAs.

D. Similar as (B), but for5S rRNAs.

E. Similar as (A), but for U6 spliceosomal RNAs.

F. Similar as (B), but for U6 spliceosomal RNAs.

The pausing profiles of other noncoding RNAs are shown in Figure 3.7. 7SK RNAs and Y RNAs are also transcribed by Pol III, and the clear pattern of peak enrichment at TESs is also seen. Unlike 5S rRNAs and U6 spliceosomal RNAs, TES pausing of these two types of RNAs lasts longer (Figure 3.7A, B). The snRNAs (Small nuclear RNAs, including 7SK RNA, U7 small nuclear RNA and various spliceosomal RNAs) and snoRNAs (Small nucleolar RNAs, including Small Cajal body specific RNAs, Small nucleolar RNA U3, SNORD12/SNORD106) also show an enrichment of pausing sites at TESs. Pol II and Pol III carry out the transcription of these genes[94]. As histone genes which are transcribed by Pol II also show the pausing at TES[38], the TES pausing is a common mechanism of both Pol II and Pol III. Even though the pausing positions of these genes are all concentrated at the TES, their pausing times show significant differences. TES related pausing of snRNAs are much shorter than the snoRNAs' (the peak density of TES related pausing of snRNAs is also lower, but this might be caused by a detection rate bias in favour of long pausing over short pausing). Also, a standard rDNA repeat containing 18S, 5.8S and 28S rRNA was used for aligning the reads. Thus, an aggregate pausing profile for rRNA transcription carried out by Pol I is shown (Figure 3.7E). Higher peak density was found in the 18S rRNA and 28S rRNA regions. Furthermore, I found a long pausing region at the TR (tandem repeat) which corresponds to the 3' region of the rDNA.

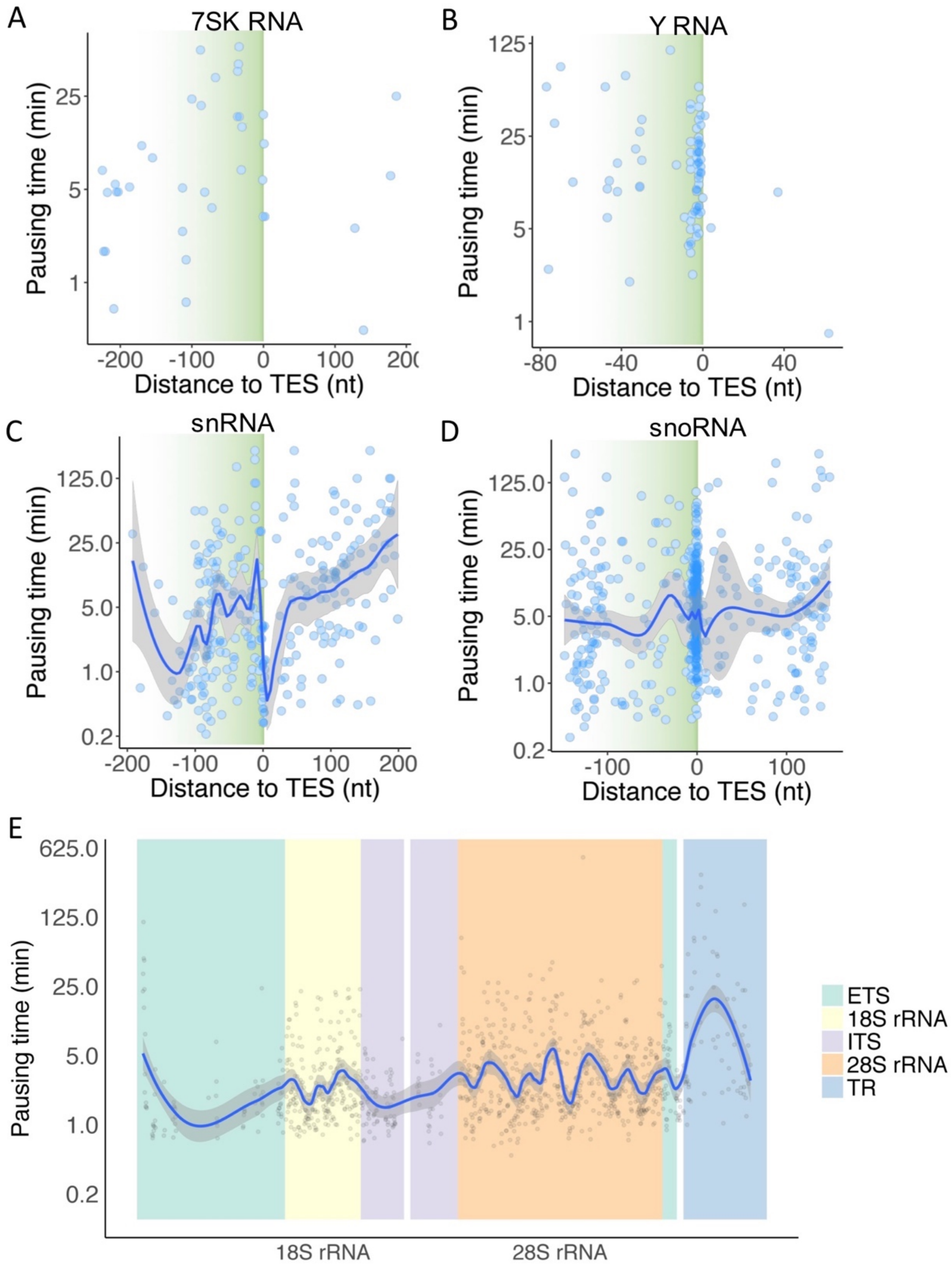


Figure 3.7 Pausing time around TES of noncoding RNAs

A. Pausing time of pausing sites close to TES of 7SK RNAs. Each dot corresponds to a pausing peak.

B. Similar as (A), but for Y RNAs.

C. Similar as (A), but for all snRNAs. The blue line corresponds to the moving average with the gray shading indicating the 0.95 confidence interval (LOESS fit).

D. Similar as (C), but for all snoRNAs.

E. Similar as (C), for ribosomal RNA genes. ETS & ITS, external & internal transcribed spacers, respectively (5.8S not shown). TR, tandem repeat.

3.3.2 Short pausing in the promoter proximal region

The current genome wide analysis about pausing time is mainly built on Trp treatment followed by ChIP-seq or GRO-seq[3, 4, 84]. These researches suggest an average pausing time of polymerases in the PPR for several minutes. However, the slow uptake of Trp[50] challenges the notion of long pausing as it might be a secondary effect of Trp uptake instead. *In vivo* experiments using FRAP show that 'promoter proximal pausing' lasts only about 42s.

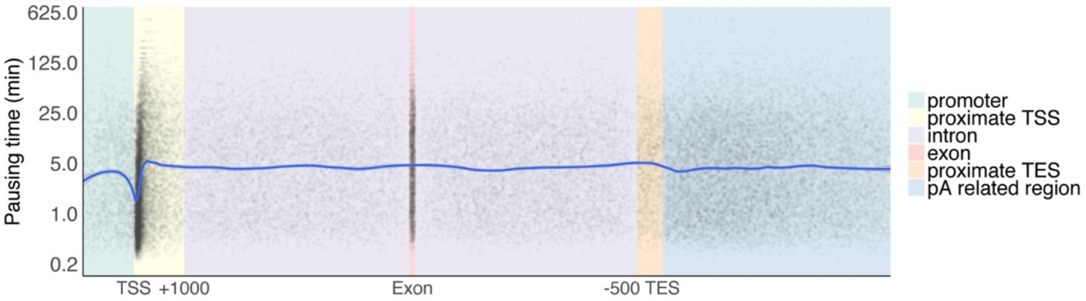


Figure 3.8 Pausing times at mRNA-transcribing metagene

Pausing times at mRNA-transcribing metagene. Each gray dot represents a pausing peak, with corresponding pausing time given by its y-axis value. The x-axis values corresponds to absolute position within -/+ 1kb of TSS (green and yellow tinged regions, respectively), 500bp upstream and 4.5kb downstream of TES (orange and blue, respectively), or relative position within the other genic sections (color code as indicated). The blue line corresponds to the moving average (LOESS fit). The gray shading indicates the confidence interval and is negligible on this scale, hence invisible over most of the graph. The widths of exons and introns have been scaled to their relative average lengths.

The TV-PRO-seq result for HEK293 cells shown in Figure 3.8 demonstrates that even though the polymerases are more likely to pause in the PPR, each pausing time is shorter than in other regions. In fact, the pausing in the first 100nt downstream of TSSs are extremely short compared to other regions (Or sarkosyl facilitates the pausing release in the PPR); however, polymerases do indeed tend to pause at a region slightly downstream of the short pausing region.

Polymerases have previously been suggested to stay longer in exons of genes[38]. TV-PRO-seq shows that Pol II does not pause longer for each pausing site in exons, but is likely to pause with higher frequency. Also, the pausing right after TESs is slightly shorter than pausing within genes and distal to TES. This shorter pausing may function in accelerating the maturation of mRNAs.

I then focused on the pausing events close to TSS. I computationally extracted pausing sites close to TSS and ordered these by their pausing time. I then displayed the positional distributions of the sites with 10% longest pausing time and 10% shortest pausing time, respectively (Figure 3.9A). The short pausing positions are concentrated around +70 downstream of TSS, which has been suggested to be the centre of the promoter proximal pausing region[2, 13]. The longer pausing is concentrated further downstream at +160, and continues with a higher density further downstream. I verified this pattern in the different cell line KBM7 (Figure 3.9B).

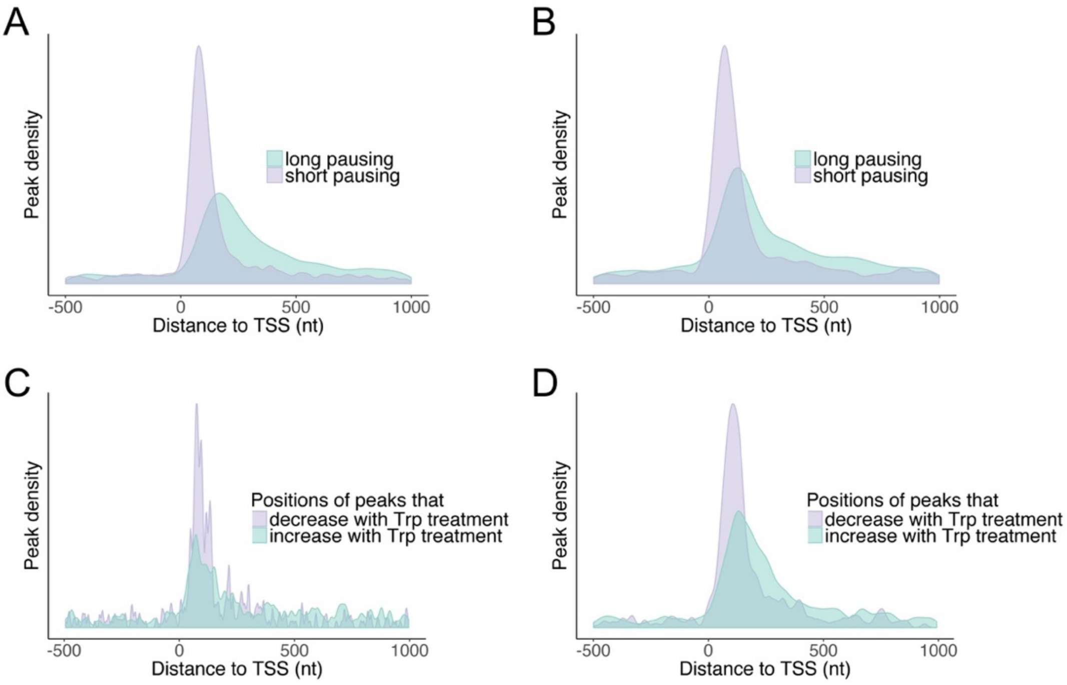


Figure 3.9 Pausing close to TSS tends to be short

A. *Peaks within -500 to +1000 of TSS were classified into 'long' and 'short' according to their pausing times and were displayed as distributions regarding their distances to TSS, P < 10-100, Mann-Whitney U test.*

B. *Similar as (A) from KBM7 data. $P < 10^{-23}$, Mann-Whitney U test.*

C. *Pre-treatment of HEK293 cells with Triptolide (Trp) to block transcription initiation leads to differential vacation of pausing sites near TSS; peaks with increased relative sizes after Trp treatment (green) are further downstream from TSS than decreasing peaks (purple), $P < 10^{-4}$, Mann-Whitney U test.*

D. *Similar as (C) from KBM7 data. $P < 10^{-11}$, Mann-Whitney U test.*

I performed Trp treatment as an additional test. HEK293 cells were treated with culture medium containing 500nM Trp following permeabilization. The permeabilizated cells then were used for PRO-seq with 8min run-on time. As the elongation rate of polymerase ranges from 0.5kb/min to 8kb/min[3, 14, 35], 10min is long enough for polymerase to vacate the first 1000nt of genes. The polymerase occupancy of short pausing sites should decrease more than long pausing sites. The ratio of reads before and after Trp treatment can thus reflect the pausing time. Figure 3.9C shows that the Trp treatment is consistent with the TV-PRO-seq result. I again verified this result in KBM7 cells. In contrast to the HEK293 cells, I performed the Trp treatment for KBM7 after permeabilization and followed with 6min run-on time. As shown in Figure3.9D, the KBM7 Trp treatment thus yielded an even clearer pattern than the HEK293's. The Trp treatment experiment is consistent with TV-PRO-seq. It suggests that pausing in the PPR is not the longest (Or sarkosyl facilitates the pausing release of polymerase in +60 to +100).

3.3.3 FP treatment shows Pol II bound to NELF / DSIF pauses shorter

Three different transcription factors have been suggested to be involved in polymerase enrichment in PPR[2, 34]. NELF and DSIF repress productive elongation while P-TEFb facilitates polymerase escaping from the PPR by phosphorylation of NELF, DSIF and Ser2 of Pol II CTD. As TV-PRO-seq shows that pausing times of polymerase in the PPR tend to be shorter, this short pausing could either be an artefact of its smaller distance to the promoter or an actual effect owing to special mechanisms of promoter proximal pausing. To resolve this question, I used FP treatment data for further analysis.

FP can inhibit P-TEFb thus keeping Pol II bound to the NELF and DSIF and in turn increasing the polymerase occupancy. I performed 300nM FP treatment followed by PRO-seq with 8min run-on time and compared reads with 8min run-on samples from TV-PRO-seq. I considered the top 10% peaks with largest increases in reads after FP treatment as representative of the region where Pol II is likely bound to NELF and DSIF, and thus named these peaks 'FP peaks' (Figure 3.10A). The FP peaks commonly have shorter pausing times . Since the FP treatment increases polymerase occupancy in the PPR, the FP peaks are more enriched in the PPR region (Figure 3.10B). If short pausing in the PPR just results from a distance effect of polymerase towards promoter, we can also expect the FP peaks to pause for short times. To prove that the short pausing of FP peaks is not just a distance artefact, I focused on the first 200nt region downstream of TSS only (green area in Figure 3.10B). The FP peaks' pausing was shorter in this region than the average pausing time of all peaks (Figure

3.10C). As shown in Figure 3.10D, FP peaks indeed pause shorter in the PPR compared to peaks with the same distance to TSS. This result confirms that Pol II bound to NELF and DSIF indeed pauses for shorter times, ruling out a distance-related artefact.

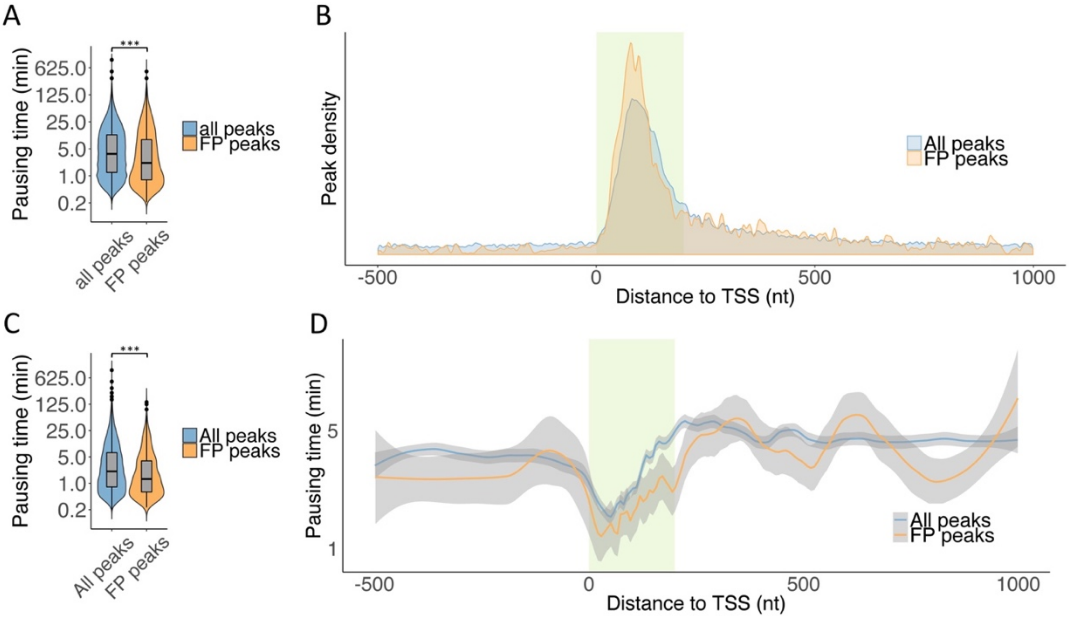


Figure 3.10 Pausing profile of FP peaks

A. *Short pausing at FP-affected peaks. Blue refers to all peaks in the genome. From these, peaks whose read counts increase at least 4.44 times (for cutoff, see Supp methods) after FP treatment were selected as 'FP peaks'* [95]*. Pausing times of FP peaks are lower than those of all peaks,* $P < 10^{-94}$*, Mann-Whitney U test.*

B. *The same groups of peaks as in (D) are shown in terms of their average densities along genes. The green region denotes the first 200nt downstream of TSSs.*

C. *Violin plots show that the pausing times of FP peaks in the green region of (E, F) are significantly shorter than all peaks,* $P < 10^{-22}$*, Mann-Whitney U test.*

D. *Pausing times of the peaks considered in (E) shown as LOESS fits as in (A).*

I note that run-on methods are influenced by technical noise and that GRO- and PRO-seq are based on permeabilized cells, thus not an optimal reflection of the situation in vivo. Therefore I examined the shorter pausing in PPR in an alternative way based on NET-seq data.

The pausing sites are always fixed to certain genome location. Even though TSS positions can vary, the pausing site does not move along with the TSS[44]. If polymerase pauses longer, the polymerase occupancy of pausing sites will increase, while the regions adjacent to the pausing sites will be influenced less. We can exploit this to test that FP treatment blocks pausing release, since the reads will become more concentrated on pausing sites after treatment. However, my result is the opposite to this suggestion. An example is shown in Figure 3.11; a distinct pausing site is located at position 170 532 209 of the chromosome 1 plus strand, which is 88nt downstream of the TSS of the gene GORAB. After FP treatment, the total reads of PPR (the green area) increased dramatically. However, when we look at the relative density of reads, we will find the reads are less concentrated on the pausing sites. It means that Pol II bound to NELF / DSIF remains shorter at this pausing site or stay longer at the other sites of the PPR.

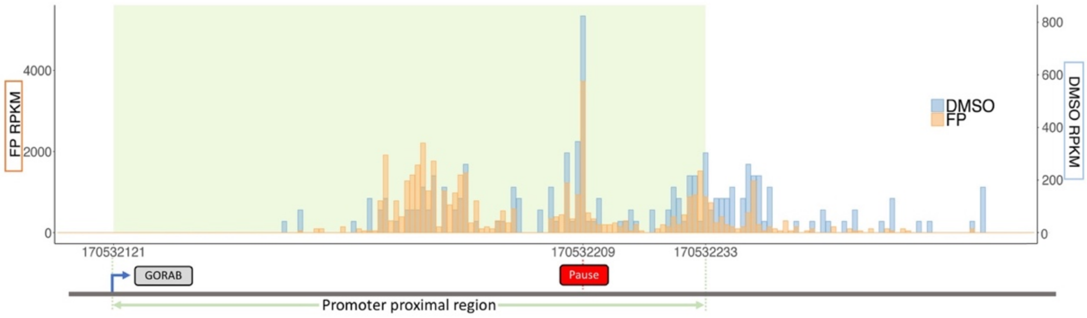


Figure 3.11 Reads at the gene GORAB from PRO-seq with FP or DMSO treatment

Orange represents FP treatment and its scale is shown on the left y-axis. The blue bars and the scale on the right y-axis refer to control (DMSO) treatment. The average read counts increased after FP treatment in the green shaded region (the boundary is defined in the methods). The blue and orange read counts are scaled to the same area within the green region.

I conducted a genome-wide analysis based on this logic (Figure 3.12). As I define the PPR end as the early termination site, I assume positions in the PPR share the same polymerase flux. The ratio of reads at pausing sites and average reads in other positions of the PPR can reflect the relative length of average residence time of polymerase on pausing sites. Considering GORAB for example, the ratio means reads of position 170 532 209 divided by the mean of reads in 170 532 121 to 170 532 233 except 170 532 209 (in other words, 170 532 121 to 170 532 208 and 170 532 210 to 170 532 233). The ratio of all peaks located in the PPR of NET-seq data[38] is shown in Figure 3.12A; the ratio significantly decreased after FP treatment. I also derived this result in an alternative way: I calculated fold change, that is the reads after FP treatment divided by those after DMSO (control) treatment. As shown in Figure 3.12B, the fold changes of the whole PPR are significantly higher than those of pausing sites. Because NET-seq data is more sparse and has a higher background noise, I also used PRO-seq data to repeat this analysis, and the result is consistent with NET-seq data (Figure 3.12C, D from Hela cells[50], Figure 3.12E, F from HEK293 data). These results confirm my TV-PRO-seq findings suggesting that Pol II bound to NELF / DSIF tends to directly drop off before entering productive elongation and that these Pol II pause for even shorter times.

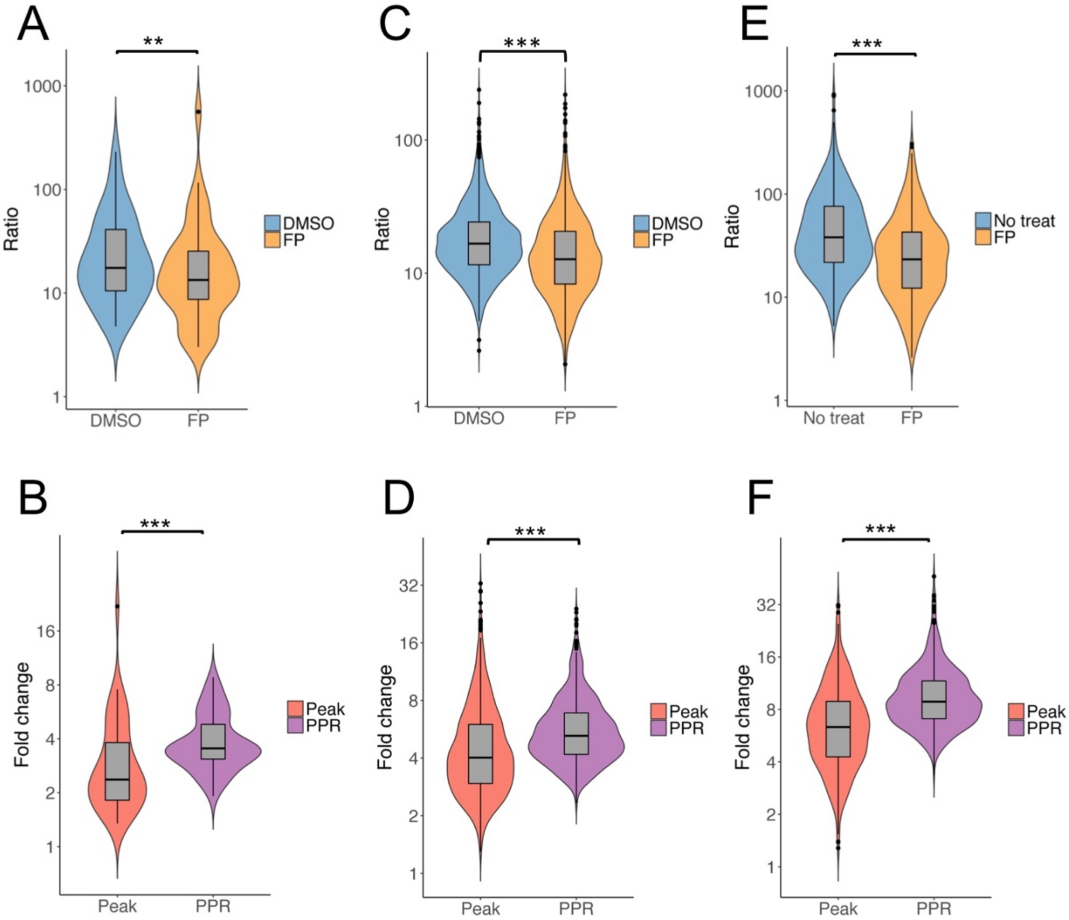


Figure 3.12 Genome-wide analysis of influence of FP treatment

A. *The ratios of reads at individual pausing sites and average reads at remaining sites within the PPR, of Hela cell NET-seq data*[50] *(potentially including other pausing sites). The blue refers to ratios after DMSO treatment and orange refers to FP treatment. P < 0.01, Mann-Whitney U test.*

B. *Fold change distribution of before and after FP treatment of Hela cell NET-seq data. Red indicates the fold changes of pausing sites in the PPR and purple indicates those of the total reads in the PPR.* $P < 10^{-3}$*, Mann-Whitney U test.*

C. *Similar to (A), data from Hela cell PRO-seq.* $P < 10^{-51}$*, Mann-Whitney U test.*

D. *Similar to (B), data from Hela cell PRO-seq.* $P < 10^{-45}$*, Mann-Whitney U test.*

E. *Similar to (A), data from Hela cell PRO-seq.* $P < 10^{-29}$*, Mann-Whitney U test.*

F. *Similar to (B), data from Hela cell PRO-seq.* $P < 10^{-31}$*, Mann-Whitney U test.*

3.3.4 Pausing profile of D. melanogaster

Core promoter architecture differs greatly between humans and D. melanogaster. D. melanogaster have a distinct pattern of motif distribution; various motifs have been found at specific distances to the TSSs. For instance, the 'pause button', which correlates with pausing, is found tobe located +26 downstream of the TSS in D. melanogaster[96,97]. Unlike D. melanogaster, few motifs have been found in core promoters of human genes[96]; no motif akin to the pausing button has emerged for human core promoters, for instance. Even motifs shared by humans and D. melanogaster, such as the TATA box and GAGA box, are distributed more widespreadly in the former [96]. This appears to be reflected by the different distributions of pausing sites in the different organisms (Figure 3.13).

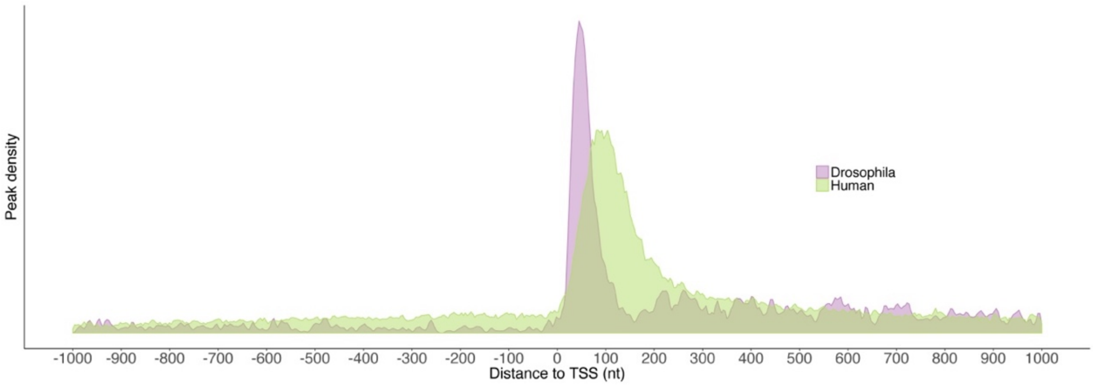


Figure 3.13 Difference of peak density between humans and D. melanogaster

The density plot shows that pausing sites are enriched in the PPRs of both D. melanogaster and human genes. However, pausing sites in D. melanogaster are closer to the TSS.

The pausing times of pausing sites close to TSS also show different profiles between humans and D. melanogaster. The pattern that polymerase remains longer in the gene body is similar (Figure 3.8 and Figure 3.14). However, the pausing time does not drop at the pausing site enriched region. On the contrary, pausing time shows a slight increase between +30 to +40, in this region.

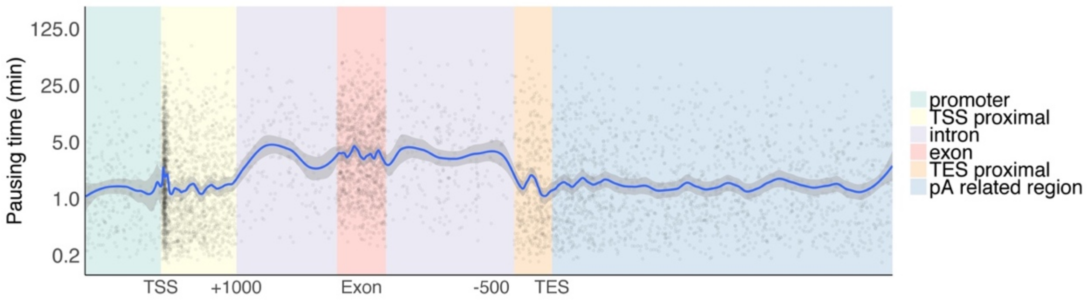


Figure 3.14 Pausing times at mRNA-transcribing metagene of D. melanogaster

Pausing times at mRNA-transcribing metagene in D. melanogaster. Definition of the region is similar to Figure 3.8.

3.4 Discussion

TV-PRO-seq provides novel insights for deepening our understanding of pausing. But it does have certain limitations:

1. As TV-PRO-seq is built on PRO-seq, it is an in vitro experiment which cannot perfectly reflect the *in vivo* state. As the run-on buffer for PRO-seq contains sarkosyl, the results can also be due to effects of sarkosyl.

2. TV-PRO-seq cannot distinguish the type of RNA polymerase[65]. For genes transcribed by both Pol II and Pol III, TV-PRO-seq can only output the compound signal for both.

3. Another limitation of TV-PRO-seq is that it cannot identify unambiguously the read source of highly repeated genes. In particular, many non-coding genes transcribed by Pol I and Pol III are highly repeated. For these, the reads are randomly assigned to repeats, thus necessarily obscuring their origins. However, this is a general problem with next generation sequencing, and prevents unambiguous alignment of reads also for assays such as ChIP-seq or RNA-seq if their sequence maps to repeat regions or pseudogenes. The pausing time of pausing sites in highly repeated genes is still meaningful, as it represents the average pausing time of pausing sites in different repeats.

4. TV-PRO-seq only produces estimated value of pausing time. It is suited well for comparisons within a dataset, but becomes less precise across independent experiments and is also subject to the priors chosen for the Bayesian estimation framework.

5. Finally, owing to its high positional resolution, the majority of PRO-seq peaks will only have a small number of reads in some time points. Thus, biological and experimental noise will influence results strongly. This limits the prospects for detailed analyses of dynamics at individual pausing sites.

Despite these limitations, TV-PRO-seq provides large amounts of highly valuable information:

1. TV-PRO-seq is the only method that can estimate pausing times at single base resolution genome-widely.

2. In contrast to previous sequencing methods, TV-PRO-seq compares reads of the same genomic position with those of different run-on times. This permits isolation of the profile of pausing sites independently of polymerase flux. Therefore we can investigate pausing times of pausing sites within genes expressed at different levels.

3. A big advance over the previously used pausing index is TV-PRO-seq's ability to produce pausing profiles of short genes. This for the first time permits analyses of pausing profiles of lncRNAs and other ncRNAs which play important roles for organisms.

My meta-analysis of TV-PRO-seq suggests that pausing of individual pausing sites in the PPR is shorter than in the gene body. I further confirmed this with analyses based on NET-seq data which can reflect the *in vivo* state. I propose that promoter proximal pausing is more akin to a toll station of a highway which stops polymerase for a certain short time. Pausing sites in the gene body, on the other hand, could be imagined as traffic lights which control the engaging speed of polymerases.

Chapter 4 Molecular mechanism of pausing

4.1 Introduction

Various mechanisms have been suggested to be involved in pausing. The most well established NELF/DSIF mediated promoter proximal pausing needs to be revisited as more and more evidence shows that the polymerase enrichment in PPR can also be caused by abortive transcription[5, 6, 50, 98]. But pausing is not restricted to the PPR, it occurs throughout the whole gene body every 20nt to 100nt[13]. This pausing is due to other mechanisms, such as: nucleosome barriers[26, 99], DNA secondary structure[100], bridge helix (DNA-RNA helix at 3' of nascent RNA)[101, 102] and nascent RNA structure[103].

The DNA template is packaged by nucleosomes. Each nucleosome core contains an octamer of histone proteins and is wrapped by 147bp of DNA[104]. In vitro experiments show that nucleosomes can enhance pausing by increasing pausing frequency and pausing time. Even without a pausing site, nucleosomes still slow down elongation as the polymerase has to wait for the unwrapping of nucleosomes to occur[51]. Both NET-seq[25, 30] and PRO-seq[26] show this effect genome-wide. H2A.Z, which is a variant of histone H2A in the histone octamer, has been linked to pausing; higher H2A.Z levels at the PPR reduce pausing by increasing turnover of the other histone types H3/H4[105]. It further increases the elongation rate in gene bodies[106], which appears consistent with the notion of short pausing in the PPR, since H2A.Z is enriched in the latter. Histone acetylation has also been suggested to enable the release of paused polymerase through loosening chromatin[51, 52, 107, 108].

The sequence of the template DNA has also been suggested to relate to pausing. For instance, the GAGA box has been reported to correlate with promoter proximal pausing[109]. However, the suggestion that polymerase enrichment in the PPR is rather due to polymerase turnover mandates a re-evaluation of links between the GAGA box and pausing. Apart from motifs related to promoter proximal pausing, the DNA template affects pausing directly by the molecular interaction of polymerase and the DNA-RNA helix of template DNA and nascent RNA. Both NET-seq[49] and PRO-seq[44] show that polymerase is more likely to pause on cytosine. This accurate pausing on cytosine is conserved from E.coli to humans[102]. Studies in E.coli also suggest that hairpin secondary structures of nascent RNA can stabilize pausing[103]. Similarly, G-quadruplexes appear to block transcription when folded[100].

Here I show that the H3K36me3 histone modification, which represses histone acetylation, correlates with long pausing. H3K9me3, a heterochromatin marker, shows a similar correlation with long pausing. I further discovered how sequence motifs can influence elemental pausing. An in-depth analysis of such a motif, APM1 (Accurate Pausing Motif 1), demonstrates that some nucleotides in the motif influence the accuracy of pausing, while others influence the pausing time.

4.2 Methods

4.2.1 Histone modification and chromatin accessibility for TV-PRO-seq data

I used existing HEK293 cell ChIP-seq data for different histone modifications from published studies and/or public depositories for the analysis. H3K4me1, H3K4me2, H3K4me3 and H3K27ac data were obtained from Gene Expression Omnibus, GSE101646[110], and H3K9me3, H3K36me3 and DNase-seq data were downloaded from ENCODE series ENCSR372WXC and ENCSR000EJR. The data were first trimmed with Trimmomatic-0.36 with options `LEADING:24 TRAILING:24 SLIDINGWINDOW:4:20 MINLEN:20`[111], then aligned to hg38 under –no-spliced-alignment condition by Hisat2[68]. The SAM files were converted to BAM files, then to BED files using Samtools[69] and Bedtools[70], respectively. The read intervals in the BED files were adjusted to the same lengths with the custom script `bed_normal_length.pl` to make sure the coverages of reads bore equal weights for each read. We then converted the data to BEDGRAPH files with the `genomeCoverageBed` command from Bedtools, using the flags `-bga` [70]. The BEDGRAPH files were annotated to TSS or pausing peaks with the custom script `Liner_bedgraph.pl`.

I then classified peaks on nuclear chromosomes into those with the longest 5% and shortest 5% pausing times, and extracted the coverage from the BEDGRAPH files within +/-1000 nt of each peak in both classes. I then removed the top 5% of these coverage intervals since these had disproportionately strong influence on the results. Finally, I averaged the coverages of each class, respectively, and displayed the results using ggplot2 in *R*.

4.2.2 Calculation of pausing index

I defined the genic regions from TSS +200bp to TES as gene body (GB) [38], and calculated a pausing index (PI) for each peak position by dividing reads in peaks by the average reads in the GB of the same gene. I considered either peaks along the whole gene or peaks within TSS +500bp only. I implemented this by processing the

UCSC mRNA gene annotation as above with the script `PI_reference_maker.pl`. I then used the script `PI_counter.pl` to count the GB reads of target genes.

4.2.3 Histone modification and chromatin accessibility for mNET-seq data

HEK293 NET-seq data was downloaded from Gene Expression Omnibus, GSE61332[30]. I used the UCSC liftOver tool to convert the BEDGRAPH file to hg38[85]. I then defined target genes for further analysis by selecting genes longer than 3000 nt, with unique TSSs and TESs. Peak selection for the mNET-seq data followed the same strategy as for TV-PRO-seq; the peak selection output file was processed with the script `Liner_bedgraph.pl` to extract histone modification states within +/−1000 nt of peaks in the same way as for TV-PRO-seq; I removed the top 5% peaks with highest average coverage of each group and plotted the average coverage of histone modification at peaks corresponding to the top and bottom 5% PI, respectively (for all peaks in target genes, or peaks within the TSS to +500 region only).

In order to compare TV-PRO-seq and mNET-seq with regards to the chromatin state results, I needed to subset the TV-PRO-seq data to the same target genes as I used for the mNET-seq data. The script `PI_TV_annotater.pl` was used to extract the coverage information of individual TV-PRO-seq peaks located in the target genes. I then selected long pausing and short pausing peaks as above. The average ChIP-seq/Dnase-seq coverages of long pausing and short pausing peaks were then used for comparison with the high PI and low PI peaks.

4.2.4 Motif analysis

The ±50bp surrounding sequence around each peak was extracted with the custom script `Peak_seq_getter.pl`, saved into a fasta file, and subjected to *de novo* motif detection. In addition, the regions from -550 to -450 and +450 to +550 at each peak were extracted to serve as control sequences. Motif detection was done with the program `findMotifs.pl` of the Homer software[112] suite with default options and by using the control sequences as background [112], which resulted in a number of position probability matrices (PPM) for enriched motifs, which I term the PPM$_e$'s. For each PPM$_e$, I used the `homer2 find` function to obtain the distances between all motif occurrences and peaks in the input sequence set. I used the parameter -strand to ensure strand-specific motif detection.

For each distance distribution resulting from a PPM$_e$, I compared the most frequently occurring distance d_1, to the second most frequently occurring distance d_2; I ranked PPM$_e$s by the relative standard error ρ in estimating the proportion $\hat{p} = n_1/(n_1 + n_2)$, based on the heuristic assumption that n_1 is binomially distributed, where n_1 and n_2 are the numbers of occurrences of d_1 and d_2, respectively,

$$\rho = \frac{1}{\hat{p}}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n_1+n_2}}.$$

After ranking by ρ, the top 6 motifs were taken for further analysis. I considered these motifs to have a unique, precise pausing site at single base resolution. I then extracted the PPM for the motifs appearing at d_1 and termed this second PPM the precision PPM, PPM$_p$. I generated sequence logos for PPM$_e$ and PPM$_p$ with the ggseqlogo R package.

I then plotted pausing times of peaks at the precise pausing sites and considered these to be related to the motifs. Peaks at distances between 20bp to 40bp with regards to the precise pausing sites were used as controls of the surrounding neighbourhood, since different genomics regions have different overall pausing characteristics/times. Box plots were used to show the pausing time distributions between motif related peaks and these adjacent controls. A Mann–Whitney U test was used to test for significant differences. I repeated this comparison for all peaks to test if the motif peaks' pausing times deviated from the genome-wide average.

The top motif output by Homer, which I termed 'Accurate pausing motif1' (APM1), and which corresponds to the sequence ACAGTCCT, was taken for further analysis. I identified 'variant motifs' from the consensus by changing individual positions of APM1 and then determined their occurrences as described above.

4.3 Results

4.3.1 Histone modification and pausing time

As polymerases have to wait for nucleosome fluctuations to be able to pass[51], different types of histone modifications influence transcription in various ways and vice versa[113]. For instance, new histone acetylation is found at many genes after a heat shock[29, 114]. Histone acetylation can also accelerate the release of paused polymerase[51, 52, 107, 108].

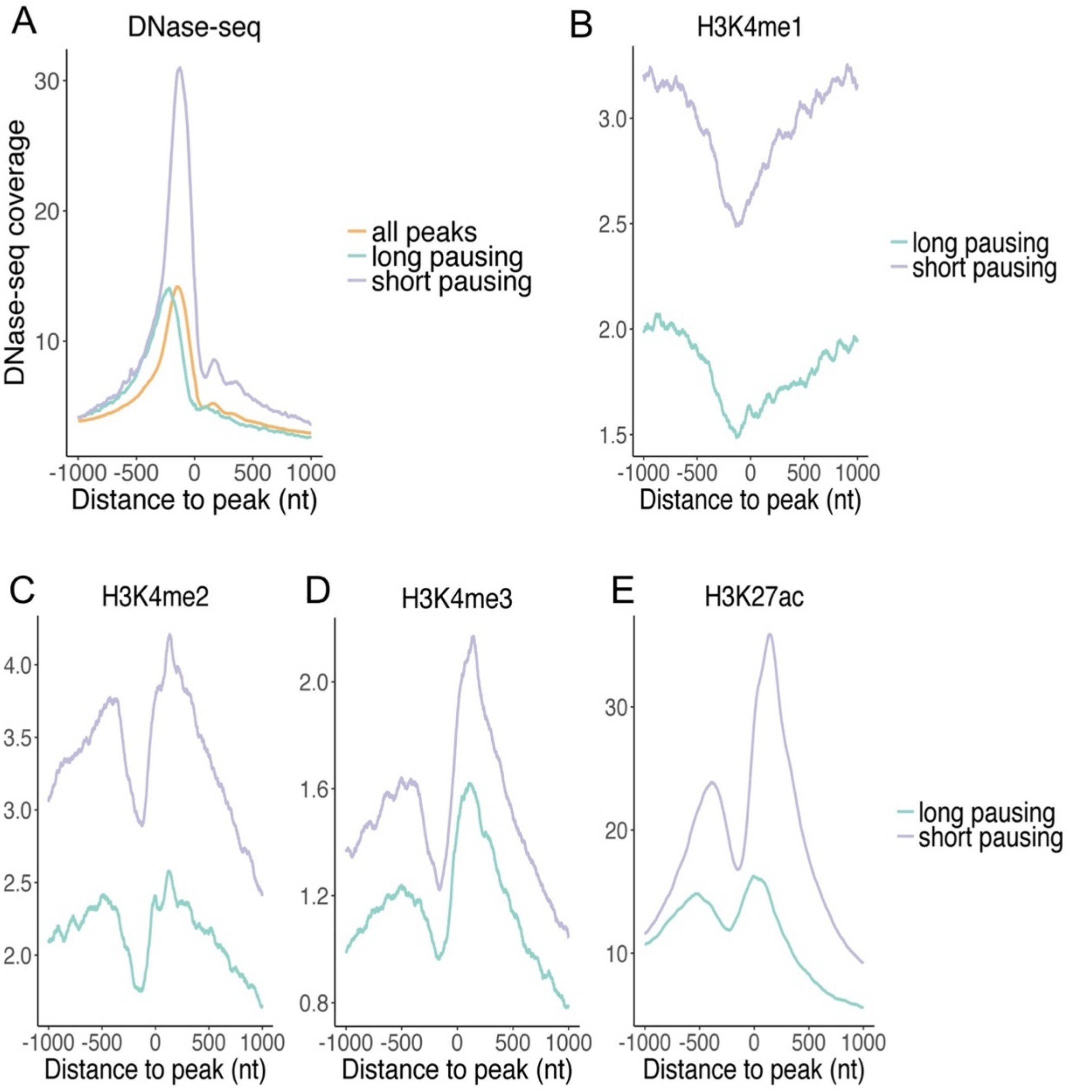


Figure 4.1 Chromatin state and pausing times

A. Peaks were classified into 'long' and 'short' according to their pausing times. The average signal of DNase-seq data is displayed in the vicinity of the two classes of peaks and all peaks.

B. Similar to (A), from H3K4me1 ChIP-seq data.

C. Similar to (A), from H3K4me2 ChIP-seq data.

D. Similar to (A), from H3K4me3 ChIP-seq data.

E. Similar to (A), from H3K27ac ChIP-seq data.

As shown in Figure 4.1A, polymerases are likely pausing in front of nucleosomes and are located upstream in an open chromatin state. This result is consistent with previous studies[115, 116]. TV-PRO-seq allowed me to investigate this further, and I classified peaks into 'long' and 'short' according to their pausing times and quantified their presence around different chromatin features. Interestingly, short pausing sites are enriched at the boundaries of open chromatin regions, while long pausing sites are shifted further downstream. As shown in Figure 3.9, long pausing sites are enriched further downstream of TSSs. This result suggests that, rather than paused polymerase maintaining the open chromatin region around the TSS, co-location of polymerases and nucleosome free region is probably a secondary effect. The other possibility is that pausing sites with longer pausing time function as regulators of elongation rates and short pausing sites have roles as checking points with general functions. Thus, a high fraction of polymerases pass the long pausing sites without pausing, while all polymerases have to pause at the short pausing sites. Activating histone modifications[113] such as H3K4 methylations and H3K27 acetylation exhibit similar profiles (Figure 4.1B-E) as the DNase data.

H3K9me3 and H3K36me3 show interesting patterns correlated to long pausing (Figure 4.2A, B). H3K9me3 is the marker of heterochromatin[117]. As shown in Figure 4.2A, long pausing is enriched in front of nucleosomes with H3K9me3, while short pausing is strongly reduced. It is reasonable to assume that packaged chromatin prevents polymerase from engaging. More surprising is the fact that H3K36me3 as an elongation marker is also found to be related to long pausing (Figure 4.2B). H3K36me3 is usually found enriched at exons of active genes[20, 118] and would thus be expected to correlate with shorter pausing as the other active expression markers. However, TV-PRO-seq yields the opposite result; in contrast to H3K9me3 which only shows sharp enrichment peak at the pausing site, H3K36me3 displays enrichment over a broader region. The potential mechanism for H3K36me3 to block polymerase engagement is its ability to reduce nucleosome turnover by facilitating histone deacetylation and remodelling of repressive chromatin[119, 120].

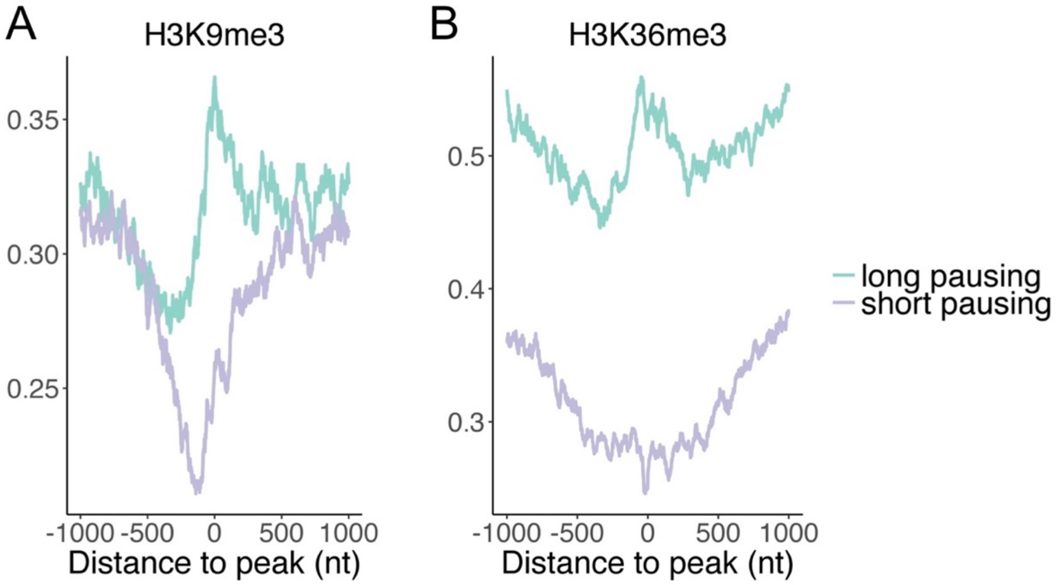


Figure 4.2 Histone modifications related to long pausing

A. Peaks were classified into 'long' and 'short' according to their pausing times. The average signal of H3K9me3 ChIP-seq data is displayed in the vicinity of the two classes of peaks and all peaks.

B. Similar to (A), from H3K36me3 ChIP-seq data.

4.3.2 Isolating pausing time by TV-PRO-seq

TV-PRO-seq evaluates pausing time at each pausing site at single nucleotide resolution, thus it is independent of the genes' expression levels. A side-by-side comparison with NET-seq data for the same cell line and chromatin states in different regions is shown in Figure 4.3. The DNase-seq signal around peaks from NET-seq and TV-PRO-seq shows similar pattern at TSSs. But for the gene body, high PI (Pausing index, calculation see 4.2.2) peaks from NET-seq data which indicate long pausing have very low signal. This is because the PI uses the average Pol II signal in gene bodies for normalization. For this reason, a high PI actually means not only that the peak tends to have longer pausing time, is also selects for location in a low expression gene. In contrast, TV-PRO-seq data demonstrates that short pausing in gene bodies is actually associated with higher H3K27ac, H3K4me2 and H3K4me3 signals. The pattern of short pausing sites in gene bodies is similar to the TSS ones, but weaker.

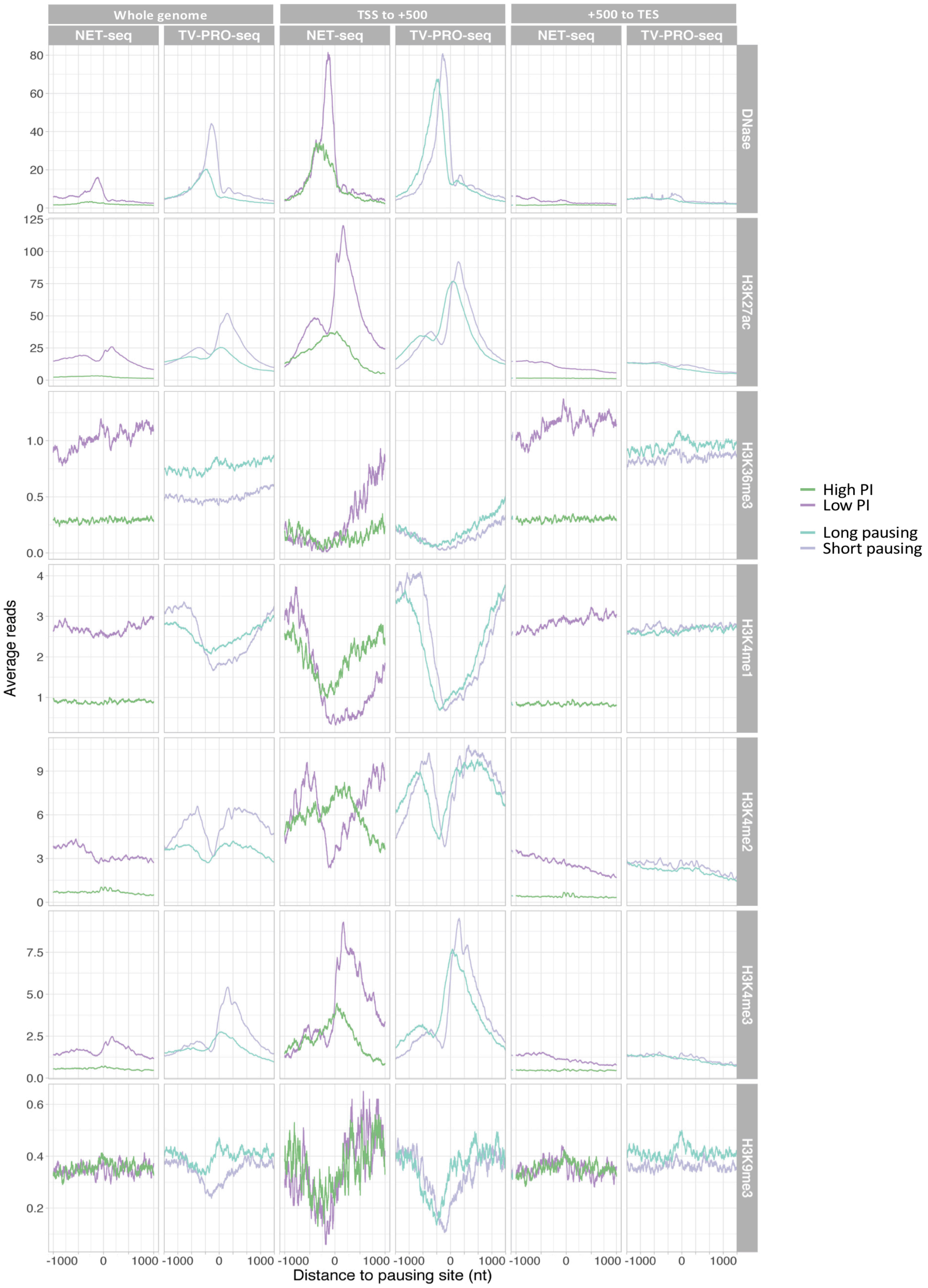


Figure 4.3 Comparison of chromatin state profiles for TV-PRO-seq and NET-seq/PI

Dark purple and dark green lines represent the low PI and high PI pausing positions from NET- seq data, respectively. Light purple and light green represent the short *pausing and long pausing positions from TV-PRO-seq data. The type of chromatin feature (as determined by DNase-seq or ChIP-seq) is shown on the right-hand side; all peaks in the gene body, peaks in the region from TSS to +500 nt or TSS +500 nt to TES are shown in the first two, middle two and last two columns, respectively. The profiles are clearer for TV-PRO-seq in many cases, and often deviate from the NET-seq profiles, suggesting that TV-PRO-seq often produces better and sometimes different information.*

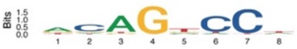
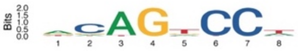
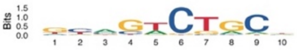
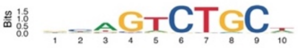
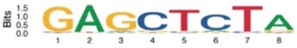
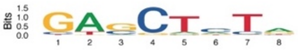
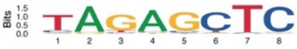
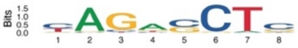
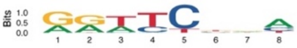
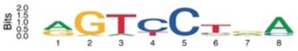
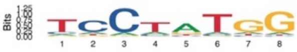
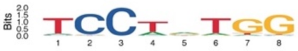
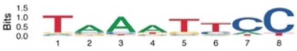
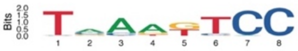
For H3K9me3, TV-PRO-seq data show a clear pattern. Even though H3K9me3 is normally absent around TSSs, its signal can still been found in front of long-paused polymerase. The pattern is clearer in gene bodies, but peaks with high PI do not show any difference with low PI ones. H3K36me3 is enriched at peaks with longer pausing times but is also found at peaks with low PI. As TV-PRO-seq measures pausing time for isolated pausing sites, the result indicates that H3K36me3 blocks polymerases, for a long time. However, using the PI produces the conflicting result that peaks with high PI have much lower H3K36me3 signal. This is because H3K36me3 as elongation marker exists in active genes which have higher polymerase occupancy in the gene body. Therefore, all peaks in these genes tend to have lower PIs.

4.3.3 Essential of APMs for elemental pause

Elemental pausing has been extensively studied in E. coli, where it has been shown that RNAP (RNA polymerase) pauses on average every 100nt due to sequence-induced RNAP active site rearrangement[121, 122]. This elemental pausing can stop polymerase and induce its backtracking and formation of RNA structure[123, 124]; the sequence of the 3’ end of the nascent RNA and the +1 position on template DNA (pausing release site) are essential in this context[102, 123]. A similar phenomenon has been found in mammalian cells[44, 49].

As TV-PRO-seq is based on PRO-seq which has single nucleotide resolution, I further analysed the function of the gene’s sequence around the pausing site towards pausing. All the peaks defined in 2.2.4 were used for this analysis. I chose several motifs found by Homer[112] for the further analysis. As the motifs are located directly at the pausing sites, we named them APMs (accurate pausing motifs). PPMs (position probability matrices) of APMs are shown in Table 4.1, the position of the pausing release site (+1 from pausing sites) has been marked as red in the consensus sequence.

Table 4.1 Accurate pausing motifs.

Motif name	Consensus sequence	Logo for PPM$_e$	Logo for PPM$_p$
Accurate Pausing Motif 1	ACAGTCCT		
Accurate Pausing Motif 2	GCAGTCTGCW		
Accurate Pausing Motif 3	GAGCTCTA		
	TAGAGCTC		
Accurate Pausing Motif 4	GRTTCBRA		
Accurate Pausing Motif 5	TCCTATGG		
Accurate Pausing Motif 6	TAAATTCC		

** Reverse complemented Accurate pausing motif3 has a different precision pausing site, i.e. the peak at a different position compared to the forward one.*

I found that nucleotides at the 3' end of the nascent RNA and the +1 template DNA are most important for the majority of elemental pausing (Shown as PPM$_P$ in the Table 4.1). In particular, polymerases are likely to be blocked during run-on when the incoming nucleotide is C. The 3' end of the nascent RNA is also likely to be essential, since C and G are common in this position. The essential nucleotides for maintaining pause are mostly located from positions -6 to +1 with respect to the pausing sites. Interestingly, I found that some APMs not only function when they are located in the sense strand, the antisense transcripts also paused on a different nucleotide of the motif. Take APM3 for example, where polymerase is paused on both APM3 and its reverse complementary sequence; both strands have a distinct peak (Figure 4.4). Even though both strands of APM3 block transcription, it is unlikely to occur in the same position. As the PPM$_p$ in Table 4.1 shows, position 4 of the forward APM3 has a high possibility to be C and position 1 is likely to be G. However, position 5 of the reverse complement APM3 does not have a high possibility to be G, and neither does position 8. APM5 also has distinct pausing sites on both strands. Overall, the common pattern of pausing on C or G and release on C are conserved for nearly all motifs.

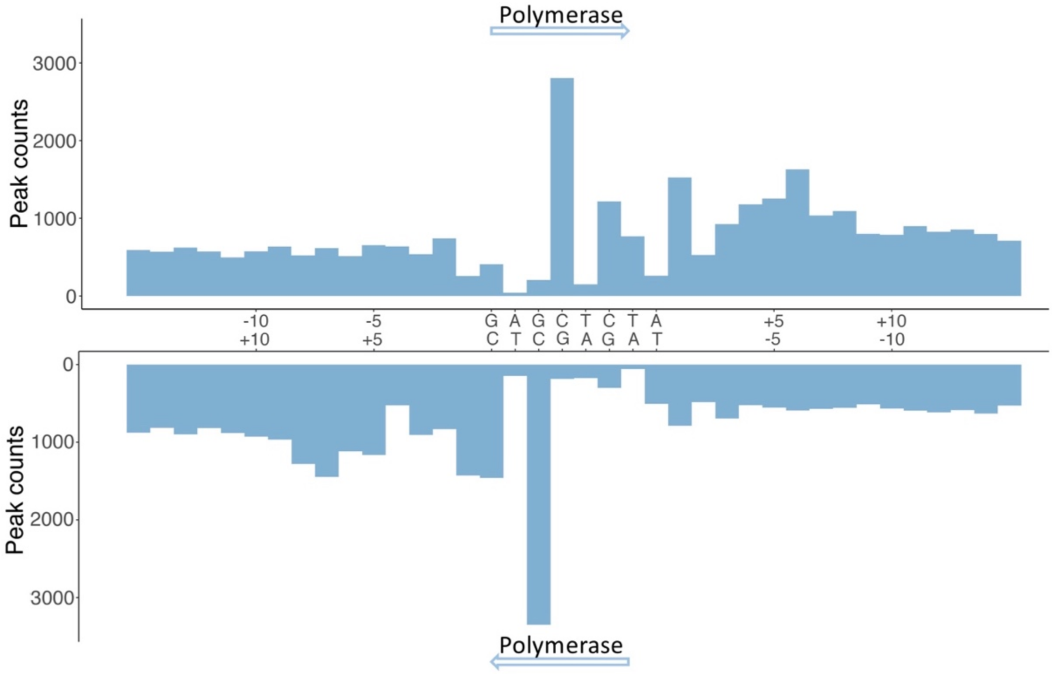


Figure 4.4 Polymerase pause on both strands of APM3

To investigate the influence of the sequences of the APMs towards pausing, I use APM1 for a more detailed analysis (Figure 4.5). The consensus sequence of APM1 is ACAGTCCT, and polymerases are likely to pause on the second C and release on the third C. Interestingly, the most important site for pausing is the release site rather than the pausing site. Because PRO-seq is a run-on based protocol, this result may in principle be due to technical reasons. However, the pausing site is not as important as we expect, since a mutant of the G at the -2 position relative to the pausing site reduces pausing more than a change of the pausing site itself. Similar patterns are seen for the other APMs (Table 4.1); the pausing release sites are always essential for making the polymerases pause at the right position, but not the pausing sites themselves. As Pol II and E. coli RNAP share all the active-site components such as trigger loop or bridge helix, the function of the +1 template DNA position towards RNAP in elemental pausing can inspire an explanation of the phenomenon. The +1 position can be trapped by an incompletely opened clamp of RNAP and lead to elemental pausing, which in turn can lead to backtracking or long pausing[125].

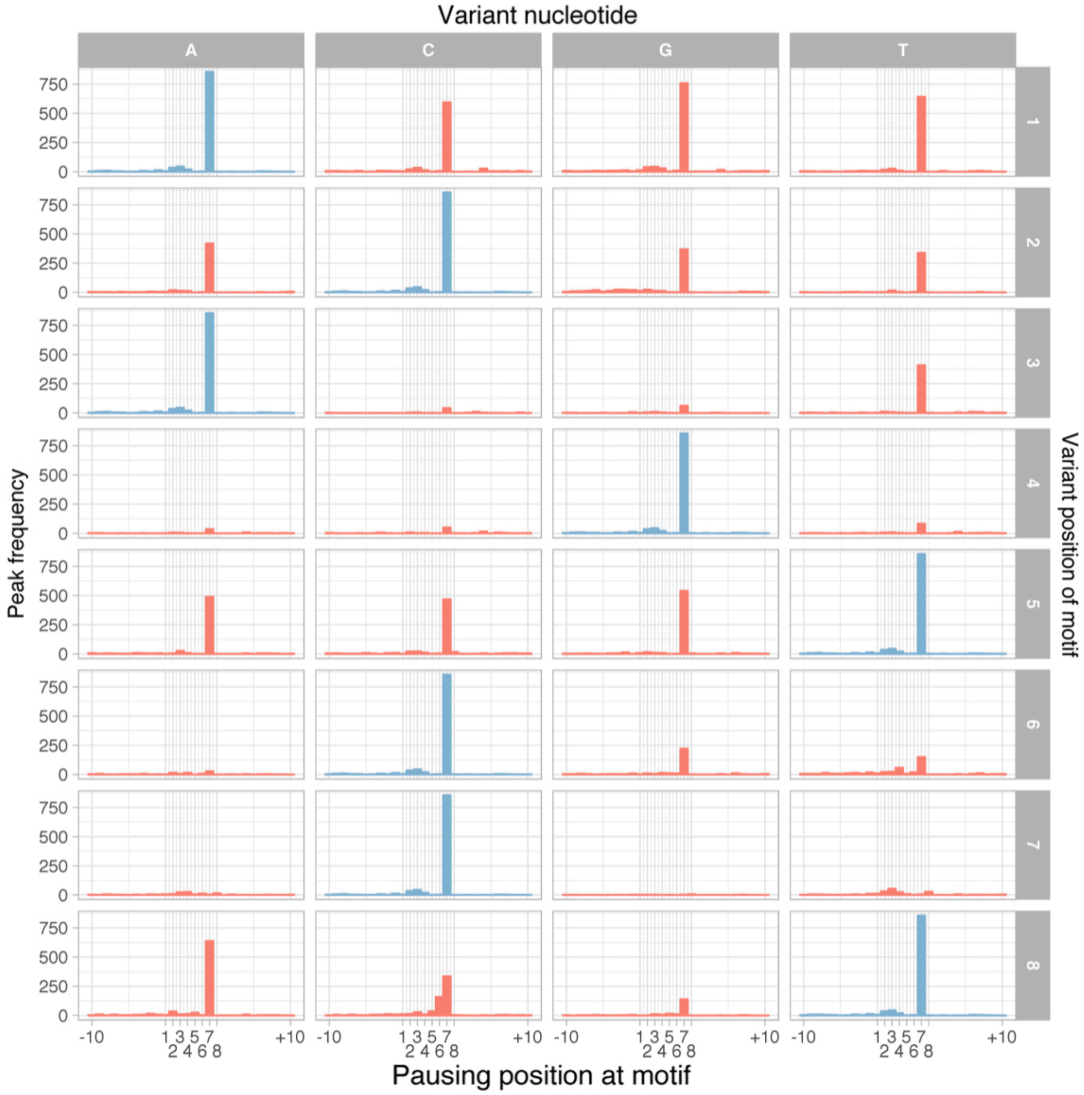


Figure 4.5 Histograms of peak frequencies at positions relative to the motif ACAGTCC and single base variants (position variants are shown in red, whereas consensus positions are shown in blue).

4.3.4 Sequence of APMs influence length of pausing

APMs' sequences not only influence the location of polymerase pausing, but also affect the duration of pausing. As shown in Figure 4.6, different APMs have different pausing times. For example, polymerases pause on APM2 significantly longer than on other peaks (Figure 4.6A). This is not a secondary effect of the motif's distribution, since pausing in adjacent regions to APM2 pause significantly shorter (Figure 4.6B).

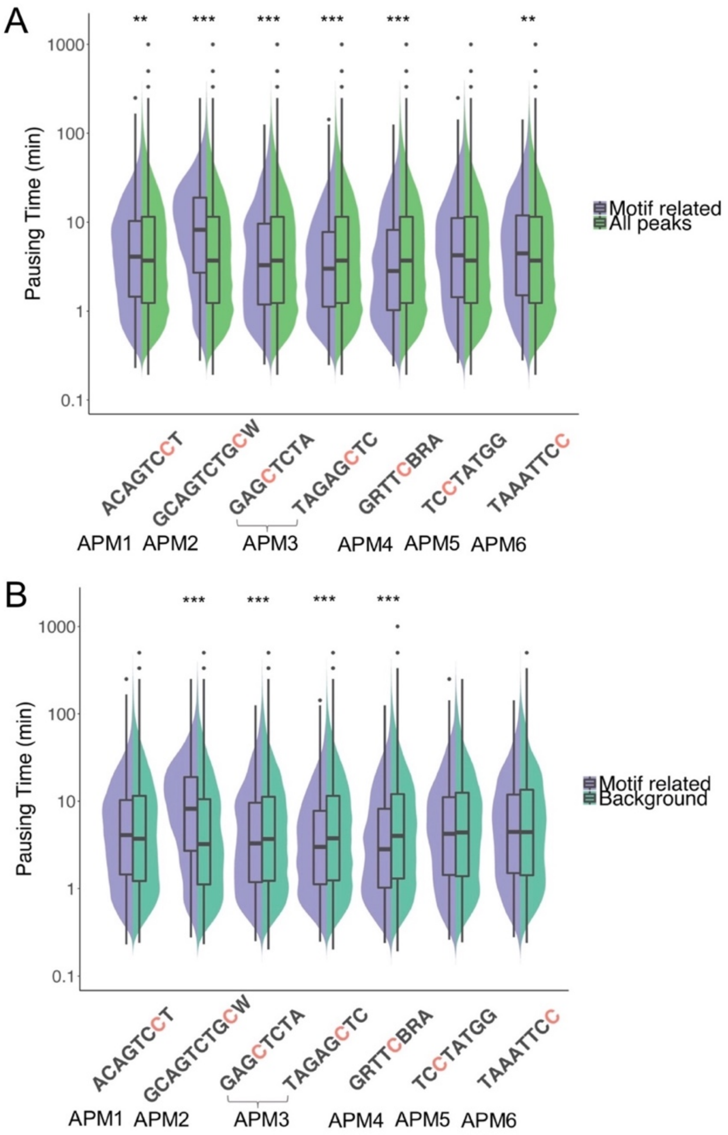


Figure 4.6 Pausing time of APMs

(**A**) Pausing time comparison for peaks at each APM and all peaks in the whole genome. **P<0.01, ***P<0.001, Mann-Whitney U test.

(**B**) Pausing time comparison for enriched motifs at peaks and nearby background sequences. **P<0.01, ***P<0.001, Mann-Whitney U test.

Again I turn to APM1 as example for a more detailed analysis of its associated pausing times. As shown in Figure 4.5, polymerase always pauses on position 7, regardless of position 1, 2, 5, and 8's nucleotides. The nucleotides of these positions do not change the accuracy of pausing, but influence the strength of pausing (Figure 4.7). For instance, polymerases pause on CCAGTCCT for significantly shorter times than on ACAGTCCT and TCAGTCCT (Mann-Whitney U test, p-value < 1×10^{-5}). GCAGTCCT pauses even shorter than CCAGTCCT (Mann-Whitney U test, p-value < 1×10^{-3}) (Figure 4.7 A). Similar to position 1, positions 2, 5, 8 also influence pausing time of APM1 without changing the enrichment site.

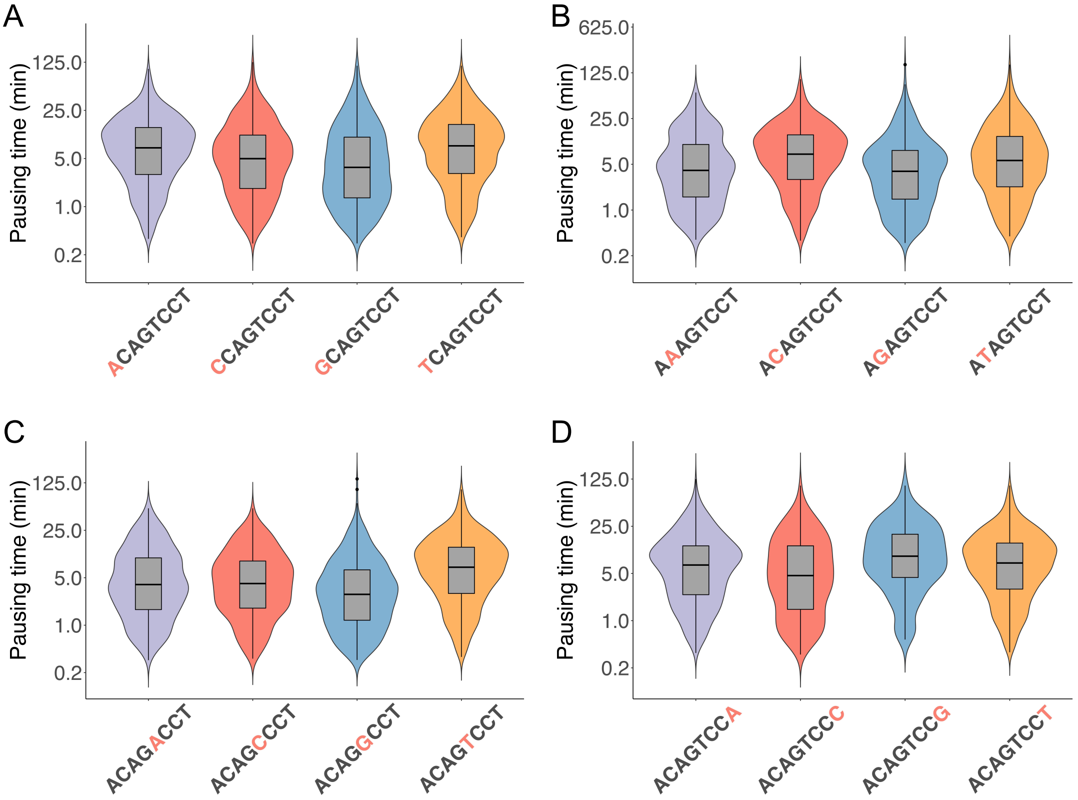


Figure 4.7 Nucleotide variants influence pausing time

A. Pausing time of APM1 with variants of the first nucleotide.

B. Pausing time of APM1 with variants of the second nucleotide.

C. Pausing time of APM1 with variants of the fifth nucleotide.

D. Pausing time of APM1 with variants of the eighth nucleotide.

Interestingly, if I consider dinucleotide variants of the first two positions, I observe systematic effects of individual bases on the pausing times of the downstream peaks (Figure 4.8 A). This pattern would be unlikely to appear by chance (Kendall tau test, all P < 10^{-6}; background pausing times do not show such a pattern, Figure 4.8 B) and agrees with elementary biochemical considerations relating affinity to lifetime of an interaction; it suggests functional relevance of the motif.

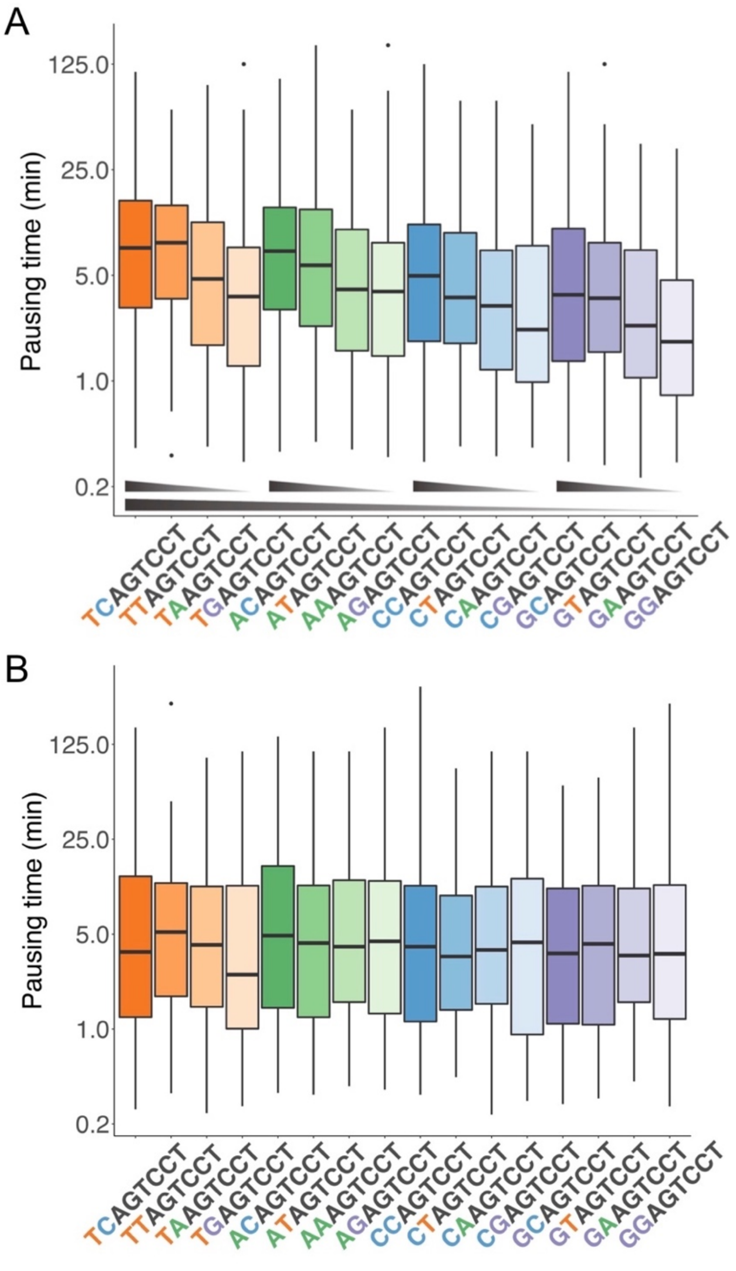


Figure 4.8 Dinucleotide variants of the APM1 show systematic effects on the pausing times

A. Black triangles were added to better illustrate the trends. Trends among all groups of four were assessed with Kendall's tau test and were found to have $P < 10^{-6}$ in all cases (H_1: tau ≠ 0).

B. The pausing times of 'Background peak' of peaks in (A), which refers to the pausing peaks within a distance of 20 to 40bp of the accurate pausing site. Trends among all groups of four were assessed with Kendall's tau test and were found to be not significant (H_1: tau ≠ 0).

4.4 Discussion

Even though pausing in the promoter proximal region is highly researched, the study of pausing mechanisms of Pol II in other regions, especially the gene body, has been limited. Here I showed that two different histone modifications, H3K9me3 and H3K36me3, can induce long pausing and how the sequence around pausing sites influences accuracy and duration of the pausing. These findings can help improve understanding of expression regulation and potentially assist in the design of BioBricks and other synthetic biology endeavors.

Pausing has been proposed to regulate gene expression[2, 12]. However, the mechanism of the regulation is still elusive. Histone acetylation has been proposed to loosen chromatin and increase nucleosome turnover, thereby helping polymerases to overcome the nucleosome barrier[29]. H3K36me3 reduces nucleosome turnover by facilitating histone deacetylation and remodelling of repressive chromatin [119], which might explain its association with long pausing. A tug of war between H3K36me3 and histone acetylation may function as speed control for elongation: paused polymerase is released by demethylation of H3K36me3 and histone acetylation after a heat shock, thus raising the elongation rate of polymerase (Figure 4.9A). I hypothesize that the reason that H3K36me3 is an active marker of expression but also associates with long pausing is because it is deposited in the wake of elongating Pol II rather than functioning as a pre-set, static marker. Methylation of H3K36 is carried out co-transcriptionally by the Set2 complex which is recruited by the carboxy-terminal domain (CTD) of Pol II[126]. H3K36me3 might thus act as a ‘speed bump’ to prevent collision with a succeeding polymerase (Figure 4.9B). This would also explain why a loss of Set2 only slightly influences expression levels of H3K36me3 positive genes[127].

The reason that H3K9me3 and H3K36me3 have not been found to be related to pausing before is maybe because these two markers are insufficiently present at the PPR (Figure 5.3). As the PPR has a much higher peak density than other regions (Figure 3.4 and 3.8), the relationship between these two markers and long pausing might be obscured by the opposite signal in the PPR. Since TV-PRO-seq can discriminate between peaks with variable pausing times, the relationship between long pausing and the two markers can be isolated.

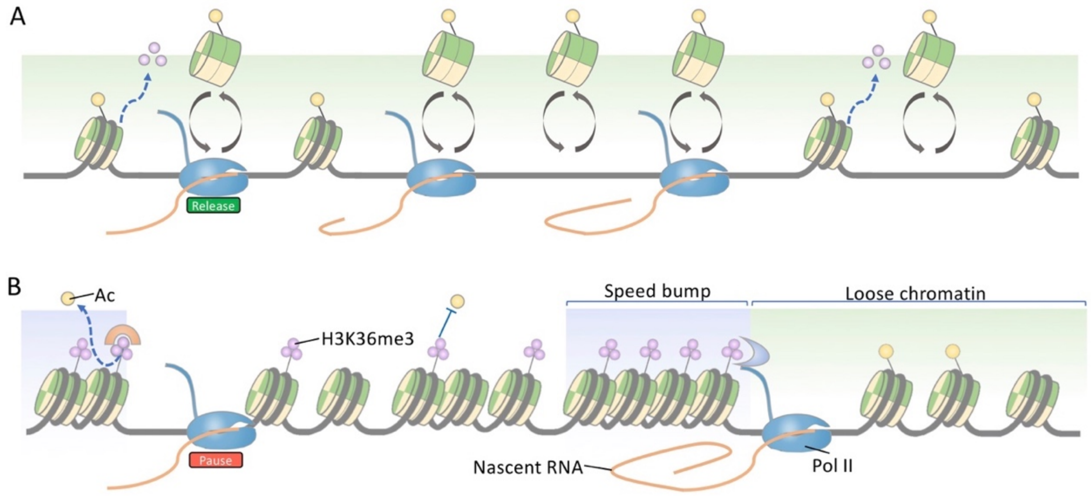


Figure 4.9 Elongation rate regulation by H3K36me3 and the dynamic equilibrium of histone acetylation

A. Histone acetylation releases paused polymerase after a heat shock.

B. Model of the dynamic equilibrium between H3K36me3 and histone acetylation under homeostasis.

The other mechanism I focused on is elemental pausing. I showed that the nucleotide of the +1 template DNA / pausing release site is essential for the accuracy of pausing. The 3' end of nascent RNA is less important compared to the pausing release site. The sequence involved in elemental pausing is mostly concentrated on the DNA-RNA helix and positions +1 to +3 of the template DNA relative to the pausing site (Table 4.1). This region also corresponds to the positions which have strong interactions with Pol II[128]. Some nucleotides in these motifs influence the accuracy of pausing (Figure 4.5), while the rest influence the strength of pausing (Figure 4.7 and 4.8).

As productive elongation consumes more than 95% of the total time transcription takes and polymerase pauses about every 100nt[13], pausing in gene bodies could potentially be the rate-limiting step of transcription. Unlike other methods, TV-PRO-seq estimates pausing time based on the pausing sites themselves, thus can evaluate pausing in gene bodies. My result illustrates the power of TV-PRO-seq and the insights into the mechanism of pausing in gene body it can produce.

Chapter 5 Pausing and gene regulation

5.1 Introduction

As the first step of gene expression, transcription is a key step for expression regulation. The consensus view is that transcriptional regulation is focused on the upstream processes of transcription, especially assembly of the pre-initiation complex[33]. Histone modifications near TSSs, such as acetylation, H3K4me2 and H3K4me3 methylations, or H2A.Z recruitment, are usually believed to mark active promoters[129]. Thus the genome coverage of these active markers would be expected to increase at the core promoters of stress responding genes after stimulation. However, recent studies show that the chromatin states stay the same in promoters of responding genes to a HS (heat shock) during a HS[130]. These results suggest that the promoter states have been pre-set for quick responses to stimuli.

Actually, it is very common that transcription becomes aborted. Only 12.7% of polymerases can be released from the promoter and enter elongation after initiation, while the rest will drop off chromatin after about 2.4 seconds; only 7.6% of polymerases continue to proceed to productive elongation. Overall, only 1% of initiation events lead to productive elongation[6]. This extremely high rate of abortive transcription suggests that failed initiation and early termination of transcription are key steps of transcriptional regulation.

Beyond these steps, the elongation rate during the productive elongation phase could potentially serve as a rate limiting step as well. The vast majority of polymerase cannot enter productive elongation; those that do, generate full length transcripts spend more than 96% of transcription's total length in productive elongation, which is 23min on average[6]. Thus regulation that happens upstream of transcription initiation will take long to take effect on the resulting mRNA numbers. Histone acetylation corresponds to a stationary state with high turnover while the nucleosome coverage remains constant. The former has been found in stress responding genes after correlated stimulation[29, 131] and facilitates the polymerase's overcoming of the nucleosome barrier thus accelerating its engagement. These findings suggest different modes of regulating expression; tuning the elongation rate in the gene body could be a way to control rapid reactions to the environment. The fast completion of semi-finished transcript can produce a rapid mRNA wave upon stimulation. Regulating expression upstream of elongation on the other hand could be responsible for longer term reactions.

In contrast to stress response genes, polymerases do not enrich in the PPR of housekeeping genes. This difference in pausing characteristics has been proposed to be responsible for the different modes of regulation of stress response genes and housekeeping genes.

In contrast to stress response genes, some studies suggest housekeeping genes might have less fluctuation of expression level[132]. Differences in chromatin state and polymerase occupancy in the PPR have been suggested to be the responsible for the variation in transcriptional dynamics[1]. Furthermore, pausing in gene bodies has also been suggested to influence transcriptional noise.

Modelling work suggests that both longer pausing time and higher pausing frequency can result in higher transcriptional noise[31, 133, 134]. Here I re-analyse a PRO-seq dataset for a heat-shock response and show that a global pausing release take place after the heat-shock. I further found that genes with higher transcriptional noise have more pausing sites along the whole gene, especially in the gene body. Compared to the extremely significant difference of pausing frequency, pausing times only show minor differences between genes with different transcriptional noise levels.

5.2 Methods

5.2.1 Calculation of local pausing index

The polymerase occupancy of a genomic position equals the product of polymerase flux and the average residence time of each polymerase (Chapter 1.2.3). If we can remove the influence of polymerase flux on polymerase occupancy, we can determine the average residence time of polymerase. I defined the 'local pausing index' (LPI) to achieve that. Polymerase pausing occurs at specific genomic positions[44]. This suggests that polymerase occupancy in a region surrounding a pausing site will not be strongly influenced by the pausing site itself or its regulation. The polymerase *flux* (Chapter 1.2.2), however, should be similar in the surrounding region and the pausing site. Thus, I defined the LPI as the average polymerase occupancy of +/−100-nt neighbourhoods around pausing peaks to normalize the occupancy of peaks.

5.2.2 Gene expression noise estimation and selection

Gene expression noise is estimated from single-cell sequencing data as[135]:

$$\eta = \mathrm{CV}^2\text{-}1/\mu,$$

where μ is the mean mRNA number for a gene, and CV is its coefficient of variation. I selected genes with the highest and the lowest noise heuristically, taking into account the dependence of η on μ as follows. I processed the single-cell sequencing dataset of [136] with the custom script `Rank_eta.pl`. This first sorts the genes into a list by their mean expression. It then moves a sliding window of size *WS* = 100 along this list and, at each position of the window, ranks the genes with regards to the value of η and records these ranks. For each gene in the list, a number *WS* of ranks results, of which the top and bottom ranks are averaged to give the 'noise score'. I refer to genes within the top and bottom 5% noise scores as 'high noise' and 'low noise' genes. For genes with equal noise scores, this procedure was repeated for *WS* = 20 and *WS* = 500, and rescaling the resulting noise scores to the range 0 to 100, followed by averaging across the three noise scores (Figure 5.1).

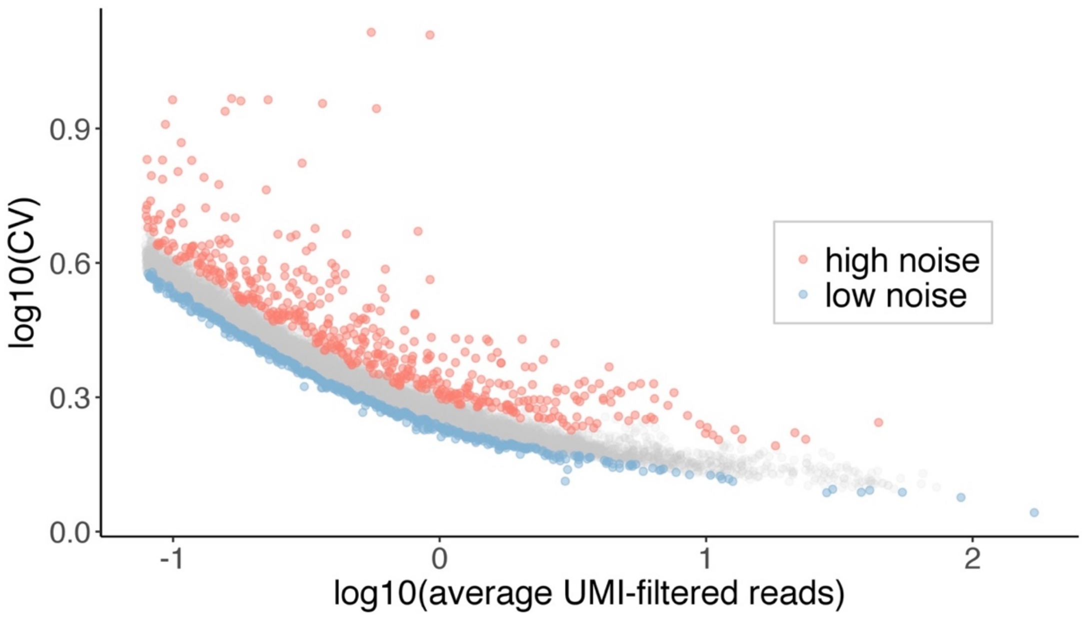


Figure 5.1 Selection of high/low-noise genes

I generated the smoothed conditional mean plot of the 'high noise' and 'low noise' genes with the same strategy as for the meta gene analysis (Figure 6B) and plotted histograms to show the absolute frequencies of peaks from 'high noise' and 'low noise' genes (Figure 6A). Density plots (Figure 6D) and split violin plots (Figure 6C) were generated with ggplot2 as before.

5.3 Results

5.3.1 Global pausing release after heat shock

NELF and DSIF mediated Pol II enrichment in the promoter proximal region has been considered to be a rate limiting step for transcription[2, 29, 34]; as I have demonstrated, this enrichment is not caused by pausing, but more likely due to polymerase turnover associated with abortive transcription (Or this effect can be removed by sarkosyl). The role of polymerase pausing in the control of expression therefore needs to be reconsidered.

Nucleosomes are regarded as barriers that can stop elongating polymerase[51, 52]. As responses to various cellular stresses such as heat shocks typically elicit widespread changes of nucleosome accessibility [29, 131], I presumed that the global pausing profile might also change. To take a closer look at this, I used mouse PRO-seq data following a heat shock for analysis[137]. Since only a single run-on timepoint was used in this study, I could only use polymerase occupancy for the analysis. I therefore normalized the size of pausing peaks to the average read densities adjacent to these, which is akin to a local pausing index (LPI) (See 5.2.1); the higher this value, the longer the average residence time.

The LPI decreases at pausing sites after 2.5min heat shock, indicating a global release of paused polymerase at this early time point (Figure 5.2 A and B). This release is rapid and soon stopped (Figure 5.2 C). The LPI then starts increasing again and approach pre-heat shock levels. It continues to grow after 60min heat shock, possibly indicating an over-compensating restoration mechanism to reset pausing to default levels (Figure 5.2 A and D). This peak release and recover surge occurs globally, and peaks in the region close to TSS (TSS to +200) show a similar pattern to peaks elsewhere (Figure 5.2 B-D).

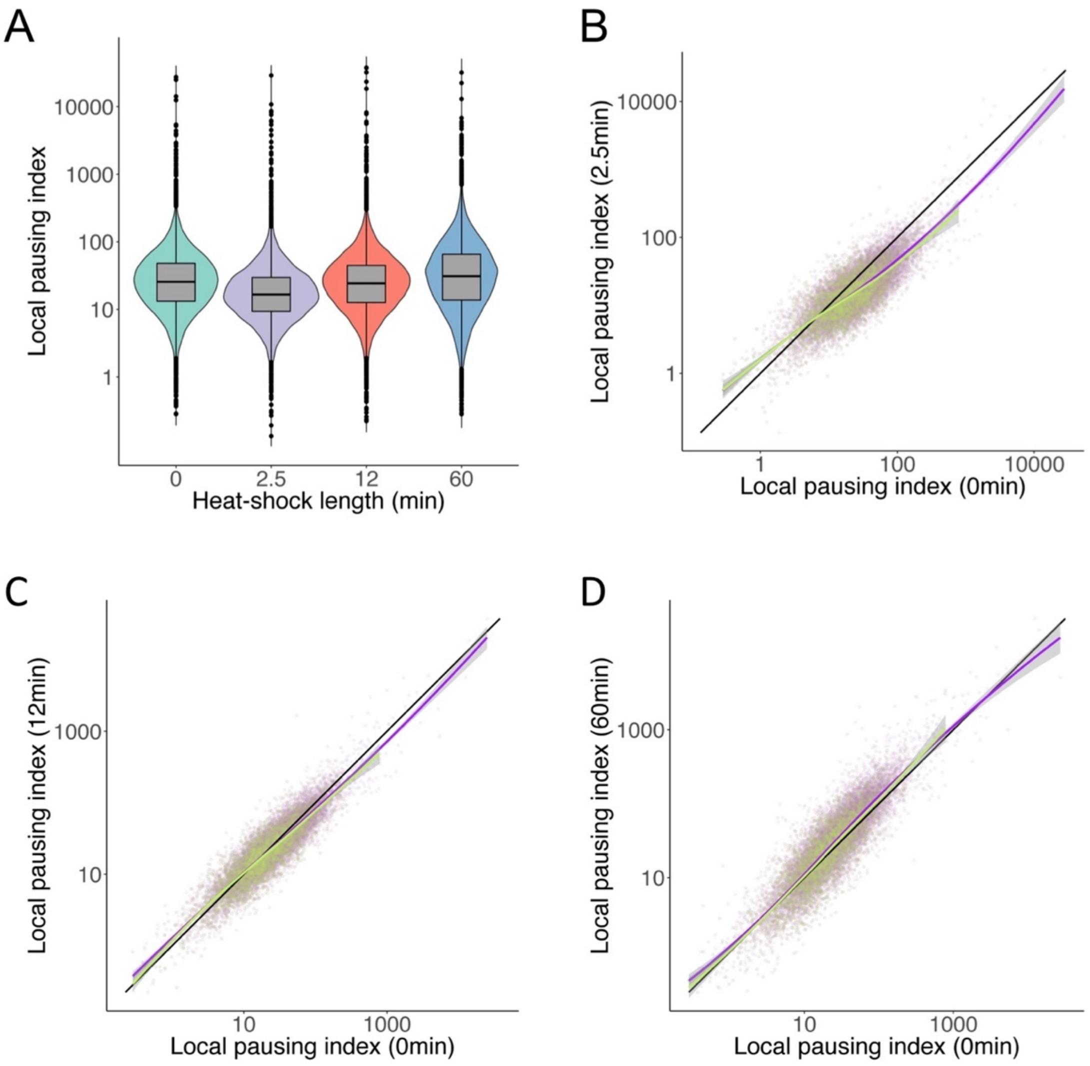


Figure 5.2 Local pausing indices change after heat shock

A. Local pausing indices of peaks without heat shock and with 2.5, 12, 60-min heat shock. All pairwise comparisons have p << 0.01, Mann-Whitney U test, Bonferroni corrected.

B. Pausing peaks at different regions behave similarly. A scatter plot shows the change of local pausing index (LPI) between no heat shock and 2.5min heat shock. The purple points represent all peaks and the green points refer to peaks within the first 200nt of genes. The black line indicates no change. The purple and green lines correspond to the moving averages of points in the same colour, the gray shading indicating the 0.95 confidence interval (LOESS fit).

C. Similar to (B), after 12min heat shock.

D. Similar to (B), after 60min heat shock.

Interestingly, this polymerase release is not only found in genes induced by 2.5min heat shock, but also in those repressed by it (Figure 5.3A). This suggests that the mechanism involved in this is general, rather than a gene-specific response mode. Other potentially rate limiting steps such as RNA processing or transcription termination might be involved in orchestrating the heat shock responses of different classes of genes instead.

By repeating this analysis for the longer heat shock of 60 min and separately considering repressed and induced genes again, we obtain a different picture; the LPI increases for genes that are repressed, but remains unchanged for induced genes (Figure 5.3B). This suggests that long/more pausing of polymerase plays a role in the repression of genes upon a long-term heat shock. For induced genes, the up-regulation might act upstream of transcription initiation, potentially for energy saving purposes.

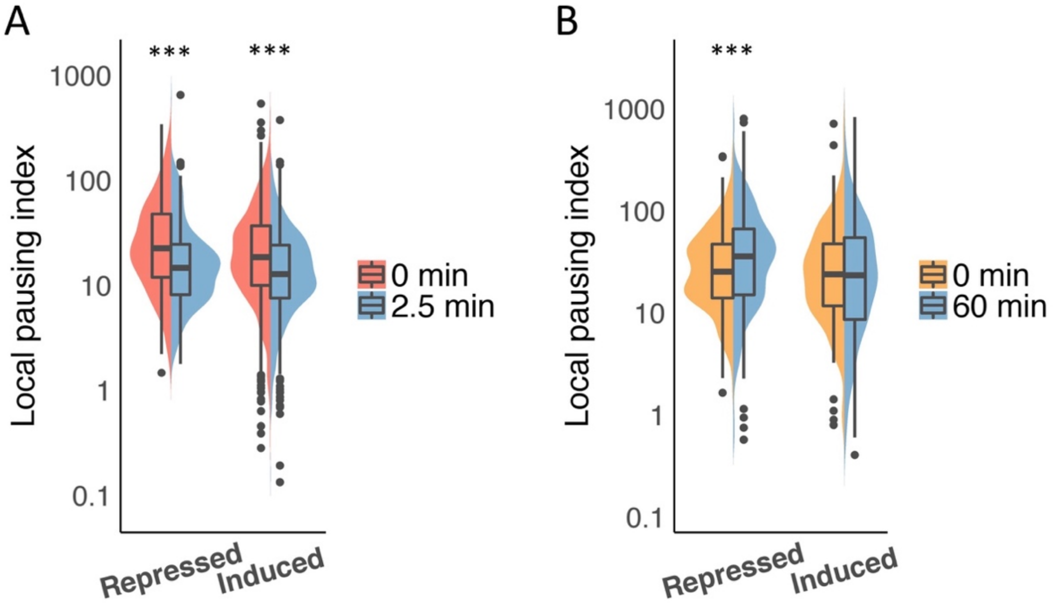


Figure 5.3 Local pausing index change of heat shock induced and repressed genes

A. Genes with the top 10% of read increases in their gene bodies after 2.5-min heat shock were classified as ‘induced’ genes, and the bottom 10% as ‘repressed’. The LPI difference between no heat shock and 2.5-min heat shock is shown for the two groups *of genes. For induced genes, $P < 10^{-13}$; for repressed genes, $P < 10^{-6}$, Mann-Whitney U test.*

B. Genes were classified as in (A), this time for 60-min heat shock. The LPI difference of no heat shock and 60-min heat shock is shown for the two groups of genes. For repressed genes, $P < 10^{-3}$ Mann-Whitney U test.

5.3.2 Polymerase pausing and transcriptional noise

A gene's expression level is determined by its initiation rate, the fraction of nascent RNA that is turned into matured RNA, and the latter's degradation rate. Polymerase pausing that adjusts the elongation rate therefore will not influence the expression level; however, it will result in the dispersed distribution of mRNAs among individual cells [31]. This dispersion, or 'noise', is quantified by the CV^2 and can be obtained in genome-wide fashion from single-cell RNA-seq (e.g. Drop-seq) data. To study the relation of noise and pausing, we used Drop-seq data for HEK293 cells[136] and classified genes based on their CV^2 for a moving average of mean expression levels. This reduces influence of the latter, which the noise depends on[135, 138] (Figure 5.1).

We assigned genes to 'low-', and 'high noise' classes. We find that, overall, noisy genes have significantly higher pausing frequency (the number of pausing peaks in a given region) throughout gene bodies (Figure 5.4A), while pausing times in most genic regions are similar (Figure 5.4B). An exception is the region following the promoter proximal dip in pausing times, where Pol II pauses significantly longer in noisy genes (Figure 5.4C). We term this region the variable pausing region. Unlike the minor difference of pausing *times* between low and high noise genes, pausing *frequency* shows a significant difference (Figure 5.4A). This suggests that the control of transcriptional dynamics relies on a joint effect exerted by multiple positions rather than the variation of individual pausing sites' characteristics. If we consider the relative distributions of pausing peaks within genes, we observe a mild shift of pausing positions away from the promoter proximal region to other parts, including the variable pausing region and exons (Figure 5.4D). These results shift the focus of potential links between polymerase pausing and transcriptional noise away from promoters[4] and towards internal genic regions, in agreement with previous theoretical considerations[31, 133, 134].

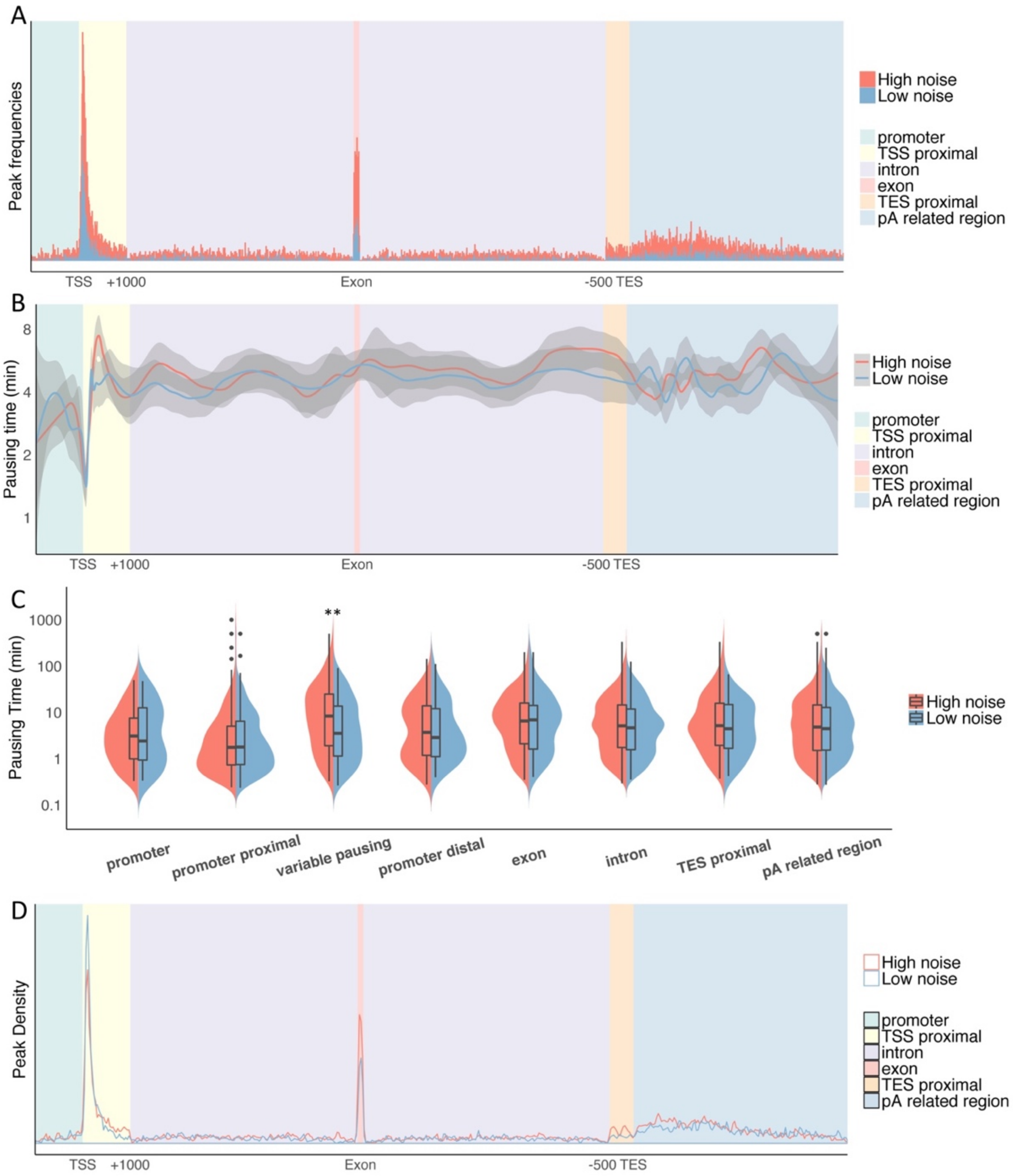


Figure 5.4 Pausing profiles and transcriptional noise.

A. Absolute peak density at mRNA- transcribing metagene as in Figure 3.8, for genes classified into different levels of transcriptional noise ('high', 'low'; red, blue, respectively).

B. Pausing times of pausing peaks among genic regions for 'low' and 'high' noise genes at the metagene as in (A) shown as LOESS fits as in Figure 3.8

C. Pausing times of different regions of high and low noise genes in (A). The promoter proximal region was defined as the first 200nt of a gene, the variable pausing region as the following 300nt, the promoter distal region as +500 to +1000nt from TSS, the TES proximal region as 500nt upstream of the TES, and the pA related region as the 4500nt downstream of TES. Finally, other regions in the gene body were classified into exon and intron. For the variable pausing region, $P < 10^{-3}$, Mann-Whitney U test.

D. Densities (so that the areas under the peaks are equal for the metagene) of pausing peaks among genic regions for 'low' and 'high' noise genes at the metagene as in (A).

5.4 Discussion

One obstacle for an in-depth dissection of transcriptional pausing is the influence of different factors that will have similar effects on polymerase occupancy (Figure 1.2). Removing the influence of polymerase flux from polymerase occupancy is a major goal. Efforts to address this issue led to development of the 'pausing index'[22, 38]. However, this index is based on the assumption that positions within the same gene share the same polymerase flux. Instead, here I used reads in positions in the neighbourhood of pausing peaks to normalize reads of the latter to obtain the local pausing index, LPI. The LPI reduces the bias introduced by the polymerase flux within a gene.

I found that the LPI reduced in genome wide fashion after 2.5min heat shock (Figure 5.2 A, B). An increase in nucleosome accessibility and/or histone acetylation after the heat shock might be responsible for this pausing release[29]. Notably, in contrast to significant changes of the nucleosome arrangements and histone acetylation in gene bodies, a recent study shows that the chromatin conformation of promoter regions of response genes remained unchanged after heat shocks[130]. This implies that a rapid response to heat shocks relies on regulation downstream of transcription initiation, more specifically, on the acceleration of elongation. An elevation of local histone acetylation and nucleosome accessibility in gene bodies is not only found after heat shock, but also with other stress response reactions, such as the unfolded protein response[131]. Pausing release mediated by loosening nucleosomes through histone acetylation[51, 52, 107, 108] might not only serve as rapid reaction to heat shock, but could be involved in a modulation or fine tuning of stress responses in general.

Since I showed that polymerase enrichment in the PPR is not caused by a single or small number of long pausing position(s), a re-examination of the links between pausing and bursting seemed prudent. Upon integrating my TV-PRO-seq data with single cell sequencing results, I found that the pausing times of individual pausing sites exhibit only minor differences between high and low noise genes (Figure 5.4B). However, the pausing frequency for these two groups of genes varies substantially (Figure 5.4A, D). I also see a difference in pausing frequency between exons and introns (Figure 5.4A, D). This implies that Pol II dynamics as reflected in noise and/or varying intronic/exonic elongation rates are subject to an influence exerted in a similar way by multiple positions, instead of the modulation of individual pausing sites' characteristics.

As Figure 5.5 shows, different pausing related TFs can be imagined to bind various genes for elongation rate control. Each TF corresponds to a different pathway, and a single gene can contain motifs binding to multiple TFs (Gene1 and Gene5). These TF binding sites ensure that Gene1 and Gene5 can respond to multiple pathways. Gene2 and Gene3 only have TF2 bound to their gene bodies. As Gene2 has more pausing sites, its elongation rate is lower than Gene3's. This also means that, while Gene2 and Gene3 have the same expression level and gene length, more polymerase will be located on Gene2. Thus, Gene3 can generate bursts of larger sizes upon correlated stimulations. Gene4 can be regarded as a housekeeping gene which is likely to have fewer TF binding motifs in their gene body. In contrast to pausing in the gene body, pausing in the PPR also can influence the expression level[4]. Thus, a high frequency of pausing can be found in the PPR of housekeeping genes. These TFs maybe not directly bind to Pol II, but may be associated with methylation of H3K36 and H3K9. This is why pausing times in exons and introns do not show significant differences (Figure 3.8), as is the case with high noise and low noise genes (Figure 5.4B).

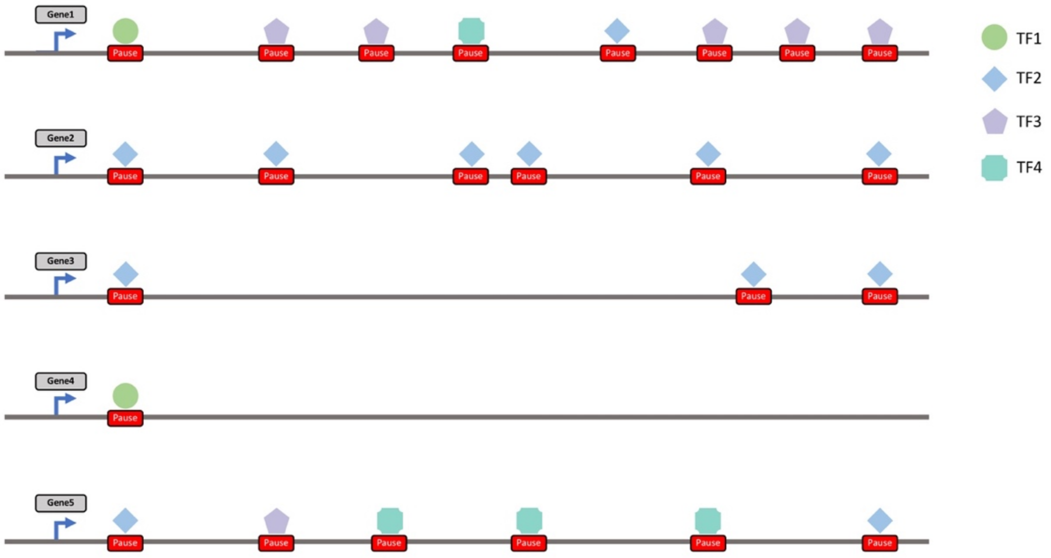


Figure 5.5 Complex elongation regulation system formed by multiple pausing sites.

Numbers, types, and positions of bound TFs and their interactions are expected to influence the pausing profiles of genes, as illustrated.

My result suggests the importance of pausing frequency towards temporal expression of genes. In addition, my data highlight a role for pausing in gene bodies. Overall, these findings provide new perspectives for the research of gene regulation.

Chapter 6 Conclusion

6.1 Introduction

Appropriate spatial and temporal expression of genes is required for various biological processes, including development, stress response, differentiation and adaptability in organisms[2, 12, 13]. As pausing functions in nearly all actives genes [3, 30, 83], together with PIC assembly, pausing has been considered as the rate-limiting step of gene expression in metazoan[34, 48].

Various sequencing methods have been designed for investigating promoter proximal pausing from different aspects, including Start-seq, NET-seq (mNET-seq), GRO-seq (PRO-seq, coPRO) and ChIP-seq (ChIP-nexus) of Pol II[139]. To evaluate promoter proximal pausing levels of each gene, the pausing index[22, 38] was developed. By using average polymerase occupancy in gene bodies to normalize polymerase occupancy in the PPR, genes with higher pausing index have been suggested to have strong pausing. However, evidence has been accumulating that a substantial fraction of transcription events terminate early in the promoter proximal region[5, 6, 50, 98]. As both strong pausing and abortive transcription will lead to a higher pausing index, a method that can remove the influence of abortive transcription from pausing is required.

Trp, which blocks transcription initiation, has been introduced to investigate pausing time by treating cells with it prior to sample preparation[3, 4, 49]. According to experiments based on Trp treatment, genes have on average 2 to 8 min promoter proximal pausing. Some genes even have long pausing times in the PPR that can exceed half an hour[3, 4, 49]. However, a recent in vivo experiment based on FRAP suggests that pausing in the PPR lasts only about 42s[6], which is approximately 1/5 of previous suggestions from Trp treatment. The difference might be due to the slow uptake and function of Trp[50].

Pausing does not only occur in the PPR, but also happens in the gene body. RNA polymerase has been found to pause every 20-100 bp in bacteria and yeast[13]. However, there does not exist a method that can measure the pausing time of pausing sites in the gene body in genome wide fashion. Sequencing-based methods following Trp treatment can only detect the overall pausing time of PPR, and FRAP has low resolution and cannot show the genome location of the pausing site. For investigations of the pausing time differences between pausing sites in the PPR and the gene body, a high-resolution method which can reveal pausing times genome widely is required.

6.2 Overview of TV-PRO-seq

I developed TV-PRO-seq, a PRO-seq based method, which enables one to evaluate pausing time of single pausing sites across the whole genome. TV-PRO-seq does not only allow us to compare pausing time differences of pausing sites in the PPR and the gene body, but also provides a route towards deeper understanding of pausing profile from various aspects. For instance, the influences of epigenetic modification, pausing related TFs and consensus sequences of pause sites towards pausing time can be identified by TV-PRO-seq. TV-PRO-seq provides a way to investigate pausing of each pausing site only by its pausing time rather than the polymerase occupancy.

The general overview of the principle of TV-PRO-seq is shown in Figure 6.1 A-C. Eight parallel PRO-seq samples with individual run-on reactions are required for TV-PRO-seq. To minimize the differences in the distribution of RNA polymerases between samples, cells for run-on should be prepared under the same conditions. Thus I mixed cells for permeabilization and then separated them into 8 tubes for run-on reactions (Figure 6.1A). After biotin-NTP is incorporated into the 3’ end of nascent RNA, further incorporation of NTP is inhibited. Therefore nascent RNAs carried by active RNA polymerases will be labelled with biotin-NTP on their 3’ ends. Reads in longer pausing sites will reach the threshold later as the latter have lower release rates (Figure 6.1 B-C). Based on a simple Bayesian model, pausing release rates are calculated and pausing times are further identified as the reciprocal of the release rates.

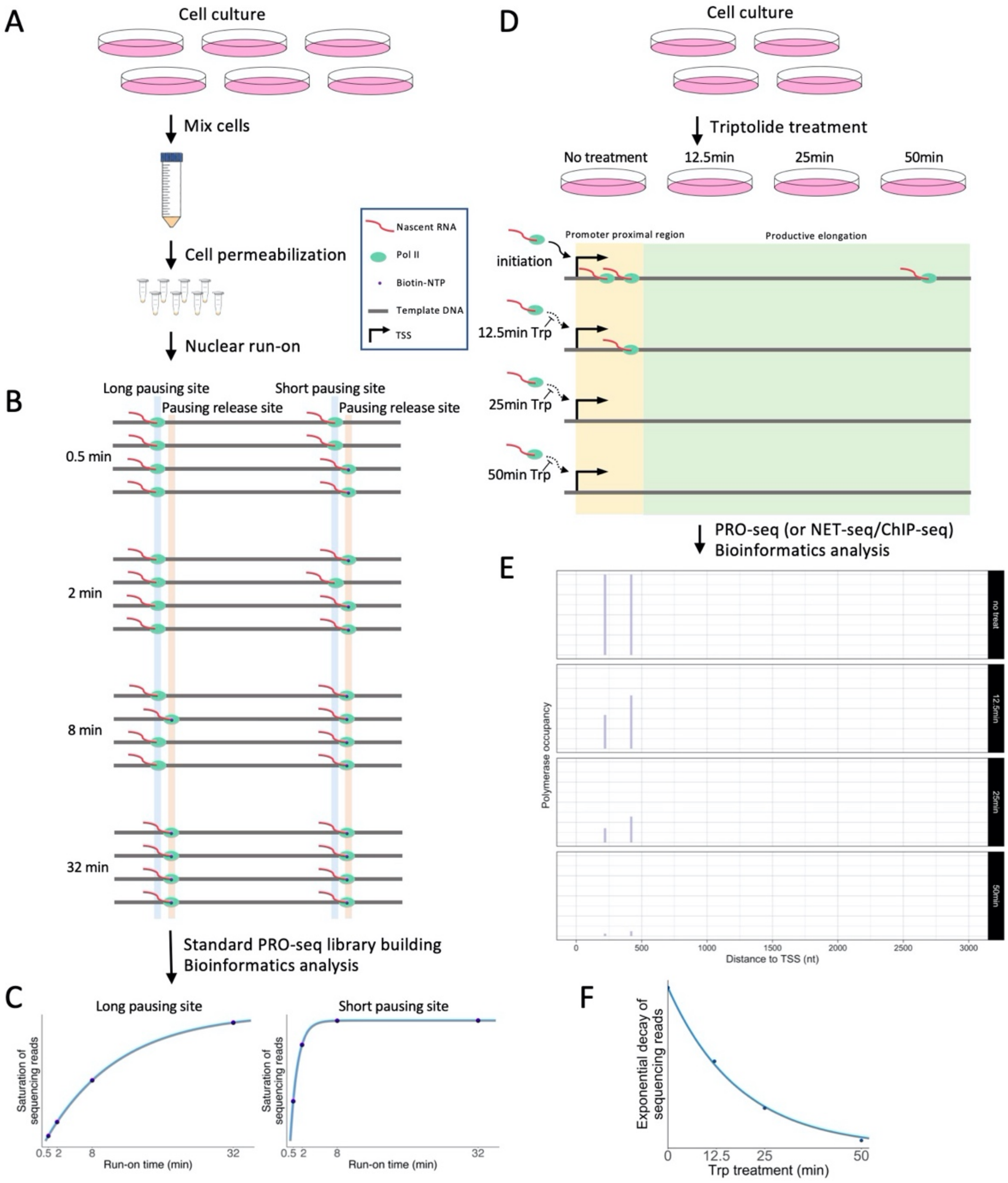


Figure 6.1 Schematic diagram of TV-PRO-seq and Trp treatment followed by sequencing

A. Permeabilizated cells preparation for TV-PRO-seq.

B. Diagrammatic explanation of different pausing release rates of peaks with different pausing times when using variable run-on times.

C. Fitting of reads of ideal peaks reveal different pausing times.

D. Schematic diagram of experimental procedure of Trp treatment followed by sequencing.

E. Ideal result of polymerase occupancy of a gene which has two pausing sites with 10min pausing in the PPR.

F. Curve fitting to reads of the PPR reveals the total pausing time of this region.

6.3 Comparison with sequencing following Trp treatment

Traditionally, Trp treatment followed by sequencing has been applied to study pausing time in the PPR[3, 4, 49]. By adding the TFIIH inhibitor Trp, initiation of transcription is blocked. Sequencing methods for nascent RNA (ChIP-seq/GRO-seq/PRO-seq/NET-seq) are then applied to reveal changes in the polymerase occupancy after initiation inhibition. An exponential decay typically results for the reads in the PPR (Figure 6.1 D-F). Based on this, pausing in the PPR has been suggested to be as long as 2min to even more than half an hour[3, 4, 49]. However, this estimation actually accounts for the sum of the Trp uptake time and promoter proximal pausing. Based on the assumption that Trp uptake is relatively quick comparing to pausing in the PPR, the half-life of the polymerase occupancy's decrease in the PPR is taken as the pausing time.

Recently, Trp has been shown to have a slow uptake[50], and pausing in PPR is only about 1min[6]. Both of these facts indicate that the pausing time of PPR measured by Trp treatment has been overestimated. TV-PRO-seq is based on the incorporation of biotin-NTP which allows the method to function independently of Trp treatment. Thus, TV-PRO-seq results will not be influenced by the Trp uptake time.

Pausing time measurements based on the inhibition of transcription initiation also limit the potential to gain insights further downstream. As the block of incoming polymerases happens at the TSS, Trp treatment prior to sequencing will only work for the region adjacent to the TSS. As show in Figure 6.1E, the upstream pausing peak will serve as a reservoir that supplies the downstream peaks with a polymerase flux for some time. This will lead to an overestimation at downstream peaks if we evaluate pausing times individually. The Biotin-NTPs will block polymerases from moving downstream of pausing sites, which also means that inhibition of incoming polymerase will not be influenced by polymerase upstream of pausing sites. This allows TV-PRO-seq to be used for peaks regardless of their distance to the TSS.

Even though sequencing methods such as PRO-seq and NET-seq have single base resolution, they lose it when they are used for measuring pausing time based on Trp treatment. In contrast, TV-PRO-seq maintains the high resolution which enable it to reveal pausing times of motifs or epigenetic modifications related to pausing.

6.4 Application of TV-PRO-seq

TV-PRO-seq is the first method which can measure RNA polymerase pausing genome-wide with single base resolution. This allows it be applied in various analyses.

TV-PRO-seq not only can reveal pausing times of peaks in the PPR, but also in other regions. Even though pausing occurs frequently in the gene body[13], TV-PRO-seq is the only methodology available for estimating pausing times of peaks in the gene body. TV-PRO-seq highlights the importance of H3K36me3 and H3K9me3 for pausing in the gene body as it enables a systematic meta-analysis of pausing in various genic regions. Also, it can help understanding of the pausing profiles of genes transcribed by Pol I and Pol III.

Its single base resolution enables TV-PRO-seq's application to motif analysis. This revealed that some nucleotides close to pausing sites are not essential for establishing pausing function but rather for controlling pausing time.

TV-PRO-seq results can be integrated with other datasets such as those derived from ChIP-seq and single cell RNA-seq assays. Such analyses are critical for dissecging the relationships between pausing time and histone modification/transcriptional noise.

Overall, with different treatments or cell lines, TV-PRO-seq has great potential to investigate various topics, including, for example, systems to induce heat shocks or to knock down TFIIS. I expect TV-PRO-seq's strength in directly studying pausing time rather than polymerase occupancy to be very fruitful in several areas, eventually leading to a much deep understanding of pausing.

Bibliography

1. Gilchrist, D.A. et al. Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell* **143**, 540-551 (2010).

2. Adelman, K. & Lis, J.T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13**, 720-731 (2012).

3. 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 (2014).

4. Shao, W. & Zeitlinger, J. Paused RNA polymerase II inhibits new transcriptional initiation. *Nat Genet* **49**, 1045-1051 (2017).

5. 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 (2017).

6. Steurer, B. et al. Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. *Proc Natl Acad Sci U S A* **115**, E4368-E4376 (2018).

7. Erickson, B., Sheridan, R.M., Cortazar, M. & Bentley, D.L. Dynamic turnover of paused Pol II complexes at human promoters. *Genes & development* **32**, 1215-1225 (2018).

8. Crick, F. Central dogma of molecular biology. *Nature* **227**, 561-563 (1970).

9. Kim, J.H. et al. Variation in human chromosome 21 ribosomal RNA genes characterized by TAR cloning and long-read sequencing. *Nucleic Acids Res* **46**, 6712-6725 (2018).

10. Khatter, H., Vorlander, M.K. & Muller, C.W. RNA polymerase I and III: similar yet unique. *Curr Opin Struct Biol* **47**, 88-94 (2017).

11. Sims, R.J., 3rd, Mandal, S.S. & Reinberg, D. Recent highlights of RNA-polymerase-II-mediated transcription. *Curr Opin Cell Biol* **16**, 263-271 (2004).

12. Levine, M. Paused RNA polymerase II as a developmental checkpoint. *Cell* **145**, 502-511 (2011).

13. Mayer, A., Landry, H.M. & Churchman, L.S. Pause & go: from the discovery of RNA polymerase pausing to its functional implications. *Curr Opin Cell Biol* **46**, 72-80 (2017).

14. Jonkers, I. & Lis, J.T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16**, 167-177 (2015).

15. Porrua, O. & Libri, D. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nature reviews Molecular cell biology* **16**, 190 (2015).

16. Pavri, R. et al. Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell* **143**, 122-133 (2010).

17. Weake, V.M. & Workman, J.L. Inducible gene expression: diverse regulatory mechanisms. *Nat Rev Genet* **11**, 426-437 (2010).

18. Henriques, T. et al. Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. *Mol Cell* **52**, 517-528 (2013).

19. Sawarkar, R., Sievers, C. & Paro, R. Hsp90 globally targets paused RNA polymerase to regulate gene expression in response to environmental stimuli. *Cell* **149**, 807-818 (2012).

20. Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R. & Young, R.A. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**, 77-88 (2007).

21. Muse, G.W. et al. RNA polymerase is poised for activation across the genome. *Nature genetics* **39**, 1507 (2007).

22. 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 (2008).

23. 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 (2010).

24. Carrillo Oesterreich, F., Preibisch, S. & Neugebauer, K.M. Global analysis of nascent RNA reveals transcriptional pausing in terminal exons. *Mol Cell* **40**, 571-581 (2010).

25. Churchman, L.S. & Weissman, J.S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368-373 (2011).

26. 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 (2013).

27. Nudler, E. RNA polymerase backtracking in gene regulation and genome instability. *Cell* **149**, 1438-1445 (2012).

28. Cheung, A.C. & Cramer, P. Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature* **471**, 249-253 (2011).

29. Vihervaara, A., Duarte, F.M. & Lis, J.T. Molecular mechanisms driving transcriptional stress responses. *Nature Reviews Genetics*, 1 (2018).

30. Mayer, A. et al. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell* **161**, 541-554 (2015).

31. Rajala, T., Hakkinen, A., Healy, S., Yli-Harja, O. & Ribeiro, A.S. Effects of transcriptional pausing on gene expression dynamics. *PLoS Comput Biol* **6**, e1000704 (2010).

32. Proudfoot, N.J., Furger, A. & Dye, M.J. Integrating mRNA processing with transcription. *Cell* **108**, 501-512 (2002).

33. Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386**, 569-577 (1997).

34. 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 (2015).

35. Danko, C.G. et al. Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol Cell* **50**, 212-222 (2013).

36. Gilchrist, D.A. et al. NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. *Genes & development* **22**, 1921-1933 (2008).

37. Oesterreich, F.C. et al. Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell* **165**, 372-381 (2016).

38. Nojima, T. et al. Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161**, 526-540 (2015).

39. Ip, J.Y. et al. Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. *Genome Res* **21**, 390-401 (2011).

40. Shukla, S. et al. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74-79 (2011).

41. Proudfoot, N.J. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* **352**, aad9926 (2016).

42. Fong, N. et al. Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Mol Cell* **60**, 256-267 (2015).

43. 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 (2019).

44. 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 (2018).

45. Pugh, B.F. & Venters, B.J. Genomic Organization of Human Transcription Initiation Complexes. *PLoS One* **11**, e0149339 (2016).

46. Chu, T. et al. Chromatin run-on and sequencing maps the transcriptional regulatory landscape of glioblastoma multiforme. *Nat Genet* **50**, 1553-1564 (2018).

47. Darzacq, X. et al. In vivo dynamics of RNA polymerase II transcription. *Nat Struct Mol Biol* **14**, 796-806 (2007).

48. Bartman, C.R. et al. Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation. *Mol Cell* **73**, 519-532 e514 (2019).

49. Gressel, S. et al. CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* **6** (2017).

50. Nilson, K.A. et al. Oxidative stress rapidly stabilizes promoter-proximal paused Pol II across the human genome. *Nucleic Acids Res* **45**, 11088-11105 (2017).

51. Hodges, C., Bintu, L., Lubkowska, L., Kashlev, M. & Bustamante, C. Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science* **325**, 626-628 (2009).

52. Bintu, L. et al. Nucleosomal elements that control the topography of the barrier to transcription. *Cell* **151**, 738-749 (2012).

53. Maizels, N.M. The nucleotide sequence of the lactose messenger ribonucleic acid transcribed from the UV5 promoter mutant of Escherichia coli. *Proc Natl Acad Sci U S A* **70**, 3585-3589 (1973).

54. Gilbert, W., Maizels, N. & Maxam, A. Sequences of controlling regions of the lactose operon. *Cold Spring Harb Symp Quant Biol* **38**, 845-855 (1974).

55. 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 (1981).

56. Saldi, T., Cortazar, M.A., Sheridan, R.M. & Bentley, D.L. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J Mol Biol* **428**, 2623-2635 (2016).

57. 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).

58. 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).

59. Strobl, L.J. & Eick, D. Hold back of RNA polymerase II at the transcription start site mediates down-regulation of c-myc in vivo. *The EMBO Journal* **11**, 3307-3314 (1992).

60. 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).

61. Law, A., Hirayoshi, K., O'Brien, T. & Lis, J.T. Direct cloning of DNA that interacts in vivo with a specific protein: application to RNA polymerase II and sites of pausing in Drosophila. *Nucleic acids research* **26**, 919-924 (1998).

62. Rhee, H.S. & Pugh, B.F. ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Current protocols in molecular biology* **100**, 21.24. 21-21.24. 14 (2012).

63. He, Q., Johnston, J. & Zeitlinger, J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. *Nature biotechnology* **33**, 395 (2015).

64. McHaourab, Z.F., Perreault, A.A. & Venters, B.J. ChIP-seq and ChIP-exo profiling of Pol II, H2A.Z, and H3K4me3 in human K562 cells. *Sci Data* **5**, 180030 (2018).

65. Mahat, D.B. et al. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat Protoc* **11**, 1455-1476 (2016).

66. Ochs, S.M. et al. Activation of archaeal transcription mediated by recruitment of transcription factor B. *J Biol Chem* **287**, 18863-18871 (2012).

67. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17** (2011).

68. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360 (2015).

69. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

70. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).

71. Grothendieck, G. Non-linear regression with brute force. *R package version 0.2* (2013).

72. Polanski, K. et al. Bringing numerous methods for expression and promoter analysis to a public cloud computing service. *Bioinformatics* **34**, 884-886 (2018).

73. Salvatier, J., Wiecki, T.V. & Fonnesbeck, C. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* **2** (2016).

74. Al-Rfou R, A.G., Almahairi A, et al. Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint* (2016).

75. Fuller, C.W. et al. The challenges of sequencing by synthesis. *Nat Biotechnol* **27**, 1013-1023 (2009).

76. Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P. & Kocher, J.A. Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. *Brief Bioinform* **18**, 973-983 (2017).

77. Posse, V., Shahzad, S., Falkenberg, M., Hallberg, B.M. & Gustafsson, C.M. TEFM is a potent stimulator of mitochondrial transcription elongation in vitro. *Nucleic Acids Res* **43**, 2615-2624 (2015).

78. Barshad, G., Marom, S., Cohen, T. & Mishmar, D. Mitochondrial DNA Transcription and Its Regulation: An Evolutionary Perspective. *Trends Genet* (2018).

79. Yamaguchi, Y., Shibata, H. & Handa, H. Transcription elongation factors DSIF and NELF: promoter-proximal pausing and beyond. *Biochim Biophys Acta* **1829**, 98-104 (2013).

80. Fujita, T., Piuz, I. & Schlegel, W. The transcription elongation factors NELF, DSIF and P-TEFb control constitutive transcription in a gene-specific manner. *FEBS Lett* **583**, 2893-2898 (2009).

81. Rahl, P.B. et al. c-Myc regulates transcriptional pause release. *Cell* **141**, 432-445 (2010).

82. Patel, M.C. et al. BRD4 coordinates recruitment of pause release factor P-TEFb and the pausing complex NELF/DSIF to regulate transcription elongation of interferon-stimulated genes. *Molecular and cellular biology* **33**, 2497-2507 (2013).

83. Chao, S.H. & Price, D.H. Flavopiridol inactivates P-TEFb and blocks most RNA polymerase II transcription in vivo. *J Biol Chem* **276**, 31793-31799 (2001).

84. 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 (2015).

85. Kent, W.J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. The human genome browser at UCSC. Genome research. *Genome Res* **12**, 996-1006 (2002).

86. Kalvari, I. et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* **46**, D335-D342 (2018).

87. Oler, A.J. et al. Human RNA polymerase III transcriptomes and relationships to Pol II promoter chromatin and enhancer-binding factors. *Nat Struct Mol Biol* **17**, 620-628 (2010).

88. Dapprich, J. et al. The next generation of target capture technologies - large DNA fragment enrichment and sequencing determines regional genomic variation of high complexity. *BMC Genomics* **17**, 486 (2016).

89. Willis, I.M. & Moir, R.D. Signaling to and from the RNA Polymerase III Transcription and Processing Machinery. *Annu Rev Biochem* **87**, 75-100 (2018).

90. Dumay-Odelot, H., Durrieu-Gaillard, S., El Ayoubi, L., Parrot, C. & Teichmann, M. Contributions of in vitro transcription to the understanding of human RNA polymerase III transcription. *Transcription* **5**, e27526 (2014).

91. Schramm, L. & Hernandez, N. Recruitment of RNA polymerase III to its target promoters. *Genes Dev* **16**, 2593-2620 (2002).

92. Teichmann, M., Dieci, G., Pascali, C. & Boldina, G. General transcription factors and subunits of RNA polymerase III: Paralogs for promoter- and cell type-specific transcription in multicellular eukaryotes. *Transcription* **1**, 130-135 (2010).

93. Abascal-Palacios, G., Ramsay, E.P., Beuron, F., Morris, E. & Vannini, A. Structural basis of RNA polymerase III transcription initiation. *Nature* **553**, 301-306 (2018).

94. Lobo, S.M. & Hernandez, N. A 7 bp mutation converts a human RNA polymerase II snRNA promoter into an RNA polymerase III promoter. *Cell* **58**, 55-67 (1989).

95. Tregouet, D.A. & Morange, P.E. What is currently known about the genetics of venous thromboembolism at the dawn of next generation sequencing technologies. *Br J Haematol* **180**, 335-345 (2018).

96. FitzGerald, P.C., Sturgill, D., Shyakhtenko, A., Oliver, B. & Vinson, C. Comparative genomics of Drosophila and human core promoters. *Genome Biol* **7**, R53 (2006).

97. 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 (2008).

98. Zhang, J., Cavallaro, M. & Hebenstreit, D. Timing Polymerase Pausing with TV-PRO-seq. *bioRxiv*, 461442 (2018).

99. Kireeva, M.L. et al. Nature of the nucleosomal barrier to RNA polymerase II. *Mol Cell* **18**, 97-108 (2005).

100. Szlachta, K. et al. Alternative DNA secondary structure formation affects RNA polymerase II promoter-proximal pausing in human. *Genome Biol* **19**, 89 (2018).

101. Vvedenskaya, I.O. et al. Interactions between RNA polymerase and the "core recognition element" counteract pausing. *Science* **344**, 1285-1289 (2014).

102. Saba, J. et al. The elemental mechanism of transcriptional pausing. *Elife* **8** (2019).

103. Kang, J.Y. et al. RNA Polymerase Accommodates a Pause RNA Hairpin by Global Conformational Rearrangements that Prolong Pausing. *Mol Cell* **69**, 802-815 e805 (2018).

104. Richmond, T.J. & Davey, C.A. The structure of DNA in the nucleosome core. *Nature* **423**, 145-150 (2003).

105. Weber, C.M., Ramachandran, S. & Henikoff, S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol Cell* **53**, 819-830 (2014).

106. Santisteban, M.S., Hang, M. & Smith, M.M. Histone variant H2A.Z and RNA polymerase II transcription elongation. *Mol Cell Biol* **31**, 1848-1860 (2011).

107. Stasevich, T.J. et al. Regulation of RNA polymerase II activation by histone acetylation in single living cells. *Nature* **516**, 272-275 (2014).

108. Galvani, A. & Thiriet, C. Nucleosome dancing at the tempo of histone tail acetylation. *Genes* **6**, 607-621 (2015).

109. Li, J. & Gilmour, D.S. Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor. *EMBO J* **32**, 1829-1841 (2013).

110. Morgan, M.A.J. et al. A cryptic Tudor domain links BRWD2/PHIP to COMPASS-mediated histone H3K4 methylation. *Genes Dev* **31**, 2003-2014 (2017).

111. Bolger, A.M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).

112. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).

113. Li, B., Carey, M. & Workman, J.L. The role of chromatin during transcription. *Cell* **128**, 707-719 (2007).

114. Vihervaara, A. et al. Transcriptional response to stress is pre-wired by promoter and enhancer architecture. **8**, 255 (2017).

115. Gilchrist, D.A. & Adelman, K. Coupling polymerase pausing and chromatin landscapes for precise regulation of transcription. *Biochim Biophys Acta* **1819**, 700-706 (2012).

116. Studitsky, V.M., Walter, W., Kireeva, M., Kashlev, M. & Felsenfeld, G. Chromatin remodeling by RNA polymerases. *Trends in biochemical sciences* **29**, 127-135 (2004).

117. Nakayama, J., Rice, J.C., Strahl, B.D., Allis, C.D. & Grewal, S.I. Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly. *Science* **292**, 110-113 (2001).

118. Kolasinska-Zwierz, P. et al. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet* **41**, 376-381 (2009).

119. Venkatesh, S. et al. Set2 methylation of histone H3 lysine 36 suppresses histone exchange on transcribed genes. *Nature* **489**, 452-455 (2012).

120. Wan, Y. et al. Role of the repressor Oaf3p in the recruitment of transcription factors and chromatin dynamics during the oleate response. *Biochem J* **449**, 507-517 (2013).

121. Larson, M.H. et al. A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science* **344**, 1042-1047 (2014).

122. Neuman, K.C., Abbondanzieri, E.A., Landick, R., Gelles, J. & Block, S.M. Ubiquitous transcriptional pausing is independent of RNA polymerase backtracking. *Cell* **115**, 437-447 (2003).

123. Weixlbaumer, A., Leon, K., Landick, R. & Darst, S.A. Structural basis of transcriptional pausing in bacteria. *Cell* **152**, 431-441 (2013).

124. Hein, P.P. et al. RNA polymerase pausing and nascent-RNA structure formation are linked through clamp-domain movement. *Nat Struct Mol Biol* **21**, 794-802 (2014).

125. Zhang, J. & Landick, R. A Two-Way Street: Regulatory Interplay between RNA Polymerase and Nascent RNA Structure. *Trends Biochem Sci* **41**, 293-310 (2016).

126. Venkatesh, S. & Workman, J.L. Histone exchange, chromatin structure and the regulation of transcription. *Nature reviews Molecular cell biology* **16**, 178 (2015).

127. Zentner, G.E. & Henikoff, S. Regulation of nucleosome dynamics by histone modifications. *Nat Struct Mol Biol* **20**, 259-266 (2013).

128. Bernecky, C., Herzog, F., Baumeister, W., Plitzko, J.M. & Cramer, P. Structure of transcribing mammalian RNA polymerase II. *Nature* **529**, 551-554 (2016).

129. Zhou, V.W., Goren, A. & Bernstein, B.E. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* **12**, 7-18 (2011).

130. Ray, J. et al. Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. *bioRxiv*, 527838 (2019).

131. Mueller, B. et al. Widespread changes in nucleosome accessibility without changes in nucleosome occupancy during a rapid transcriptional induction. **31**, 451-462 (2017).

132. Zenklusen, D., Larson, D.R. & Singer, R.H. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat Struct Mol Biol* **15**, 1263-1271 (2008).

133. Ribeiro, A.S., Hakkinen, A., Healy, S. & Yli-Harja, O. Dynamical effects of transcriptional pause-prone sites. *Comput Biol Chem* **34**, 143-148 (2010).

134. Dobrzynski, M. & Bruggeman, F.J. Elongation dynamics shape bursty transcription and translation. *Proc Natl Acad Sci U S A* **106**, 2583-2588 (2009).

135. Klein, A.M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).

136. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).

137. 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 (2016).

138. Dar, R.D. et al. Transcriptional Bursting Explains the Noise-Versus-Mean Relationship in mRNA and Protein Levels. *PLoS One* **11**, e0158298 (2016).

139. Wissink, E.M., Vihervaara, A., Tippens, N.D. & Lis, J.T. Nascent RNA analyses: tracking transcription and its regulation. *Nat Rev Genet* **20**, 705-723 (2019).