From Medical Biochemistry and Biophysics
Karolinska Institutet, Stockholm, Sweden

TRANSCRIPTION KINETICS IN PLURIPOTENT CELLS: RNA TURNOVER, TRANSCRIPTION VELOCITY, AND EPIGENOMIC REGULATION

Rui Shao

Karolinska Institutet

Stockholm 2022

Published by Karolinska Institutet.
Printed by Universitetsservice US-AB, 2022

ISBN 978-91-8016-768-0
Cover illustration: © Rui Shao. Transcription crossing nucleosomes.

Transcription kinetics in pluripotent cells: RNA turnover, transcription velocity, and epigenomic regulation

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Rui Shao

The thesis will be defended in public at Air&Fire, Scilifelab, Stockholm, October 3rd 2022

Principal Supervisor:

Associate Prof. **Simon Elsässer**

Karolinska Institutet
Department of Medical Biochemistry and Biophysics
Division of Genome Biology

Co-supervisor(s):

Dr. **Michael Lidschreiber**

Max Planck Institute for Biophysical Chemistry
Department of Molecular Biology

Prof. **Patrick Cramer**

Max Planck Institute for Biophysical Chemistry
Department of Molecular Biology

Opponent:
Assistant Prof. **Anniina Vihervaara**
KTH Royal Institute of Technology
Department of Gene Technology

Examination Board:
Prof. **Eckardt Treuter**
Karolinska Institutet
Department of Biosciences and Nutrition

Associate Prof. **Qi Dai**
Stockholm University
Department of Molecular Biosciences

Associated Prof. **Peter Svensson**
Karolinska Institutet
Department of Biosciences and Nutrition

ABSTRACT

Transcriptional regulation is one of the primary steps in gene expression control. It is now appreciated that a large fraction of coding genome is transcribed in concert of other functional RNAs. A quantitative method for transient transcriptome sequencing (TT-seq) allows profiling of entire transcriptional activities, *de novo* transcription unit (TU) annotation, and estimation of transcription kinetics from initiation to termination.

In **Paper I**, we showed the establishment of TT-seq method in mouse embryonic stem cells (mESCs) to understand transcriptome plasticity for both coding and non-coding RNAs. With external references in form of a spike-in RNA mix, we were able to estimated RNA synthesis and turnover rates, which consolidated the attenuation under inhibitor-induced pluripotent states (naïve 2i and paused mTORi). We also extended the estimation of transcription velocity to each annotated TU, by integration of RNA polymerase II (Pol II) quantitative profiles from MINUTE-ChIP (quantitative multiplexed ChIP). After explaining transcription velocity with chromatin features, we also evaluated its genome-wide contribution to termination distance.

In **Paper II**, we mapped endogenous genomic G-quadraplex structures (G4) with CUT&Tag in HEK293T and mESCs. We verified the high signal-to-ratio G4 peaks to reflect the DNA motifs of both canonical and trans-strand putative quadraplex sequences (PQS), which enriched on both gene and active enhancer TSSs (transcription start sites). After stabilizing G4 with the small molecule PDS, we observed a genome-wide reduction of RNA synthesis (by TT-seq). The co-occupancy of G4 and R-loop was further verified at transcribed promoters and enhancers. However, promoter G4s could consistently form after transcription inhibition, which suggests an intricate cause-consequence relationship between G4 and transcription activity.

In **Paper III**, we evaluated the regulatory role of repressive histone modifications, H2AK119 ubiquitination and H3K27 tri-methylation. We introduced a rapid H2Aub depletion by BAP1 pulse expression with the amber-suppression system, and observed a wide Polycomb target genes de-repression, especially in the bivalent chromatin state (H3K4me3 + H3K27me3). Further, we observed that H2Aub-mediated repression strength was associated with H3K27me3 occupancy. However, double depletion of H3K27me3 by Ezh2 inhibition with ectopic BAP1 failed to enlarge Polycomb genes de-repression. We also measured transcriptional responses with TT-seq and observed that H2Aub depletion immediately triggered transcription activation before the redistribution of Polycomb proteins and their associated nucleosomes decompaction. Together, our results indicate that H2Aub directly mediates Polycomb integrity and nucleosome barrier that limits early transcription check-points.

LIST OF SCIENTIFIC PAPERS

I. **Shao Rui**, Kumar Banushree, Lidschreiber Katja, Lidschreiber Michael, Cramer Patrick, Elsässer J. Simon. Distinct transcription kinetics of pluripotent cell states. Mol Syst Biol. 2022 Jan;18(1):e10407. doi: 10.15252/msb.202110407. PMID: 35020268; PMCID: PMC8754154.

II. Lyu Jing, **Shao Rrui**, Kwong Yung Yuk Philip, Elsässer J. Simon. Genome-wide mapping of G-quadruplex structures with CUT&Tag. Nucleic Acids Res. 2022 Feb 22;50(3):e13. doi: 10.1093/nar/gkab1073. PMID: 34792172; PMCID: PMC8860588.

III. **Shao Rui***, Yung Yuk Philip*, Elsässer J. Simon. H2A de-ubiquitinylation reverses Polycomb-mediated transcription repression. 2022. (manuscript) *Co-first author.

CONTENTS

LIST OF ABBREVIATIONS

4sU	4-thiouridine
AcKRS	Acetyl-lysyl-tRNA synthetase
asRNA	cis-antisense RNA
bp	base pair
BrdU	5-Bomo-2-deoxyUridine
ButKRS	Butyryl--lysyl-tRNA synthetase
ChIP	Chromatin Immuno-Precipitation
CKO	Conditional Knock-Out
CoA	Coenzyme A
CTD	C-terminal domain
cPRC1	Canonical Polycomb Rrepressive Complex 1
DHS	DNase I hypersensitive site
DRB	5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole
ERCC	External RNA Controls Consortium
eRNA	Enhancer RNA
FACT	FAcilitates Chromatin Transcription
G4	G-quadraplex
GEO	Gene Expression Omnibus
GRO-seq	Global Run-On Sequencing
H2Aub	Histone H2A lysine 119 mono-ubiquitination
HAT	Histone Acetyltransferase
HDAC	Histone deacetylase
HDAC	Histone Deacetylase
HibKRS	β-Hydroxyisobutyry-lysyl-tRNA synthetase
HMM	Hidden Markov Model
IP	Immuno-Precipitation
KAT	Lysine Acetylation Transferase
kb	kilo base-pair
lincRNA	Long intergenic non-coding RNA
lncRNA	Long non-coding RNA
LLPS	Liqid-Liqid Phase Separation
m7G	7-methylGuanosine
mESC	mouse Embryonic Stem Cell
MNase	Micrococcal Nuclease
MINUTE-ChIP	Multiplexed INdexed Unique barcoded T7 paired-End ChIP
MS	Mass Spectrometry
ncRNA	Non-Coding RNA
NDR	Nucleosome Depleted Region
NET-seq	Native Elongation Transcript Sequencing
NGS	Next Generation Sequencing
nt	Nucleotide
pA	polyAdenylation
P-TEFb	Positive Transcription Elongation Factor
PcG	Polycomb Group
PIC	Pre-Initiation Complex
POI	Protein Of Interest
PRC	Polycomb Rrepressive Complex
PRC1	Polycomb Rrepressive Complex 1
PRC2	Polycomb Rrepressive Complex 2

PRO-seq	Precision nuclear Run-On Sequencing
PTM	Post-Translational Modification
PylRS	Pyrrolysyl-tRNA synthetase
RNP	Ribonucleoprotein
scRNA	single-cell RNA
SVM	Support Vector Machine
TAD	Topologically Associated Domain
TCA	Tri-Carboxylic Acid
TES	Transcription End Site
TEVp	Tobacco Etch Virus protease
TPM	Transcripts Per Million
Trp	Triptolide
TSA	TrichoStatin A
TSS	Transcription Start Site
TT-seq	Transient Transcriptome Sequencing
TTS	Transcription Termination Site
TU	Transcription Unit
UAA	Unnatural amino acid
uaRNA	Upstream antisense RNA
vPRC1	Variant Polycomb Rrepressive Complex 1

1 INTRODUCTION

Mouse embryonic stem cells (mESC), derive from inner cell mass of a blastocysts at the pre-implantation stage of early embryo development, are widely applied for gene regulation studies. While pluripotent stem cells exist only transiently in the embryo, mESC retain pluripotency indefinitely in *in vitro* cell culture. Hence, mESC are not only a model system for molecular mechanistic studies, but have provided a model for embryonic development and a platform for multidisciplinary techniques developments, from single-cell to population, from transcription to translation, from RNA to protein, from DNA double-helix to genome architecture, and from pluripotent states to developmental fates (Figure 1.1).

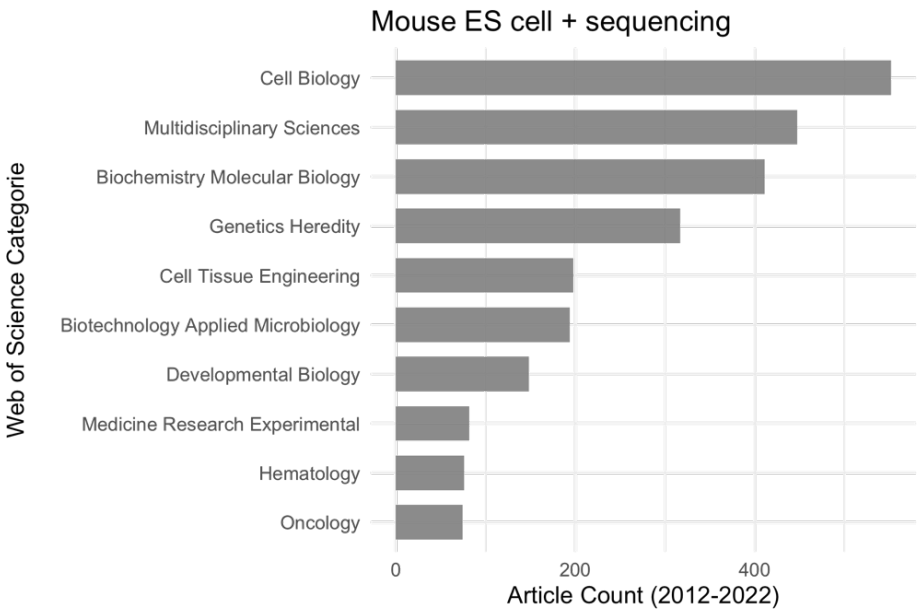

Figure 1.1 A bar-plot of article numbers by the keywords "mouse embryonic stem cell + sequencing" in the Web of Science from 2012 to 2022.

Owing to the low cost of the mESC model, its rich biology has often been explored with pioneer techniques that unveiled gene-regulatory mechanisms of development. The rapid growth in new sequencing methods provides a tremendous amount of data, in both breadth and depth. But for mechanistic implications and hypothesis testing, multi-omics datasets crucially require bioinformatic frameworks that address the underlying biological question in quantitative and statistical terms, for instance, dynamics, kinetics, stability, similarity, localization, covariance, correlation, causality, and feature importance.

In cell, transcription controls the information flow from DNA to RNA, as the first step of gene expression. In genome, gene positions like a raft, where transcription travels and unwraps numerous nucleosome packed DNA that organizes genome into higher orders. Inside the histone octamer, different variants and modifications overlay additional characteristics of nucleosome, which endorses gene selective expression and potential epigenomic memory. The widely studies histone modifications are closely related with gene activity, but their causal relation with transcription is just emerging from several case-studies of enzymatic knock-out [12] and transcription inhibition [3].

Therefore, this doctoral thesis will explore transcriptional regulation mechanism in mESC, including transcription kinetics estimation, DNA secondary structure mapping, and histone modification modulation.

2 LITERATURE REVIEW

2.1 EUKARYOTIC TRANSCRIPTION

2.1.1 Transcription stages of RNA Polymerase II

Transcription copies protein-coding information from DNA to mRNA with three major stages: initiation, elongation, and termination (Figure 2.1). Briefly, RNA Polymerase II (Pol II) contacts with promoter sequence and interacts with the general transcription factors (GTFs), TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH[4]. The pre-initiation complex (PIC) assembles upon DNA motifs, e.g. TATA box, with mediators stimulation[5]. The largest subunit of Pol II, Rbp1, contains a C-terminal domain (CTD) of 52 YSPTSPST repeats in vertebrate. A hallmark of initiation-elongation transition is the phosphorylation of CTD serine 5 by TFIIH subunit Cdk7[67]. RNA Pol II then travels a short distance then stalls at a small range of promoter-proximal pausing sites (~50 bp)[8,9]. At the same time, 7-methylguanosine (m7G) cap appears at the 5' end of nascent RNA to coordinate mRNA processing and nuclear export[10]. The positive elongation factor (P-TEFb) subunit Cdk9 kinase catalyzes Pol II CTD serine 2 and transitions Pol II pause-release to productive elongation. The other subunit of P-TEFb, cyclin T1, engages with Pol II CTD hyperphosphorylation and forms liqid-liqid phase separation (LLPS)[11]. Recently, RNA binding protein PSPC1 is also found to promote transcription condensates, therefore increases elongation efficiency[12]. Transcription termination occurs with Pol II complex disassembly and nascent RNA cleavage, mostly after the polyadenylation (pA) sequence. But the new 5' end emerged in termination is not capped and digested by 5'-3' exonuclease Xrn2[13,14].

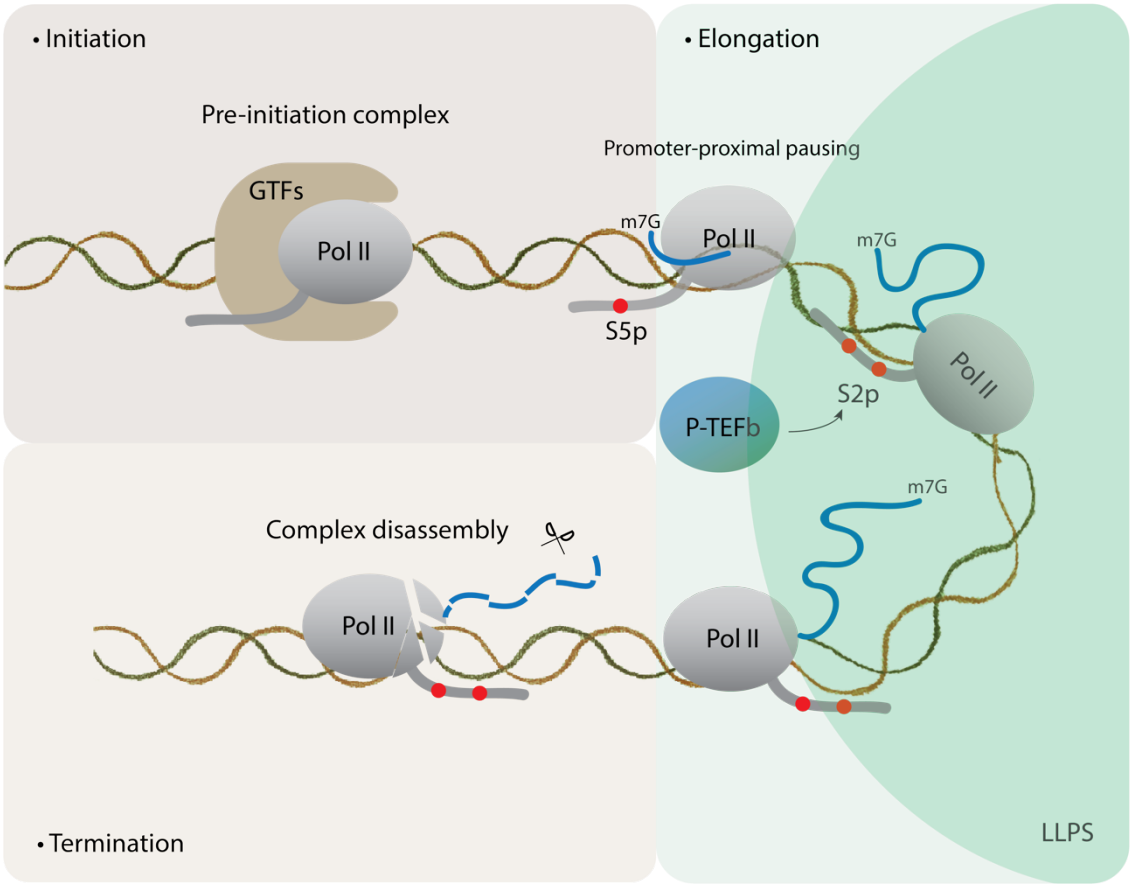

Figure 2.1 A diagram of RNA Pol II transcriptional stages. Pol II complex with CTD tail is colored in grey. Nascent RNA is in blue. Core of the transcription factory[15] is in a green sphere. The shattered Pol II complex in termination is with nascent RNA nucleolytic degradation.

2.1.2 Transcription profiling methods

After PIC forms at TSS, most RNA Pol II stalls shortly before release into the elongation phase[16]. Therefore, Pol II chromatin engagement can be independent of the nascent RNA production.

Many techniques have been developed to measure the transcription activity, primarily by cellular fractionation and the newly synthesized RNA purification (Table 2.1). These methods use a variety of detection principles and produce qualitatively and quantitatively different reads-outs. Global Run-On (GRO) is the first method that labels nascent RNA in nuclei extracts after restoration to 30°C from ice[17]. Productive Pol II incorporates 5-bromouridine triphosphate (Br-UTP) and enables nascent RNA purification with the antibody against 5-bromo-2-deoxyuridine (BrdU). After combining next-generation sequencing, GRO-seq can determine genome-wide transcription activity[18]. Precision nuclear Run-On sequencing (PRO-seq) achieves higher nucleotide resolution with four biotin-NTPs[19]. But the incorporation of a biotinylated nucleotide may stall the nascent RNA strand.
In contrast, living cells can directly take 4-thioutidine (4sU), a precursor of 4s-UTP, and label the newly synthesized RNA in continuous elongation. To clean the pre-existing RNA, 4sU pulse labeling methods uses biotinylation and streptavidin purification (4sU-seq[20–22]). Alternatively, 4sU saturated labeling methods rely on alkylation and quantification of T->C conversion (SLAM-seq[23,24] and TimeLapse-seq[25]). TT-seq (transient transcriptome sequencing) inherits from 4sU-seq, with an extra step of RNA fragmentation before biotinylation, can capture the newly synthesized RNA with high confidence[26]. Since, the RNA fragmentation and extensive wash steps guarantee the purity of 4sU labeled RNA from a large pre-existing RNA pool. Also the biotin-purification approach lowers the number of 4sU incorporation thereby a shorter labeling time, as 2 hours 4sU treatment can just competitively label <1% of the total RNA reads by T->C conversion in MCF-7 cells[27].

Table 2.1 Nascent RNA sequencing methods comparison.

Method	Material	Labeling	Purification	Variation	Year	Ref.
GRO-seq	· Nuclei	· Br-UTP · 5 min	· Anti-BrU beads	GRO-cap fastGRO	2008	[18]
PRO-seq	· Nuclei	· Biotin-NTP · 3-5 min	· Streptavidin beads	PRO-cap	2013	[19]
4sU-seq	· Cell	· 4-thiouridine · 10 min	· Biotin · Streptavidin beads	4tU-seq 4sUBRD-seq 4sU Chase-seq	2011	[28]
Bru-seq	· Cell	·Bromouridine · 10 min	· Anti-BrdU	BRIC-seq BruChase-seq	2014	[29]

TT-seq	· Cell	· 4-thiouridine · 5 min	· Biotin · Streptavidin beads	-	2016	[26]
SLAM-seq	· Cell	· 4-thiouridine · 45 min - 24 h	· T->C conversion	TimeLapse-seq scSLAM-seq NASC-seq	2017	[25,30]
NET-seq	·Cell lysis	-	· Tagged Rpb3 · Anti-RNA Pol II · CTD antibody	mNET-seq POINT-seq	2011	[31,32]

2.1.2.1 Spike-in normalization

Transcription kinetics estimation requires sample normalization to the absolute scales. External RNA spike-in references have been developed in many different ways (Table 2.2). For instance, a GRO-seq study added luciferase RNAs in the BrdU purification step to control sample size and background noise[33]. While the whole-genome RNA as the spike-in could decrease library complexity. A study adds 4sU labeled *Drosophila* and unlabeled *S. cerevisiae* total RNA to ~30% of the library[34], which means 70% of reads will remain after sample scaling.

In contrast, TT-seq mixes three 4sU labeled and three unlabeled ERCC (External RNA Controls Consortium) spike-ins to the TRIzol cell lysis to allow a coherent control throughout the procedures[26]. And these short spike-in sequences take ~2% of the labeling library and ~0.2% of the unlabeled library (in this study with serum-naïve mESCs). In addition, the *in vitro* synthesized ERCC transcripts can pre-mix in the weight and the labeled ratios, as the external standards for kinetics estimation[35].

Table 2.2 Examples of spike-in reference of the metabolic labeling RNA-seq methods.

Method	Spike-in	Labeled	Unlabeled	Reference
TT-seq	ERCC RNA	3 transcripts	3 transcripts	[26]
TT-seq (this study)	ERCC RNA	4 transcripts	4 transcripts	[35]
PRO-seq	Total RNA	NA	*Drosophila*	[36 37]
GRO-seq	Rluc RNA	1 transcript	NA	[33]
fastGRO	Total RNA, nuclei	4sU 5min *Drosophila* RNA	*Drosophila*	[38]
SLAM-seq	Total RNA	NA	*Arabidopsis*	[30]
s^4U Chase-seq	Total RNA	NA	*S. pombe*	[39]

2.1.3 Transcription unit

Most transcription events occur in the known gene regions, besides intergenic intervals produce abundant types of pervasive non-coding transcripts[40].

2.1.3.1 TU annotation

To understand the regulatory role of intergenic sequence, the genomic "dark matter," transcribed genomic regions are identified by *de novo* transcripts annotation, which requires high quality and genome-wide mapping of transcription activities. In 2001, FANTOM1 (Functional ANnoTation Of the Mammalian genome) found that the non-coding RNAs

(ncRNAs) out-numbered mRNAs in the complementary DNA (cDNA) libraries[41]. With next-generation sequencing (NGS), the ENCODE project found two critical messages in the massive RNA-seq from various cell types. Most unannotated ncRNAs are cell type-specific, and nearly half of ncRNAs emerged from gene neighbors[42].

Today, biochemical enrichment methods of newly transcribed RNA help to identify different classes of novel transcripts. The often unstable ncRNAs can be found located in the nuclear and chromosomal compartment due to their high sensitivity to ribonucleolytic RNA exosome digestion[43,44]. However, cell fractionation alone is insufficient for quantifying transcription frequency since chromosomal RNAs represent nascent transcripts and RNAs interacting with DNA in trans[45]. Often, metabolic labeling-purification methods can provide direct benchmarks of transcription activity, as previously described with TT-seq[26] and GRO-seq[46].

After TU annotation, the downstream analysis is specific to the studies. GRO/PRO-seq annotation has multiple downstream processing approaches, for instance, transcripts calling from local peak-to-gene fold cutoff (HOMER)[47], regulatory transcript discovery with hidden Markov model (HMM)[48], regulatory intergenic TSSs annotation with support vector machine (SVM)[46], and a follow-up work with support vector regression (dREG method) to monitor initiation regions[49]. TT-seq signal is mainly processed for TU annotation with HMM, enhancer annotation from epigenomic marks[50], estimations of transcription frequency, elongation velocity[51], RNA turnover, and termination[35,52].

2.1.3.2 Active enhancer localization

Enhancer-gene shows transcriptional co-regulation[35], besides the non-transcriptional but connected activities, for example, non-coding promoter elements and the ncRNA splicing, which also affect the nearby gene expression[53]. Enhancers boost gene expression by chromatin looping to gene promoters. The physical contact to adjacent gene promoters in a negative logarithm relation by distance[54], so enhancer influence follows a power law decrease[55]. And additive activation allows multiple enhancers to amplify the target gene output in adjacent neighborhood[50,56].

When enhancer acts as spatial-temporal gene expression regulators, it exhibits multifaceted characteristics. Besides the chromatin characteristics H3K27ac and H3K4me1, transcription is also a proxy of enhancer activity[50,57]. However, a larger repertoire of enhancers has potential but lacks eRNA production in the native context[58]. So a broader definition can also include poised and primed enhancers in addition to the active enhancers with H3 lysine-27 acetylation (H3K27ac) (Table 2.3). For instance, the poised enhancers' emergence is developmental stage dependent[59], with a hallmark of repressive H3 lysine-27 tri-methylation (H3K27me3) and Polycomb-mediated chromatin interaction[60].

Table 2.3 Three enhancer types defined by histone modifications in mouse ES cell[61]. The respective marks are set with 1 kb threshold.

Enhancer	p300	H3K27me3	H3K27ac	H3K4me1
Poised	< 1 kb	< 1 kb	-	-
Active	< 1 kb	-	< 1 kb	-
Primed	-	-	-	< 1 kb

2.1.4 Transcription kinetics

2.1.4.1 Transcription frequency

RNA synthesis rate equals transcription frequency, in a combination of the distinct RNA Pol II dynamics in initiation, elongation, and termination stages. Regarding the rate-limiting pause-release step, transcription frequency has a synonym of "initiation frequency"[22]. However, elongation rate and transcription rate are in the velocity taxonomy[62]. To further clarify, single-cell RNA (scRNA) transcriptomics has the ability to describe stochasticity of transcription events with burst size and burst frequency parameters[63,64]. Therefore, scRNA-seq transcription burst frequency does not represent the populational average as transcription frequency from bulk nascent RNA-seq. Because the labeled RNA enrichment method directly measures RNA synthesis. Differently, transcription burst frequency is a conjugated term with the burst size estimated from total scRNA distribution, which can be subject to the sequencing method.

2.1.4.2 Pausing duration

After the transcription PIC formation (Figure 2.1), RNA Pol II shortly moves till it encounters the pausing regulators (e.g., TFIID, NELF, DSIF, and Integrator), the well-phased +1 nucleosome barrier, and specific DNA motifs[9,65–67]. In one transcription cycle, promoter-proximal pausing is the second longest event, as a case study reports that pausing occupies 23% of Polymerase and a median of 42 seconds[16]. Under a microscope, the paused Pol II form into clusters. And the pause-release inhibitor DRB (5,6-dichloro-1-β-D-ribofuranosylbenzimidazole) can enlarge Pol II aggregates[68]. However, this phenomenon is unclear whether DRB accelerates premature Pol II recycling or induces additional Pol II pause. Transcription frequency might not increase after a prolonged Pol II pausing but will decrease if blocks release. In GRO-seq, Cdk9 inhibition (Flavopiridol) results in lower transcriptional outputs of almost all protein-coding genes and several enhancers (by reanalyzing)[69]. Additionally, the pausing index is a conventional and straightforward ratio of Pol II density at TSS pausing interval and gene body. It shows a weak anti-correlation with transcription frequency[70], and suggests Pol II pause-release to be a rate-limiting step in transcription[71–73].

2.1.4.3 Elongation velocity

In mammalian cells, gene-level transcription velocity varies in a wide range, from 1 to 4 kb/min revealed by inhibition-release[74]. Many genomic features correlate with the elongation rate, such as histone modifications, DNA/RNA motifs, and the density of exons[74]. Interestingly, the correct pause-release checkpoint also impacts elongation velocity in the gene body, as Cdk9 inhibition leaked Pol II travels significantly slower[75]. Typically, the local velocity after the pause-release process immediately increases towards the gene body, known as "getting up to speed"[76]. And velocity slows down after pA signal to facilitate transcription termination.

2.1.4.4 Termination distance

At the last stage of transcription, slow Pol II is vulnerable to stop, which provides a window opportunity for exonuclease-mediated RNA cleavage and Pol II co-factors disassociation. The first event supports the "torpedo model," and the second event is known as the "allosteric model." Compared to the malicious translational read-through, prolonged transcription termination is benevolent with flexible distances. An ultimate termination site

lacks a sharp peak as Pol II at TSS but can be estimated from the gradual decrease of newly synthesized RNA coverage. In human K562 cells, the ultimate termination distance is a median of 3.3 kb[26].

2.1.4.5 RNA turnover and half-life

Compared to proteins ~9 hours half-lives[77], RNA lifespan is shorter, especially for the non-coding types. RNA turnover from metabolic pulse labeling represents a current tendency to replace the pre-existing RNA. While RNA half-life from pulse-chase labeling measures the first-order degradation kinetics without considering RNA replacement. So RNA degradation rates exclude the dilution effect of cell growth, which confounds in RNA turnover rates. Hence, in terms of half-life, turnover is faster than decay, as TT-seq and SLAM-seq/TimeLapse-seq reveal (Figure 2.2)[23,26].

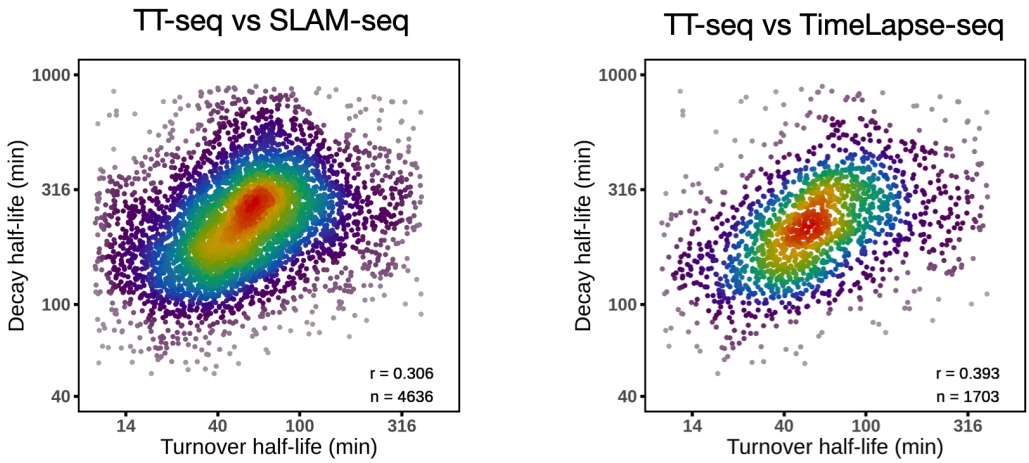

Figure 2.2 Scatter plots of the RNA turnover and the decay half-life. Published results in **Paper I** (Author response Figure 1D-E). TT-seq turnover half-life is measured in naïve state mESC, compared with decay half-lives in SLAM-seq[23] and TimeLapse-seq[25]. Of note, TT-seq measures turnover half-lives of 10537 genes. The plots above show the intersected gene sets. Pearson's correlation is calculated after log transformation.

2.2 NUCLEOSOMAL REGULATION OF GENE TRANSCRIPTION

2.2.1 Active histone code

It has been almost 60 years since the first post-translational modification (PTM) histone lysine acetylation was characterized[78]. Histone modifications exist in different chromatin states and compositions of "histone code." Lysine acetylation is of particular interest in transcription activation, since it neutralizes the positive charge on histone lysine ε-amino group, leads to a decompaction of the nucleosome fiber, and recruits co-activators that initiate the transcription machinery. The acetyl-Lys 'readers' (such as Brd4), the 'writers'---histone acetyltransferases (HATs), and the 'erasers'---histone deacetylases (HDACs) together fine-tune chromosomal regulation.

Nevertheless, histone acetylation is not a single cause or consequence of gene expression. In *S. cerevisiae*, transcription activates HATs and changes histone acetylation patterns[79]. A study with HDAC (histone deacetylases) inhibition (TSA, Trichostatin A) increased histone

acetylation and instantaneously activated a few genes by releasing RNA Pol II promoter-proximal pausing, but further transcriptional inhibition failed to reverse the elevation of the histone acetylation[80]. Histone acetylation generally bookmarks the functional genomic areas related to enhancers[59] and genes for rapid activation[81]. However, a recent study suggests that depletion of H3.3K27 acetylation cannot rewire the transcriptional program[82], although the antagonistic H3K27 methylation limits enhancer activation[83].

2.2.1.1 Histone acylation manipulation with genetic code expansion

The endogenous TCA (tricarboxylic acid) cycle generates acetyl-coenzyme A (CoA) as the source of histone acetyl groups, for lysine acetylation transferases (KATs) installation. Nonetheless, acetylation is under competition with other acyl-CoAs, especially when metabolic shifts the acyl-CoAs availability. Acylation can occur on almost every amine residue of histone lysine with a mixture of possibilities (Figure 2.3 A-B). So histone acylation could record and correspond to a transcriptional adaptation of specific physiological changes. However, the regulatory mechanism of histone acylation is far from clear.

To answer this question and compare each histone acylation, the site-specific installation of acylation marks by genetic code expansion provides a possible route to this challenge. Pyrrolysyl-tRNA synthetase (PylRS) and $tRNA^{Pyl}$ pair bio-orthogonally incorporate unnatural amino acid (UAA) to the $tRNA^{Pyl}$ matched mRNA codon during protein translation (Figure 2.3 C)[84]. This method has accomplished pulse expressions of the pre-modified H3K27ac, H3K56ac, and H3K64ac, by switching lysine codons into the amber stop codon. Of note, aminoacyl-tRNA synthetase has variants of AcKRS, CrKRS, HibKRS, and ButyKRS, corresponding to the substrates of acetylated, crotonylated, hydroxybutyrylated, and butyrylated lysines. Therefore, the genetic code expansion approach allows the manipulation of a large panel of histone acylation (Figure 2.3 D).

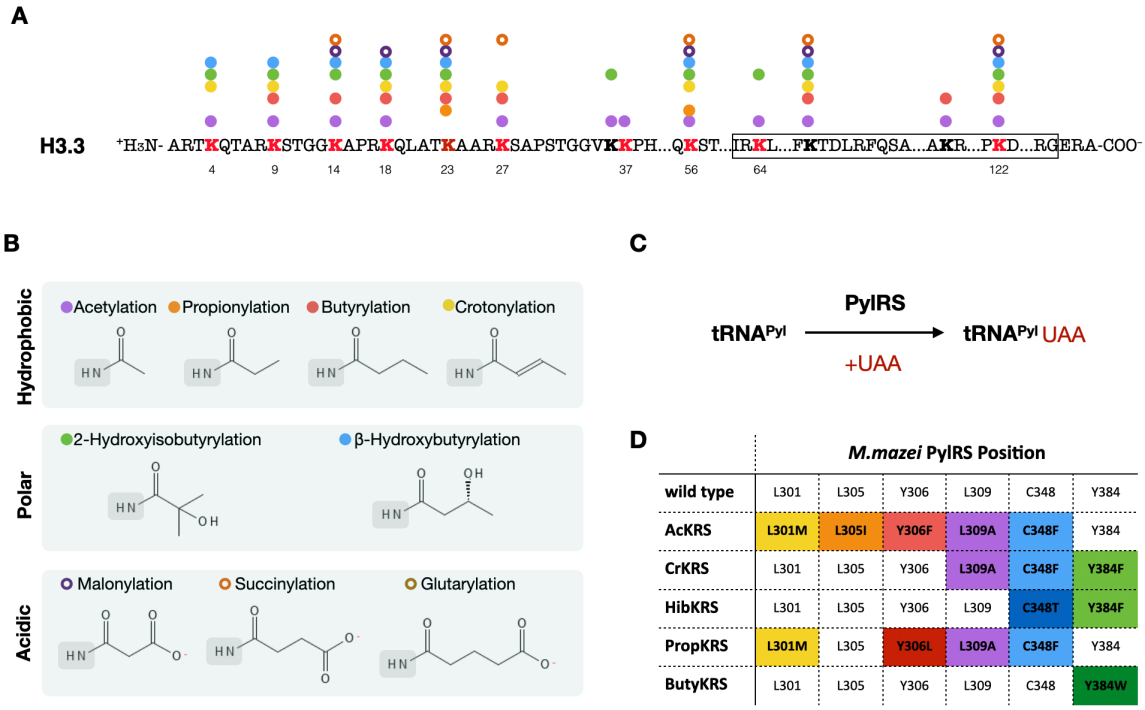

Figure 2.3 Scheme of histone acylation and genetic code expansion. (A) An example of the known acylation sites on histone H3.3. (B) Classes of acyl groups that have been identified on histone

H3.3[85]. (C) The process of Pyrrolysyl-tRNA reacts with unnatural amino acids. (D) PylRS mutants with the known acyl groups selectivity[86].

2.2.2 Repressive histone code

The repressive chromatin states control gene expression equally critical as the active states. Histone H3K27me3 and H2A119ub1 (H2Aub) establish the repressive chromatin vicinity that poises developmental gene expression[87]; whereas H3K9me3 is found in the constitutive inactive regions associated with heterochromatin protein 1 (HP1)[88], and in a small number of cases with H3.3[89]. From the transcription aspect, H3K27me3 is essential for the long-term transcriptional memory in mESC differentiation[90]; in contrast, H2Aub has a temporal role in transcription repression, loss of which deprives mESC self-renewal[91] and viability[92].

2.2.2.1 H2A119ub1 and H3K27me3 interplay

H2A119ub1 and H3K27me3 are catalytic products of the Polycomb repressive complex 1 (PRC1) and 2 (PRC2). PRC has many compositions and plays different roles in gene repression. Briefly, variant PRC1 (vPRC1) binds to unmethylated CpG islands on gene promoters via its sub-unit KDM2B[93], and *de novo* establishes H2Aub. The Rybp in vPRC1 and Jarid2 in PRC2.2 recognize H2Aub and propagate H2Aub[94–97] and H3K27me3[98–101].

H2Aub and H3K27me3 interplay are mostly uni-directional. By enzymatic modulations, H2Aub can promote H3K27me3[102], H2Aub passive depletion reduces H3K27me3[103–105], and H2Aub passive accumulation increases H3K27me3[1,106,107]. So their colocalization exists in a circuit where H2Aub is *de novo* catalyzed by vPRC1 and assists PRC2's tethering via Jarid2[93,108], then condensates Polycomb domains via H3K27me3-cPRC1-PHC1/2/3 interaction[109]. On the other hand, H2Aub is reported to be stable after the H3K27me3 depletion[96], or minorly increases on PcG binding genes in the Ezh1 KO - Ezh2 Y726D or Eed -/-, PRC2 null conditions[90,94]. This irreversible relation explains that H2Aub is prior to repressive chromatin formation before H3K27me3 emergence on developmental genes[110,111].

2.2.2.2 Transcriptional control by H2A119ub and H3K27me3

The repressive histone marks participate in many transcription regulatory stages, from initiation to termination[109]. Recently, with scRNA-seq and PRC1/PRC2 knock-out, Polycomb and H2Aub are found to control transcription burst frequency[104,112]. Non-catalytic PRC1 depletes H2Aub and widely de-represses Polycomb target genes[92,105]. However, H3K27me3 has not been reported to exhibit an equivalent role in mESC, since its depletion moderately rewires gene expression[90] and promoter-enhancer interactions[60]. Furthermore, the transcriptional regulation of H2Aub also ascribes to its nucleosomal DNA compaction ability[113,114], which is in line with the co-occurrence of H2A.Z at +1 nucleosome and Polycomb factors binding on the developmentally important genes[115–118].

2.2.2.3 Asymmetric effect of H2A119ub1 increase and decrease

H2Aub is described as a "rheostat" that fine-tunes PcG target gene expression by restricting chromatin potential[109]. However, this analogy may not be accurate since the increase and decrease of H2Aub fail to show reversible effects. The first evidence is observed in the marginal overlap between the de-ubiquitinase BAP1 conditional knock-out (CKO) down-regulated genes and PRC1 CKO up-regulated genes[106]. In agreement with the general rule

of H2Aub, BAP1 CKO represses global transcription, increases H3K27me3, and compacts chromatin. However, H2Aub excessive accumulation also paradoxically activates PcG target genes[119,120]. This effect is suggested to the pervasive H2Aub that relocates Polycomb factors and dilutes their bindings on designated PcG targets[119]. Another possibility could be that the new PRC-associated domains form into topological clusters and activate genes in the new proximity[121]. Reversely, H2Aub depletion de-represses genes with Polycomb binding, but not for every PcG target gene. Intriguingly, H2Aub accumulation and depletion can activate the same set of PcG target genes[120]. This phenomenon suggests this subset of PcG target genes to be responsive to H2Aub's either changes, unlikely in a rheostatic tuning.

In sum, the repressive histone modifications, especially H2Aub, regulate the transcription for mESC pluripotency maintenance and developmental program.

2.3 MOUSE EMBRYONIC PLURIPOTENT STATES

To mimic *in vivo* pre-implantation pluripotency, mESC has been subject to various empirical conditions of feeder cells, growth factors, and inhibitors in culture media, since it was identified four decades ago[122]. The standard mES cell culture medium contains leukemia inhibitory factor (LIF) and fetal bovine serum (FBS) to counteract cell differentiation tendency. In the serum+LIF (SL) medium, mESCs exhibit the "**naïve**" pluripotency. Mitogen-activated protein kinase (ERK1/2) blockage stops the intrinsic auto inductive differentiation and keeps mESCs in the naïve "**ground**" pluripotency[123]. In contrast, the epiblast-derived stem cells (EpiSC) from post-implantation epiblasts present the "**primed**" pluripotency. Moreover, the inhibition of nutrient and energetic sensing mTOR pathway pauses mESC proliferation that resembles diapaused blastocysts *in vivo*, so it induces mESC into a "**paused**" state[124].

2.3.1 Epigenomic characteristics

The usage of mitogen-activated protein kinase kinase (MEK ½) and glycogen-synthase kinase 3 (GSK3β) inhibitors in the serum-/LIF+ medium (2i-LIF) establishes distinct epigenomic profiles compared to SL cells.

Interestingly, FGF signal blockage in 2i cells induces global hypomethylation[125], similar to the demethylation in pre-implantation blastocysts and primordial germ cells (PGCs). SL cells sustain DNA hypermethylation by expressing the DNA methyltransferases, Dnmt3a and Dnmt3b; but 2i cells suppress Dnmt3a/b expression[126], render the global methylation decrease.

In the histone PTM aspect, an early study found a decline of H3K27me3 in the 2i-ground state, although with an incomplete ChIP-seq normalization[127]. However, 2i cells decrease gene expression in both bulk and single-cell RNA-seq[124,128]. Soon, mass spectrometry (MS) confirmed 2i cells to have a global increase of H3K27me3[129]. H3K27me3 level also can be feedback to reconcile DNA demethylation in 2i cells by Eed knock-out[129]. So transcription attenuation is not the only reason for the H3K27me3 increase in 2i cells, but as a consequence of cell signaling inhibition together with DNA demethylation[130]. Furthermore, the gene promoter H3K4me3+H3K27me3 bivalency has decreased due to H3K4me3 decline and H3K27me3 neighbor-diffusion in 2i cells[130]. But their chance of co-occurrence still correlates well between SL and 2i cells in co-ChIP[131].

The significant epigenomic alternation in mTORi paused cells is mainly associated with transcription attenuation and translation deprivation[132]. Accordingly, mTORi decrease active histone marks H3K4me3 and acetylation on H3K9 and H3K27 in mESC[132].

2.3.2 Transcriptomic characteristics

Due to cellular RNA abundance being a balance of RNA synthesis and degradation, it is essential to trace the source of gene differential expression during the pluripotent states transitions. In the previous image-based 5-Ethynyl Uridine (EU) labeling, 2i and mTORi cells have shown a global reduction of transcription[124]. Protein translation inhibition in mTORi cells is suggested to induce the pronounced transcriptional reduction[132]. Accordingly, total RNA decreases in both of the inhibitory conditions.

2i cultured mESCs have uniform morphology compared to SL cells (Figure 2.4) since serum is previously hypothesized to increase mESC gene expression heterogeneity[127]. Intriguingly, the single-cell RNA-seq studies reveal that 2i cells resemble closer blastocyst cells *in vivo*[128], and the cell populational distribution in serum is slightly more heterogenous[133]. The deprivation of serum-mediated stimulation in 2i culture medium may explain the reduction of stochastic transcription burst[134], and slowdown elongation velocity[135], in addition to the blockage of differentiation tendency by cell signaling inhibitors.

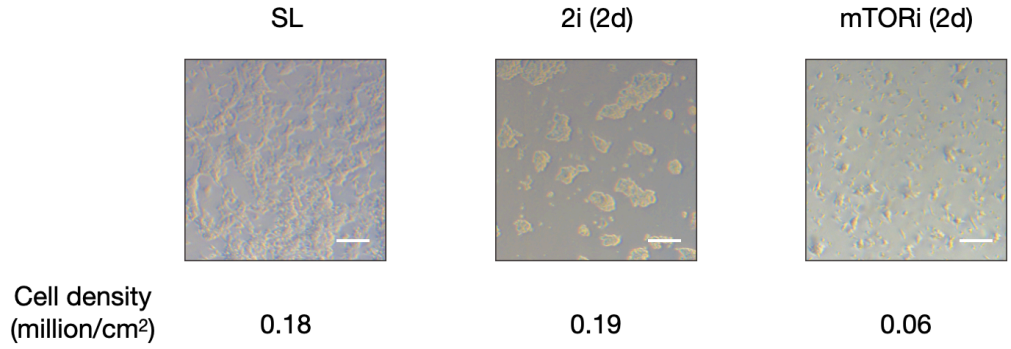

Figure 2.4 The morphology of mESC under the three pluripotent states in this study. RW4 (male, 129X1/SvJ) cells exhibit different colony sizes and shapes in response to the culture media shifts. Image scale bar is 100 μm.

3 RESEARCH AIMS

The aim of this thesis is to assess the epigenomic regulation of transcription in mouse embryonic stem cells. Estimate and evaluate the transcription kinetics in different pluripotent states.

Paper I: Measurement of newly synthesized RNA and evaluation of transcription kinetics in three mESC pluripotent states.

Paper II: Development of quantitative genome-wide G4 mapping method to assess endogenous G4 regulatory function.

Paper III: Examination of histone H2A K119 mono-ubiquitination mediated Polycomb repression mechanisms.

4 MATERIALS AND METHODS

4.1 MOLECULAR CLONING AND CELL CULTURE

4.1.1 Mouse embryonic stem cell

Mouse embryonic stem cell RW4 (male, 129X1/SvJ) were cultured in 0.1% gelatin-coated dish with Knock-out DMEM medium, 15% FBS (Sigma, F7524), 0.1mM ESGRO LIF (Sigma, ESG1107), 2 mM GlutaMAX (ThermoFisher, 10565018), 0.1 mM Non-Essential Amino Acid (Sigma, M7145), 0.1 mM β-mercaptoethanol (Sigma, M3148). 2i medium contains ESGRO Complete Basal Medium (Millipore, SF002), 3 μM GSK3β inhibitor CHIR99021 (Sigma, SML1046), 1 μM Mek ½ inhibitor PD0325901 (Sigma, PZ0162), 0.1 mM LIF. Inhibition of mTOR was in serum-LIF (SL) medium supplemented with 200nM INK128 (CAYM11811-1).

4.1.2 Genetic code expansion

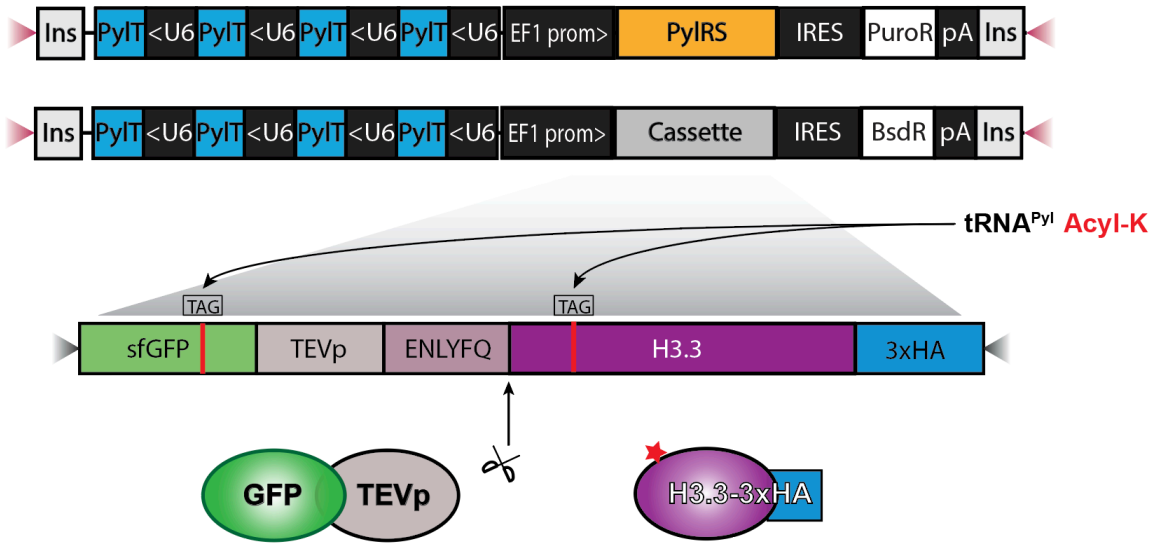

Figure 4.1 The design of the amber suppression system for site-specific histone H3.3 acylation in mESCs.

Genetic code expansion was utilized M.mazei PylRS-tRNA^{Pyl} pair, with the protein of interest (POI) on a separate vector to allow the resistance selection for piggyback-mediated genome insertion to achieve stable *in vivo* expression. To enlarge the availability of tRNA^{Pyl}, four copies of the u6-promoter-driven tRNA^{Pyl} were designed in the upstream position of PylRS/POI. Inside the POI cassette (Figure 4.1), a guide-and-clip gadget was cloned for reporting and limiting the downstream histone expression. The fusion of Tobacco Etch Virus protease (TEVp) and its recognition peptide as a guide protein, GFP-TEVp readily self-clips in translation and releases the full-length histone H3.3 with a triple HA tag (Figure 4.1).

For studying histone acylation, four PylRS variants (CroKRS, HibKRS, PropKRS, and ButyKRS) were sub-cloned from wild-type PylRS and AcKRS templates. And the amber codon substitutes of histone H3.3 at K4, K9, K14, K18, K23, K27, K37, K56, K64, K122, and histone H4 at K5+K12, K8, K16, K20 were cloned for the acyl groups installation. 2mg/mL of acetyl-L-lysine (Sigma-Aldrich, A4021), butyral lysine (Okeanos), crotonyl

lysine (Fluoro Chem), propynyl lysine (Okeanos), succinyl lysine (Okeanos) were supplemented to the culture median to initiate POI expression.

4.1.3 Amber suppression expression verification

For the verification of POI expression, HEK 293t cells (per 10 cm dish) were transfected with 10 μg histone plasmid and 1 μg PylRS plasmid by Lipofectamine 2000 (ThermoFisher). The stable mES cell lines were constructed with the same transfection condition and selected with 10mg/mL Blasticidin (Invivogen) and Puromycin (VWR).

Western blot assay was performed after the subcellular fractionation in the following steps:

1. Once PBS wash of cell pellet after trypsin collection (10 cm dish).
2. Buffer A 100 μL (10 mM Tris pH 8.0, 0.32 M sucrose, 3 mM $CaCl_2$, 2 mM magnesium acetate, 1 mM dithiothreitol (DTT), 0.1 mM ethylenediaminetetraacetic acid (EDTA), 0.5% NP-40 and freshly added protease inhibitors (Roche)), pipette 30 times. Vortex for 30 s, centrifuge for 1 min at 1400 g, and collect the supernatant as cytoplasmic fraction.
3. NE (nuclear extraction) buffer 100 μL (20 mM Hepes, pH 7.9, 25% v/v glycerol, 420 mM KCl, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 1 mM DTT and freshly added protease inhibitors), pipette 20 times. Vortex for 30 s, 1400 g centrifuge for 1 min, collect the supernatant as the nuclear fraction.
4. Twice NE buffer wash: 10 s vortex, 10 s spin down. When the precipitate becomes a transparent gel (chromatin fraction), cook with 1x SDS loading buffer at 95ºC for 10 minutes.

Anti-HA magnetic beads (ThermoFisher, 88837) were applied to nuclear extraction lysis (step 2 pellet) after genomic DNA digestion (1U DnaseI at RT for 5 min) and Protein G Dynabeads (ThermoFisher) pre-clear. A Pull-down reaction was carried out for 4 h at 4ºC, then beads were triple washed with the wash buffer (20 mM Hepes, pH 7.9, 20% v/v glycerol, 0.2 mM EDTA, 0.2% Triton X-100, 150 mM KCl and freshly added protease inhibitors). Histone H3.3 (Millipore, 09-838), Histone H3 (Abcam, ab1791), and Histone H3K27Ac (Abcam, ab4729) antibodies were used for the western blot.

4.2 SEQUENCING PREPARATION

4.2.1 TT-seq

TT-seq labeling followed the original protocol as described before[26] with minor modifications (Figure 4.2). Briefly, mouse ES cells were cultured for 2 days in four 15 cm dishes; sparing one dish for cell number counting, and the rest were supplemented with 500 μM 4-thiouridine (4sU) (Sigma-Aldrich, T4509) for 5 min at 37°C and 5% CO2, then immediately quenched with TRIzol (ThermoFisher, 15596018). Total RNA was extracted after mixing with spike-in RNAs (0.4 ng / million cells). Extracted total RNA was fragmented to an average of 1000 nt with Bioruptor (Diagenode, 3 cycle: 30 sec ON / 30 sec OFF at HIGH power), then incubated with HPDP-Biotin (ThermoFisher, 21341) dissolved in dimethylformamide (VWR, 1.02937.0500). A small aliquot was saved as fragmented total RNA (FRNA), and the rest was subjected to μMACS streptavidin beads (Miltenyi Biotec, 130-074-101) purification. After HPDP linker uncoupling with 100 mM DTT, the eluted labeled RNA (LRNA) was subjected to DNase digestion (Qiagen, 79254). The fragmented total RNA and labeled RNA were prepared with Ovation Universal RNA-

Seq kit (NuGEN, 0348). The pooled DNA library was cleaned and size-selected by Ampure XP beads (Beckman Coulter, A63881) before sequencing on Illumina NextSeq® 500/550 platform with High Output Kit v2 (Illumina, FC-404-2005, 75 cycles).

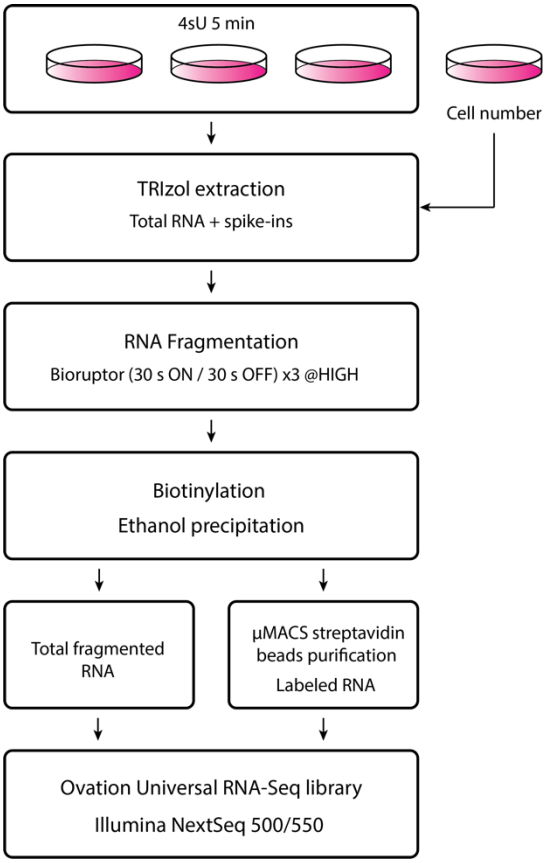

Figure 4.2 TT-seq experimental workflow of this study.

4.2.2 MINUTE-ChIP

The culture of $1x10^6$ mESCs in each condition was collected and processed with the MINUTE-ChIP protocol[130]. In brief, the native cells were subjected to micrococcal nuclease (MNase, New England BioLabs, M0247S) to fragment genomic DNA into mono- to tri- nucleosomes in Lysis buffer (100 mM Tris-HCL [pH 8.0], 0.2% Triton X-100, 0.1% sodium deoxycholate, 10 mM CaCl2 and 1x PIC). For each condition, dsDNA adaptors (containing T7 promoter, 8 bp sample barcode, and a 6 bp unique molecular identifier (UMI)) were ligated to the DNA fragments in the same pool with blunting and ligation reagents. The barcoded samples were then mixed thoroughly and aliquoted for the ChIP procedure (sparing 5% as Input) 4 h at 4°C with the target antibody pre-coupled Protein G magnetic beads (BioRad, 161-4023). Next, ChIPed and input DNA were collected for the libraries' construction through sequential steps of in vitro transcription, RNA 3' adapter ligation, reverse transcription, and PCR amplification. After Ampure XP beads (Beckman Coulter, A63881) clean-up and quantification with BioAnalyzer, the DNA libraries were pooled at 4 nM concentration and sequenced on the Illumina NextSeq500 platform. So, the resulting Illumina reads demultiplexed by each sample barcode, enabling quantitative comparison amongst all samples in the pool.

4.3 DATA ANALYSIS

4.3.1 Read alignment

4.3.1.1 RNA-seq and TT-seq

Paired-end short reads were aligned to mouse mm9 and mm10 genome references (GENCODE) by STAR 2.7.3a with settings: --outFilterMismatchNoverReadLmax 0.02 –outFilterMultimapScoreRange 0 –alignEndsType EndToEnd.

Gene level read counts were obtained from Kallisto (0.46.2) estimated counts with GENCODE vM21 transcriptome and aggregation by gene. For the differential expression analysis of non-coding TUs, the STAR-aligned bam files were subject to featureCounts (Rsubread 1.34.7) and analyzed by DESeq2 (1.24.0). Gene coverage extraction from bam files for elongation velocity analysis was performed with Bioconductor package "Rsamtools" and "GenomicAlignments."

4.3.1.2 ChIP-seq

MINUTE ChIP reads were processed with the "minute" pipeline available on GitHub (https://github.com/NBISweden/minute). In short, each sample was de-multiplexed from the pool of libraries with the indicated sample barcodes and carried to reads alignment, global input normalization, and bigwig coverage generation. The output genomic coverage in the bigwig format is ready for downstream analysis. An optional fine-tune step can fine-tune small global ChIP fold changes ($|\log_2 FC| < 1$), as described in the section 4.3.10.

4.3.2 Transcription unit annotation

Mapped reads were subjected to a three-step TU annotation as described before[26]. "TU filter" R shiny app (https://github.com/shaorray/TU_filter) was developed to allow multi-samples processing in a reproducible manner (Figure 4.3).

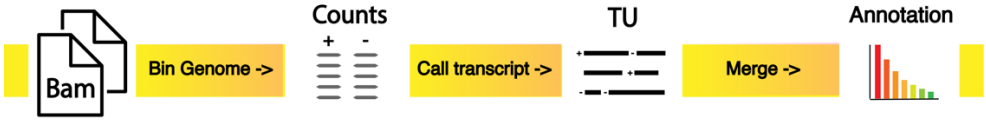

Figure 4.3 The flow chart of the "TU filter" app processing nascent RNA-seq reads with three main steps: bins and combines replicates, calls transcribed regions with HMM, and annotates TU intervals with the gene reference.

Briefly, the paired-end reads mid-points were binned into 200 bp genome coverage matrices by both strands (multiple replicates will be summed), then subjected to the HMM binary state calling by R package "GenoSTAN" with "PoissonLog" method. Next, the active states were treated as the raw TUs and joined by exons for each gene. Non-coding TU locations were named by their relative position to the nearby genes (Figure 4.4).

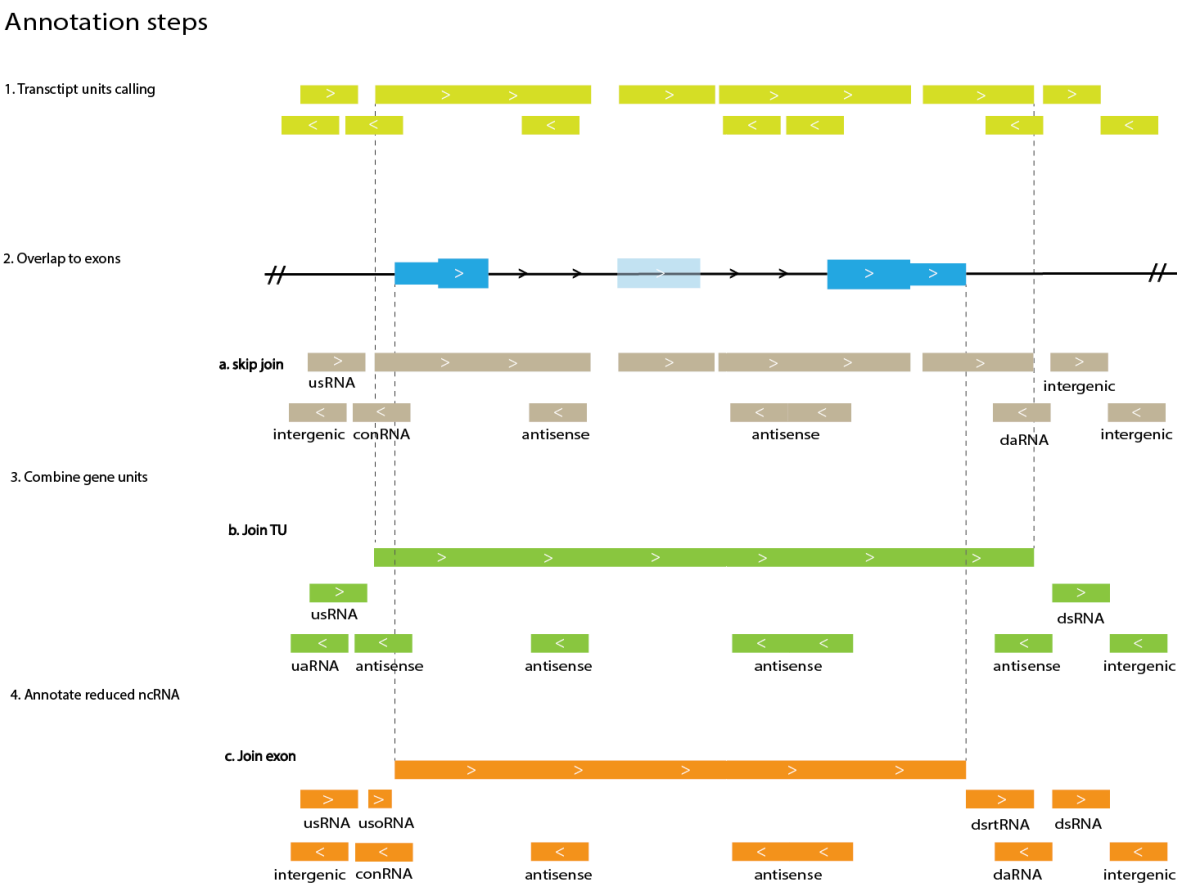

Figure 4.4 TU annotation steps. Non-coding TUs are named according to their relative location to the neighbored gene.

4.3.3 Spike-in RNA design

To achieve coherent normalization of total RNA and labeled RNA, 4sU labeled and unlabeled ERCC synthetic spike-in RNAs were used as the external references described before[26]. The six pairs of spike-in RNAs were prepared in the labeled/unlabeled mixture (Table 4.1):

Table 4.1 Design of spike-in RNA mix, prepared with unlabeled and 4sU labeled *in vitro* transcripts.

ERCC spike-in RNA	Concentration (ng/μL)	Labeled rate (%)
Sp2 (ERCC-00043)	1	100
Sp4 (ERCC-00136)	0.1	100
Sp5 (ERCC-00145)	1	10
Sp8 (ERCC-00092)	0.1	10
Sp9 (ERCC-00002)	1	0
Sp12 (ERCC-00170)	0.1	0

After BioAnalyzer (Agilent, 2100) verification of the mix, 0.4 ng/million cell spike-in was added into the TRIzol (ThermoFisher, 15596018) cell lysis to track the technical errors through biotinylation, RNA purification, and library preparation steps. The RNA molecular number and labeled rate were estimated from the designed standards of spike-in RNAs.

Cross-contamination rate was estimated from the ratio between the observed LRNA reads of unlabeled spike-ins (Sp9 and Sp12) and the expected LRNA reads given their FRNA reads, assuming the labeled rates were 100% in the same linear model trained on the labeled rate and spike-ins TPM (transcripts per million reads).

Two linear models for RNA copy number and turnover rate prediction were illustrated with the colored pseudo-scales (Figure 4.5), in agreements with the spike-ins weights and labeled rates as the parallel dashed lines indicated. Nevertheless, the tilted lines suggest that 4sU labeled RNAs could be moderately over-represented in the total RNA library. The unlabeled spike-in #12 showed a technical error of reads counting in the labeled samples, so this spike-in was omitted for model training.

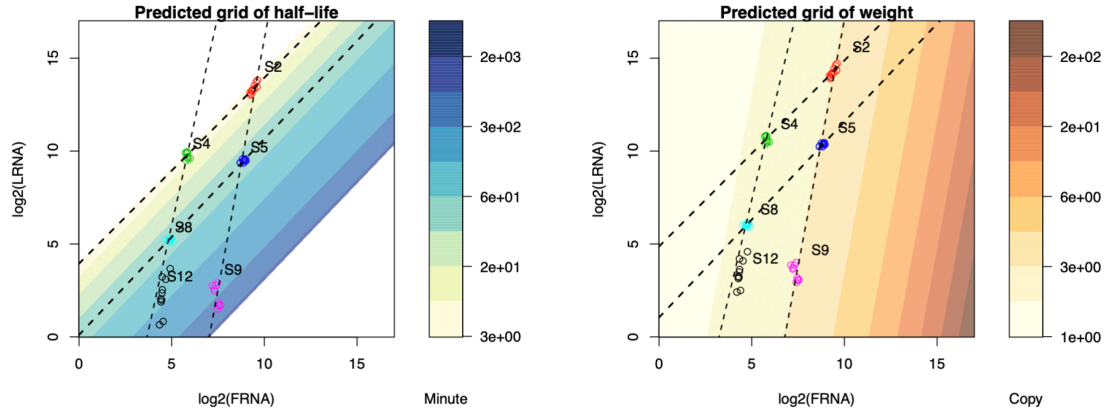

Figure 4.5 The test of spike-in reproducibility in different samples. Each dot indicates a spike-in's tpm (labeled RNA and fragmented total RNA) in a sample, and the overlaps of dots indicate the spike-ins' internal scales are well preserved across the 12 samples. The dashed lines connect the spike-in RNAs with the same mole number and the same labeled rate. The color scales are generated from spike-ins trained linear models for copy number and turnover half-life predictions.

4.3.4 TT-seq sample size estimation

Sample relative sizes were derived from the DESeq's size factors. In detail, the alignment-free mapper (Kallisto 0.46.2) was applied with the indexed transcriptome of GENCODE transcripts, *de novo* annotated ncRNAs, and six spike-ins sequences. The relative sizes of total and labeled samples were calculated separately with DESeq's method[136], the size factors were obtained from spike-ins normalized transcriptomes (total RNA libraries normalized with all spike-ins, labeled libraries normalized with labeled spike-ins). The size factors of spike-ins for sample normalization were calculated with DESeq's method.

4.3.5 TT-seq RNA synthesis rate estimation

After spike-in normalization, the gene-level estimated labeled/total RNA read counts of GENCODE transcripts were subjected to the linear model trained for labeled rate, which used labeled spike-ins (Sp2, Sp4, Sp5, and Sp8) $\log_2$ labeled (X_L) and total (X_F) read counts in response to the respective label rates r: $\log_2(r) \sim X_F + X_L$. After obtaining the predicted

labeled rates, the transcript copy number per cell was predicted with a second model trained on all spike-ins weight per cell, w: $\log_2(w) \sim X_F + X_L$. Then the RNA synthesis rate ($cell^{-1}$ $minute^{-1}$, or copy/min per cell) was converted by multiplying the labeled rate and copy number.

4.3.6 TT-seq RNA turnover half-life estimation

4.3.6.1 RNA half-life definition

First, the RNA half-life term by TT-seq represents neither RNA stability nor RNA decay rate, but the turnover rate as the reasons in Box 4.1.

Box 4.1 The short pulse metabolic labeling, below 10 minutes, yields the average turnover momentum in the current cell population, which is confounded by cell volume, cell growth rate, RNA co-transcriptional processing rate, and pre-existing RNA degradation rate. While the long-term pulse and chase strategy, such as SLAM-seq, provides the average stability of individual RNA at the cell population level, irrespective of the cell growth dilution effect and co-transcriptional processes. Sometimes these two perspectives are ambiguous in terms of "half-life," but the two concepts generate distinct scales of minutes. Therefore, in this thesis, TT-seq RNA half-life will be written as "turnover half-life" to avoid misunderstanding.

Next, RNA turnover half-life estimation would not require the steady state hypothesis or coerce degradation rate to synthesis rate but can be transformed into the same equation (2) under classic assumptions.

The steady-state model fixes total RNA abundance with equal synthesis to supplement degradation:

$$\mathrm{X(t)} = Y_{st}(1 - e^{-\lambda t}) \quad (1)$$

Where Y_{st} is the steady state total RNA, λ is the degradation rate, X is the RNA synthesis rate. Then the classic RNA half-life $t_{1/2}$ equals to:

$$t_{1/2} = \frac{\ln(2)}{\lambda} \quad (2)$$

4.3.6.2 Assumptions based estimation

Nonetheless, direct evidence is required that the cellular RNA content Y_{st} is a constant given any moment, to support the steady-state model. In pulse metabolic labeling, RNA half-life can be approximated under **two assumptions,** without calculating degradation:

1. no labeled RNA decays.
2. total RNA abundance is stable after the pulse labeling.

Since assumption 1, newly synthesized RNA is a linear accumulation of synthesis rate μ with the pulse Δt:

$$X(\Delta t) = \mu \Delta t$$

(3)

After Δt, total RNA $\mathrm{Y}(\Delta \mathrm{t})$ will be a mix of pre-existing RNA $\mathrm{Y}'(\Delta \mathrm{t})$ and newly synthesized RNA $X(\Delta t)$:

$$\mathrm{Y}(\Delta \mathrm{t}) = \mathrm{Y}'(\Delta \mathrm{t}) + X(\Delta t)$$

(4)

Since the 5 minutes labeled rate r can be directly predicted from the spike-ins trained linear model:

$$r = \frac{X(\Delta t)}{Y(\Delta t)} = \frac{\mu \Delta t}{\mathrm{Y}'(0) e^{-\lambda \Delta t} + \mu \Delta t}$$

(5)

Due to assumption 2, the initial total RNA abundance $\mathrm{Y}'(0) \approx Y(\Delta \mathrm{t})$. Then after simplification, the pseudo-degradation rate λ can be wrote as:

$$\lambda = \frac{-\log(1 - r)}{\Delta \mathrm{t}}$$

(6)

So the turnover half-life can be calculated with only the labeled rate:

$$t_{1/2} = -\Delta \mathrm{t} * \frac{\log(2)}{\log(1 - r)}$$

(7)

Of note, this equation (7) is identical to the previous TT-seq half-life calculation in the "SpikeinNormalization" package, which assumed a steady-state model. But the difference is that the new spike-ins design in this study has wider linear space to adjust the dynamic range of RNA synthesis, as labeled spike-ins can occupy 2% and 20% of SL and mTORi cells libraries. Therefore, a single spike-in concentration may have an over-fitting issue when weak transcription deviates away from the spike-in scale, in both cell-level and transcript-level estimation.

4.3.6.3 Simplified turnover half-life

Another well-known RNA half-life approximates with the ratio between labeled RNA and total RNA read count, $\frac{L}{T}$, to represent the turnover tendency. So the simplified RNA half-life becomes:

$$t_{1/2} = \frac{\ln(2)}{\frac{L}{T}} = \frac{\ln(2)}{\frac{X}{Y}} = \frac{\ln(2)}{r}$$

(8)

The equation (7) denominator by the Taylor series transformation can be written as:

$$-\ln(1 - r) = \sum_{n=0}^{\infty} \frac{r^n}{n} = r + \frac{r^2}{2} + \frac{r^3}{3} + \cdots$$

(9)

Equation (9) has the identical term r with equation (8) denominator, therefore explaining why equation (8) can be used to approximate RNA half-life. Since the small magnitude of labeled rate r, the rest terms of equation (9), $\frac{r^2}{2} + \frac{r^3}{3} + \cdots$, are small. In practice, the results from equation (7) and (8) are highly correlated (Pearson's r > 0.95). Only caveat of using this equation (8) is that it can enlarge unstable RNAs' half-lives, in exchange of easy calculation.

4.3.7 Transcription elongation velocity estimation

The principle of velocity estimation behind is that TT-seq labeled RNA measures initiation frequency, and Pol II ChIP represents the molecular number of transcription machinery.

Let P_0 be the number of Pol II initiate per minute (a cell populational average), then for the next i th minute, with any travel length L_i, the Pol II P_i in the i th minute will equal to P_0,

$$P_i = P_0, \forall i \in \{1, \dots, m\} \quad (10)$$

Even if backtracking and pre-mature termination occur, the average number of Pol II per kb is:

$$\bar{P} = \frac{\sum_1^m P_i}{l} = \frac{mP_0}{l} \quad (11)$$

given a gene with l kb and m minutes transcription. Then, the average velocity is:

$$\bar{v} = \frac{\sum_1^m L_i}{m} = \frac{l}{m} \quad (12)$$

$$\bar{v} = \frac{P_0}{\bar{P}} \quad (13)$$

Since TT-seq LRNA RPK represents RNA synthesis copy per minute or transcription initiation frequency, it is proportional to P_0. The number of Pol II initiated per unit of time is available from the synthesis rate. Therefore,

$$\hat{v} = \frac{RPK_{TT-seq\ LRNA}}{RPK_{Pol\ II\ S5p}} \quad (10)$$

This velocity estimation method has been applied in the previous TT-seq studies[51,73,137]. The difference in this study is that we combined TT-seq LRNA coverage with Pol II S5p MINUTE ChIP-seq, which represents the movement of RNA Pol II rather than the synthesis events as mNET-seq describes.

4.3.8 Pausing index and pausing duration

TSS pausing intervals were obtained from the gene TSS Start-seq peaks, aligned by STAR 2.7.3a, and called peaks by HOMER with the following parameters: *findPeaks -style groseq -size 20 -fragLength 20 -inputFragLength 40 -tssSize 5 -minBodySize 30 -pseudoCount 1*. Pol II S5p MINUTE-ChIP density in the TSS Start-seq peaks and the gene body intervals, (+500, +1500 bp), were divided and yielded the pausing index. Pausing duration was obtained from the pausing interval length and the estimated velocity in respective interval.

4.3.9 Termination site detection

To handle coverage skewness, sparsity, and local fluctuation, a global detection method was developed that enables reproducible sample comparison. This method is a lightweight version of the previous termination site calling by a local segmentation method with an

external R package dependancy[26]. The termination evaluation has another weighted signal method for comparison between samples, but calling of termination site is not available[138].

The new method here detects the max density contrast in the termination window, before and after at the termination site i:

$$Arg\max\ (d_1 - d_2) \tag{11}$$

The average density of the coverage X given the termination window length L,

$$\bar{d} = \frac{\sum X}{L} \tag{12}$$

The contrast at the site i, can be simplified as follows:

$$\begin{aligned} d_1 - d_2 &= \frac{\sum_i x_i}{i} - \frac{\sum X - \sum_i x_i}{L - i} \\ &= \left(\sum_i x_i - \frac{i\sum X}{L}\right) * \frac{L}{(L-i)*i} \\ &= \sum_i \left(x_i - \bar{d}\right) * \frac{L}{(L-i)*i} \end{aligned} \tag{17}$$

Of note, the normalized cumulative sum of coverage (left term) is multiplied by the sliding weight (right term) which exaggerates beginning and end positions irrespective of the body coverage (Figure 4.6).

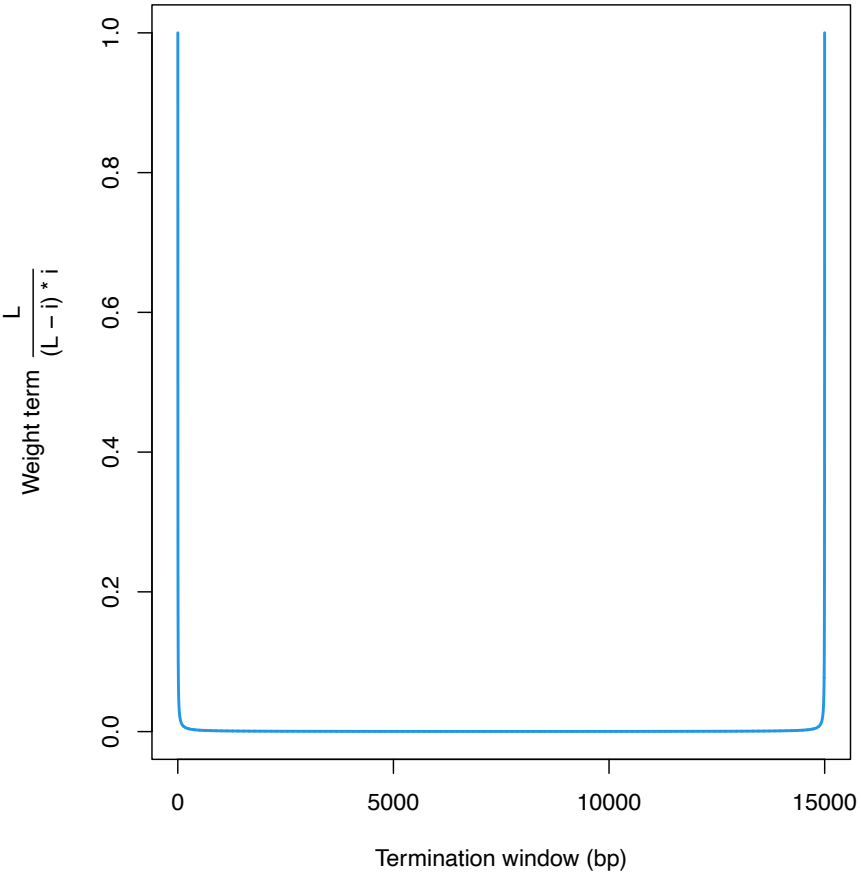

Figure 4.6 A line plot of the weight function that introduces a position bias on the termination site calling.

Therefore, removing this weight term can stabilize the result and increase robustness:

$$Arg\max\ \Delta d\ \approx\ \sum_i \left(x_i - \bar{d}\right) \quad (13)$$

This simplified algorithm also decreases time complexity by skipping an iteration loop for each base-pair in the 15 kb window. In practice, only three steps are required:

1. Normalize the coverage in the termination window to its mean,
2. Calculate the cumulative sums,
3. Call the max position[35].

4.3.10 Multiplexed ChIP spike-in-free normalization

The background normalization method[139] was adapted to correct the conventional ChIP sample size, and now can be used to stabilize the technical fluctuation that interferes with biological effect interpretation in MINUTE-ChIP. For example, time series transition can have less than 10% global difference, which is close to the input normalization inborn error. Because a global scaling against total input reads only vertically corrects the sample size, the horizontal technical errors still exist and require a background normalization, especially for the datasets that show systematic abnormalities for any ChIP target from that sample (Figure 4.7). Of note, this method was only applied to ChIP with genuine backgrounds, for example, Nanog, H3K4me3, and Pol II, if the global change was larger than technical fluctuation.

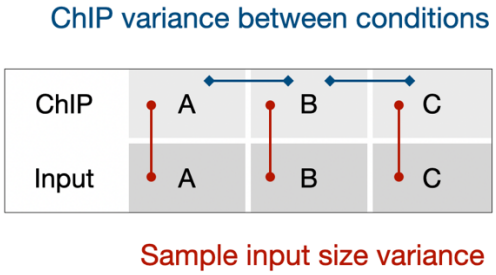

Figure 4.7 Multi-sample normalization with background control. An example of three conditions (A, B, and C), with vertical scaling to correct the sample size. But the technical variance still exists and introduces small fluctuation.

4.3.11 G-quadruplex pattern match

The reference genomes mm9 and hg19 were used for G-quadruplex (G4) sequence annotation. The inter-strand motifs and intra-strand motifs were matched, by expanding the canonical G4 motif $G_{3+}L_{1\text{-}7}$ to the opposite strand. Let A = G_{3+} and B = C_{3+}, 8 G4 combinations (AAAA, AAAB, AABA, AABB, ABAA, ABAB, ABBA, and ABBB) were detected with the regular expression. Beyond the canonical intra-strand G4 pattern AAAA, the extended intra-strand PQS (Putative G-Quadruplex Sequences) $G_{3+}L_{1\text{–}12}$, and two-tetrads $G_2L_{1\text{–}12}$, were assigned. The genome coverages of matched motifs were produced with the Bioconductor package “rtracklayer 1.46.0”.

4.3.12 Data availability

TT-seq data in mESC pluripotent states were uploaded in GEO (GSE168378). Transcription kinetics analysis and figure generation scripts are available on Github (https://github. com/shaorray/TT-seq_mESC_pluripotency).

5 RESULTS

5.1 PROJECT ONE: TRANSIENT TRANSCRIPTOME IN MOUSE ES CELL

5.1.1 TU annotation in mESC

An embryonic stem cell has a transcription permissive genome configuration[40]. To perform reproducible TU annotation of both the known genes and *de novo* annotated non-coding transcripts, a shiny app "TU filter" was developed following the annotation steps in the first TT-seq study [26].

In the new annotation pipeline, several minor changes have been made. The original procedure joins gene TUs including the weak 5' region from the alternative TSSs and the diminishing 3' termination region. To calculate an accurate transcription level, TU filter disjoins the flanking non-coding regions (Figure 4.2). This step helps to remove the influence of alternative TSSs and the decay in the termination window but creates usRNA and dsRNA that are from the gene transcription (Figure 5.1).

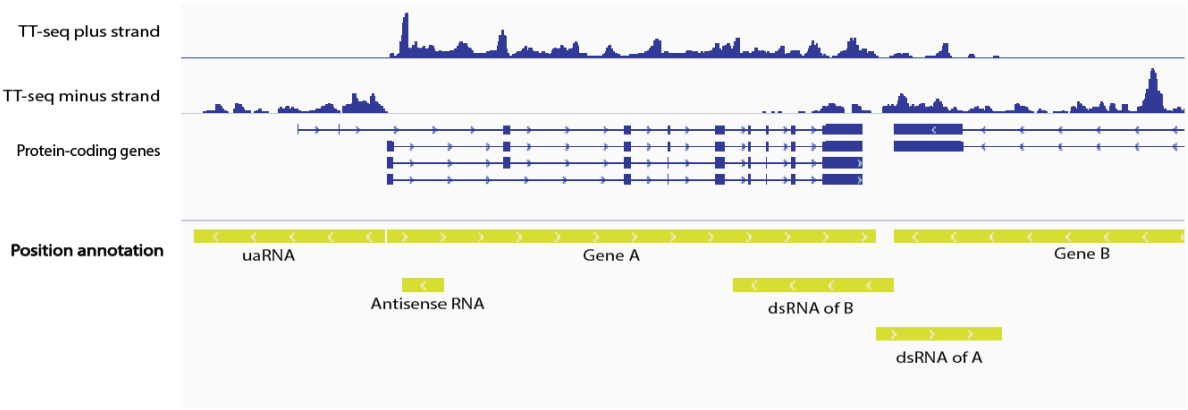

Figure 5.1 An example of TU annotation with overlapped transcription termination regions of two convergent genes.

In this study, non-coding RNAs may have slightly different names. Often, in the gene-dense regions, the relative location of ncRNA can be in multiple cases depending on the neighbored gene references. So ncRNA naming is according to its position to the nearest gene by promoter/termination regions (2 kb) overlaps. The pre-defined order will assign a unique name to each ncTU:

1. intergenic RNA
2. asRNA (gene cis-antisenses RNA)
3. uaRNA (upstream-antisense RNA)
4. conRNA (convergent RNA)
5. dsRNA (downstream-senses)
6. usRNA (upstream-sense RNA)
7. daRNA (downstream-antisense RNA)

Therefore, a later name assignment could override an earlier assignment in the multi-gene-neighbor situation. So a small number of ambiguous naming might exist in the closely related ncRNA types, e.g., uaRNA and conRNA (Figure 5.2).

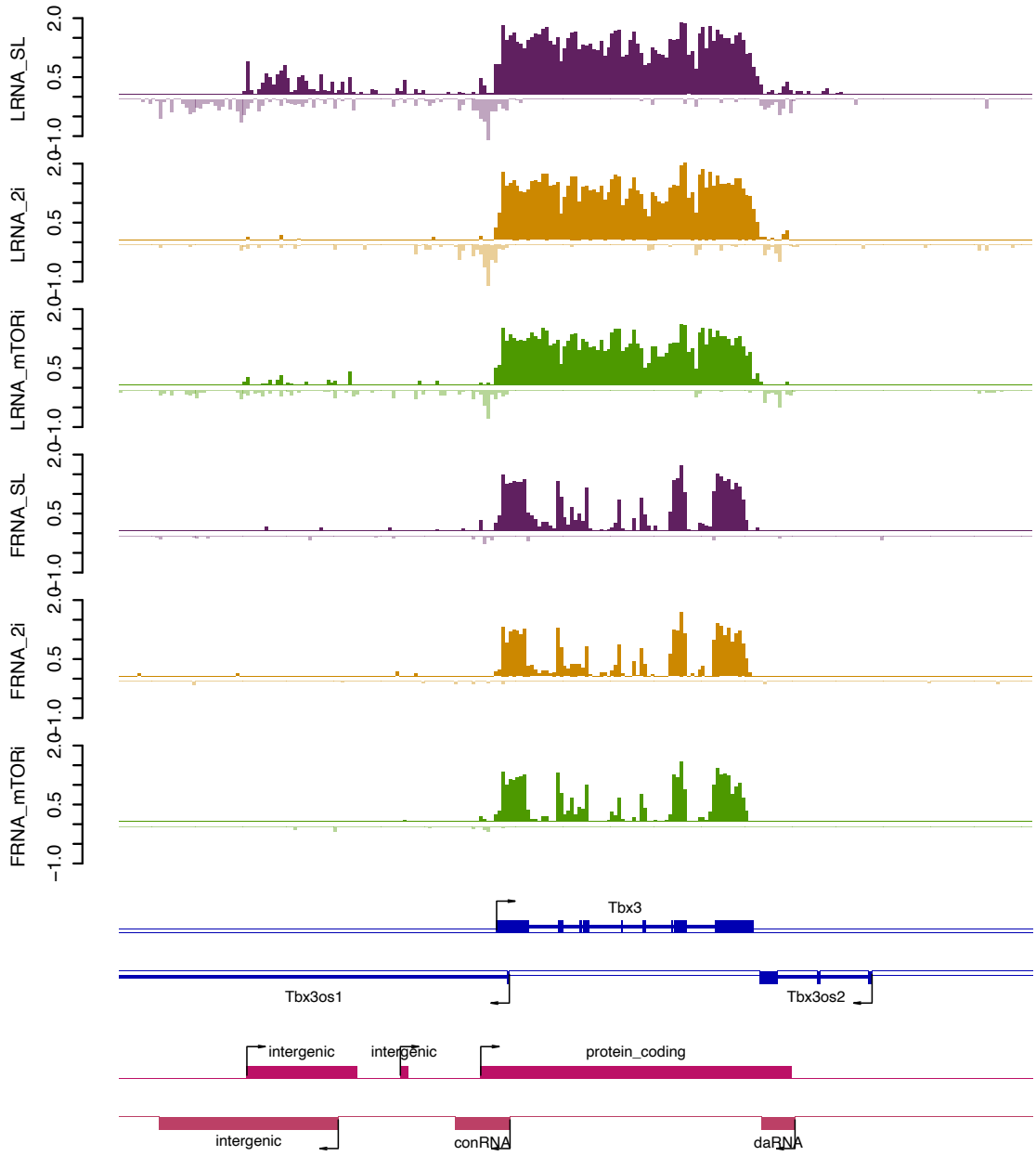

Figure 5.2 An example of TU annotation results in the Tbx3 gene neighborhood. TT-seq labeled RNA (LRNA) and total fragmented RNA (FRNA) are spike-in normalized. GENCODE reference (blue) and the annotated TUs (red) are indicated.

5.1.2 RNA turnover in a living cell

Estimating RNA turnover half-life requires two assumptions (4.3.6.2) of TT-seq short pulse labeling. Integration RNA degradation rate is available from SLAM-seq[23] resulted in an interesting saturation parameter. The current cellular RNA abundance Y' compared to theoretical steady-state Y_{st} is a ratio of saturation:

$$\rho = \frac{Y'}{Y_{st}} \tag{14}$$

Using equation (1), then:

$$\rho = \frac{1 - e^{-\lambda t}}{r} \quad (20)$$

For the pulse metabolic labeling, the turnover half-life can also be written as:

$$t_{1/2} = \frac{\ln(1+\rho)}{\lambda} \quad (21)$$

In the equation (21), the saturation parameter ρ adjusts to the shorter turnover half-life below the steady-state since $\rho < 1$. The empirical distribution of ρ, with TT-seq labeled rate r and SLAM-seq degradation rates λ, centers at 0.31 in mouse ES cells (Figure 5.3). The unsaturated turnover suggests an individual cell accumulates total RNA abundance with continuous cell growth.

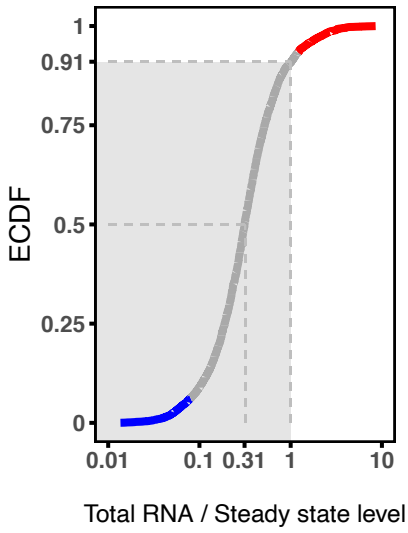

Figure 5.3 The empirical cumulative density curve of the saturation parameter ρ, as a ratio between the observed total RNA abundance and the theoretical steady-state capacity, from the TT-seq synthesis rate (this study) and SLAM-seq decay rate[23].

5.1.3 RNA labeling efficiency verification

Technical verification of RNA labeling efficiency was addressed from three aspects, 4sU incorporation rate, *bona fide* labeling time, and cross-contamination rate.

First, without alkylation, 4sU has a natural mismatch rate of 10% in the complementary DNA (cDNA) synthesis step[30]. This T->C rate can be recapitulated by comparing the mismatch frequency between labeled RNA and total RNA reads (Figure 5.4 A). The four samples showed no significant difference in 4sU incorporation rate.

Second, 4-thiouridine incorporation has many steps. After supplementing in the culture medium, 4sU experiences the free diffusion into the cell converts to 4sUTP and competitively incorporates into newly synthesized transcripts. Any step slowdown could confound into labeled RNA quantification. Since RNA splicing occurs co-transcriptionally[140], the splicing rate of labeled RNA may represent the effective labeling time. By extracting the splicing rate across different conditions, only a tiny difference

appeared among the pluripotent states (Figure 5.4 B). Hence, cell morphology and cell colony size will not impede 4sU incorporation or alter adequate labeling time.

Third, the mix of labeled/unlabeled spike-in RNA allows cross-contamination estimation. The cross-contamination rates were predicted from the spike-in-trained linear model and appeared to consistently have low values for all biological replicates (Figure 5.4 C). Due to the spike-in #12 having a technical error of Kallisto's read count (Figure 4.5), its actual cross-contamination could be lower.

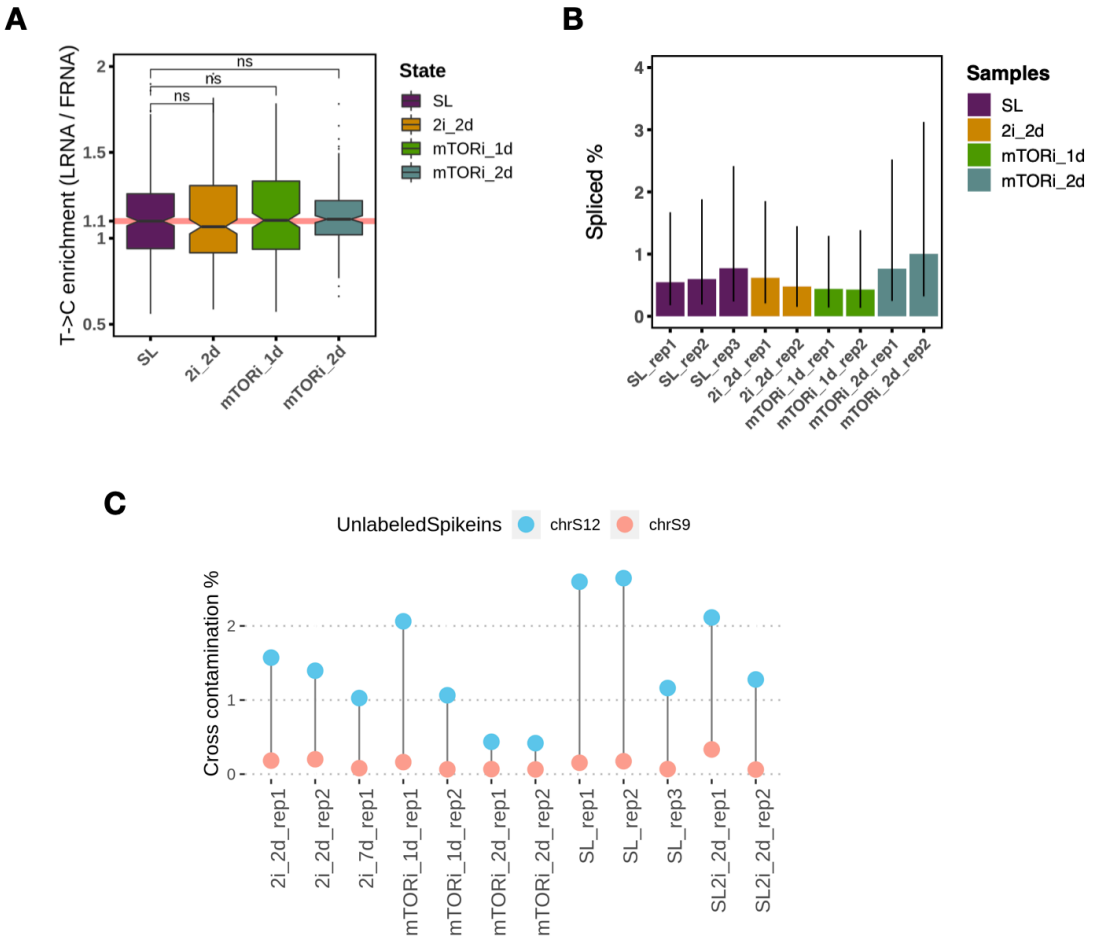

Figure 5.4 TT-seq 4sU labeling efficiency verification. (A) T->C mutation frequency rates across the conditions in comparison. The Student's two-tail t-test is performed (n.s., no significance). (B) Labeled reads splicing rates are compared across the samples. The error bar indicates the (0.25, 0.75) quantiles. (C) Cross-contamination rates of the unlabeled spike-in RNA across the biological replicates.

5.1.4 Transcription kinetics change in the pluripotent state transitions

RNA transcription contributes a large portion of the total RNA variation[141]. And mRNA abundance explains 40%-84% of the protein translation variance in mammalian cells[28]. Hence transcription is the main determinant of gene expression and is especially significant for unstable non-coding RNAs.

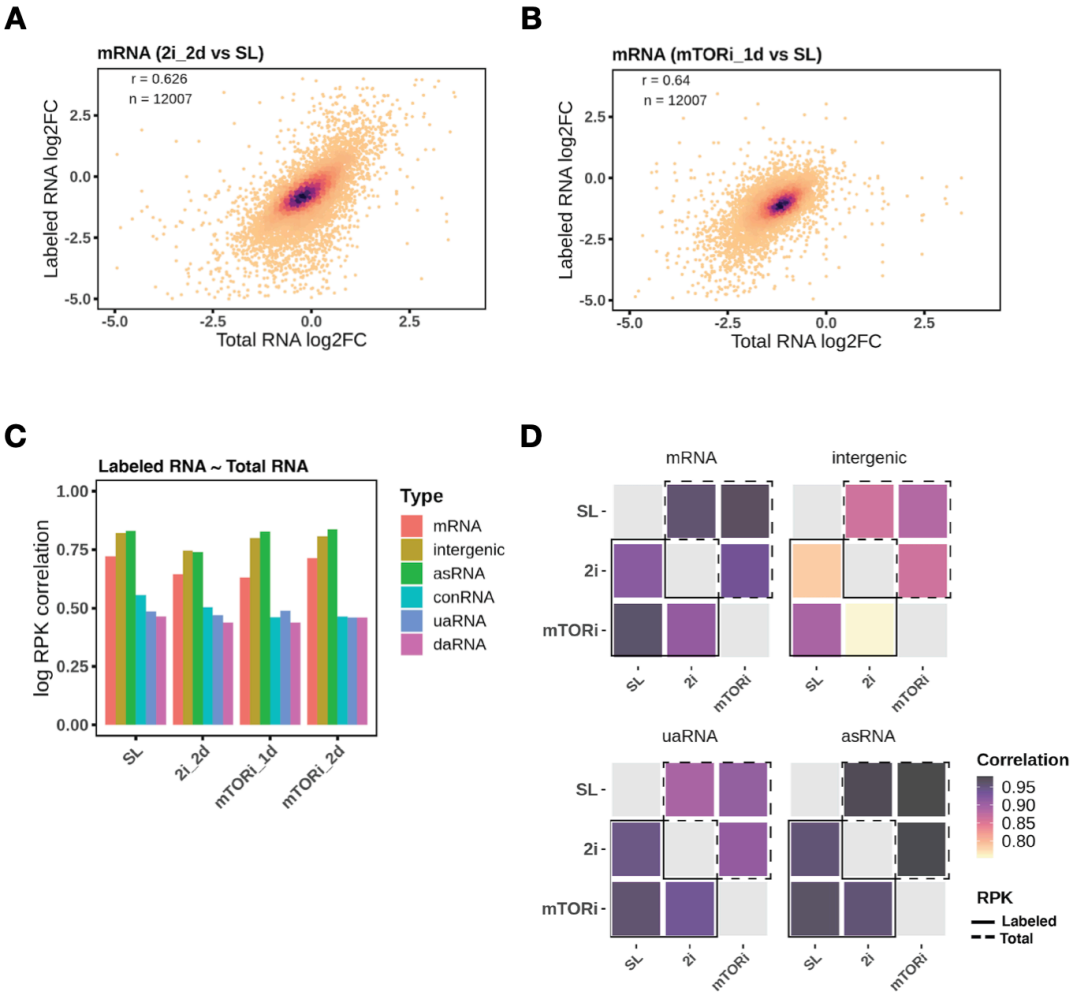

Figure 5.5 Transcription explains the majority of RNA abundance on the TT-seq annotated TUs. (A-B) Protein-coding RNAs log$_2$ fold-changes comparisons between transcription and total RNA in the pluripotent state transitions. (C) Pearson's correlation of log TU RPK between transcription and total RNA abundance by coding and non-coding TU types in the pluripotent states. (D) The pair-wise Pearson's correlation of log TU RPK between the pluripotent states by four TU types.

Newly synthesized RNA can evaluate transcription's contribution to RNA abundance on TT-seq *de novo* annotated TU types. For the 12007 protein-coding TUs, the changes in transcription correlated well with the changes in RNA abundance in the pluripotent state transitions (Figure 5.5A-B). In addition, the levels between transcription and total RNA abundance also correlated well for mRNAs, intergenic RNAs, and cis-antisense RNAs (Figure 5.5C). However, three non-coding RNA types (conRNA, uaRNA, and daRNA), derived from the opposite strand of the main genes, showed disassociated total RNA abundance with transcription. This result suggests a location-specific degradation of ncRNAs that overrides transcription's contribution. After post-transcriptional processing, mRNAs and intergenic ncRNAs abundance converged better among the pluripotent states than transcription, but slightly not for the unstable uaRNAs (Figure 5.5D). Together, transcription contributes a large portion of total RNA variance under cell physiological buffering.

5.1.5 Transcription neighboring effect

Enhancer elements have been found with Pol II occupancy and eRNA products[142]. As a proxy of enhancer activity, enhancer transcription shows an additive regulation of neighboring genes[50,54,56,143]. And enhancer-promoter contact frequency has been recently shown to follow a logarithmic decrease by genomic distance[55]. Before this physical evidence, ncRNA-mRNA co-expression in TT-seq also revealed a similar reverse relation by distance in SL-2i transition (Figure 5.6A).

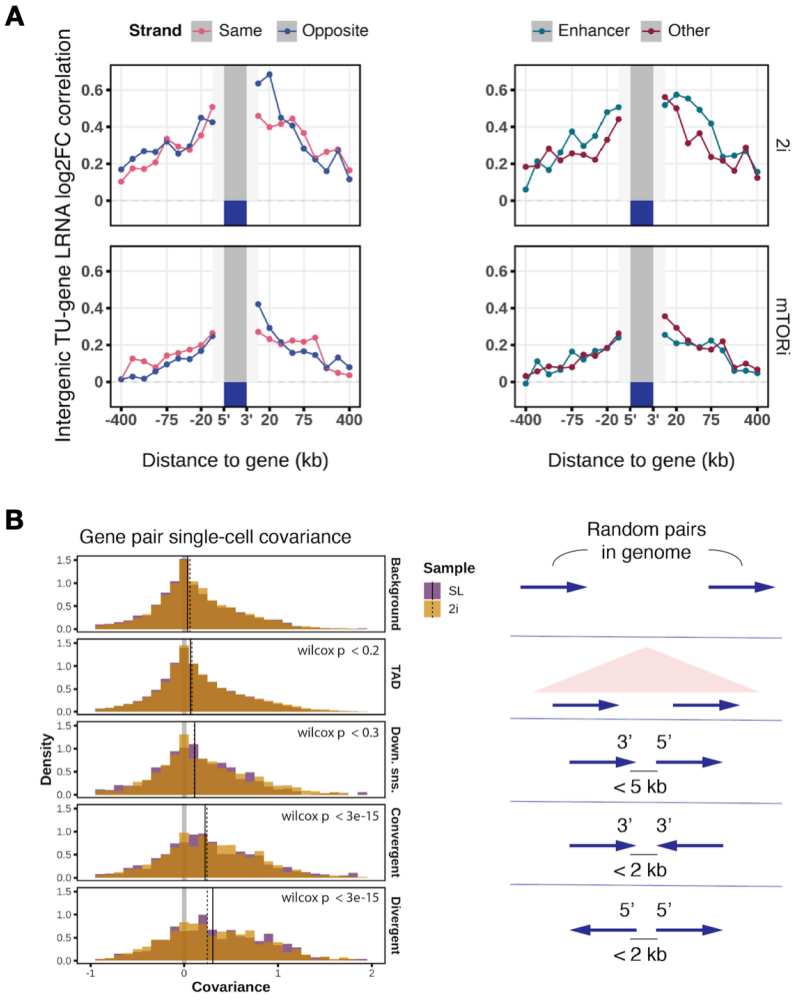

Figure 5.6 Transcription neighboring co-expression in mESC pluripotency transitions. (A). Labeled RNA $\log_2$ fold-change correlation between intergenic TUs (n = 14,978) and genes (n = 7,087) by the binned distance, the strand, and the enhancer annotations (FANTOM5, ChromHMM, and STARR-seq[58]). (B). Gene pairs covariance in single-cell RNA-seq, by random genome pairs, pairs in the same topological associated domain (TAD), adjacent sense-strand pairs, downstream convergent pairs, and divergent promoter pairs. Vertical lines are drawn at the medians of SL and 2i covariance. Wilcoxon signed-rank tests are performed against zero covariance.

Of note, this distance-dependent co-expression appeared to have no preference for strand or enhancer annotation, except for downstream convergent TU pairs in closer associations (Figure 5.6A top). For cross-validation, gene-pairs co-expression test was performed in an SL and 2i single-cell RNA-seq dataset[128]. The promoter-divergent gene pairs showed the highest covariance, and the downstream convergent gene pairs ranked the second. These results manifest co-expression of ncRNAs with the neighbored genes.

5.1.6 Transcription velocity estimation with TT-seq and Pol II S5p coverage

TT-seq labeled RNA, and RNA Pol II S5p provide elongation velocity estimation at both gene level and local intervals (pausing window, gene body, and termination window) (Materials and Methods, 4.3.7). In this study, the estimated velocity has been cross-validated. The GRO-seq dataset with a time series Cdk9 inhibition was reprocessed[69]. The ongoing transcription distances from gene TSS were annotated by the "TU filter" tool. For each gene, the distances and times were subjected to a linear model in response to the Cdk9 inhibition duration, which provided the elongation velocity from the slope coefficient term. The resulted velocity measurements (n=1944) were used as the "measured elongation velocity" and appeared to correlate with the TT-seq estimated velocity (Figure 5.7 B).

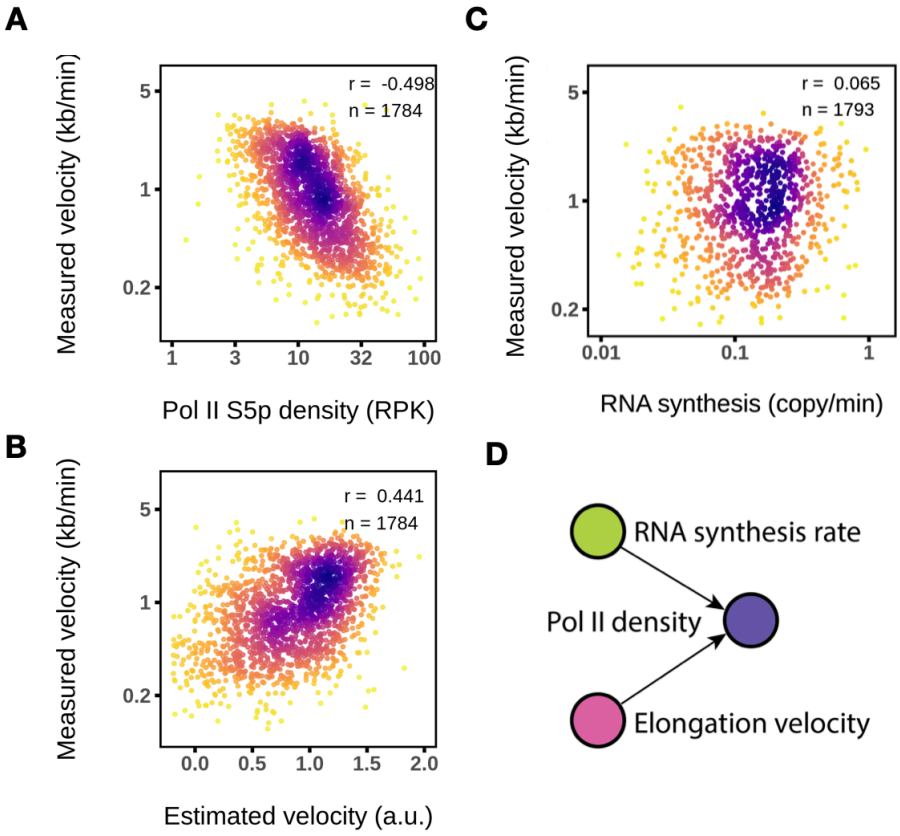

Figure 5.7 Transcription velocity estimation. (A) A scatter plot with RNA Pol II gene body density and GRO-seq externally measured velocity. Pearson's correlation is performed after log transformation. (B) A scatter plot between TT-seq estimated velocity and GRO-seq measured velocity. Pearson's correlation is indicated. (C) A scatter plot between TT-seq RNA synthesis rate and GRO-seq measured velocity. Pearson's correlation is shown. (D) A diagram of RNA synthesis, elongation velocity, and Pol II density relations.

Fast Pol II dilutes its occupancy and condenses Pol II chromatin binding. Accordingly, Pol II gene body density reversely correlated well with GRO-seq measured velocity (Figure 5.7 A). However, transcription initiation frequency (or RNA synthesis rate) association with elongation velocity appeared insignificant (Figure 5.7 C), which confirms that the observed Pol II coverage is subjected to transcription initiation and elongation velocity (Figure 5.7 D). Nevertheless, velocity modulation would help to evaluate if elongation velocity is a rate-limiting parameter.

5.1.7 Transcription velocity interpretation

The three transcription stages appear on the estimated velocity profile (Figure 5.8 C). Pol II movement showed a deep stall at TSS and a quick velocity restoration after pause-release. Pol II traveled through the gene body at a steady velocity for all three pluripotent states. Beyond the transcript end site (TES), Pol II S5p accumulated, RNA synthesis diminished, and the elongation slowed down in the termination process.

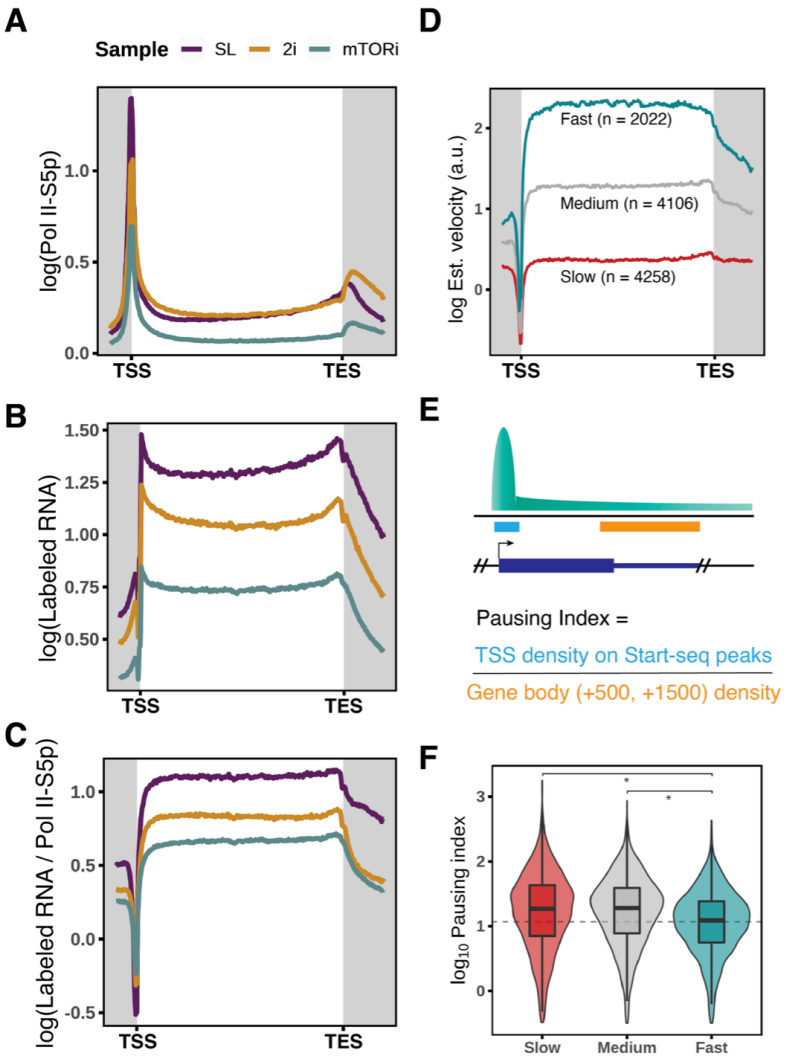

Figure 5.8 Estimated transcription velocity profile. (A-C) Average gene coverage profile of quantitative MINUTE-ChIP of Pol II S5p, spike-in-normalized TT-seq-labeled RNA, and the estimated elongation velocity (n=10,447, a.u.). (D) K-means clustering of the estimated velocity in the SL condition (E) A diagram of the pausing index calculation method. (F) Boxplot of pausing index distribution by three elongation velocity groups. Student's t-test is perform with log-transformed pausing index, *P < 2.2e-16.

The transcription pause-release dynamics can be represented by the pausing index, a ratio of Pol II densities between the TSS pausing interval and gene body (Figure 5.8E). To test whether the successful Pol II release related to elongation velocity, we clustered average gene velocity into three groups and found less pausing in the fast elongation group (Figure 5.8F). This result implies that promoter-proximal pausing is not an independent event but a connected step for a transcription cycle.

Further, to understand the estimated elongation velocities, various public genomic features were collected and compared (Figure 5.9A). The closest associations appeared from the repressive histone variant and modification (H2A.Z, H3K7me3, and H2Aub), elongation-related H3K36me3, H3K79me2, and chromatin remodelers (Chd2 and Chd9). DNA sequence motifs, DNA/RNA modifications, general transcription factors, and histone acetylation exhibited marginal correlations with elongation velocity. Chromatin opening (DHS and FAIRE) and looping (Med1 and CTCF) also moderately indicated the velocity extent. Therefore, the chromatin determinant of velocity might also reside in the gene body that controls the flexibility of chromatin fiber.

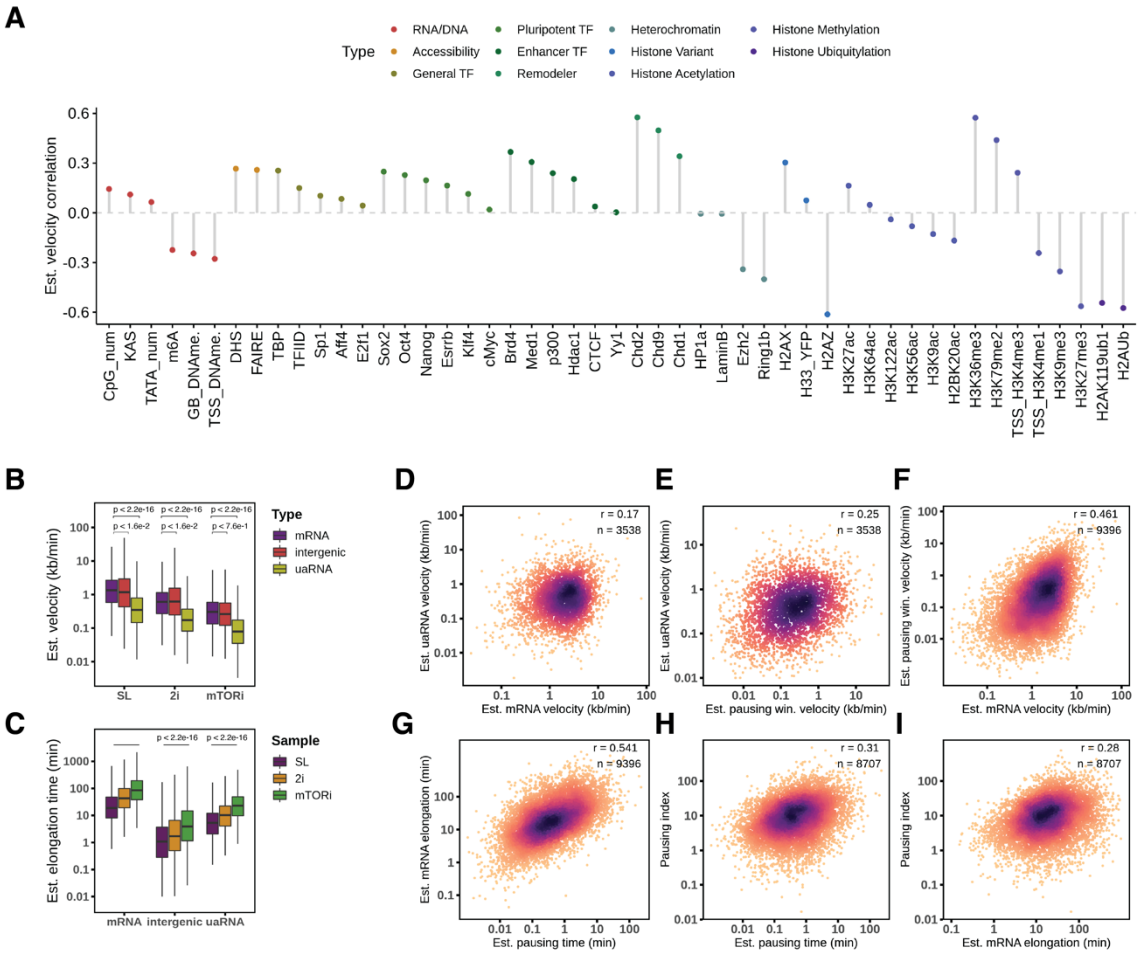

Figure 5.9 Transcription velocity interpretation. (A) A dot plot of Pearson's correlation coefficients between the estimated gene elongation velocity (n = 10,611, SL state) and the genomic features in mESC. (B-C) Estimated velocity scaled by GRO-seq measured velocity plotted across the culture conditions and TU types. Boxplots are with central bands at the median, 0.25 and 0.75 quartiles box area, and 1.5× interquartile range (IQR) whiskers; outliers are hidden. A two-tailed unpaired Student's t-test is performed on the log scale. (D-F) Estimated velocity correlation between mRNA, paired uaRNA, and mRNA TSS pausing interval (Start-seq peaks). The estimated elongation dynamic parameters are in SL state, the same as below. (G-H) Pearson's correlations of estimated pausing time (in Start-seq peaks) with gene body elongation time and pausing index. (I) Pearson's correlation between the estimated mRNA gene body elongation time and the pausing index.

This inhibition-free velocity estimation relied on the external scale from GRO-seq measurements and provided velocities in non-coding regions. For uaRNA-gene pairs, we found the non-coding direction transcribed slower, which might explain the lack of active chromatin marks (Figure 5.9B-C). However, uaRNA velocity appeared almost independent

of the main gene velocity (Figure 5.9D), and uaRNA elongation was marginally associated with the pausing dynamics. The velocity in the short pausing window might be limited with TT-seq and Pol II S5p ChIP. Nevertheless, at the pausing interval, velocity and pausing time appeared to correlate better with gene body elongation velocity and time, above the gene length confounding of two random variables (Figure 5.9F-G). The evidence of pause-release dynamics with non-random velocities (Figure 5.8F, 5.9H-I) again suggests that promoter-proximal pausing is a connected step in a transcription cycle.

5.1.8 Transcription termination site estimation

Transcription termination involves DNA motif, RNA secondary structure, and exoribonuclease digestion[26]. In this study, termination distance supports Pol II elongation slowdown[144,145], as manifested in the binned TT-seq labeled RNA coverages by the GRO-seq externally measured velocities (Figure 5.10 A). The inhibitor-induced states (2i and mTORi) appeared with slower elongation and shorter termination distances (**Paper I**, Fig 5G)[145]. Furthermore, the multi-feature comparison confirmed the top importance of velocity in the termination window that determines the termination distance (Figure 5.10 B).

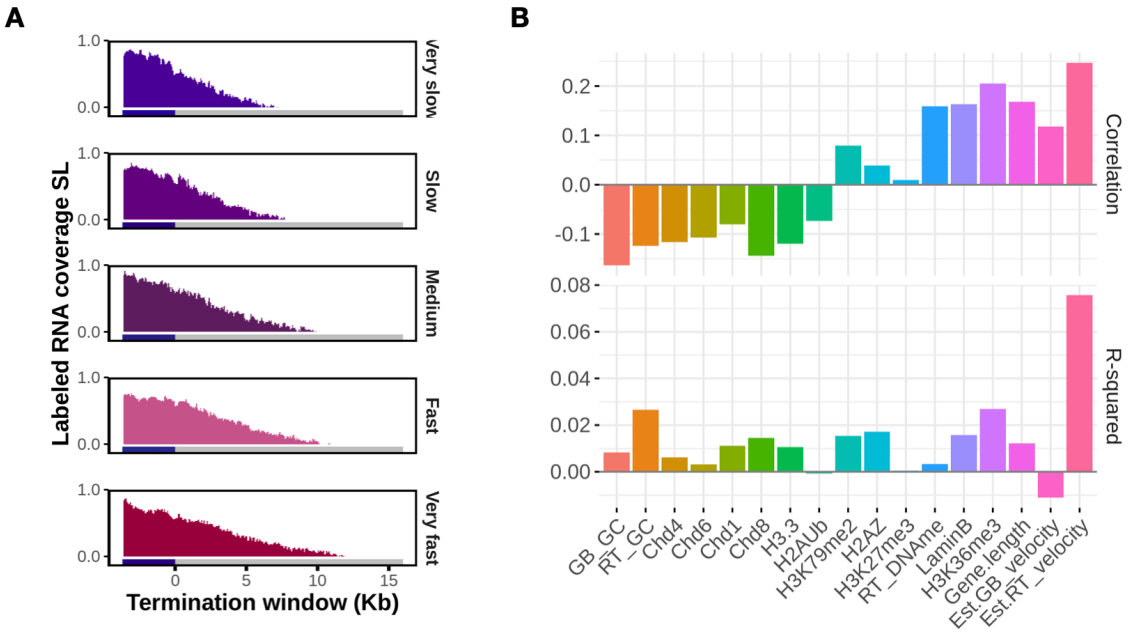

Figure 5.10 Termination distance is associated with velocity. (A) The median coverage of TT-seq labeled RNA (SL) in the 15 kb termination window grouped by the GRO-seq measured elongation velocity classes. (B) The bar plots show the Person's correlation coefficients (upper) and R-squared values (lower) of termination distances with the velocity and chromatin features as the explanatory variables.

5.1.9 Epigenome modulation of transcription kinetics in ES cells

To explain transcription frequency, turnover half-life, and total RNA abundance from an epigenomic perspective, the protein-coding RNAs and intergenic ncRNAs in mESC were separately subject to Pearson's correlations and R-squared decompositions with a collection of genomic features (Figure 5.11A). The combined features explained 72.4% of mRNA transcription variance and 46.5% of mRNA abundance variance. However, the intergenic ncRNAs transcription and RNA abundance variances were only explained by 16.9% and 30.8%. Intriguingly, chromatin accessibility and histone acetylation are anti-correlated with

intergenic ncRNA abundance. Although RNA metabolic turnover might be indirectly associated with the chromatin features, 16.1% of intergenic RNA turnover half-life variance was explained compared to the 5.2% explanation of mRNA half-life. Of note, H3K27me3's weakly positive correlation with intergenic RNA half-life might suggest that the repressive chromatin vicinity protects ncRNA from degradation.

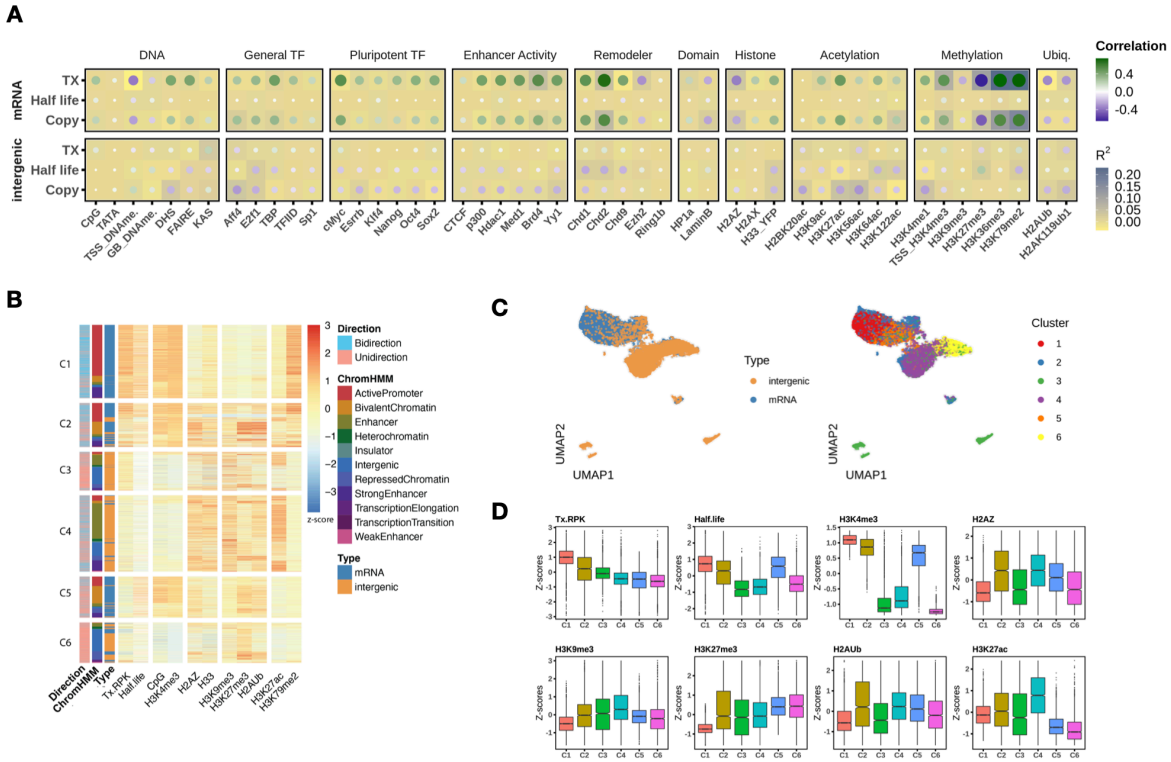

Figure 5.11 Epigenomic features distinguish mRNA from intergenic ncRNA**.** (A) The genomic features' heatmaps of Pearson's correlation (dot) and decomposed R-squared values (color tile) explain transcription frequency, turnover half-life, and total RNA abundance. (B) Mclust groups separate the mix of mRNA and intergenic ncRNA with 25 genomic features. The clusters are sorted by transcription frequency and annotated with promoter directionality, ChromHMM states, and several selected histone modifications. (C) UMAP two-dimensional space with the 25 genomic features separates mRNAs from intergenic ncRNAs. (D) Boxplots contrast the difference between mRNAs and intergenic ncRNAs with the indicated features after log-Z-transformation.

Next, to test the connectivity between transcriptome and epigenome, we kept the 25 non-redundant features (with positive R-squared values in mRNA transcription explanation) and clustered mRNAs and intergenic RNAs together with a Gaussian mixture method. After ranking by transcription levels, the 6 clusters separated mRNAs (C1, C2, C5) and intergenic RNAs (C3, C4, C6) (Figure 5.11 B). With the predefined ChromHMM states[146], active promoter states (red) were prominent in the highest transcribed cluster (C1) and lowly transcribed clusters (C4 and C5) enriched with enhancer and bivalent states. And active mRNA cluster (C1) and enhancer cluster (C4) were predominant with bidirectional promoters[147], compared to the uni-directional bivalent promoters (C5). The clustering result was further integrated with the selected chromatin features. mRNA TSSs distinguished from ncRNAs by CpG dinucleotide and H3K4me3 enrichment; lowly transcribed mRNAs (C2) were coated with higher levels of H2A.Z, H3.3, and repressive histone modification; in addition to H3K27ac; the enhancer cluster (C4) also enriched with H3K9me3, H2A.Z, and H3.3 (Figure 5.11 B). The 25 chromatin features were further projected to the UMAP two-dimensional space and separated mRNAs and ncRNAs with a clear gap (Figure 5.11

C). These results suggest that although lowly transcribed mRNAs (C2) and a small fraction of ncRNAs (C3) share repressive chromatin features, their epigenome compositions were intrinsically distinguishable in terms of active marks and turnover kinetics (Figure 5.11 D).

5.2 PROJECT TWO: HISTONE ACYL-MODIFICATION WITH GENETIC CODON EXPANSION

5.2.1 Pre-modified protein acylation with genetic code expansion

The hydrophobic pocket of PylRS catalytic cavity has UAA selectivity[86]. To engineer bio-orthogonal PylRS catalysis of aminoacyl-tRNA^{Pyl} as a histone acylations tool, we tested the affinities of five PylRSs with six acyl-lysine substrates.

The respective constructs of PylRS variants and histone templates were cloned as described (Material and Methods, 4.1.2). Our preliminary test showed that acetyl-lysine, crotonyl-lysine, and propinyl-lysine could efficiently be incorporated into the GFP reporter with the AcKRS (Figure 5.12 A). And ButKRS specifically incorporated butyryl-lysine with high efficiency. β-hydroxyisobutyry-lysine only showed a weak reactivity with its enzyme HibKRS (Figure 5.12 B). Thus the combination of AcKRS and ButKRS could allow histone acylation incorporation to a wide extent (Figure 5.12 C).

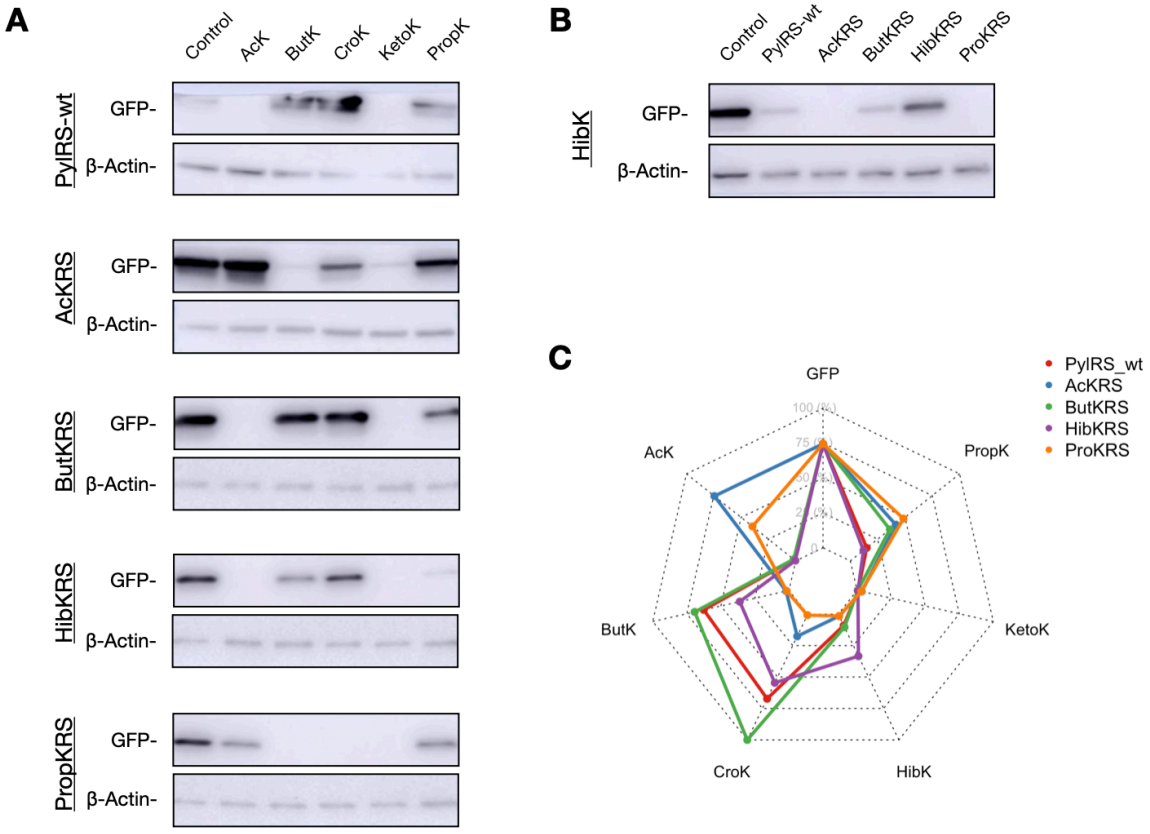

Figure 5.12 Acylation incorporation reactivity profiling. (A-B) A single amber stop codon (150) GFP reports the acyl-tRNA^{Pyl} reactivity in HEK293t cells with the amber suppression system (Materials and Methods 4.1.2). Different mmPylRS variants were treated with 2 mg/mL acyl-lysines for 24 hours. Western blots indicate the GFP expression levels of each combination. (C) A radar plot of the relative ratios of GFP fluorescence emission values scaled to total protein weight (BCA) and the wild-type GFP florescence as a reference. Axis labels are acyl-lysines, and line groups are the specified PylRS variants.

5.2.2 Install histone acylation in vivo

Next, histone H3.3 were expressed with acetyl-lysine and butyryl-lysine supplementation to test endogenous activity in HEK293t cells. After 24 hours, the expression of H3.3 reached ~10% of the total H3.3 amount (Figure 5.13 A, left, input lane). However, HA-tagged H3.3K27ac were incomparable to total endogenous H3K27ac (Figure 5.13 A, middle). Moreover, the missing signal of butyryl-lysine treated cells suggests that anti-H3K27ac antibody cannot cross-react with the butyryl group. In the HA-IP lanes, HA-tagged H3.3 overwhelmed the endogenous H3.3, but this gap diminished for H3K27ac (Figure 5.13 B). So these results suggest that the pre-modified H3.3K27ac is deacetylated to more than 40% (Figure 5.13 C). Hence, the *in vivo* expressed histone H3.3 confronts a challenge for further functional readouts (e.g., TT-seq and mass spectrometry). Because:

1. H3K27ac is highly unstable with the endogenous expression system.

2. 24 hours pulse expressed H3.3 only takes up 10% of total histone H3.3. Proteomic signal could be difficult to detect under total H3 background with mass spectrometry.

3. The nucleosomal histone H3 non-specific binding on beads mixes in a large fraction of H3K27ac signal and predominates HA-tagged H3.3 (Figure 5.13 A, right, anti-H3 WB).

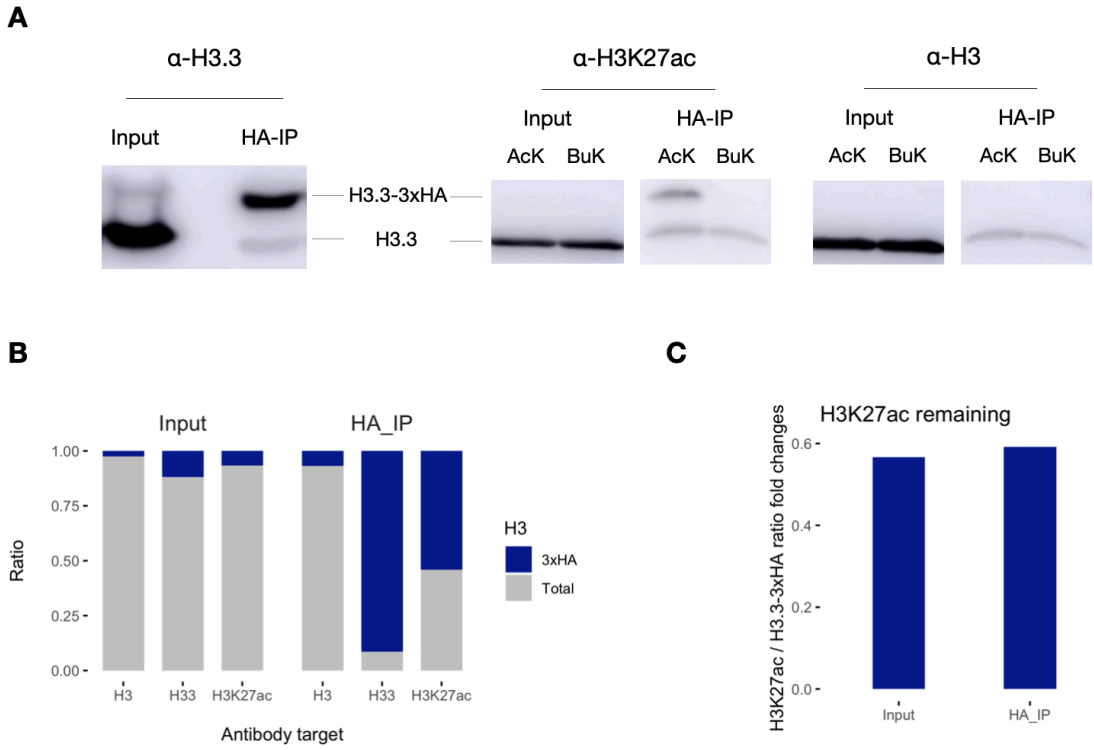

Figure 5.13 Measurement of in vivo expression H3.3 with K27-acylation. (A) HA-IP western blots of 24 hours pulse expression of HA-tagged H3.3 in HEK293t. Input and pull-down lanes are indicated. (B) The proportion bar-plot of relative HA-tagged H3.3 amount in the input and HA-IP histone H3 pool. (C) The relative amount of HA-tagged H3.3 with K27ac modification.

Given the facts above, histone acylation pre-modified strategy could be inevitably limited to changing hardwired transcription programs. It is still inspiring to further investigate the native responses to the pre-occupation of one acyl residue on the same site that precludes the alternatives, e.g. H3K27me3, without introducing a mutant histone.

5.3 PROJECT THREE: RAPID H2A DE-UBIQUITINATION BY BAP1 PULSE EXPRESSION

Active or repressive histone modification is named with gene expression states, but the causal relation is still largely unclear. Here, we evaluated the direct role of the repressive histone modification H2Aub (Histone H2A lysine 119 mono-ubiquitination) in mESC, by pulse expressing the H2Aub-specific de-ubiquitinase BAP1 (Materials and Methods, 4.1.2). The results confirmed the central role of H2Aub in Polycomb-mediated gene silencing.

5.3.1 Active H2Aub depletion reverses Polycomb mediated repression

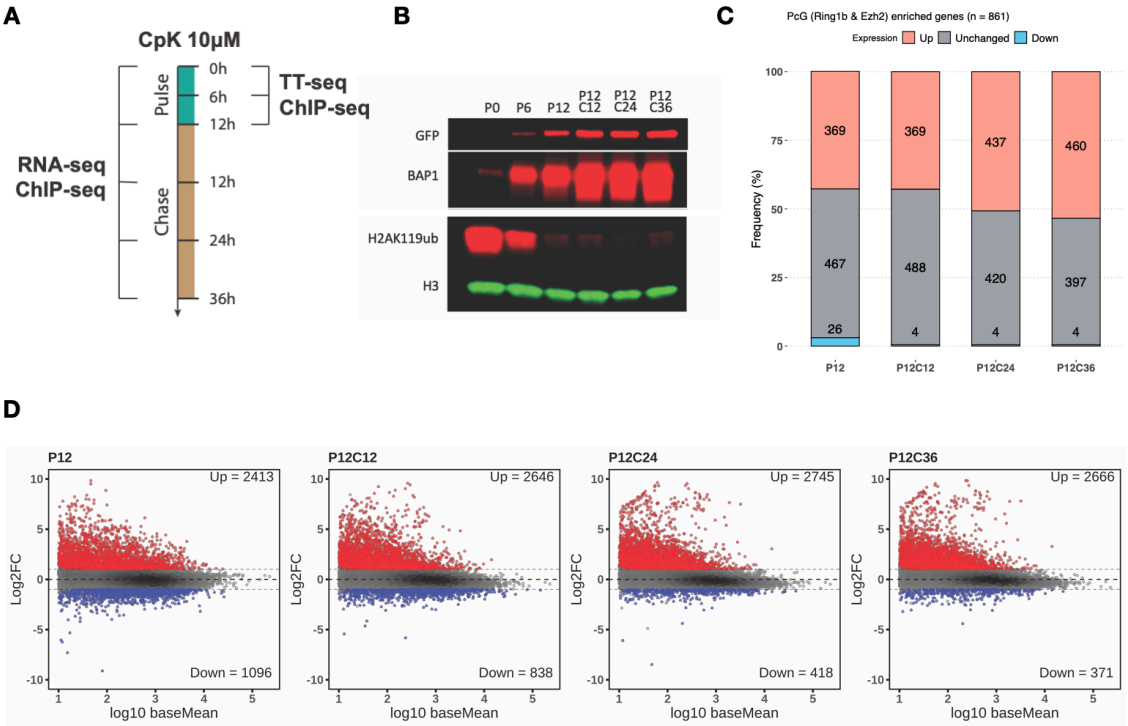

Figure 5.14 BAP1 pulse expression depletes endogenous H2Aub and reverses PcG-mediated repression. (A) Experimental design after BAP1 induction with CpK (Nε-(1-methylcycloprop-2-enecarboxamido)-lysine), after 12 hours pulse and 36 hours chase with two batches of sequencing readouts, batch 1 (right) and batch 2 (left). (B) Western blots of GFP, BAP1, H2AK119ub, and total histone H3 in the pulse-chase periods. (C) PcG enriched genes (intersection of Ring1b and Ezh2 enriched genes) are labeled for the gene responses in the BAP1 pulse-chase RNA-seq series. (D) RNA-seq MA plots show the differential gene expression at the indicated time points.

The previous studies designed with H2Aub writer PRC1 (Polycomb repressive complex 1) conditional knock-out (CKO) or Ring1b catalytic-null mutants inevitably affect the assembly of Polycomb group proteins, confounding with repressive effect from PRC1-mediated chromosomal compaction[92,104,105]. Due to Ring1b catalytic-null mutants attenuated in chromatin binding, whether gene de-repression requires PRC1 relocation is unknown. With BAP1 mRNA microinjection, a recent study successfully depleted H2Aub in the early mouse embryo and de-repressed half of PcG target genes[110]. To record H2Aub depletion responses (Figure 5.1.4 A), we expressed the BAP1-complex with the amber-suppression scaffold PylRS-tRNA^{Pyl} (**Paper III**, Fig 1A). Ectopic BAP1 rapidly removed endogenous H2Aub, and stably de-repressed half of PcG enriched genes in the pulse-chase periods (Figure 5.1.4 B-D). So this method provides a platform for evaluating H2Aub in transcription repression.

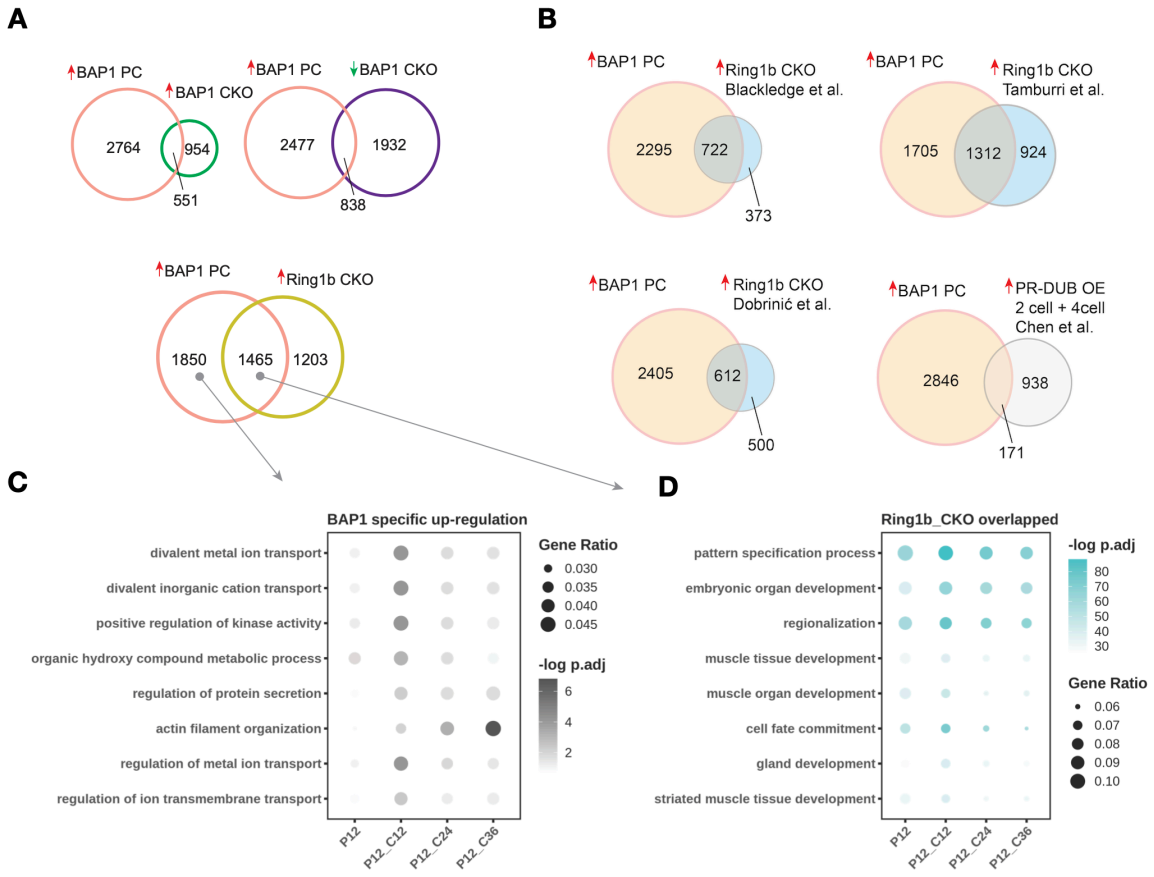

Figure 5.15 Differential expression gene set analysis. (A) Venn diagrams of the BAP1 pulse-chase up-regulated genes intersection with the combined gene list of BAP1 CKO up and down-regulated genes[107,120,148], and the combined PRC1 CKO up-regulated genes[92,105]. (B) Venn diagrams BAP1 pulse-chase up-regulated genes intersection with 3 days PRC1 CKO[92,105], 8 hours PRC1 CKO[104], and BAP1 mRNA over-expression in pre-implantation mouse embryos[110]. (C-D) Top gene ontology terms for BAP1 pulse-chase up-regulated genes specific to the ectopic BAP1 expression (C) and shared with Ring1b CKO (D), which suggest that BAP1 induces both cytoplasmic Ca^{2+} flux and H2Aub depletion.

Interestingly, H2Aub direct depletion appeared to have higher convergence with PRC1 CKO responses from H2Aub passive depletion rather than with the BAP1 CKO that passively accumulates H2Aub (Figure 5.15 A). In the early PRC1 CKO, gene responses were less reproducible with H2Aub depletion (Figure 5.15 B). H2Aub direct depletion in the early mouse embryos also showed cell-type specific responses. Of note, BAP1 pulse expression in mESC induced two parallel pathways, IP3R3 de-ubiquitinating induced Ca^{2+} signaling[149] and H2Aub repressed developmental pathways (Figure 5.15 C-D). Together, these data consolidate the critical role of H2Aub in Polycomb-mediated transcription repression.

5.3.2 H2Aub is required for Polycomb assembly

Non-catalytic Ring1b mutant leaves a question of whether H2Aub is required for PRC1 chromatin binding[92,105]. Therefore we examined the Polycomb factors genome occupancy after H2Aub depletion. Consistently, H2Aub removal caused the loss of Polycomb factors (Ring1b and Ezh2) enrichment on their targets, which emerged from P12 and remained stable decreases (Figure 5.15 A). In addition, Rybp showed a minimal decline, confirming cPRC1 and H2Aub association as previously reported[105]. Moreover, H3K27me3 manifested

a passive dilution due to cell division on PcG-enriched genes. However, at the global level, Polycomb factors exhibited stable genome-wide occupancy, implying Polycomb factors' loss of assembly but not the elimination of chromatin binding due to H2Aub depletion.

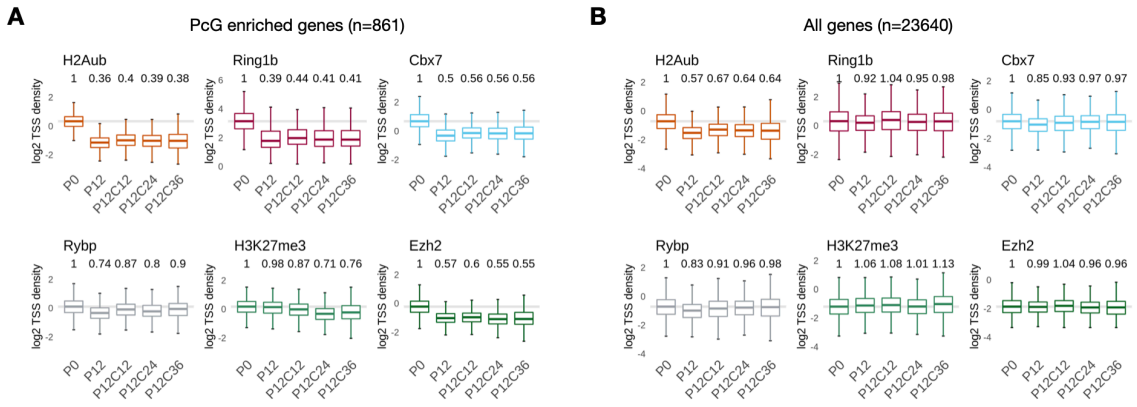

Figure 5.15 Polycomb domain disassembly after H2Aub depletion. (A) Boxplots of log_2-transformed reads density in ±1 kb TSS regions of PcG enriched genes, along the BAP1 pulse-chase series. Relative median ratios to the control are indicated. (B) Boxplots of log_2-transformed reads density in ±1 kb TSS regions of all genes. The same dataset and style as A.

5.4 PROJECT FOUR: CUT&TAG MAPS G-QUADRUPLEX IN MOUSE ES CELLS

G-quadruplex (G4) forms a secondary nucleic acid structure associated with gene expression regulation[150,151]. G4s have dynamic nature and sequence-dependent antigenic profile; the endogenous detection is often challenging even though many G4-specific antibodies have been developed[152–156]. Previous endogenous genome-wide G4 capturing methods by BG4 antibody and G4 probe were performed with formaldehyde cross-linked cells[157,158], which left a caveat of masking or denaturing G4 profiles. To solve these issues, CUT&Tag method was applied to map native G4 structures and achieved highly sensitive readouts[151,159].

In this study, BG4 antibody was used as the primary antibody in CUT&Tag for G4 mapping in mESC, following the procedure as described before[160]. As expected, most G4 CUT&Tag peaks formed on the gene promoters with considerable overlap with the canonical G4 sequence pattern (Figure 5.16 A-C). Although most bona fide canonical G4s appeared at gene TSSs, surprisingly, G4 stabilizer PDS immediately inhibited transcription elongation only after 5 minutes treatment (Figure 5.16 D).

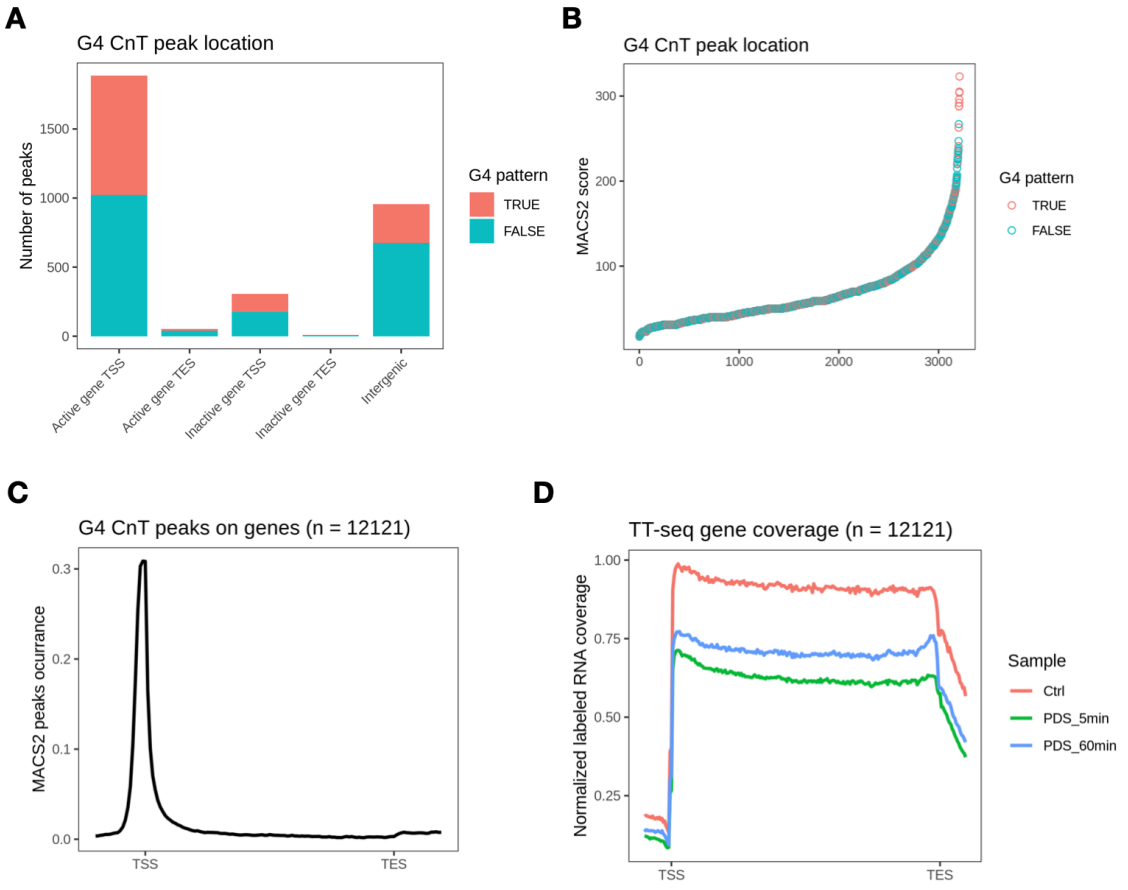

Figure 5.16 Most G4s occur at promoters. (A) mESC G4 CUT&Tag MACS2 peaks are annotated with gene references and G4 sequence patterns. (B) The rank of MACS2 scores of G4 CUT&Tag peaks is color-labeled by G4 sequence pattern matches. (C) MACS2 called G4 peaks coverage on the gene regions. (D) TT-seq spike-in normalized labeled RNA coverage before and after PDS (50 μM) treatments.

However, G4 is ambiguous in transcriptional regulation. Due to the high CG content at the promoter regions, G4 formation has high chances with G-rich single-strand sequences of melted DNA double helix at active gene promoters[161]. But it is more likely that G4 stabilization by PDS limits RNA Pol II elongation efficiency as TT-seq shows (Figure 5.16D), besides a recent study claimed that TMPyP4 G4 stabilizer precludes initiation[151]. Intriguingly, transcription inhibition does not decrease G4 formation[159]. So the stable promoter G4s is unlikely sensitive or linked to downstream gene transcription. Given that G4 can also form with RNA structures that recognized by PRC2, RNA G4 mediates PRC2 enzymatic inhibition and decreases H3K27me3 with nascent RNA precursing[162–164]. So future study might need to distinguish DNA and RNA G4 with either short-term local-response or long-term transcriptional memory across cell development stages.

6 DISCUSSION

6.1 TRANSCRIPTION KINETICS WITH A MULTI-OMICS APPROACH

6.1.1 Transcription frequency measurement

As explained earlier (2.1.3.1), transcription frequency represents the number of full-length RNA synthesis per unit of time. To this end, nascent RNA metabolic labeling is required to exclude the pre-existing RNA fraction, otherwise, the unlabeled nascent RNA will carry a bias from elongation velocity[22] (Figure 6.1). Hence, a flat gene body coverage and high intronic reads density are benchmarks of *bona fide* transcription output[38], as described in **Paper I** (Figure EV2 B-C). In addition, TT-seq can keep cross-contamination rates below 1% (Figure 5.4 C), which again ensures accurate transcription frequency estimation.

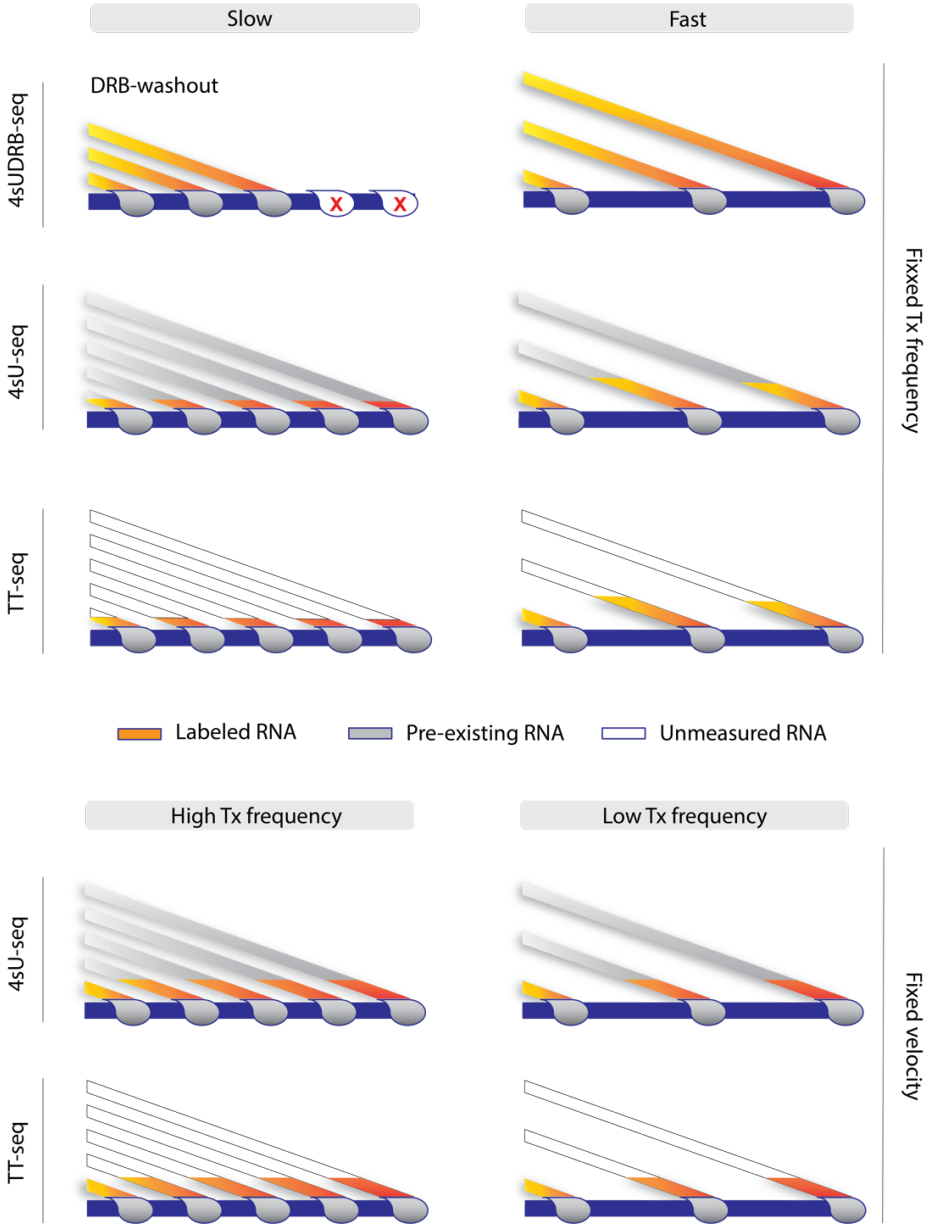

Figure 6.1 RNA fragmentation assists transcription frequency estimation. Elongation rate (kb/min) can be estimated with the pause-release inhibition (DRB) and washout approach. But a previous study using 4sUDRB-seq found that the nascent RNA reads gene body distribution is subject to

both transcription velocity and initiation frequency[22]. As the authors observed a skewed coverage of labeled reads at gene 5' end. They also found if transcription frequency is fixed, the slope is negatively associated with elongation velocity (Figure 6.1, top). Later Bru-seq recapitulated this skewed reads distribution with the same experimental design[165]. Also, a sloped coverage can occur with Pol II-associated RNAs in a labeling-free nascent RNA-seq approach[32]. After RNA fragmentation, labeled RNA reads coverage will anchor to transcription frequency for any elongation velocity (Figure 6.1, top), therefore providing a uniform mapping of transient transcription (Figure 6.1, bottom).

6.1.2 Limits of transcription velocity estimation

An accurate velocity estimation in TSS pausing window is challenging, which requires both RNA synthesis and Pol II occupancy at base-pair resolution. Compared to RNA synthesis, Pol II TSS occupancy is more obscure, since miscellaneous states could exist in the short pausing window. A study with exonuclease reveals the Pol II stall footprint between 20-51 nt and can slide back along DNA template[166]. With MNase (micrococcal nuclease), a study refines mNET-seq by preliminary size-selection of captured RNAs and discovers that Pol II has a 20-30 nt footprint from TSS for the short RNA library (20-60 nt), but in the longer RNA library (60-160 nt) Pol II shows ~60 nt footprints without pausing peaks[167]. Since a nascent RNA has to be 15 nt long to reach the Pol II surface and 23 nt long to be protected by the capping enzymes from MNase digestion[168], the nucleotide oligo mixtures inside Pol II imply disparate fates of early RNA production. Therefore, Pol II S5p and mNET-seq may measure different mixtures of Pol II TSS activities, although the estimated gene body velocities from TT-seq/mNET-seq align with TT-seq/Pol II-S5p estimates (**Paper I**, Fig EV4 C).

Promoter-proximal pausing might not be simply stationary, as non-productive Pol II exists. Interfering pause-release with P-TEFb inhibitors prohibits new rounds of Pol II initiation[169] and increases Pol II TSS occupancy[69,170,171]. However, both the short RNA production in PRO-seq[75] and the fraction of paused Pol II in "methyltransferase footprinting"[172] are stable after pause-release inhibition. Accordingly, a short-capped RNA 4sU labeled method reports non-productive turnover of early RNAs from the paused Pol II, in a median half-life of 5 minutes and an average of ~80% premature termination rates[173]. Hypothetically, the early elongation stage may have many flimsy steps that can be distinguished by different transcription measurements (Figure 6.2). The similar TSS peaks in Pol II S5p ChIP, PRO-seq, and mNET-seq actually represent a series of decisions before a full-length transcription.

TT-seq profile is absent of TSS peak because promoter high GC content disfavors 4sU labeling in addition to the removal of 5' end pre-existing RNA fraction. Plus the isopropanol precipitation after biotinylation in TT-seq might also filter out the short paused RNA (<50 nt), as benchmarked for micro-RNA recovery[174]. These features could make TT-seq devoid of premature terminated RNA. Since TSS short RNAs are readily lost after Triptolide-mediated initiation block, again suggesting a fast turnover of the non-productive nucleotides from the paused Pol II[172].

In addition to pause-release dynamics, the low resolution of estimated local velocity may compromise the explanation of other local transcription mechanisms, for example, backtracking. Transcription can spontaneously discontinue when confronted with elongation obstacles and backtrack a few base-pair[175,176]. Backtracked Pol II generates a peak pattern in the gene body in NET-seq[31]. To alleviate backtracking and stimulate elongation, RNA folds into secondary structures[62], requires the cleavage of the short

flanking 3' end by Pol II complex subunit TFIIS[177], and with the aids of elongation factors (PAF1, RTF1, and STP6) to pass nucleosomal roadblocks[51,178,179]. Of note, the DNA melting energy per se can predict Pol II backtracks[180]. TT-seq labeled RNA coverages frequently decline at the high GC content loci throughout gene body and termination window. So a base-pair precision evaluation of the backtracking is challenging with our MNase digested Pol II ChIP, but it might be feasible with the published NET-seq and TT-seq datasets[51,73,137,181]. Nevertheless, Pol II backtracking frequently appears near the pause-release check-point[165], to what extent backtracking participates in the termination slowdown is unclear. Hence, the short reads sequencing is limited to address whether RNA cleavage in the termination window accompanies by a discontinuous elongation velocity. This question could be answered by the long-read nascent RNA sequencing methods[138,182], which might unveil the termination cleavage site at base-pair resolution.

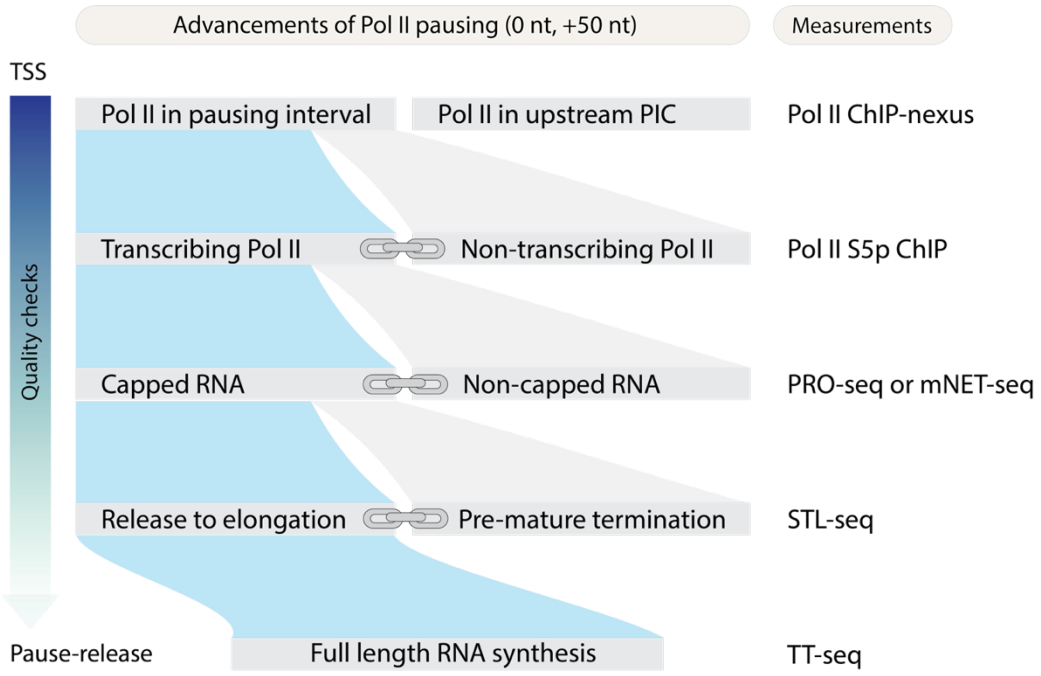

Figure 6.2 Pol II pause-release steps revealed in different sequencing methods. A high-resolution Pol II ChIP-nexus allows for dissecting the initiation and pausing engagements[169]. Pol II S5p captures the initiated Pol II, but might inevitably contain physically stalled non-productive Pol II. In support, RNA synthesis in the pausing window can decrease with stable Pol II occupancy by backtracking enforcement[165]. PRO-seq and mNET-seq may capture non-capped early RNAs. Even after capped RNA enrichment, a fast turnover nature of paused Pol II might contribute a considerable fraction of non-elongated transcripts as STL-seq reveals[173]. While the pre-mature terminated RNA is neglectable in TT-seq.

Albeit elongation velocity has not been widely recognized as a rate-limiting factor for gene expression, it can respond to cell-intrinsic and extrinsic signals. An appropriate transcription velocity provides RNA binding factors "window opportunity," mediating the exon junction usage during co-transcription splicing[183], which might be valuable for the alternative splicing study.

6.1.3 Non-steady-state RNA turnover in a living cell

For simplicity, the steady-state hypothesis is widely applied for RNA turnover estimation[25,184,185]. But verification of it is rare in publications. In the TT-seq mESC dataset, we observed highly variable cellular RNA abundance of mESC biological replicates.

Although the assumptions of short pulse metabolic labeling are compatible with the steady-state hypothesis (section 4.3.6.2), a recent study found that only half of the genes can meet steady-states in Hela cells up to 12 hours of RNA metabolic labeling[186].

RNA degradation rate is presumably a constant in mESC, as the agreement between SLAM-seq and TimeLapse-seq half-life units. Moreover, total RNA abundance is anchored with cell size to achieve mRNA concentration homeostasis[187]. So in our mESC, RNA synthesis is expected to be larger than total RNA decay and pushes total RNA in a non-steady-state (Figure 6.3 A). This merits cellular RNA to accumulate before a subsequent cell division (Figure 6.3 B). More importantly, it also allows transcription to shape the non-steady-state total RNA pool faster in response to environmental stimuli.

So far, many single-cell RNA-seq methods has been adapted with 4sU labeling (scNT-seq[188] / scSLAM-seq[189] / sci-fate[190] for 2 hours, NASC-seq[191] for 30 minutes), and with 5-ethynyluridine labeling (scEU-seq[192] for 2 hours). Importantly, scEU-seq pulse-chase experiment shows that the non-steady-state assumption can achieve a better fitting of synthesis and degradation rates[192], which again instantiates a possible non-steady-state turnover at single-cell level.

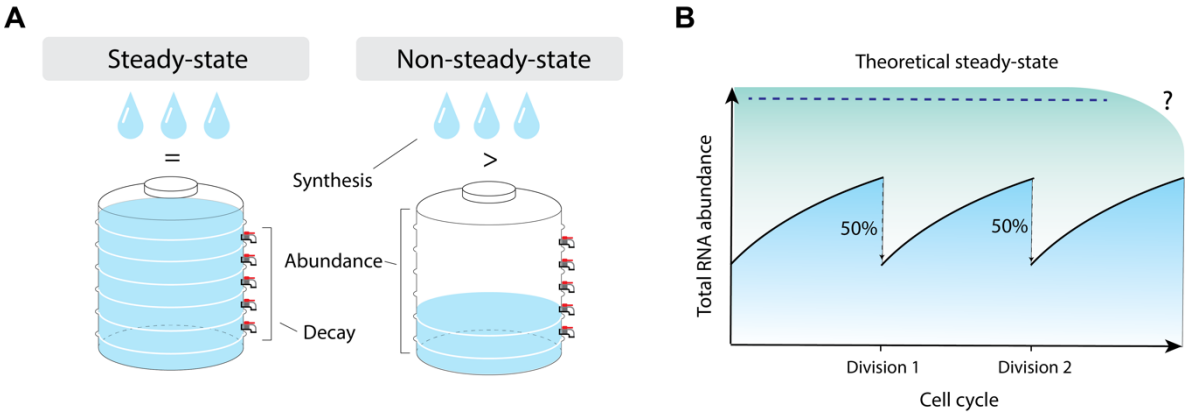

Figure 6.3 Turnover strategy in a proliferating cell. (A) For an individual gene or a cell under the state-state, total RNA degradation equals RNA synthesis. In contrast, a smaller pool of total RNA with the same synthesis rate is not only able to accumulate total RNA abundance but also ready to respond to environmental cues. (B) With cell cycle progression, every division splits 50% of total RNA. To sustain cellular RNA abundance, the populational average of growing cells may retain total RNA beneath the theoretical steady-state. But the ceiling capacity may decrease when RNA synthesis attenuates in high cell confluence or serum starvation, manifesting a cell volume shrinkage.

The pulse labeling estimated turnover kinetics has many drawbacks. First, the RNA turnover represents only the current tendency to replace pre-existing RNA, which is specific to the cell culture condition (Figure 6.3 B, Figure 2.4). Second, the labeling time must be short enough to avoid nascent RNA loss before post-transcriptional processing. A 5 minutes 4sU labeling increases the requirement of cell number, especially low cost-effective for weakly transcribed cell types. Third, the turnover estimation of extremely unstable RNA species requires a deep sequencing of both newly synthesized RNA and total RNA to assist TU annotation and reads quantification precision.

6.2 TRANSCRIPTIONAL RESPONSE TO REPRESSIVE HISTONE MODIFICATION LOSS

Transcription activation can also obtain from repressive histone marks ablation. In this thesis, H2Aub direct depletion broadly activates Polycomb repressed genes that as observed in Ring1b CKO[92,104,105]. The consequential loss of cPRC1-PRC2.2 aggregation occurs after H2Aub depletion suggests that Polycomb disassembly is a prerequisite for every gene de-repression. On contrary, several Hox genes up-regulation accompany by minimal Ring1b binding changes. So H2Aub direct depletion can unveil H2Aub-mediated Polycomb integrity from many aspects.

First, H2Aub is associated with nucleosome compaction[113] and precludes FACT (FAcilitates Chromatin Transcription) mediated nucleosome disassembly[193]. Efficient transcription elongation needs to overcome nucleosome barriers[51]. So the loss of bulky ubiquitin on H2A could facilitate nucleosome reassembly and remodeling during transcription.

Second, most de-repressed Polycomb target genes can maintain up-regulation through BAP1 pulse-chase periods in response to global H2Aub decrease (Figure 5.14 B-D). In contrast, BAP1 pulse expression elicits Ca^{2+} signaling and transiently up-regulates non-PcG genes. The Ca^{2+}-induced gene expression disturbance readily disappeared in BAP1 chase periods. Also, the cellular RNA abundance showed a ~25% decrease at BAP1 12 hours pulse and recovered after 12 hours chase. Albeit BAP1 pulse expression stressed the global transcription, a few Polycomb target genes increased in elongation velocity (**Paper III**, Figure S8B).

Third, Polycomb factors relocation may self-reinforce via nucleosome decompaction initiated transcription activation. Since transcription machinery can evict PRC2 chromatin binding[194,195], but not H2Aub. So H2Aub initiates transcriptional response upstream to Polycomb disassociation, as well as nucleosome decompaction (Figure 6.4). In support, H3K27me3 depletion fails to establish activate transcription, even if it is required for long-range chromatin interaction and nucleosome compaction. Therefore, H3K27me3 is downstream to both Polycomb occupancy and transcription impermissive Polycomb-associated nucleosomes.

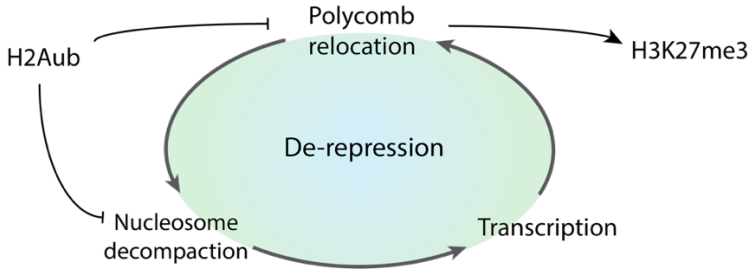

Figure 6.4 Hypothetical diagram of H2Aub-mediated gene repression.

Last but not least, H2Aub and H3K27me3 deficiency have cell-type-specific gene responses. For example, drosophila requires H2Aub writer (Sce) in embryogenesis but not H2Aub per se in epidermal differentiation[103]. In embryonic stem cell development, H3K27me3 depletion shows a more profound impact on the gene expression program in human pluripotency than in mouse background[196]. Hence, our observation supports a pivotal role of H2Aub in Polycomb-mediated immediate transcription repression, which is critical for mESC pluripotency and self-renewal[91].

7 CONCLUSIONS

In summary, TT-seq maps the newly synthesized RNA in mESCs (**Paper I**). With a careful test of 4sU labeling efficiency, the inhibitor-induced pluripotent states have lower transcription frequencies than the serum-naïve state. The SL-2i transition has a widespread gene differential expression compared to the homogenous decrease in mTORi cells. Moreover, the gene neighbor co-regulation with adjacent intergenic ncRNAs shows a power law decrease with distance. This neighbor co-expression mechanism is stably hardwired for both gene-ncRNA and gene-gene pairs in the SL-2i transition. With the new spike-in standards, we estimate RNA synthesis and turnover rates and hypothesize that total RNA abundance is in a non-steady-state for a growing cell. Transcription velocity can be estimated with transcription frequency and Pol II occupancy. And the shorter termination distances in 2i and mTORi cells instantiate the decreases in termination velocity.

Using TT-seq, we also found that G4 stabilization inhibits transcription elongation, although most G4s form at gene promoter regions (**Paper II**). The new CUT&Tag method established by Jing Lyu performs with native chromatin, recovers G4 localization with melted DNA duplex, and enriches higher signal to noise than cross-linked G4 mapping methods.

In addition, H2A lysine 119 mono-ubiquitination is pivotal to the Polycomb-mediated repression in mESCs. The direct depletion of H2Aub by its de-ubiquitinase BAP1 induces Polycomb target genes activation measured by TT-seq (**Paper III**). The mechanism behind the H2Aub-mediated repression is hypothetically arising from Polycomb-associated nucleosome decompactions that facilitate Pol II crossing nucleosomes.

8 POINTS OF PERSPECTIVE

The projects included in this thesis focus on transcription measurement in many scenarios, mouse embryonic stem cell pluripotent states, histone modifications, and G4 quadruplexes. The new biological insights benefit from improvements in breadth and depth, and experimental/analytical integration. Owing to the popularity of mESC, this thesis can investigate transcription regulation in a concert of multi-omics data.

First, the priority is integrating homemade sequencing data with correct and reproducible workflows. This requirement will appear from the experimental design step. For example, MINUTE-ChIP allows dozens of protein targets to be profiled in a couple of conditions with technical replicates. The epigenomic features and transcription machinery were profiled altogether during the establishment of the MINUTE-ChIP method. The original plan did not include the velocity integration with TT-seq, but it is graceful for one colleague's (Banushree) sharing of her data. So a careful experimental design would save many efforts in the downstream analysis.

Moreover, the integration with public data for method benchmarking and cross-validation is increasingly necessary. Over the past decade, published datasets explosively grew with new sequencing methods, since sequencing became a low-cost experimental readout. So making a sufficient claim will inevitably require re-analysis and cross-validation of the published data. Above this, meta-analysis provides bird-eye's views upon a domain of knowledge, and the data integration work can deserve an independent project if the cost is beyond the budget. So a small exploration stage often detours before finishing a specific task which is to answer a biological question while collecting evidence for ten new questions. A specific and novel question can circumvent the dimensional curse that leads to high repetition and costs. For example, predicting mRNA half-life from RNA sequence and epigenetic features might be achievable by rebuilding a wheel with deep learning methods[197–199], but the cost of data collection and method adaptation is overkill for a descriptive task of transcription kinetics in mESC pluripotent states.

Finally, the divergence of techniques, rather than convergence to a universal standard, expand the scope of knowledge. If data integration is a dimension reduction process, developing new methods is the birthplace of dimensionality. For instance, single-cell sequencing methods introduce many dimensions, by cell, allele, and spatial position, to diversify measurements of a particular gene expression. Bulk sequencing lacks cell information, but specifying a single read can also achieve a dimension on single molecule precision. As "methyltransferase footprinting" can describe populational TSS architecture from individual Pol II occupancies. In the future, a "master technology" might not win out and abolish the rest. But ascribing to the development of methods, the enlarging boundaries of knowledge depict impossibility that aids the formation of complexity, as an emergent universe with an obscure number of dimensions. Stephen Hawking described in *The Grand Design*, "The Feynman sum allows for all of these, for every possible history for the universe, but the observation that our universe has three large space dimensions selects out the subclass of histories that have the property that is being observed." Perhaps a hundred years later, RNA-seq becomes an unfamiliar technique, while every cell in the body has an ID.

9 ACKNOWLEDGEMENTS

“Life has a limit, but knowledge has no edge --- Chuang Tzu.”

This work has been practiced in Science for Life Laboratory (Scilifelab) and Division of Genome Biology, Department of Biochemistry and Biophysics, Karolinska Institutet. I am grateful for the platform in the scilifelab and the doctoral education at Karolinska Institutet.

I would like to thank my supervisor, Simon Elsässer, who provides me with the research opportunity technically and financially resourceful in the well-equipped laboratory and the friendly scientific community. I cannot remember how many times I discussed the technical details with you, but I learned from you that patience and decisiveness often lead to good results. In the first paper revision, your generosity in time made the scattered sentences and figures into a scientific article. Also, the critical reading of this thesis owes you many thanks. Your collaboration also made me meet my co-supervisors.

I would like to thank my co-supervisors, Michael Lidschreiber and Patrick Cramer, who introduced me to the TT-seq method. I can recall the first time when I came to the Novum building to learn about the TT-seq experiment in the winter of 2016. The technical issues were readily solved with the kindness of Katja Lidschreiber’s help. I keep the memory of Stockholm’s spring when Michael and Patrick came right off the plane to join my half-time seminar and review the TT-seq results. Michael, your generosity in sharing the TT-seq scripts helped me quickly gain the ability to do transcriptomic analysis. And Patrick, your revision of the manuscript helped to formalize my casual writing into an academic tone.

I also would like to thank the Swedish Bioinformatic Advisory Program where I met my mentor, Jakub Westholm. With your support, I have a chance of updating with a broader bioinformatics community, know many friends, and play the games organized by Björn Nystedt.

I would like to thank the members of my thesis committee, Prof. Dr. Eckardt Treuter, Dr. Qi Dai, and Dr. Peter Svensson, my opponent Dr. Anniina Vihervaara, and the chairperson Dr. Mikael Lindström for sharing your time and expertise in the defense.

And I would like to thank the Elsässer group members, for spending a good time together, not only on the office or bench or Uppmax server, but also in the Djurgården, in the Ballbreaker, in Chania-Greece, in the Laserdome, in the Escape Room, and at many places with beers. The memorable time we also spent together in the Alpha 5 meeting room, in the gene expression journal club, in the Epichrome conference, in the Solvik retreat, and in the floor’s kitchen ☺.

Finally, I would like to thank my parents, your support in spirit and living that help me spend the time pursuing a doctorate in Sweden. “Love is the one thing we're capable of perceiving that transcends dimensions of time and space --- *Interstellar*.”

10 REFERENCES

1. Wang, L. *et al.* Resetting the epigenetic balance of Polycomb and COMPASS function at enhancers for cancer therapy. *Nat. Med.* **24**, 758–769 (2018).
2. Douillet, D. *et al.* Uncoupling histone H3K4 trimethylation from developmental gene expression via an equilibrium of COMPASS, Polycomb and DNA methylation. *Nat. Genet.* **52**, 615–625 (2020).
3. Wang, Z. *et al.* Prediction of histone post-translational modification patterns based on nascent transcription data. *Nat. Genet.* **54**, (2022).
4. Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**, 295–301 (2012).
5. Lin, J. J. *et al.* Mediator coordinates PIC assembly with recruitment of CHD1. *Genes Dev.* **25**, 2198–2209 (2011).
6. Hsin, J. P. & Manley, J. L. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev.* **26**, 2119–2137 (2012).
7. Wong, K. H., Jin, Y. & Struhl, K. TFIIH Phosphorylation of the Pol II CTD Stimulates Mediator Dissociation from the Preinitiation Complex and Promoter Escape. *Mol. Cell* **54**, 601–612 (2014).
8. 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).
9. Core, L. & Adelman, K. Promoter-proximal pausing of RNA polymerase II: A nexus of gene regulation. *Genes Dev.* **33**, 960–982 (2019).
10. Ramanathan, A., Robb, G. B. & Chan, S. H. mRNA capping: Biological functions and applications. *Nucleic Acids Res.* **44**, 7511–7526 (2016).
11. Lu, H. *et al.* Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature* **558**, 318–323 (2018).
12. Shao, W. *et al.* Phase separation of RNA-binding protein promotes polymerase binding and transcription. *Nat. Chem. Biol.* **18**, 70–80 (2022).
13. 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).
14. Eaton, J. D. *et al.* Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity. *Genes Dev.* **32**, 127–139 (2018).
15. Sutherland, H. & Bickmore, W. A. Transcription factories: gene expression in unions? *Nat. Rev. Genet.* **10**, 457–466 (2009).
16. 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.* **115**, E4368–E4376 (2018).
17. Patrone, G. *et al.* Nuclear Run-On Assay Using Biotin Labeling, Magnetic Bead Capture and Analysis by Fluorescence-Based RT-PCR. *Biotechniques* **29**, 1012–1017 (2000).
18. 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).
19. 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).
20. Eser, P. *et al.* Determinants of RNA metabolism in the Schizosaccharomyces pombe genome . *Mol. Syst. Biol.* **12**, 857 (2016).
21. Tani, H. & Akimitsu, N. Genome-wide technology for determining RNA stability in mammalian cells. *RNA Biol.* **9**, 1233–1238 (2012).
22. Fuchs, G. *et al.* 4sUDRB-seq: measuring genomewide transcriptional elongation

rates and initiation frequencies within cells. *Genome Biol.* **15**, R69 (2014).
23. Herzog, V. A. *et al.* Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods* **14**, 1198–1204 (2017).
24. Muhar, M. *et al.* SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis. *Science* **360**, 800–805 (2018).
25. Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C. & Simon, M. D. TimeLapse-seq: Adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat. Methods* **15**, 221–225 (2018).
26. Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352**, 1225–1228 (2016).
27. Boileau, E., Altmüller, J., Vries, I. S. N. & Dieterich, C. A comparison of metabolic labeling and statistical methods to infer genome-wide dynamics of RNA turnover. *Briefings in Bioinformatics* **00**, 1–14 (2021).
28. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
29. Paulsen, M. T. *et al.* Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. *Methods* **67**, 45–54 (2014).
30. Herzog, V. A. *et al.* Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods* **14**, 1198–1204 (2017).
31. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368–373 (2011).
32. Sousa-Luís, R. *et al.* POINT technology illuminates the processing of polymerase-associated intact nascent transcripts. *Mol. Cell* **81**, 1935-1950.e6 (2021).
33. Roberts, T. C. *et al.* Quantification of nascent transcription by bromouridine immunocapture nuclear run-on RT-qPCR. *Nat. Protoc.* **10**, 1198–1211 (2015).
34. Lugowski, A., Nicholson, B. & Rissland, O. S. DRUID: A pipeline for transcriptome-wide measurements of mRNA stability. *Rna* **24**, 623–632 (2018).
35. Shao, R. *et al.* Distinct transcription kinetics of pluripotent cell states. *Mol. Syst. Biol.* **18**, 1–19 (2022).
36. Dukler, N. *et al.* Nascent RNA sequencing reveals a dynamic global transcriptional response at genes and enhancers to the natural medicinal compound celastrol. *Genome Res.* **27**, 1816–1829 (2017).
37. Judd, J. *et al.* A rapid, sensitive, scalable method for Precision Run-On sequencing (PRO-seq). *bioRxiv* 2020.05.18.102277 (2020) doi:10.1101/2020.05.18.102277.
38. Barbieri, E. *et al.* Rapid and Scalable Profiling of Nascent RNA with fastGRO. *Cell Rep.* **33**, (2020).
39. Duffy, E. E., Canzio, D., Maniatis, T. & Simon, M. D. Solid phase chemistry to covalently and reversibly capture thiolated RNA. *Nucleic Acids Res.* **46**, 6996–7005 (2018).
40. Efroni, S. *et al.* Global Transcription in Pluripotent Embryonic Stem Cells. *Stem Cell* **2**, 437–447 (2008).
41. Team, T. R. G. E. R. G. P. I. I. & Consortium, T. F. Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–690 (2001).
42. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
43. Andersson, R. *et al.* Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nat. Commun.* **5**, 5336 (2014).
44. Statello, L., Guo, C. J., Chen, L. L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).
45. Bell, J. C. *et al.* Chromatin-associated RNA sequencing (ChAR-seq) maps genome-wide RNA-to-DNA contacts. *Elife* **7**, 1–28 (2018).

46. Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat. Methods* **12**, 433–438 (2015).
47. 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).
48. Nagari, A., Murakami, S., Malladi, V. S. & Kraus, W. L. Computational Approaches for Mining GRO-Seq Data to Identify and Characterize Active Enhancers. *Methods Mol. Biol.* **1468**, 121–138 (2017).
49. Wang, Z., Chu, T., Choate, L. A. & Danko, C. G. Identification of regulatory elements from nascent transcription using dREG. *Genome Res.* **29**, 293–303 (2019).
50. Choi, J. *et al.* Evidence for additive and synergistic action of Mammalian enhancers during cell fate determination. *Elife* **10**, 1–27 (2021).
51. Žumer, K. *et al.* Two distinct mechanisms of RNA polymerase II elongation stimulation in vivo. *Mol. Cell* **81**, 3096-3109.e8 (2021).
52. Schwalb, B. *et al.* TT-seq maps the human transient transcriptome. *Science* **352**, 1225–1228 (2016).
53. Engreitz, J. M. *et al.* Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* **539**, 452–455 (2016).
54. Michel, M. *et al.* TT-seq captures enhancer landscapes immediately after T-cell stimulation. *Mol. Syst. Biol.* **13**, 920 (2017).
55. Zuin, J. *et al.* Nonlinear control of transcription through enhancer–promoter interactions. *Nature* **604**, (2022).
56. Lidschreiber, K. *et al.* Transcriptionally active enhancers in human cancer cells. *Mol. Syst. Biol.* **17**, e9873 (2021).
57. Arnold, P. R., Wells, A. D. & Li, X. C. Diversity and Emerging Roles of Enhancer RNA in Regulation of Gene Expression and Cell Fate. *Front. Cell Dev. Biol.* **7**, 1–14 (2020).
58. Peng, T. *et al.* STARR-seq identifies active, chromatin-masked, and dormant enhancers in pluripotent mouse embryonic stem cells. *Genome Biol.* **21**, 243 (2020).
59. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci.* **107**, 21931–21936 (2010).
60. Crispatzu, G. *et al.* The chromatin, topological and regulatory properties of pluripotency-associated poised enhancers are conserved in vivo. *Nat. Commun.* **12**, 1–17 (2021).
61. Cruz-Molina, S. *et al.* PRC2 Facilitates the Regulatory Topology Required for Poised Enhancer Function during Pluripotent Stem Cell Differentiation. *Cell Stem Cell* **20**, 689-705.e9 (2017).
62. Turowski, T. W. *et al.* Nascent Transcript Folding Plays a Major Role in Determining RNA Polymerase Elongation Rates. *Mol. Cell* **79**, 488-503.e11 (2020).
63. Larsson, A. J. M. *et al.* Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).
64. Peccoud, J. & Ycart, B. Markovian Modeling of Gene-Product Synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).
65. Gajos, M. *et al.* Conserved DNA sequence features underlie pervasive RNA polymerase pausing. *Nucleic Acids Res.* **49**, 4402–4420 (2021).
66. Beckedorff, F. *et al.* The Human Integrator Complex Facilitates Transcriptional Elongation by Endonucleolytic Cleavage of Nascent Transcripts. *Cell Rep.* **32**, 107917 (2020).
67. Tatomer, D. C. *et al.* The Integrator complex cleaves nascent mRNAs to attenuate transcription. *Genes Dev.* **33**, 1525–1538 (2019).
68. Cho, W. K. *et al.* RNA Polymerase II cluster dynamics predict mRNA output in living cells. *Elife* **5**, 1–31 (2016).

69. 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–e02407 (2014).
70. Sanchez, G. J. *et al.* Genome-wide dose-dependent inhibition of histone deacetylases studies reveal their roles in enhancer remodeling and suppression of oncogenic super-enhancers. *Nucleic Acids Res.* **46**, 1756–1776 (2018).
71. Liu, L. *et al.* Transcriptional Pause Release Is a Rate-Limiting Step for Somatic Cell Reprogramming. *Stem Cell* **15**, 574–588 (2014).
72. Krumm, A., Hickey, L. B. & Groudine, M. Promoter-proximal pausing of RNA polymerase II defines a general rate-limiting step after transcription initiation. *Genes Dev.* **9**, 559–572 (1995).
73. Gressel, S., Schwalb, B. & Cramer, P. The pause-initiation limit restricts transcription activation in human cells. *Nat. Commun.* **10**, 3603–3612 (2019).
74. Muniz, L., Nicolas, E. & Trouche, D. RNA polymerase II speed: a key player in controlling and adapting transcriptome composition. *EMBO J.* **40**, 1–21 (2021).
75. Booth, G. T., Parua, P. K., Sansó, M., Fisher, R. P. & Lis, J. T. Cdk9 regulates a promoter-proximal checkpoint to modulate RNA polymerase II elongation rate in fission yeast. *Nat. Commun.* **9**, (2018).
76. Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Publ. Gr.* **16**, 167–177 (2015).
77. Chen, W., Smeekens, J. M. & Wu, R. Systematic study of the dynamics and half-lives of newly synthesized proteins in human cells. *Chem. Sci.* **7**, 1393–1400 (2016).
78. PHILLIPS, D. M. The presence of acetyl groups of histones. *Biochem. J.* **87**, 258–263 (1963).
79. Martin, B. J. E. *et al.* Transcription shapes genome-wide histone acetylation patterns. *Nat. Commun.* **12**, 1–9 (2021).
80. Vaid, R., Wen, J. & Mannervik, M. Release of promoter-proximal paused Pol II in response to histone deacetylase inhibition. *Nucleic Acids Res.* **48**, 4877–4890 (2020).
81. Pelham-Webb, B. *et al.* H3K27ac bookmarking promotes rapid post-mitotic activation of the pluripotent stem cell program without impacting 3D chromatin reorganization. *Mol. Cell* **81**, 1732-1748.e8 (2021).
82. Zhang, T., Zhang, Z., Dong, Q., Xiong, J. & Zhu, B. Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biol.* **21**, 45 (2020).
83. Ferrari, K. J. *et al.* Polycomb-Dependent H3K27me1 and H3K27me2 Regulate Active Transcription and Enhancer Fidelity. *Mol. Cell* **53**, 49–62 (2014).
84. Elsässer, S. J., Ernst, R. J., Walker, O. S. & Chin, J. W. Genetic code expansion in stable cell lines enables encoded chromatin modification. *Nat. Methods* **13**, 158–164 (2016).
85. Sabari, B. R., Zhang, D., Allis, C. D. & Zhao, Y. Metabolic regulation of gene expression through histone acylations. *Nat. Rev. Mol. Cell Biol.* **18**, 90–101 (2017).
86. Wan, W., Tharp, J. M. & Liu, W. R. Pyrrolysyl-tRNA synthetase: An ordinary enzyme but an outstanding genetic code expansion tool. *Biochim. Biophys. Acta - Proteins Proteomics* **1844**, 1059–1070 (2014).
87. Stock, J. K. *et al.* Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. *Nat. Cell Biol.* **9**, 1428–1435 (2007).
88. Jacobs, S. A. & Khorasanizadeh, S. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science* **295**, 2080–2083 (2002).
89. Elsässer, S. J., Noh, K.-M., Diaz, N., Allis, C. D. & Banaszynski, L. A. Histone H3.3 is required for endogenous retroviral element silencing in embryonic stem cells. *Nature* **522**, 240–244 (2015).

90. Lavarone, E., Barbieri, C. M. & Pasini, D. Dissecting the role of H3K27 acetylation and methylation in PRC2 mediated control of cellular identity. *Nat. Commun.* **10**, 1–16 (2019).
91. Zepeda-Martinez, J. A. *et al.* Parallel PRC2/cPRC1 and vPRC1 pathways silence lineage-specific genes and maintain self-renewal in mouse embryonic stem cells. *Sci. Adv.* **6**, 1–16 (2020).
92. Blackledge, N. P. *et al.* PRC1 Catalytic Activity Is Central to Polycomb System Function. *Mol. Cell* **77**, 857-874.e9 (2020).
93. Blackledge, N. P. *et al.* Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. *Cell* **157**, 1445–1459 (2014).
94. Zhao, J. *et al.* RYBP/YAF2-PRC1 complexes and histone H1-dependent chromatin compaction mediate propagation of H2AK119ub1 during cell division. *Nat. Cell Biol.* **22**, 439–452 (2020).
95. Moussa, H. F. *et al.* Canonical PRC1 controls sequence-independent propagation of Polycomb-mediated gene silencing. *Nat. Commun.* **10**, 1–12 (2019).
96. Tavares, L. *et al.* RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3. *Cell* **148**, 664–678 (2012).
97. Gao, Z. *et al.* PCGF Homologs, CBX Proteins, and RYBP Define Functionally Distinct PRC1 Family Complexes. *Mol. Cell* **45**, 344–356 (2012).
98. Healy, E. *et al.* PRC2.1 and PRC2.2 Synergize to Coordinate H3K27 Trimethylation. *Mol. Cell* **76**, 437-452.e6 (2019).
99. Cooper, S. *et al.* Jarid2 binds mono-ubiquitylated H2A lysine 119 to mediate crosstalk between Polycomb complexes PRC1 and PRC2. *Nat. Commun.* **7**, 1–8 (2016).
100. Oksuz, O. *et al.* Capturing the Onset of PRC2-Mediated Repressive Domain Formation. *Mol. Cell* **70**, 1149-1162.e5 (2018).
101. Kasinath, V. *et al.* JARID2 and AEBP2 regulate PRC2 in the presence of H2AK119ub1 and other histone modifications. *Science* **371**, (2021).
102. Kalb, R. *et al.* Histone H2A monoubiquitination promotes histone H3 methylation in Polycomb repression. *Nat. Struct. Mol. Biol.* **21**, 569–571 (2014).
103. Pengelly, A. R., Kalb, R., Finkl, K. & Müller, J. Transcriptional repression by PRC1 in the absence of H2A monoubiquitylation. *Genes Dev.* **29**, 1487–1492 (2015).
104. Dobrinić, P., Szczurek, A. T. & Klose, R. J. PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency. *Nature Structural and Molecular Biology* vol. 28 (Springer US, 2021).
105. Tamburri, S. *et al.* Histone H2AK119 Mono-Ubiquitination Is Essential for Polycomb-Mediated Transcriptional Repression. *Mol. Cell* **77**, 840-856.e5 (2020).
106. Campagne, A. *et al.* BAP1 complex promotes transcription by opposing PRC1-mediated H2A ubiquitylation. *Nat. Commun.* **10**, 1–15 (2019).
107. Kolovos, P. *et al.* PR-DUB maintains the expression of critical genes through FOXK1/2- A nd ASXL1/2/3-dependent recruitment to chromatin and H2AK119ub1 deubiquitination. *Genome Res.* **30**, 1119–1130 (2020).
108. Cooper, S. *et al.* Targeting Polycomb to Pericentric Heterochromatin in Embryonic Stem Cells Reveals a Role for H2AK119u1 in PRC2 Recruitment. *Cell Rep.* **7**, 1456–1470 (2014).
109. Blackledge, N. P. & Klose, R. J. The molecular principles of gene regulation by Polycomb repressive complexes. *Nat. Rev. Mol. Cell Biol.* **0123456789**, (2021).
110. Chen, Z., Djekidel, M. N. & Zhang, Y. Distinct dynamics and functions of H2AK119ub1 and H3K27me3 in mouse preimplantation embryos. *Nat. Genet.* **53**, 551–563 (2021).
111. Hickey, G. J. M. *et al.* Establishment of developmental gene silencing by ordered polycomb complex recruitment in early zebrafish embryos. *Elife* **11**, 1–24 (2022).

112. Ochiai, H. *et al.* Genome-wide kinetic properties of transcriptional bursting in mouse embryonic stem cells. *Sci. Adv.* **6**, 1–20 (2020).
113. King, H. W., Fursova, N. A., Blackledge, N. P. & Klose, R. J. Polycomb repressive complex 1 shapes the nucleosome landscape but not accessibility at target genes. *Genome Res.* **28**, 1494–1507 (2018).
114. Xiao, X. *et al.* Histone H2A Ubiquitination Reinforces Mechanical Stability and Asymmetry at the Single-Nucleosome Level. *J. Am. Chem. Soc.* **142**, 3340–3345 (2020).
115. Creyghton, M. P. *et al.* H2AZ Is Enriched at Polycomb Complex Target Genes in ES Cells and Is Necessary for Lineage Commitment. *Cell* **135**, 649–661 (2008).
116. Surface, L. E. *et al.* H2A.Z.1 Monoubiquitylation Antagonizes BRD2 to Maintain Poised Chromatin in ESCs. *Cell Rep.* **14**, 1142–1155 (2016).
117. Wang, Y. *et al.* Histone variants H2A.Z and H3.3 coordinately regulate PRC2-dependent H3K27me3 deposition and gene expression regulation in mES cells. *BMC Biol.* **16**, 1–18 (2018).
118. Chen, P. *et al.* H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin. *Genes Dev.* **27**, 2109–2124 (2013).
119. Conway, E. *et al.* BAP1 enhances Polycomb repression by counteracting widespread H2AK119ub1 deposition and chromatin condensation. *Mol. Cell* **81**, 3526-3541.e8 (2021).
120. Fursova, N. A. *et al.* BAP1 constrains pervasive H2AK119ub1 to control the transcriptional potential of the genome. *Genes Dev.* **35**, 749–770 (2021).
121. Gentile, C. & Kmita, M. Polycomb Repressive Complexes in Hox Gene Regulation: Silencing and Beyond: The Functional Dynamics of Polycomb Repressive Complexes in Hox Gene Regulation. *BioEssays* **42**, 1–12 (2020).
122. Martin, G. R. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 7634–7638 (1981).
123. Ying, Q.-L. *et al.* The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).
124. Bulut-Karslioglu, A. *et al.* Inhibition of mTOR induces a paused pluripotent state. *Nature* **540**, 119–123 (2016).
125. Ficz, G. *et al.* FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell* **13**, 351–359 (2013).
126. Sim, Y. J. *et al.* 2i Maintains a Naive Ground State in ESCs through Two Distinct Epigenetic Mechanisms. *Stem Cell Reports* **8**, 1312–1328 (2017).
127. Marks, H. *et al.* The transcriptional and epigenomic foundations of ground state pluripotency. *Cell* **149**, 590–604 (2012).
128. Kolodziejczyk, A. A. *et al.* Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Stem Cell* **17**, 471–485 (2015).
129. van Mierlo, G. *et al.* Integrative Proteomic Profiling Reveals PRC2-Dependent Epigenetic Crosstalk Maintains Ground-State Pluripotency. *Cell Stem Cell* **24**, 123-137.e8 (2019).
130. Kumar, B. & Elsässer, S. J. Quantitative Multiplexed ChIP Reveals Global Alterations that Shape Promoter Bivalency in Ground State Embryonic Stem Cells. *Cell Rep.* **28**, 3274-3284.e5 (2019).
131. Weiner, A. *et al.* Co-ChIP enables genome-wide mapping of histone mark co-occurrence at single-molecule resolution. *Nat. Publ. Gr.* **34**, 1–12 (2016).
132. Bulut-Karslioglu, A. *et al.* The Transcriptionally Permissive Chromatin State of Embryonic Stem Cells Is Acutely Tuned to Translational Output. *Cell Stem Cell* **22**, 369-383.e8 (2018).

133. Guo, G. *et al.* Serum-Based Culture Conditions Provoke Gene Expression Variability in Mouse Embryonic Stem Cells as Revealed by Single-Cell Analysis. *Cell Rep.* **14**, 956–965 (2016).
134. Molina, N. *et al.* Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 20563–20568 (2013).
135. Kirkconnell, K. S. *et al.* Gene length as a biological timer to establish temporal transcriptional regulation. *Cell Cycle* **16**, 259–270 (2017).
136. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
137. Caizzi, L. *et al.* Efficient RNA polymerase II pause release requires U2 snRNP function. *Mol. Cell* **81**, 1920-1934.e9 (2021).
138. Arnold, M., Bressin, A., Jasnovidova, O., Meierhofer, D. & Mayer, A. A BRD4-mediated elongation control point primes transcribing RNA polymerase II for 3′-processing and termination. *Mol. Cell* **81**, 3589-3603.e13 (2021).
139. Liang, K. & Keleş, S. Normalization of ChIP-seq data with control. *BMC Bioinformatics* **13**, 199 (2012).
140. Wachutka, L., Caizzi, L., Gagneur, J. & Cramer, P. Global donor and acceptor splicing site kinetics in human cells. *Elife* **8**, 1251 (2019).
141. Li, J. J., Bickel, P. J. & Biggin, M. D. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* **2**, e270 (2014).
142. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).
143. Xiong, L. *et al.* Oct4 differentially regulates chromatin opening and enhancer transcription in pluripotent stem cells. *Elife* **11**, 1–28 (2022).
144. Hazelbaker, D. Z., Marquardt, S., Wlotzka, W. & Buratowski, S. Kinetic competition between RNA Polymerase II and Sen1-dependent transcription termination. *Mol. Cell* **49**, 55–66 (2013).
145. Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161**, 526–540 (2015).
146. Pintacuda, G. *et al.* hnRNPK Recruits PCGF3/5-PRC1 to the Xist RNA B-Repeat to Establish Polycomb-Mediated Chromosomal Silencing. *Mol. Cell* **68**, 955-969.e10 (2017).
147. Duttke, S. H. C. *et al.* Human Promoters Are Intrinsically Directional. *Mol. Cell* **57**, 674–684 (2015).
148. Conway, E. *et al.* BAP1 enhances Polycomb repression by counteracting widespread H2AK119ub1 deposition and chromatin condensation. *Mol. Cell* **81**, 3526-3541.e8 (2021).
149. Bononi, A. *et al.* BAP1 regulates IP3R3-mediated Ca 2+ flux to mitochondria suppressing cell transformation. *Nature* **546**, 549–553 (2017).
150. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J. & Hurley, L. H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 11593–11598 (2002).
151. Li, C. *et al.* Ligand-induced native G-quadruplex stabilization impairs transcription initiation. *Genome Res.* **31**, 1546–1560 (2021).
152. Schaffitzel, C. *et al.* In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 8572–8577 (2001).
153. Fernando, H., Rodriguez, R. & Balasubramanian, S. Selective recognition of a DNA G-quadruplex by an engineered antibody. *Biochemistry* **47**, 9365–9371 (2008).
154. Biffi, G., Tannahill, D., McCafferty, J. & Balasubramanian, S. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.* **5**, 182–

186 (2013).
155. Henderson, A. *et al.* Detection of G-quadruplex DNA in mammalian cells. *Nucleic Acids Res.* **42**, 860–869 (2014).
156. Liu, H. Y. *et al.* Conformation Selective Antibody Enables Genome Profiling and Leads to Discovery of Parallel G-Quadruplex in Human Telomeres. *Cell Chem. Biol.* **23**, 1261–1270 (2016).
157. Hänsel-Hertsch, R. *et al.* G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.* **48**, 1267–1272 (2016).
158. Zheng, K. W. *et al.* Detection of genomic G-quadruplexes in living cells using a small artificial protein. *Nucleic Acids Res.* **48**, 11706–11720 (2020).
159. Lyu, J., Shao, R., Kwong Yung, P. Y. & Elsässer, S. J. Genome-wide mapping of G-quadruplex structures with CUT&Tag. *Nucleic Acids Res.* **50**, E13 (2022).
160. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1–10 (2019).
161. Lago, S. *et al.* Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome. *Nat. Commun.* **12**, 1–13 (2021).
162. Guo, Y., Zhao, S. & Wang, G. G. Polycomb Gene Silencing Mechanisms: PRC2 Chromatin Targeting, H3K27me3 'Readout', and Phase Separation-Based Compaction. *Trends Genet.* **37**, 547–565 (2021).
163. Zhang, Q. *et al.* RNA exploits an exposed regulatory site to inhibit the enzymatic activity of PRC2. *Nat. Struct. Mol. Biol.* **26**, 237–247 (2019).
164. Beltran, M. *et al.* G-tract RNA removes Polycomb repressive complex 2 from genes. *Nat. Struct. Mol. Biol.* **26**, 899–909 (2019).
165. 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.e4 (2019).
166. Samkurashvili, I. & Luse, D. S. Structural Changes in the RNA Polymerase II Transcription Complex during Transition from Initiation to Elongation. *Mol. Cell. Biol.* **18**, 5343–5354 (1998).
167. Nojima, T. *et al.* RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing. *Mol. Cell* **72**, 369-379.e4 (2018).
168. Martinez-Rucobo, F. W. *et al.* Molecular Basis of Transcription-Coupled Pre-mRNA Capping. *Mol. Cell* **58**, 1079–1089 (2015).
169. Shao, W. & Zeitlinger, J. Paused RNA polymerase II inhibits new transcriptional initiation. *Nat. Genet.* **49**, 1045–1051 (2017).
170. Cheng, B. *et al.* Functional association of gdown1 with RNA polymerase II poised on human genes. *Mol. Cell* **45**, 38–50 (2012).
171. Rahl, P. B. *et al.* c-Myc Regulates Transcriptional Pause Release. *Cell* **141**, 432–445 (2010).
172. Krebs, A. R. *et al.* Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol. Cell* **67**, 411-422.e4 (2017).
173. Zimmer, J. T., Rosa-Mercado, N. A., Canzio, D., Steitz, J. A. & Simon, M. D. STL-seq reveals pause-release and termination kinetics for promoter-proximal paused RNA polymerase II transcripts. *Mol. Cell* **81**, 4398-4412.e7 (2021).
174. Kim, Y. K., Yeo, J., Kim, B., Ha, M. & Kim, V. N. Short Structured RNAs with Low GC Content Are Selectively Lost during Extraction from a Small Number of Cells. *Mol. Cell* **46**, 893–895 (2012).
175. Nudler, E., Mustaev, A., Lukhtanov, E. & Goldfarb, A. The RNA-DNA hybrid maintains the register of transcription by preventing backtracking of RNA polymerase. *Cell* **89**, 33–41 (1997).

176. Shaevitz, J. W., Abbondanzieri, E. A., Landick, R. & Block, S. M. Backtracking by single RNA polymerase molecules observed at near-base-pair resolution. *Nature* **426**, 684–687 (2003).
177. Izban, M. G. & Luse, D. S. Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. *J. Biol. Chem.* **267**, 13647–13655 (1992).
178. Farnung, L., Ochmann, M., Garg, G., Vos, S. M. & Cramer, P. Structure of a backtracked hexasomal intermediate of nucleosome transcription. *Mol. Cell* 1–9 (2022) doi:10.1016/j.molcel.2022.06.027.
179. Hou, L. *et al.* Paf1C regulates RNA polymerase II progression by modulating elongation rate. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 14583–14592 (2019).
180. Lukačišin, M., Landon, M. & Jajoo, R. Sequence-specific thermodynamic properties of nucleic acids influence both transcriptional pausing and backtracking in yeast. *PLoS One* **12**, 1–16 (2017).
181. Gressel, S. *et al.* CDK9-dependent RNA polymerase II pausing controls transcription initiation. *elife* **6**:e29736 (2017)
182. Herzel, L., Straube, K. & Neugebauer, K. M. Long-read sequencing of nascent RNA reveals coupling among RNA processing events. *Genome Res.* **28**, 1008–1019 (2018).
183. Fong, N. *et al.* Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev.* **28**, 2663–2676 (2014).
184. Russo, J., Heck, A. M., Wilusz, J. & Wilusz, C. J. Metabolic labeling and recovery of nascent RNA to accurately quantify mRNA stability. *Methods* **120**, 39–48 (2017).
185. Furlan, M. *et al.* Genome-wide dynamics of RNA synthesis, processing, and degradation without RNA metabolic labeling. *Genome Res.* **30**, 1492–1507 (2020).
186. Kawata, K. *et al.* Metabolic labeling of RNA using multiple ribonucleoside analogs enables the simultaneous evaluation of RNA synthesis and degradation rates. *Genome Res.* **30**, 1481–1491 (2020).
187. Berry, S. & Pelkmans, L. Mechanisms of cellular mRNA transcript homeostasis. *Trends Cell Biol.* **32**, 655–668 (2022).
188. Qiu, Q. *et al.* Massively parallel and time-resolved RNA sequencing in single cells with scNT-seq. *Nat. Methods* **17**, 991–1001 (2020).
189. Erhard, F. *et al.* scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature* **571**, 419–423 (2019).
190. Cao, J., Zhou, W., Steemers, F., Trapnell, C. & Shendure, J. Sci-fate characterizes the dynamics of gene expression in single cells. *Nat. Biotechnol.* **38**, 980–988 (2020).
191. Hendriks, G. J. *et al.* NASC-seq monitors RNA synthesis in single cells. *Nat. Commun.* **10**, 1–9 (2019).
192. Battich, N. *et al.* Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. *Science* **367**, 1151–1156 (2020).
193. Wang, Y. Z. *et al.* H2A mono-ubiquitination differentiates FACT's functions in nucleosome assembly and disassembly. *Nucleic Acids Res.* **50**, 833–846 (2022).
194. Riising, E. M. *et al.* Gene silencing triggers polycomb repressive complex 2 recruitment to CpG islands genome wide. *Mol. Cell* **55**, 347–360 (2014).
195. Holoch, D. *et al.* A cis-acting mechanism mediates transcriptional memory at Polycomb target genes in mammals. *Nature Genetics* (2021). doi:10.1038/s41588-021-00964-2.
196. Kumar, B. *et al.* Polycomb repressive complex 2 shields naïve human pluripotent cells from trophectoderm differentiation. *Nat. Cell Biol.* **24**, (2022).
197. Agarwal, V. & Shendure, J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep.* **31**, 107663 (2020).

198. Angenent-Mari, N. M., Garruss, A. S., Soenksen, L. R., Church, G. & Collins, J. J. A deep learning approach to programmable RNA switches. *Nat. Commun.* **11**, 5012–5057 (2020).
199. Sahu, B. *et al.* Sequence determinants of human gene regulatory elements. *Nat. Genet.* **54**, 283–294 (2022).

11 APPENDIX

11.1 SPIKE-IN DESIGN

Spike-in 2	ERCC-00043	1023 nt	33% GC	1 ng/µL	100% labeled
AATACCTTTACAAATGCTTTAACAAGAGGAAATTGTGTTTTTGCCAATTTAAGACCTAATTTAATAGTTAAACCATTAA CCTTAGTTGTTCCAAGGCATAATATAGAGAGTGAGATACAGGATGAGCTATTTCAGGGAGTTATTCAGTATGCAGTTGC CAAGGCAGTTGCTGATTTAGATTTAGATGAAGATTTAAAGGTTGTTGTCTCTGTTAATGTCCCAGAGGTTCCAATAACC AATTTAAATAAAAGAAAACTCTTCCAATACTTCTATGCCTCAGCAAAGTTAGCTATAAACAGAGCTTTAAATGAATATC CTTCAAAAGAGAAGGTAAAGAAAGAGAAATATAGAGCTTTGCATCCATTAGTTGGATTTAGGGATGTTAGATTGGAGT ATCCTCCATATCTACAAATTGCTTTGGATGTCCCAACTATGGAGAATTTGGAATTTTTGTTACAAACAATTCCAAATAG CGACCACATCATCTTAGAGGCTGGAACACCACTAATTAAAAAGTTTGGTTTAGAGGTTATTGAAATAATGAGAGAATA TTTTGATGGCTTTATTGTTGCTGATTTAAAAACCTTAGACACTGGAAGGGTTGAGGTAAGATTGGCATTTGAAGCAACA GCTAATGCAGTGGCAATAAGTGGAGTAGCACCAAAATCAACAATAATTAAAGCTATCCACGAATGTCAAAAATGTGGT TTAATCAGCTATTTGGATATGATGAACGTCTCTGAACCTCAAAAATTATATGATTCATTAAAATTAAAGCCAGATGTTG TTATCTTGCATAGAGGGATTGATGAGGAGACATTTGGAATTAAAAAGGAATGGAAATTTAAGGAAAACTGCTTATTAG CAATTGCTGGAGGAGTTGGTGTGGAGAATGTTGAAGAGCTTTTAAAAGAATATCAAATATTAATCGTTGGTAGAGCAA TTACAAAATCAAAAGACCCAGGAAGAGTAATTAGGATTTTATAAACAAGATGGGTTAAAAAAAAAAAAAAAAAAAAA AAA					
Spike-in 4	ERCC-00136	1033 nt	42% GC	0.1 ng/µL	100% labeled
TTTCGACGTTTTGAAGGAGGGTTTTAAGTAATGATCGAGATTGAAAAACCAAAAATCGAAACGGTTGAAATCAGCGAC GATGCCGAATTTGGTAAGTTTGTCGTAGAGCCACTTGAGCGTGGATATGGTACAACTCTGGGTAACTCCTTACGTCGTA TCCTCTTATCCTCACTCCCTGGTGCCGCTGTAACATCAATCCAGATAGATGGTGTACTGCACGAATTCTCGACAATTGA AGGCGTTGTGGAAGATGTTACAACGATTATCTTACACATTAAAAAGCTTGCATTGAAAATCTACTCTGATGAAGAGAA GACGCTAGAAATTGATGTACAGGGTGAAGGAACTGTAACGGCAGCTGATATTACACACGATAGTGATGTAGAGATCTT AAATCCTGATCTTCATATCGCGACTCTTGGTGAGAATGCGAGTTTCCGAGTTCGCCTTACTGCTCAAAGAGGACGTGGG TATACGCCTGCTGACGCAAACAAGAGAGGCGATCAGCCAATCGGCGTGATTCCGATCGATTCTATCTATACGCCAGTTT CCCGTGTATCTTATCAGGTAGAGAACACTCGTGTAGGCCAAGTTGCAAACTATGATAAACTTACACTTGATGTTTGGAC TGATGGAAGCACTGGACCGAAAGAAGCAATTGCGCTTGGTTCAAAGATTTTAACTGAACACCTTAATATATTCGCTGGT TTAACTGACGAAGCTCAACATGCTGAAATCATGGTTGAAGAAGAAGAAGATCAAAAAGAGAAAGTTCTTGAAATGAC AATTGAAGAATTGGATCTTTCTGTTCGTTCTTACAACTGCTTAAAGCGTGCGGGTATTAACACGGTTCAAGAGCTTGCG AACAAGACGGAAGAAGATATGATGAAAGTTCGAAATCTAGGACGCAAATCACTTGAAGAAGTGAAAGCGAGACTAGA AGAACTTGGACTCGGACTTCGCAAAGACGATTGACTAGTTTCCCTTGTGAACTAGGATTTTCCCGGGTACAAAAAAAA AAAAAAAAAAAAA					
Spike-in 5	ERCC-00145	1042 nt	44% GC	1 ng/µL	10% labeled
ACTGTCCTTTCATCCATAAGCGGAGAAAGAGGGAATGACATTGTTCTTACACGGCACAAGCAGACAAAATCAACATGG TCATTTAGAAATCGGAGGTGTGGATGCTCTCTATTTAGCGGAGAAATATGGTACACCTCTTTACGTATATGATGTGGCT TTAATACGTGAGCGTGCTAAAAGCTTTAAGCAGGCGTTTATTTCTGCAGGGCTGAAAGCACAGGTGGCATATGCGAGC AAAGCATTCTCATCAGTCGCAATGATTCAGCTCGCTGAGGAAGAGGGACTTTCTTTAGATGTCGTATCCGGAGGAGAG CTATATACGGCTGTTGCAGCAGGCTTTCCGGCAGAACGCATCCACTTTCATGGAAACAATAAGAGCAGGGAAGAACTG CGGATGGCGCTTGAGCACCGCATCGGCTGCATTGTGGTGGATAATTTCTATGAAATCGCGCTTCTTGAAGACCTATGTA AAGAAACGGGTCACTCCATCGATGTTCTTCTTCGGATCACGCCCGGAGTAGAAGCGCATACGCATGACTACATTACAA CGGGCCAGGAAGATTCAAAGTTTGGTTTCGATCTTCATAACGGACAAACTGAACGGGCCATTGAACAAGTATTACAAT CGGAACACATTCAGCTGCTGGGTGTCCATTGCCATATCGGCTCGCAAATCTTTGATACGGCCGGTTTTGTGTTAGCAGC GGAAAAAATCTTCAAAAAACTAGACGAATGGAGAGATTCATATTCATTTGTATCCAAGGTGCTGAATCTTGGAGGAGG TTTCGGCATTCGTTATACGGAAGATGATGAACCGCTTCATGCCACTGAATACGTTGAAAAAATTATCGAAGCTGTGAAA GAAAATGCTTCCCGTTACGGTTTTGACATTCCGGAAATTTGGATCGAACCGGGCCGTTCTCTCGTGGGAGACGCAGGCA CAACTCTTTATACGGTTGGCTCTCAAAAAGAAGTGGATAAGCTGTACAATCGTTTCATCATTCGGCGTGCGAATTAAAA AAAAAAAAAAAAAAAAAAAAAA					
Spike-in 8	ERCC-00092	1124 nt	50% GC	0.1 ng/µL	10% labeled
AGATGTATATATGATGTCCTTGGACGGGGTGGCGCAGTATTACTGCAAGAGAGCGGACAGATTAGTGTGTTGGAGCCG ACACATCAAAGGTTCGTCCGGGGACCGATCTGCAGCCTACGGGACATTTATCCGTAAAAGCATGGCGCTGTTTCGTACT TATCGGAGGCCAGGTATCGTCGCGGCGAGTCTCCCCGACGACGGAGATGGGCGTTACTATCTGGGCCGTCTCGTACTCT GTTACTTGGCACAGATGCGAGCCCTCGTAATGTGCATCAGCTAAGGGCGATATTATAATGCGACGTTTGTACGGATTCG TTACTAACGTGTTGGACGCTAGTGGAATATGTGTCGTTGGTTAGCCTACCCATGGCTTTCGCGGCGACACATGCTTAGA CTCTTTCAAAACTTCGGTGAAGTTCACTCAAGCCGCGGAGCGCCGTCGTAATTCACTAGGGATGGCGGTACCCGTGCCC GTCCGATTCGTAGCAACCTGCATCACGATTTTGTCTTCGGGCGACTTATCAGATACGGTAATGTAAATACCTGGCATTT GGGCACTTCTTGCGTTTAAGCGGGAAAGATCGCGAGGGCCCGCTATTTGCGATACTTCCCATGTCGGTGCCGTCGCCTC TATGTACTCGGAGACGTTAATGCAGAGGCTAAGGACAATTTACCATGACTCGGTAATCCGTTCGTCAAGCAGGTAGCT CGAGTCTCCCCACGGACACGTAGTGGGTTTGTAACGATCGATACCGAGTCTTTTTGTCTAGTAGAACCAACCAACCATT AAGGAGTTCACTAGCACATCTTTGCGACCCGATCGTCCGTGTGTCGCGTAATACTTTTGTTATGACGAGACATACGCTC AAGCCCTGGGTAGCTAGTCGCGGAGGCACGTTACCGCGCACAACCCCTATTCGTTTACATGTACATCGCATCTGAGGTA GTACACTTCCGGCGTACGTGAGTATTTGCGCGTAATAAGCGCGTGTTTAGCTGATCCCCTCTCGTATCGAGGTTAAGGC AGATTAGTGCCCAGTAATTGCGTTTTTTTGTCGTTGTCGCAGAACGCGATTTGCTCCGAAAGCTTTAAGCCGTGGAAAA AAAAAAAAAAAAAAAAAAAA					
Spike-in 9	ERCC-00002	1061 nt	51% GC	1 ng/µL	0% labeled
TCCAGATTACTTCCATTTCCGCCCAAGCTGCTCACAGTATACGGGCGTCGGCATCCAGACCGTCGGCTGATCGTGGTTT TACTAGGCTAGACTAGCGTACGAGCACTATGGTCAGTAATTCCTGGAGGAATAGGTACCAAGAAAAAAACGAACCTTT GGGTTCCAGAGCTGTACGGTCGCACTGAACTCGGATAGGTCTCAGAAAAACGAAATATAGGCTTACGGTAGGTCCGAA TGGCACAAAGCTTGTTCCGTTAGCTGGCATAAGATTCCATGCCTAGATGTGATACACGTTTCTGGAAACTGCCTCGTCA TGCGACTGTTCCCCGGGGTCAGGGCCGCTGGTATTTGCTGTAAAGAGGGGCGTTGAGTCCGTCCGACTTCACTGCCCCC TTTCAGCCTTTTGGGTCCTGTATCCCAATTCTCAGAGGTCCCGCCGTACGCTGAGGACCACCTGAAACGGGCATCGTCG					

CTCTTCGTTGTTCGTCGACTTCTAGTGTGGAGACGAATTGCCAGAATTATTAACTGCGCAGTTAGGGCAGCGTCTGAGG AAGTTTGCTGCGGTTTCGCCTTGACCGCGGGAAGGAGACATAACGATAGCGACTCTGTCTCAGGGGATCTGCATATGTT TGCAGCATACTTTAGGTGGGCCTTGGCTTCCTTCCGCAGTCAAAACCGCGCAATTATCCCCGTCCTGATTTACTGGACTC GCAACGTGGGTCCATCAGTTGTCCGTATACCAAGACGTCTAAGGGCGGTGTACACCCTTTTGAGCAATGATTGCACAAC CTGCGATCACCTTATACAGAATTATCAATCAAGCTCCCCGAGGAGCGGACTTGTAAGGACCGCCGCTTTCGCTCGGGTC TGCGGGTTATAGCTTTTCAGTCTCGACGGGCTAGCACACATCTGGTTGACTAGGCGCATAGTCGCCATTCACAGATTTG CTCGGCAATCAGTACTGGTAGGCGTTAGACCCCGTGACTCGTGGCTGAACGGCCGTACAACTCGACAGCCGGTGCTTG CGTTTTACCCTTAAAAAAAAAAAAAAAAAAAAAAA					
Spike-in 12	ERCC-00170	1023 nt	34% GC	0.1 ng/µL	0% labeled
TATTGGTGGAGGGGCACAAGTTGCTGAAGTTGCGAGAGGGGCGATAAGTGAGGCAGACAGGCATAATATAAGAGGGG AGAGAATTAGCGTAGATACTCTTCCAATAGTTGGTGAAGAAAATTTATATGAGGCTGTTAAAGCTGTAGCAACTCTTCC ACGAGTAGGAATTTTAGTTTTAGCTGGCTCTTTAATGGGAGGGAAGATAACTGAAGCAGTTAAAGAATTAAAGGAAAA GACTGGCATTCCCGTGATAAGCTTAAAGATGTTTGGCTCTGTTCCTAAGGTTGCTGATTTGGTTGTTGGAGACCCATTGC AGGCAGGGGTTTTAGCTGTTATGGCTATTGCTGAAACAGCAAAATTTGATATAAATAAGGTTAAAGGTAGGGTGCTAT AAAGATAATTTAATAATTTTTGATGAAACCGAAGCGTTAGCTTTGGGTTATGAAACTCCATGATTTTCATTTAATTTTTT CCTATTAATTTTCTCCTAAAAAGTTTCTTTAACATAAATAAGGTTAAAGGGAGAGCTCTATGATTGTCTTCAAAAATAC AAAGATTATTGATGTATATACTGGAGAGGTTGTTAAAGGAAATGTTGCAGTTGAGAGGGATAAAATATCCTTTGTGGA TTTAAATGATGAAATTGATAAGATAATTGAAAAAATAAAGGAGGATGTTAAAGTTATTGACTTAAAAGGAAAATATTT ATCTCCAACATTTATAGATGGGCATATACATATAGAATCTTCCCATCTCATCCCATCAGAGTTTGAGAAATTTGTATTA AAAAGCGGAGTTAGCAAAGTAGTTATAGACCCGCATGAAATAGCAAATATTGCTGGAAAAGAAGGAATTTTGTTTATG TTGAATGATGCCAAAATTTTAGATGTCTATGTTATGCTTCCTTCCTGTGTTCCAGCTACAAACTTAGAAACAAGTGGAG CTGAGATTACAGCAGAGAATATTGAAGAACTCATTCTTTAGATAATGTCTTAGGTTAAAAAAAAAAAAAAAAAAAAA AA					