Investigation Of The Nascent Human Transcriptome

Philipp Lorenz



A thesis presented for the degree of

Doctor of Philosophy

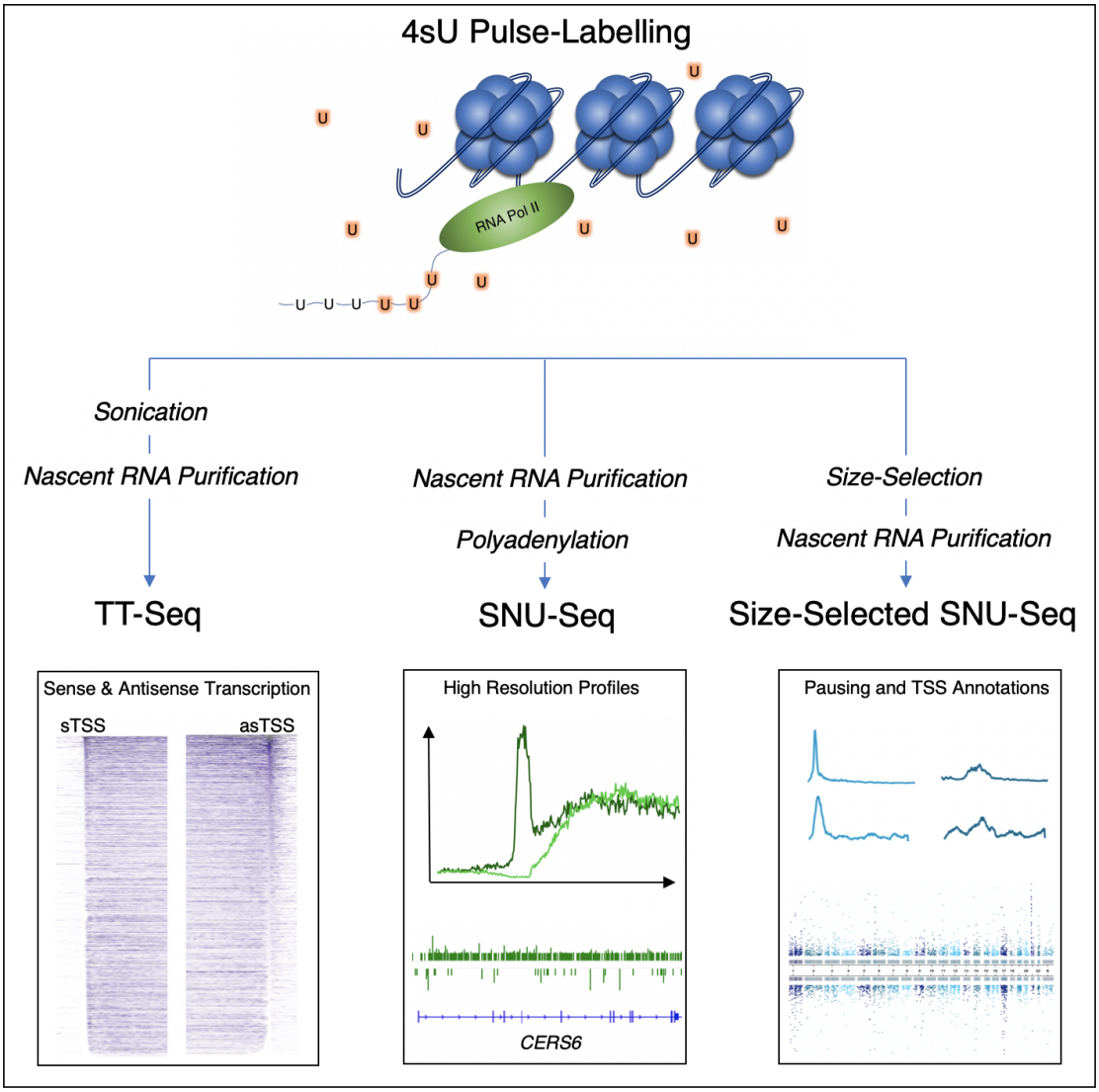
Supervised by Professor Jane Mellor

University College

University of Oxford

September 2020

Graphical Abstract



Abstract

High-throughput sequencing of transcripts through RNA-Seq has revolutionised the way gene expression can be understood. However, RNA-Seq only measures the steady-state levels of stably-expressed transcripts rather than the act of transcription. To measure nascent transcription, a range of sequencing methods have been developed, all with a range of different benefits and disadvantages. This study aims to advance the already established transient transcriptome sequencing (TT-Seq) approach and develop two novel nascent transcriptomic methods in human cells that utilise metabolic labelling and deliver high resolution profiles, called SNU-Seq (single nucleotide resolution 4sU-Seq), and size-selected SNU-Seq. The former is particularly well-suited to measure elongation of transcription throughout the gene body whereas the latter is used for studying transcription at the very 5' end of genes. This study therefore includes a comprehensive analysis of promoter-proximal pausing behaviour and the development of a new size-selected SNU-Seq-based pipeline for annotations of actively-transcribed transcription start sites.

COVID Statement

In accordance with the University of Oxford regulations and guidelines implemented on 01/07/2020 regarding thesis submissions during the COVID19 pandemic, a statement addressing this matter is provided below:

Part of the results presented in this study, as well as the writing of this thesis, was generated during the global COVID19 pandemic. A decision was made not to ask for a thesis deadline extension and instead write up all results and perform additional computational analyses without any further wet lab work. Additional experiments that were planned could not be performed due to the closure of the Biochemistry Department. These experiments would have included additional repeats of SNU-Seq and size-selected SNU-Seq in different cell lines (HeLa and HEK293) as well as a SNU-Seq experiment with oligo-dT magnetic beads to remove already polyadenylated RNAs. Furthermore, the lack of access to departmental computational facilites prohibited parts of additional comparative analysis between different sequencing methods as well as further genome-wide analysis of size-selected SNU-Seq. A more detailed description of the additional approaches and added value of these experiments is discussed in the 'Future Work' section of chapter 8.

Declaration

I confirm that the work presented in this study is my own. Where appropriate, contributions from colleagues and fellow scientists are indicated and their work and contribution is credited accordingly.

Acknowledgements

This thesis could not have been written without the support and help from my brilliant colleagues, namely Anna Lamstaes, Shidong Xi, Ulku Uzun, Silvia Raineri, and Harry Fischl. Thanks go out to Struan Murray and Charlotte George for teaching me a range of bioinformatic skills, and Tom Brown not only for his continued help throughout my project, but also for his initial work on k-means clustering of nascent transcriptomic sequencing data, on which part of this study is based. Most importantly, I would like to thank my supervisor Jane Mellor for being such a brilliant mentor and providing so much insightful advice over the past four years. Finally, I want to thank my parents and sister for their incredible support throughout my entire academic career.

Publications

Brown T, Howe F, Murray S, Wouters M, Lorenz P, Seward E, Rata S, Angel A, Mellor J. (2018) Antisense transcription-dependent chromatin signature modulates sense transcript dynamics. *Mol Sys Biology* 14:2.

Murray S, Lorenz P, Howe F, Wouters M, Brown T, Xi S, Fischl H, Khushaim W, Rayappu J, Angel A, Mellor J. (2020) H3K4me3 is neither instructive for, nor informed by, transcription. bioRxiv doi.org/10.1101/709014

List of Abbreviations

ACF1	ATP-Dependent Chromatin Assembly and Remodeling Factor 1
AFF1 1/4	AF2/FMR2 Family Member 1/4
AF9	ALL1-fused Gene from Chomosome 9 Protein
AMIDA	INO80 Complex Subunit F
ARP4/5/6	Actin Related Protein 4/5/6
asTSS	antisense Transcription Start Site
ATAC-Seq	Assay for Transposase-Accessible Chromatin using Sequencing
ATP	Adenosine Triphosphate
BAF	Brg1/BRM Assocated factor
BPTF	Bromodomain PHD Finger Transcription Factor
BRCA1/2	Breast Cancer Type 1/2 Susceptibility Protein
BRD4	Bromodomain-containing Protein 4
BRE	TFIIB Recognition Element
Brd1	Brahma-Related Gene 1
Brm	SWI/SNF ATPases Brahma
CAGE (-Seq)	Cap Analysis of Gene Expression (followed by sequencing)
CBC	Cap-binding Complex

CDK	Cyclin-dependent Kinase
CHD	Chromodomain Helicase DNA Binding Protein
ChIP-Seq	Chromatin Immunoprecipitation followed by Sequencing
CHRAC	CHromatin Remodeling and Assembly Complex
CoPRO	Coordinated Precision Run-On Sequencing
CPSF	Cleavage and Polyadenylation Specificity Factor
CSTF	Cleavage Stimulation Factor
CTD	C-Terminal Domain
CTIF	CBP80/20-dependedn Translation Initiation Factor
DDR	DNA-Damage Response
DMAP1	DNMT1-assocaited Protein 1
DNMT1/3	DNA Methyltransferase 1/3
DOT1	Disruptor Of Telomeric Silencing / H3K79 Methyltransferase
DRB	Dichloro-1-beta-D-Ribofuransylbenzimidazole
DRE	Downstream Recognition Element
DSB	Doublestrand Break
DSIF	DRB Sensitivity Inducing Factor
EAF1/2	ELL-associated Factor 1/2
ENL	Eleven-Nineteen Leukemia
FACT	Facilitates Transcription
FCP1	TFIIF-Associating CTD Phosphatase 1
FISH	Fluorescence *in situ* Hybridization
GAF	GAGA Factor

GAS41	Glioma-amplified Sequence 41
GDOWN1	GRINL1A Downstream Protein 1
GRO-Seq	Global Run-On Sequencing
GTase	Guanylyltransferase
GTF	General Transcription Factor
GTP	Guanosine Triphosphate
hnRNP	Heterogeneous Ribonucleoprotein Particle
HSC	Hematopoietic Stem Cell
HSP	Heat-Shock Protein
HP1	Heterochromatin Protein 1
IES6	INO80 Complex Subunit C
IGV	Integrated Genome Viewer
INO80	Inositol-Requiring 80
Inr	Initiator Sequence
ISWI	Imitation Switch
lincRNA	Long Intervening/Intergenic Noncoding RNA
MBD	Methyl-binding Domain
MCRS1	Microspheryle Protein 1
MeCP2	Methyl CpG binding Protein 2
MRG15	MORF-related Gene 15
mRNA	Messenger RNA
MTA1/2	Metastasis-assocated Protein 1/2
N7 MTase	Guanine-N7-Methyltransferase

NASC-Seq	New Transcriptome Alkylation-dependent Single-Cell Sequencing
NDR	Nucleosome-Depleted Region
NELF	Negative Elongation Factor
NET-Seq	Native Elongating Transcript Sequencing
NFRKB	Nuclear Factor Related to Kappa B Binding Protein
NoRC	Nucleolar Remodeling Complex
NuRD	Nucleosome Remodeling Deacetylase
NURF	Nucleosome Remodeling Factor
obsTSSs	Observed TSSs
P-TEFb	Positive Transcription Elongation Factor b
PAP	Poly(A)-Polymerase
PAF1	Polymerase-associated Factor 1
PHD	Plant Homeodomain
PIC	Pre-Initiation Complex
PIN1	Peptidylprolyl Cis/Trans Isomerase, NIMA-Interacting 1
Pol I	RNA Polymerase I
Pol II	RNA Polymerase II
Pol III	RNA Polymerase III
PRC 1/2	Polycomb-Repressive Complex 1/2
PRO-Seq	Precision Run-On Sequencing
PROMPT	Promoter Upstream Transcript
PTB	Polypyrimidine-Tract Binding Protein
PTM	Post-Translational Modification

PTT	Premature Transcription Termination
RBBP4/7	Retinoblastoma-Binding Protein 4/7
REF	RNA Export Factor
RNAi	RNA Interference
RNA TPase	RNA 5’ Triphosphatase
RPAP2	RNA Polymerase II Associated Protein 2
RPB1	RNA Polymerase II Subunit B1
RSF1	Remodeling and Spacing Factor 1
RVB1/2	RuvB Like AAA ATPase 1/2
SAM	S-adenosyl Methionine
SAGA	Spt-Ada5-Gcn5 Acetyltransferase
SCAF4/8	SR-Related CTD Associated Factor 4/8
SNF2H	SNF2 Homologue
SNF2L	SNF2-Like
SNU-Seq	Single-Nucleotide Resolution 4-sU Sequencing
SEC	Super Elongation Complex
SET1	Su(var)3-9, Enhancer-of-Zeste and Trithorax - Containing Protein 1
SETX	Senataxin
snRNA	Small Nuclear RNA
snRNP	Small Nuclear Ribonucleoprotein
SPT	Suppressor of Ty
SSU72	SSU72 Homolog, RNA Polymerase II CTD Phosphatase
sTSS	sense Transcription Start Site

SWI/SNF	Switch/Sucrose Nonfermentable
SWR	SWI2/SNF2-Related 1 Chromtain Remodeling Complex
TAF	TBP-associated Factor
TBP	TATA-binding Protein
TFII (A - H)	General Transcription Factor II (A - H)
TGF	Transforming Growth Factor
TIP5	TTF-I Interacting Peptide 5
TOP1	Topoisomerase 1
TREX	Transcription Export Complex
TSS	Transcription Start Site
TT-Seq	Transient Transcriptome Sequencing
4sU	4-thiouridine
UCH37	Ubiquitin C-Terminal Hydrolase 37
uTSSs	Un-annotated TSSs
WSTF	Williams Syndrome Transcription Factor
XRN2	Exoribonuclease 2
YL1	YY1-like Protein
YY1	Yin and Yang 1 Protein

List of Figures

1.1 The Pre-Initiation Complex (PIC). 23

1.2 Phosphorylation States of the RNA Pol II CTD 25

1.3 Transcription Initiation and Pausing 31

1.4 P-TEFb-containing Complexes during Promoter Proximal Pausing . . . 32

1.5 Overview of Mammalian Chromatin Remodellers 35

1.6 Metagene Profiles of RNA Pol II States and Epigenetic Modifications . 38

1.7 5'-Capping of RNA in Eukaryotes 41

1.8 Splicing in Eukaryotes 43

1.9 Termination of Transcription 48

1.10 Promoter Architecture and RNA Pol II Pausing in *Drosophila*. 51

1.11 Comparison of Sequencing Methods used for Genome-Wide Transcriptomics 57

1.12 mNET-Seq Workflow 58

1.13 TT-Seq Workflow 59

1.14 Venn Diagram of Nascent RNA Sequencing Methods and their Benefitis 66

2.1 Genome-Wide Normalised Coverage of Fragmented RNA-Seq and TT-Seq 69

2.2 Metagenes and Positional Read Ratios 70
2.3 IGV Genome Browser Screenshot at the *ANRIL* Locus 72
2.4 TT-Seq Scatter Plots and Alignment Percentages 74
2.5 TT-Seq IGV Genome Browser Screenshot at the *FBL* Locus 75
2.6 Genome-Wide RNA Processing Assessed by TT-Seq 77
2.7 Genome-Wide RNA Processing Assessed by TT-Seq 78
2.8 Genome-wide Assessment of Antisense Transcription by TT-Seq and NET-Seq 81
2.9 TT-Seq of HeLa Cells with or without alpha-amanitin Treatment 84

3.1 SNU-Seq Workflow 89
3.2 SNU-Seq Technical Details 90
3.3 Oligo-dT Cleanup of SNU-Seq 91
3.4 Transcription Profiles of SNU-Seq, TT-Seq, PRO-Seq, and NET-Seq (Large Scale) 93
3.5 Transcription Profiles of SNU-Seq, TT-Seq, PRO-Seq, and NET-Seq (Small Scale) 94
3.6 Metagene Analysis of SNU-Seq Approach 97
3.7 SNU-Seq, TT-Seq, PRO-Seq, and NET-Seq profiles at the *SSR3* locus. 99
3.8 Correlative Comparison of SNU-Seq with TT-Seq and RNA-Seq 100

4.1 Size-Selected SNU-Seq Workflow 107
4.2 Size-Selection of Nascent RNA 108
4.3 Size-selected SNU-Seq Library Preparation and Sequencing Stats . . 109

4.4 IGV Screenshot of size-selected SNU-Seq, TT-Seq, and SNU-Seq around the *CHEK2* locus. 110
4.5 Metagene Analysis of Size-Selected SNU-Seq 112

5.1 k-Means Clustering Approach 116
5.2 k-Means Clusters of Human NET-Seq and Size-selected SNU-Seq . . 117
5.3 k-Means Cluster Profiles of Human SNU-Seq, TT-Seq and TT-Seq after alpha-amanitin treatment 119
5.4 k-Means Cluster Profiles of NELF, INTS3, and MED26 ChIP-Seq . . . 120
5.5 Statistical Analysis of NELF, Intgretor, and Mediator Occupancy within Gene Clusters 122
5.6 Classification Approach of NELF, Integrator, and Mediator Levels into NET-Seq based Clusters 124
5.7 k-Means Cluster Profiles of Size-selected SNU-Seq, ATAC-Seq, and CoPRO 128
5.8 Gene Ontology Analysis of Genes with and without Promoter-Proximal Pausing 130
5.9 Transcription Factor Enrichment in Genes with High Levels of Pausing 133

6.1 SNU-Seq based TSS Annotation Workflow 136
6.2 Clustering of SNU-Seq based TSS candidates 138
6.3 Manhattan Plot of SNU-Seq-based TSS annotations 139
6.4 Characterisation of SNU-Seq-based TSS Annotations 140
6.5 Density Plots of 3'ends of Short Nascent RNAs around Different TSS Annotations 142

7.1 SNU-Seq Methods and Detection of Eukaryotic Transcription Cycle Steps148

Contents

1 Introduction **21**

1.1 Eukaryotic Transcription 22

1.1.1 The Transcription Cycle 22

1.1.2 Transcriptional Pausing 49

1.2 Next-Generation Sequencing Methods in Nascent Transcriptomics . . 56

1.2.1 NET-Seq 56

1.2.2 4sU-Seq and TT-Seq 58

1.2.3 GRO-Seq and PRO-Seq 60

1.2.4 CAGE 61

1.2.5 START-Seq 62

1.2.6 Other Approaches in Nascent Transcriptomics 62

1.2.7 The Need for Further Method Development 64

1.3 Aims and Research Outlines 65

2 TT-Seq Generates Nascent Human Transcriptome Profiles **67**

2.1 TT-Seq Measures Nascent Transcription with High Sensitivity 68

2.2 Nascent RNA Processing Dynamics and TT-Seq 72

2.3 4sU-based Metabolic Labelling is Highly Suitable for Measuring Non-coding Transcription 80

3 Development of SNU-Seq 87

3.1 Motivation for SNU-Seq 88

3.2 SNU-Seq Workflow 88

3.3 SNU-Seq at the 5' and 3' Ends of Genes 96

3.4 SNU-Seq and Gene Expression 100

3.5 Benefits and Considerations of SNU-Seq 101

3.5.1 Resolution 101

3.5.2 5' End Biases 102

3.5.3 Practicability and Cost 103

4 SNU-Seq and Size-Selection 105

4.1 Motivation for Size-Selection 106

4.2 Size-Selected SNU-Seq Workflow 107

4.3 Detection of Pausing by Size-Selected SNU-Seq 110

4.4 Benefits and Considerations of Size-Selected SNU-Seq 113

4.4.1 Resolution 113

4.4.2 Practicability and Cost 114

5 Genome-Wide Analysis of the Promoter-Proximal Pause 115

5.1 K-Means Clustering Reveals Differential Pausing Behavior 116

5.2 Assessing The Predictability of Pausing Clusters from ChIP-Seq Data . 123

5.3 Identification of Alternative TSS Usage from Cluster Analysis 127

5.4 Gene Ontology Analysis . . . 129

5.5 Transcription Factor Enrichment Analysis . . . 132

6 A Pipeline for TSS Annotations using SNU-Seq 135

6.1 Establishing A Pipeline For SNU-Seq-based TSS Annotations . . . 136

6.2 Characterisation of Novel TSSs . . . 139

7 Discussion 144

7.1 Summary . . . 145

7.2 Why 4sU-Labelling? . . . 146

7.3 Detection of Key Steps of the Eukaryotic Transcription Cycle by SNU-Seq and Size-selected SNU-Seq . . . 148

7.4 Future and Additional Work . . . 150

8 Methods and Materials 154

8.1 Methods . . . 155

8.1.1 Cell Culture . . . 155

8.1.2 TT-Seq . . . 155

8.1.3 TT-Seq - Sequencing Analysis . . . 157

8.1.4 SNU-Seq without size-selection . . . 158

8.1.5 SNU-Seq without size-selection – Sequencing Analysis . . . 159

8.1.6 SNU-Seq with size-selection . . . 160

8.1.7 SNU-Seq with size-selection – Sequencing Analysis . . . 161

8.1.8 TSS annotations . . . 161

8.1.9 Bioinformatics and Data Analysis . . . 162

8.1.10 Cluster Analysis 164
8.1.11 Mathematical Modelling and Determination of Transcription Constants 165
8.2 Resources and Materials 167
8.2.1 Sequencing Data 167
8.2.2 Kits and Reagents 168

9 Appendix **169**

9.1 Metagene Code (MATLAB) 170
9.2 TSS Annotation Pipeline (Linux) 189

Chapter 1

Introduction

1.1 Eukaryotic Transcription

1.1.1 The Transcription Cycle

PIC Formation and Initiation

The act of transcription is heavily regulated, beginning with the assembly of the pre-initiation complex (PIC, shown in Figure 1.1). This complex consists of RNA Polymerase II (RNA Pol II), mediator, and the basal transcription factors TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH, initiating transcription after unwinding the DNA of active promoters (Holstege et al., 1996; Luse, 2013; Reinberg et al., 1998; Roeder, 1996; Vannini and Cramer, 2012). These general transcription factors (GTFs) have a range of different functions, with TFIIH being the only GTF with enyzmatic function (Conaway and Weliky Conaway, 1989; Flores et al., 1992; Oksenych and Coin, 2010). TFIID is a protein complex containing 14 subunits, amongst them TAFs (TBP-associated factors), and TBP (TATA-binding protein) which interacts with the TATA-box, a widely conserved DNA sequence located close to the initation site of transcription for a subset of genes across eukaryotes (Furukawa and Tanese, 2000; Müller et al., 2010; Näär et al., 2001; Rhee and Pugh, 2012; Smale and Kadonaga, 2003). Despite the conservation of TATA-boxes throughout eukaryotes, only a small percentage of genes contain TATA-boxes in most organisms, and promoter recognition is often mediated through TAFs rather than TBP (Müller et al., 2010; Rhee and Pugh, 2012).

Another GTF required for initiation is TFIIB, which is known to interact with TBP and direct RNA Pol II to the promoter sequence by binding to DNA elements directly

upstream and downstream of the TATA-box, termed (TFII-)B recognition element, or BRE (Lagrange *et al.*, 1998; Littlefield *et al.*, 1999; Deng and Roberts, 2005; Kostrewa *et al.*, 2009).

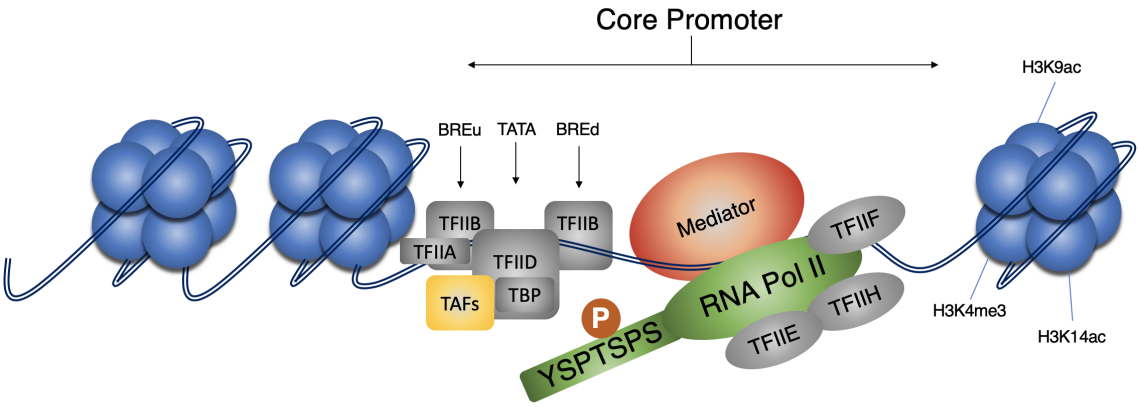


Figure 1.1: The Pre-Initiation Complex (PIC).

Simplified representation of pre-initiation complex (PIC) components bound to the eukaryotic core promoter. This includes the general transcription factors (TFIIA, B, D, E, F, H), RNA Pol II with its CTD, mediator, and a chromatin environment that includes a specific set of histone modifications. Positions of relevant DNA recognition sites (TATA-box and BREs) are marked with arrows.
TBP: TATA-binding protein. TAF: TBP-associated factor. BREu: upstream TFIIB-recognition element. BREd: downstream TFIIB-recognition element.

TFIIF was originally suggested to be responsible for loading RNA Pol II onto the promoter sequence (Sopta et al., 1985). However, TFIIF is now known not to be required for transcription initiation, but rather stabilising TFIIB and retaining it within the transcription complex after initiation, thereby promoting productive elongation after promoter escape (Čabart et al., 2011; Fishburn and Hahn, 2012).

TFIIE was originally considered to act as the loading factor for TFIIH, providing a checkpoint for PIC-assembly (Maxon et al., 1994; Ohkuma et al., 1995). However, work on the archaeal homologue to TFIIE, TFE, showed that this GTF can trigger template melting without TFIIH (Grohmann et al., 2011).

TFIIA is commonly categorized as a GTF, but is stricly speaking not required for the

assembly of the transcriptionally active RNA Pol II complex (Luse, 2013). Functionally, TFIIA has been shown to enhance TFIID interaction with the TATA-Box and stabilise TBP-binding to DNA (Cianfrocco et al., 2013; Hieb et al., 2007).

RNA Pol II itself is a highly conserved protein complex with 12 subunits termed RPB 1-12 (Armache et al., 2004; Liu et al., 2014; Myer and Young, 1998). One key subunit, RPB1, has a C-terminal domain (CTD) with a heptad repeat, YSPTSPS, that is repeated 26 times in *Saccharomyces cerevisiae* and 52 times in humans (Chapman et al., 2008). Five of the individual amino acids of this heptad repeat can have differential phosphorylation states and are crucial points of control, reflecting different stages of the transcription cycle (Harlen and Churchman, 2017; Komarnitsky et al., 2000). As shown in Figure 1.2 the CDK7 subunit of TFIIH phosphorylates serine 5 and 7 of the RNA Pol II CTD during initiation, which is commonly considered a mark for transcription at the 5' end of genes (Hengartner et al., 1998; Komarnitsky et al., 2000; Schroeder et al., 2000).

Dissecting the actual process of PIC assembly required for transcription initiation, a set of systematic order-of-addition studies using TATA-box containing templates has shed light on the PIC assembly pathway, suggesting a stepwise mechanism (Buratowski et al., 1989; Sainsbury et al., 2015): TFIID binds to its recognition element (which can be TBP binding to the TATA-Box or TAFs interacting with other promoter elements), causing a bend in the DNA, an interaction that is then stabilised by TFIIA and TFIIB. Crucially, TFIID is required for transcription of TATA-containing as well as TATA-less promoters in yeast (Warfield et al., 2017). TBP-TFIIB binds to the BRE element asymmetrically, which is crucial for the unidirectional nature of PIC assembly

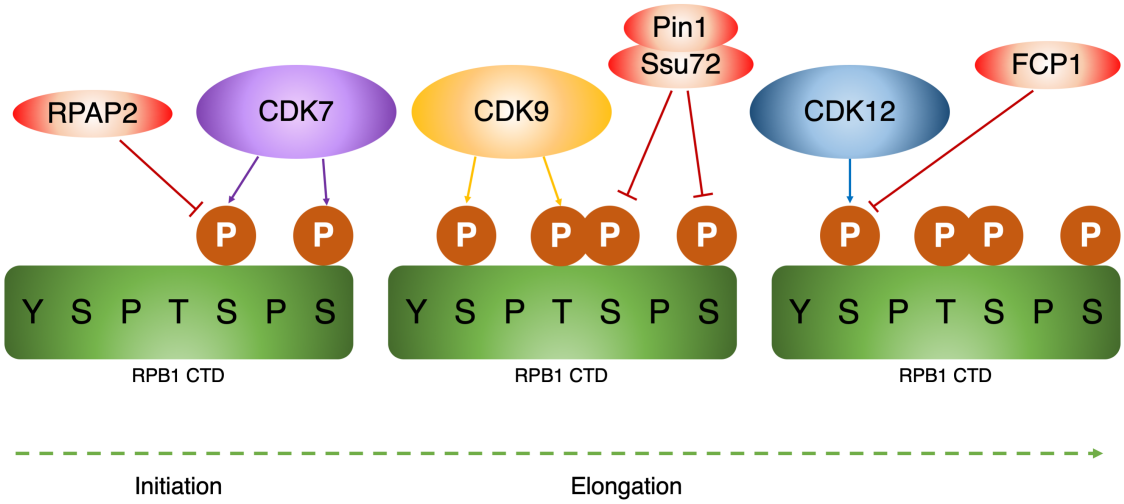


Figure 1.2: Phosphorylation States of the RNA Pol II CTD

The various phosphorylation states of the YSPTSPS heptad repeat of the RPB1 CTD are shown. Dephosphorylation steps are shown with inhibitory arrows along with the phosphatases shown in red. The CTD phosphorylation states and the respective enzymes involved in their control are reviewed by Hsin and Manley (2012). CDK7 phosphorylates serine 5 and 7 (Akhtar et al., 2009; Hengartner et al., 1998; Kim et al., 2009), whilst CDK9, a component of P-TEFb, phosphorylates serine 2 and threonine 4 (Hsin et al., 2011; Meinhart et al., 2005). CDK12, the metazoan homolog of yeast Ctk1, can also phosphorylate serine2 (Yu et al., 2015). FCP1 is a phosphatase that removes the phosphate group from serine 2, and is considered to be relevant in RNA Pol II recycling for renewed PIC assembly (Cho et al., 1999). FCP1 can also dephosphorylate serine5, but preferentially dephosphorylates serine 2 (Lin et al., 2002). Ssu72, which is stimulated by Pin1 (a peptidyl–prolyl cis/trans isomerase), dephosphorylates serine5 and serine 7, and has a distribution accross genes that peaks near the 3'end (Ganem et al., 2003; Zhang et al., 2012). RPAP2 dephosphorylates serine 5 (Egloff et al., 2012).

required to ensure transcription initiation proceeds into one direction (Francis Tsai, 2000). Subsequently, TFIIF-RNA Pol II join the complex before addition of TFIIH and TFIIE complete PIC assembly (Buratowski et al., 1989; Sainsbury et al., 2015). To complete initiation, opening of DNA into a transcription bubble and promoter escape requires TFIIE and TFIIH, which is mediated by the DNA-dependent ATPase activity of TFIIH (Conaway and Weliky Conaway, 1989; Schaeffer et al., 1993).

In this context, it is crucial to point out that, despite the prevalence of TBP-based PIC formation for transcription initiation, the majority of promoters from yeast to humans, are actually TATA-less (Yang et al., 2007). Another crucial sequence motif found

in eukaryotic promoters is the Inr sequence element present in TATA-box containing as well as TATA-less promoters (Ngoc et al., 2017; Smale and Baltimore, 1989). Inr can recruit TFIID which is the key step for PIC formation (Kaufman and Smale, 1994). Another core promoter element that is conserved from *Drosophila* to mammals, MTE (motif ten element), is located downstream of Inr, and can complement for the loss of a TATA-box (Lim et al., 2004). Finally, a downstream promoter element (DPE), is as prevalent in *Drosophila* promoters as TATA-boxes, and acts cooperatively with Inr to recruit TFIID (Kutach and Kadonaga, 2000).

Besides the general transcription factors and RNA Pol II, one factor crucial during transcription initiation is mediator, a highly conserved multimer protein complex comprised of 25 subunits in yeast and 30 in humans (Soutourina, 2018). Mediator binds to RNA Pol II through an interaction between RPB3 and mediator subunit Med17 (Soutourina et al., 2011). Deletion of several mediator subunits are lethal in yeast and embryonic lethal in mammals (Holstege et al., 1996; Ito et al., 2000, 2002; Malik and Roeder, 2010; Soutourina et al., 2011). Mediator is crucial for PIC assembly and transcription as it interacts and regulates a range of steps and components: Mutations in Med17 and Med22 affect TFIIH, RNA Pol II, and TFIIE recruitment, and mutant Med11 reduces RNA Pol II CTD Ser5 phosphorylation (Esnault et al., 2008). Furthermore, Med10 interacts with TBP and contributes towards its recruitment to the promoter (Eychenne et al., 2016).

In addition to the proximal interactions relevant for PIC formation, mediator is also involved in organising promoter-enhancer contacts relevant for transcriptional activation (Poss et al., 2013; Soutourina, 2018). This includes studies showing that mediator

is required for the activity of super-enhancers (Lovén et al., 2013) and a mechanism involving mediator forming a complex with cohesin to facilitate loop-extrusion in order to connect distal DNA elements, such as promoters and enhancers (Kagey et al., 2010).

Finally, mediator dissociation is crucial for promoter escape: Phosphorylation of the CTD by the TFHII subunit Kin28 leads to mediator dissociation from the PIC and is required for subsequent promoter escape of the RNA Pol II complex, which was concluded from Kin28-anchor away experiments *in vivo* (Guidi et al., 2004; Wong et al., 2014).

In addition to the various protein factors bound to the promoter, the chromatin landscape is crucial in the context of transcription initiation. One important aspect of chromatin promoter architecture in this regard is nucleosome occupancy: In order to sterically allow transcription factor binding to the promoter to activate transcription, promoters of active genes are commonly located within NDRs, or nucleosome-depleted regions (Clapier and Cairns, 2009). Experiments on individual genes as well as genome-wide experiments have confirmed the highly prevalent occurrence of NDRs upstream of the TSS, which is a conserved feature amongst eukaryotes from yeast to *Drosophila* to humans (Lee et al., 2007; Liu et al., 2006; Mavrich et al., 2008a,b; Morohashi et al., 2007; Ozsolak et al., 2007; Schones et al., 2008; Sekinger et al., 2005).

One of the most studied histone modification, H3K4me3, is highly enriched at the 5' end of actively transcribed genes across eukaryotes (Guenther et al., 2007; Morillon et al., 2005; Pokholok et al., 2005). This modification plays an active role in

promoter dynamics of transcription initiation at a subset of genes as it stabilises PIC through its interaction with TFIID, through the PHD finger binding of TAF3 (Lauberth et al., 2013; Vermeulen et al., 2007). Crucially, H3K4me3 strongly correlates with gene expression (Santos-Rosa et al., 2002). However, this strong correlation has frequently lead to the description of H3K4me3 as being an "activating" modification - which is misleading given the lack of evidence of H3K4me3 having an instructive or causative role for transcription (Howe et al., 2017). In fact, massive depletion of H3K4me3 has no effect on nascent transcription in yeast (Murray et al., 2019), and in the maternal-zygotic transition in zebrafish, the majority of H3K4me3 deposition occurs after transcriptional activation (Vastenhouw et al., 2010).

Additional key histone modifications that are a ubiquitous part of the epigenetic landscape of active promoters are H3K9 and H3K14 acetylation, both of which are correlated strongly with each other as well as gene expression (Karmodiya et al., 2012; Wang et al., 2008). In fact, SAGA (Spt-Ada5-Gcn5 acetyltransferase), the protein complex responsible for H2B deubiquitination and H3K9 acetylation in yeast, is required for all RNA Pol II transcription, and GCN5-containing acetyltransferase complexes contribute towards H3K9ac at promoters of actively transcribed genes, genome-wide, in humans and yeast (Bonnet et al., 2014). SAGA is considered to be a ubiquitous cofactor for RNA Pol II transcription in yeast, and is required for the recruitment of the TATA-binding protein, TBP (Baptista et al., 2017). H3K14ac is considered to be a similarly activating histone modification with a high frequency of genome-wide co-occupancy of H3K9ac and H3K14ac at active genes, and low levels of H3K14ac at inactive genes (Karmodiya et al., 2012).

Whilst most promoter elements and the corresponding protein complexes are highly conserved from yeast to humans, certain features of promoters and transcription initiation are exclusive to higher eukaryotes. One major distinction between vertebrate and non-vertebrate genomes, and promoter functionality in particular, is the presence of CpG islands in vertebrates. Although not mutually exclusive, only very few promoters have both a CpG island and a TATA-box, which agrees with the idea that CpG island promoters are commonly associated with house-keeping genes whereas TATA-boxes are more frequently associated with tissue-specific promoters (Sandelin et al., 2007; Saxonov et al., 2006). Normally, CpG dinucleotides are methylated throughout the genome, however, stretches of approximately 1000 bp long CpG islands (CGIs) with higher G/C content than on average, frequently lack methylation in vertebrates (Bird, 1980; Deaton and Bird, 2011). CpG islands are present in, or associated with, an average of 70% of vertebrate promoters (Saxonov et al., 2006).

Hypo- and unmethylated CpG islands are strongly associated with active transcription start sites, and show differential methylation patterns across cell types and during development (Kafri et al., 1992; Straussman et al., 2009). Crucially, CpG islands along with their methylation status in promoters show interactions with the local chromatin environment: Cfp1, a subunit of the H3K4 histone methyltransferase SET1, binds to unmethylated CpG islands, therefore linking H3K4me3 and CpG island methylation status in promoters (Clouaire et al., 2012). Depletion of Cfp1 leads to a stark loss of H3K4me3 levels but no significant effect on mRNA levels (Clouaire et al., 2012).

Furthermore, CpG islands are known to recruit polycomb-repressive complexes (PRCs) that deposit histone modifications H2AK119ub and H3K27me3 by PRC1, and

PRC2, respectively (Lynch et al., 2012; Mendenhall et al., 2010; Mikkelsen et al., 2007). These two modifications are critical in gene silencing during development and beyond (Blackledge et al., 2015). CpG islands are non-methylated by default, even for inactive genes, but methylation of CpG islands can lead to consistent silencing of the corresponding promoter during vertebrate development (Deaton and Bird, 2011; Mohn et al., 2008; Payer and Lee, 2008; Stein et al., 1982). Gene silencing through CpG island methylation is enacted either by excluding transcription factor binding to DNA, or by methyl-binding-domain (MDB) - containing proteins, such as MeCP2, that recognise and bind the methylated CpG island before recruiting transcriptional repressors to the promoter (Bogdanović and Veenstra, 2009; Klose and Bird, 2006).

Pausing and Release

Promoter-proximal pausing is a widely observed phenomenon during the transcription cycle for a large proportion of genes amongst eukaryotes (Adelman and Lis, 2012; Core and Adelman, 2019; Core et al., 2008; Kwak et al., 2013; Nechaev et al., 2010). Originally, pausing was discovered in *Drosophila* cells in the context of transcription of heat shock genes: RNA Pol II was recruited to the 5' end of *hsp* genes before heat shock signals were initiated (Gilmour and Lis, 1986; Rougvie and Lis, 1988).

As depicted in Figure 1.3, promoter-proximal pausing occurs in the context of transcription initiation before productive elongation in the transcription cycle, and is orchestrated by the main pausing regulators NELF, DSIF, and P-TEFb as follows: The nascent RNA that is being synthesized during the early stage of the transcription cycle recruits Spt5, which heterodimerises with its binding partner Spt4 to form the

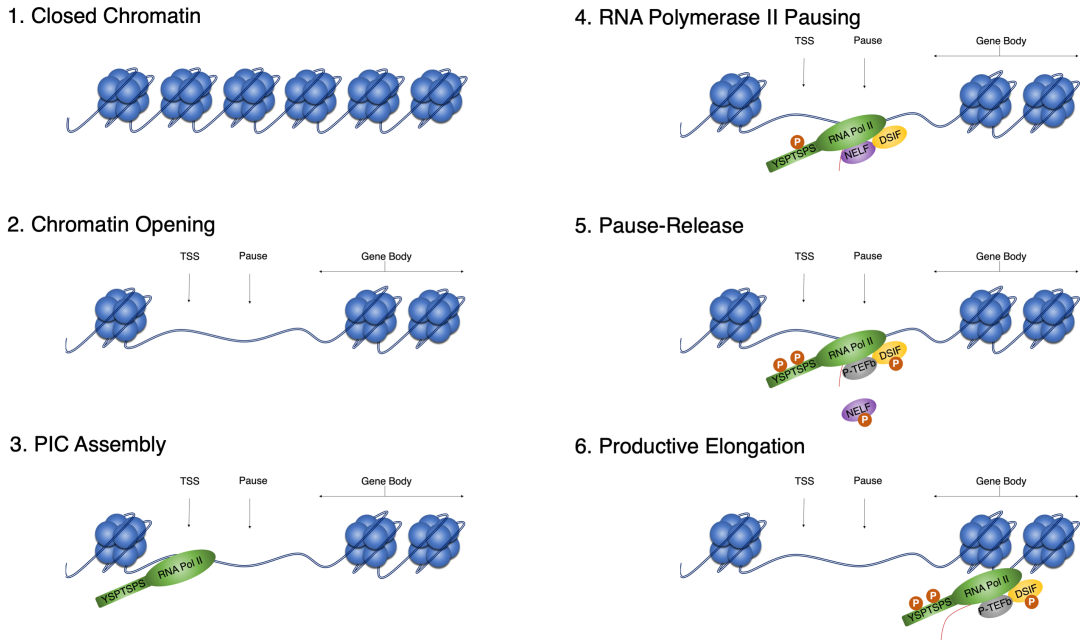


Figure 1.3: Transcription Initiation and Pausing

The steps involved in promoter-proximal pausing within the context of transcription initiation are shown. RNA Pol II pausing is orchestrated by the main pausing regulators NELF, DSIF, and P-TEFb.

dichloro-1-β-D-ribofuranosylbenzimidazole sensitivity-inducing factor, or DSIF (Missra and Gilmour, 2010). Spt5 joins the RNA Pol II complex by binding near its RNA exit channel (Bernecky et al., 2017; Vos et al., 2018a,b).

NELF (negative elongation factor) then recognises Spt5 bound to RNA Pol II and joins the complex at the pausing site (Missra and Gilmour, 2010; Yamaguchi et al., 1999, 2002). The stalled polymerase is stablised by DSIF and NELF, and crucially, NELF prevents RNA Pol II from any further transcription (Adelman and Henriques, 2018; Vos et al., 2018a,b; Yamaguchi et al., 1999). In addition to these factors, GDOWN1, sometimes also referred to as GRINL1A, has been shown to colocalise with RNA Pol II, which can prevent TFIIF and TFIIB binding, thereby promoting transcriptional repression at the 5' end of genes (Jishage et al., 2018).

The key regulator for release of the paused state is the positive transcription elongation factor P-TEFb (Brès et al., 2008). P-TEFb is a cyclin-dependent kinase (CDK)

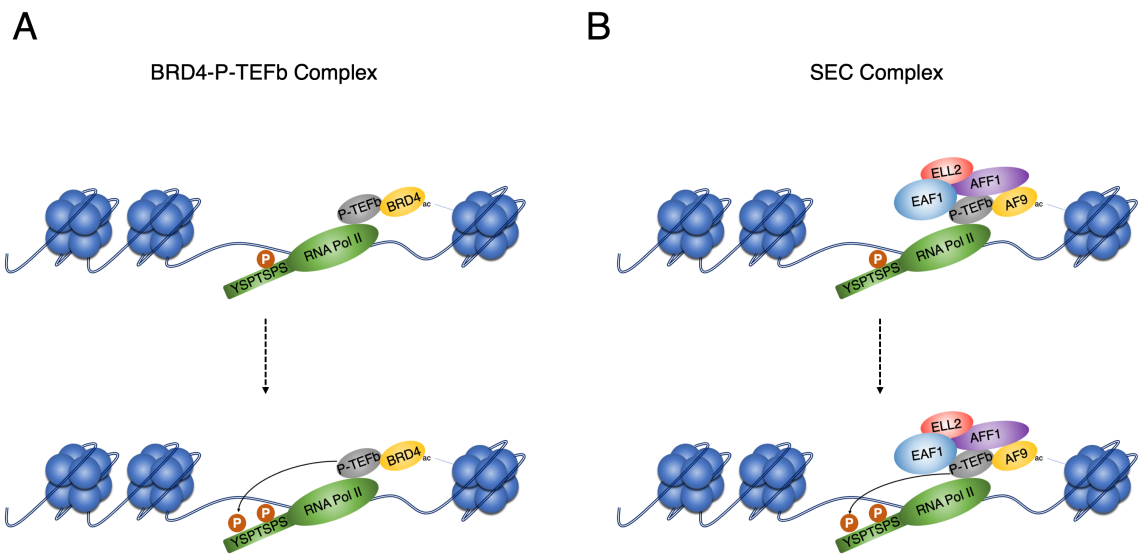


Figure 1.4: P-TEFb-containing Complexes during Promoter Proximal Pausing

A. BRD4-P-TEFb complex. The bromo-domain of BRD4 recognises acetylated histone residues. BRD4 recruits P-TEFb to the pause site. B. The super-elongation complex (SEC). AF9 recognises acetylated histones. Note that some complexes have EAF2 instead of EAF1, AFF4 instead of AFF1, and ENL instead of AF9. BRD4: Bromodomain-containing Protein 4. AF9: ALL1-fused Gene from Chromosome 9 Protein. ENL: Eleven-Nineteen Leukaemia. EAF1: ELL-associated Factor 1. EAF2: ELL-associated Factor 2. AFF1: AF2/FMR2 Family Member 1. AFF4: AF2/FMR2 Family Member 4.

that is recruited to RNA Pol II through mediator subunit med26, and a range of transcription factors (Peterlin and Price, 2006; Takahashi et al., 2011). For full functionality in promoting the transition from paused RNA Pol II to elongation, P-TEFb is usually part of one of two complexes: the Super-Elongation Complex (SEC) or in a complex with BRD4, as shown in Figure 1.4. BRD4 (Bromodomain containing protein 4) has a double-bromodomain which binds acetylated histones, and through its interaction with P-TEFb can target this key kinase to the pausing site and eventually promote release from the pause (Kanno et al., 2014; Yang et al., 2005).

In addition to P-TEFb itself, the SEC contains the protein factors AFF1/AFF4, ELL1/ELL2, ENL, and AF9 (Luo et al., 2012a,b). These factors can have a structural role to serve as scaffolds of the SEC, like AFF1 or AFF2, or, like AF9 or ENL,

can target the SEC to chromatin with acetylated histones (Core and Adelman, 2019; Erb et al., 2017; He et al., 2010; Li and Gilmour, 2013; Lin et al., 2010; Shilatifard et al., 1996; Sobhian et al., 2010; Wan et al., 2017).

In the context of pausing and pause-release, the kinase activity of P-TEFb, through its CDK9 subunit, is of particular importance: Phosphorylation of Spt5 is considered to be key for release as it causes NELF release which ultimately allows RNA Pol II to resume RNA synthesis (Cheng and Price, 2007; Guo et al., 2000; Wu et al., 2003; Yamada et al., 2006). In addition to phosphorylation of Spt5, which turns DSIF into a positive elongation factor, P-TEFb also phosphorylates NELF, which dissociates from RNA Pol II upon phosphorylation, and serine 2 of the RNA Pol II CTD, a post-transcriptional modification associated with productive elongation (Marshall et al., 1996; Ni et al., 2008; Wada et al., 1998; Yamada et al., 2006).

The release from promoter-proximal pausing is also the point at which elongation factors are recruited. For example, the dissociation of NELF after P-TEFb phosphorylation leaves an open binding site for PAF1 (polymerase-associated factor 1), which is a key regulator of elongation: PAF1 has a range of functions, and, depending on the context, it can both maintain or stabilise the pause, or aide in its release (Van Oss et al., 2017): For example, depletion of PAF1 can lead to release of RNA Pol II into the gene body of a subset of genes, and enhancer activation through pause-release can be suppressed by PAF1 (Chen et al., 2017, 2018). On the other hand, PAF1 can promote release from the pause at more than 5000 human genes and positively affects elongation rates (Hou et al., 2019; Yu et al., 2015).

In addition to this description of the mechanistic regulation of pausing followed by

the transition to productive elongation, a section on the biological context and purpose of pausing is included below (chapter 1.1.2).

Elongation

Once RNA Pol II has been released from the promoter proximal pause (for those genes that do undergo pausing), the transcription cycle enters the productive elongation stage. This step is met with different types of challenges and needs to be regulated in a different way: First and foremost, the polymerase complex needs to be able to move through nucleosomes (Petesch and Lis, 2012; Teves et al., 2014). During elongation, one of the two H2A-H2B dimers of histone octamers is removed, leaving a histone hexasome consisting of one H3-H4 tetramer and one H2A-H2B dimer (Kireeva et al., 2002). This dimer removal is mediated by FACT (facilitates chromatin transcription), a protein complex that binds directly to the H2A-H2B dimer, promotes histone de-assembly and re-assembly, and moves along the DNA engaged with RNA Pol II during elongation (Bondarenko et al., 2006; Orphanides et al., 1998; Saunders et al., 2003). Mechanistically, this is achieved by SPT16, a subunit of FACT, shielding the H2A-H2B dimer and acting as a place-holder for DNA, which stabilises the nucleosomal intermediate states during RNA Pol II motion through chromatin (Hondele et al., 2013; Kemble et al., 2015; Liu et al., 2020).

In addition to partial removal of histones, ATP-dependent chromatin remodellers can act on the entire nucleosome by using ATP-hydrolysis to remove nucleosomal barriers to transcription (Hargreaves and Crabtree, 2011; Teves et al., 2014). The four main classes of these types of chromatin remodellers are called imitation switch

(ISWI), switch/sucrose nonfermentable (SWI/SNF), chromodomain helicase DNA binding protein (CHD), and inositol-requiring 80 (INO80). An overview of the mammalian version of these complexes along with their subunits is displayed in Figure 1.5.

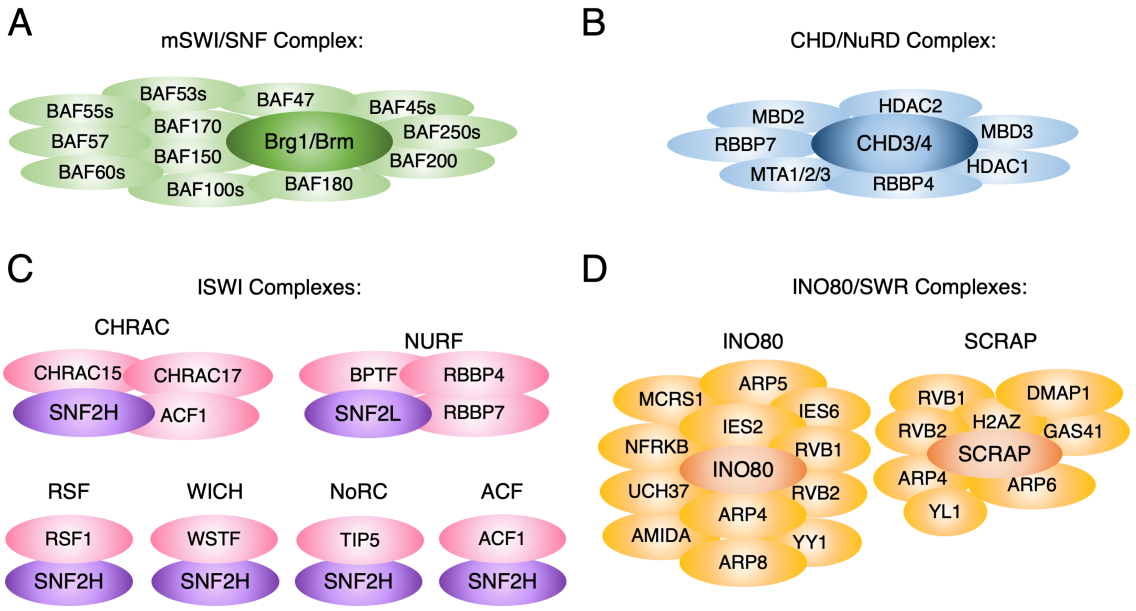


Figure 1.5: Overview of Mammalian Chromatin Remodellers

The four main types of mammalian ATP-dependent chromatin remodellers are shown. The key ATPase containing subunit is highlighted. A. Mammalian SWI/SNF complex (or BAF complex). B. Mammalian CHD complex. C. Mammalian ISWI complexes CHRAC, NURF, RSF, WICH, NoRC, and ACF. D. Mammalian INO80/SWR complexes NURF and SCRAP. Figures are partially based on Hargreaves and Crabtree (2011) and Sokpor et al. (2018). SWI/SNF: Switch/Sucrose Nonfermentable. CHD: Chromodomain Helicase DNA Binding Protein. NuRD: Nucleosome Remodelling Deacetylase. ISWI: Imitation Switch. INO80: Inositol-Requiring 80. SWR: SWI2/SNF2-Related 1 Chromatin Remodeling Complex. Brm: SWI/SNF ATPases Brahma. Brd1: Brahma-Related Gene 1. BAF: Brg1/BRM Associated Factor. MBD2/3: Methyl-CpG Binding Domain Protein 2/3. MTA1/2/3: Metastasis-Associated Protein 1/2/3. RBBP4/7: Retinoblastoma-Binding Protein 4/7. CHRAC: Chromatin Remodeling and Assembly Complex. SNF2H: SNF2 Homologue. SNF2L: SNF2-Like. ACF1: ATP-Dependent Chromatin Assembly and Remodelling Factor 1. BPTF: Bromodomain PHD Finger Transcription Factor. RSF1: Remodelling and Spacing Factor 1. WSTF: Williams Syndrome Transcription Factor. TIP5: TTF-I Interacting Peptide 5. RVB1/2: RuvB Like AAA ATPase 1/2. IES6: INO80 Complex Subunit C. ARP4/5/6: Actin Related Protein 5. MCRS1: Microspherule Protein 1. NFRKB: Nuclear Factor Related To KappaB Binding Protein. UCH37: Ubiquitin C-Terminal Hydrolase 37. AMIDA: INO80 Complex Subunit F. YY1: Yin And Yang 1 Protein. YL1: YY1-Like Protein. DMAP1: DNMT1 Associated Protein 1. GAS41: Glioma-Amplified Sequence 41.

SWI/SNF complexes are highly conserved in eukaryotes from yeast to humans, with the mammalian SWI/SNF version referred to as BAF (Brd1-associated factor, see Figure 1.5 A). SWI/SNF complexes are responsible for nucleosome eviction by

promoting the removal of the H2A-H2B dimer and further loss of the entire octamer (Dechassa et al., 2010). Despite their mechanistic conservation, mammalian BAF complexes additionally have a range of tissue-specific roles and are particularly relevant in the development of the brain (reviewed in Narayanan and Tuoc (2014), Bachmann et al. (2016), Sokpor et al. (2018)).

The chromodomain-helicase-DNA binding (CHD) family of remodelling complexes has two subfamilies - subfamily I (CHD1/2) and subfamily II (CHD3/4, Figure 1.5 B). The latter subfamily contains the ATPase-containing core subunits CHD3/4, which can associate with histone deacetylases (HDAC1 or HDAC2) and a range of DNA-binding proteins (such as MTA1/2/3, MDB2/3), forming the NuRD complex, which is predominantly involved in gene repression (Hirose et al., 2002; Le Guezennec et al., 2006; Marhold et al., 2004; Schmidt and Schreiber, 1999).

ISWI remodelling complexes are involved in nucleosome editing, (re-)assembly, and remodelling, which is carried out through a mechanism that involves ISWI-binding to the N-terminal domain of histone H4 (Clapier and Cairns, 2009; Ito et al., 1997; Tsukiyama et al., 1995). In mammals, SNF2H/SNF2L is the catalytic subunit of ISWI complexes (Cavellán et al., 2006; Ito et al., 1997). Some of the mammalian ISWI complexes are shown in Figure 1.5 C, of which ACF is the most basic one. Crucially, the ACF-complex can regulate the spacing between nucleosomes and therefore affect chromatin accessibility (Ito et al., 1997). Other variations of ISWI complexes are shown in Figure 1.5 C, including CHRAC, RSF, WICH, NoRC, and NURF. The latter can bind H3K4me3 and affect remodelling of nucleosomes in the 5'end of genes (Hargreaves and Crabtree, 2011; Li et al., 2006).

INO80-complexes are a diverse group of remodellers, of which the INO80 and SCRAP complexes are shown in Figure 1.5 D. In addition to the catalytic subunit, INO80 as well as SCRAP have two helicase subunits, RVB1 and RVB2 (Sokpor et al., 2018). The INO80 complex, which like the SCRAP complex has two helicase subunits, RVB1, and RVB2, mediates the regularly repeated nucleosome spacing of genes (Gerhold and Gasser, 2014; Yen et al., 2012).

A range of factors involved in regulating the pause can also travel with RNA Pol II and modulate elongation behaviour: PAF1, for example, is known to stay associated with RNA Pol II and provide a binding platform for a range of factors relevant in transcriptional activation and elongation, which includes the above mentioned CHD1 complex (Simic et al., 2003), and FACT (Pavri et al., 2006; Pruneski et al., 2011). Another such an example of a protein being involved in several stages of the transcription cycle is DSIF (SPT4-SPT5). As discussed above, DSIF stabilises the pause. However, upon phosphorylation by P-TEFb, DSIF becomes a positive elongation factor by stabilising the closed clamp of RNA Pol II to promote continuous passing of DNA through the RNA Pol II cleft (Doamekpor et al., 2014; Grohmann et al., 2011; Klein et al., 2011; Shetty et al., 2017).

Over the course of the transcription cycle, and depending on the relative position within the gene or transcription unit, post-translational and epigenetic modifications are changing and show characteristic (average) profiles, or metagenes, as displayed in Figure 1.6. Beginning with RNA Pol II and its associated CTD phosphorylation states, RNA Pol II, along with Ser 5 phosphorylation peaks just after the promoter, which represents pausing at a proportion of genes influencing the average profile

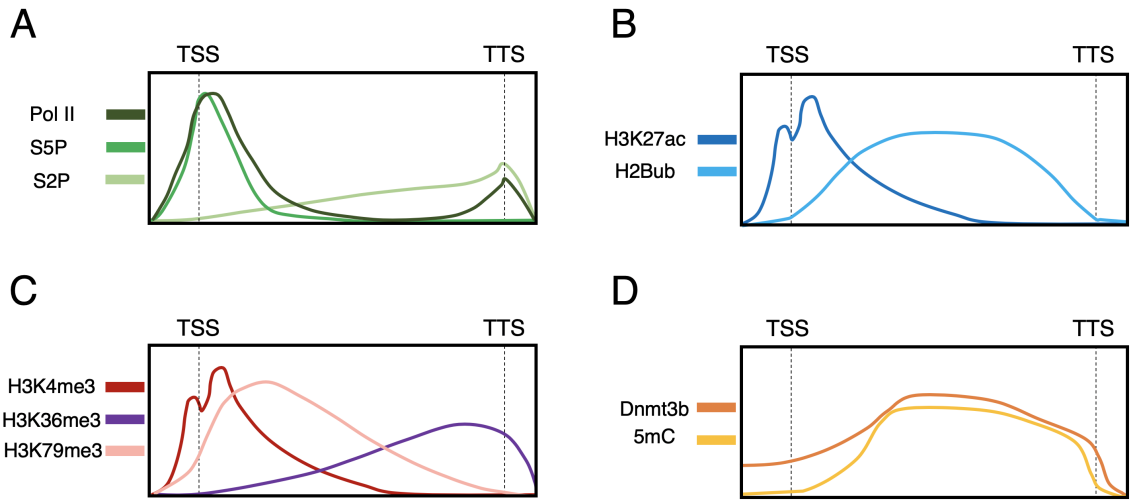


Figure 1.6: Metagene Profiles of RNA Pol II States and Epigenetic Modifications

A. Metagene profile of RNA Pol II, CTD serine 5 phosphorylation, and CTD serine 2 phosphorylation. B. Metagene profile of H3K27ac and H2Bub. C. Metagene profile of H3K4me3, H3K36me3, and H3K79me3. D. Metagene profile of Dnmt3 and 5mc. Figure partially based on Chen et al. (2018).

(Figure 1.6 A). This is followed by, as described above, an increased serine 2 phosphorylation signal after promoter-proximal pausing due to the action of P-TEFb (Ni et al., 2008).

In addition to this, those marks associated with active transcription that are peaking at the 5' end of genes tend to have gradually decreasing levels over the gene body, which includes H3K4me3 and H3K27ac (reviewed in Chen et al. (2018)). In contrast, H3K36me3, H3K79me3, and H2Bub are much more prevalent in the gene body of actively transcribed genes (Barski et al., 2007; Fuchs et al., 2014; Lee et al., 2010; Vakoc et al., 2006), as shown in Figure 1.6 B and C. Interestingly, in yeast, ubiquitinatin of H2B (H2Bub) is deposited by Bre1, an enzyme that is in turn regulated by the chromatin remodeller Chd1 (Batta et al., 2011; Lee et al., 2012; Tanny et al., 2007). The function of H3K79me3 is not very well understood, but in metazoans, a link to the super elongation complex (SEC) was established when the association between DOT1, the enzyme responsible for H3K79me3 deposition, and AF9 and ENL was

found (Mohan et al., 2010).

The role of H3K36me3, on the other hand, has been established in both yeast as well as mammals. Set2 (or SETD2 in humans), which deposits H3K36me3, has been found to associate with the CTD of RNA Pol II and prevent cryptic transcription, i.e. the initiation of new transcription within the gene body (McDaniel and Strahl, 2017). This mechanism was originally discovered in yeast, but has a similar procedure in mammals, where an additional regulatory layer involving DNA methylation is added: H3K36me3, deposited by SETD2, is recognised by the PWWP domain of Dnmt3b (DNA (cytosine-5)-methyltransferase 3B), leading to increased gene body cytosine methylation levels that can prevent cryptic transcription outside of promoters (Neri et al., 2017). This explains the metagene profiles of 5mC and Dnmt3 with higher levels over gene bodies compared to the 5' end of the gene, as seen in Figure 1.6 D.

Another aspect to consider in the context of transcription elongation is DNA supercoiling and overwound DNA that can occur as a result of RNA Pol II progression through the gene (Corless and Gilbert, 2016). Topoisomerases can relieve this torsional stress by inducing double-strand breaks (DSBs), and topoisomerase 1 (TOP1) in particular, is active during elongation through stimulation by BRD4-mediated phosphorylation of the RNA Pol II CTD (Baranello et al., 2016). The active role of TOP1 in promoting transcription elongation is supported by the observation that depletion of TOP1 leads to transcriptional defects primarily at longer genes compared to shorter genes (King et al., 2013; Solier et al., 2013).

In addition to the processivity of transcription aimed at moving RNA Pol II forward through the gene, the polymerase has to face a range of obstacles or events hin-

dering the forward movement that can trigger backtracking of the polymerase: Backtracking means that the DNA upstream and downstream of the transcription bubble is rewound whilst the catalytic site of RNA Pol II becomes disengaged from the 3'end of the nascent RNA (reviewed in Nudler (2012)). Backtracking, which is universal across eukaryotes and common across the entire gene, can be triggered by misincorporation of a nucleotide or by facing nucleosomal and DNA-binding protein barriers (Bondarenko et al., 2006; Churchman and Weissman, 2011; James et al., 2016; Thomas et al., 1998).

Crucially, this event can be resolved by TFIIS, which is known to rescue backtracked polymerases by promoting cleavage of the nascent RNA to generate a new 3'end and allow RNA Pol II to move forward again (Izban and Luse, 1992; Reinberg D and Roeder RG, 1987; Reiness, 1992; Thomas et al., 1998; Zhang and Burton, 2004). Throughout the transcription cycle, TFIIS activity is not only relevant in aiding release from the pause near the 5'end of the gene, but is required for the maintenance of rapid elongation rates across the gene body (Sheridan et al., 2019).

Processing

During the act of transcription eukaryotic transcripts generated by RNA Pol II are heavily processed. The standard trifecta of RNA processing from pre-mRNA towards mature RNA is capping, splicing and polyadenylation. The latter will be discussed in a later section dedicated to transcription termination.

During the synthesis of the first 30 bp of the nascent RNA Pol II transcript, the first major RNA modification, capping, takes place (Moteki and Price, 2002). As

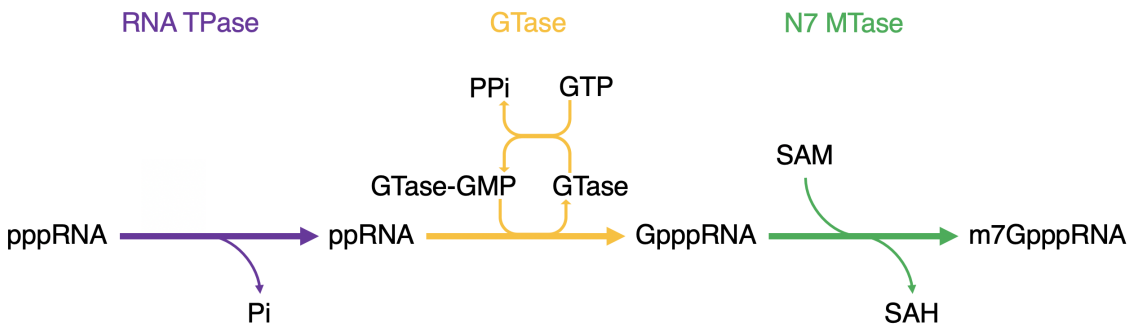


Figure 1.7: 5'-Capping of RNA in Eukaryotes

A simplified diagram depicting the procedure of the enzyme-directed nuclear capping process in eukaryotes is shown. Figure partially based on Ramanathan et al. (2016). RNA TPase: RNA 5' triphosphatase. GTase: guanylyltransferase. N7 MTase: gaunine-N7-methyltransferase. SAM: S-adenosyl methionine. SAH: S-adenosyl homocysteine.

outlined in Figure 1.7, the 5' end of the nascent RNA is capped in three steps: the triphosphate of the 5' nucleotide is dephosphorylated to a diphosphate by RNA TPase (RNA 5' triphosphatase), which is followed by a GTP-dependent addition of a guanylyl-monophosphate (reviewed in Ramanathan et al. (2016)). Finally, methylation of position 7 of the guanoside nucleotide is catalysed by N7 MTase (guanine-N7-methyltransferase), which requires SAM (S-adenosyl-methionine), resulting in the final chemical structure referred to as the m7G-cap (Moteki and Price, 2002).

Crucially, capping is closely linked to the transcription process (Moteki and Price, 2002). In fact, the capping enzymes in both yeast and humans have been shown to be recruited to RNA Pol II via the phosphorylated serine 5 residue of the heptad repeat of the Rpb1 subunit (Cho et al., 1997; Cowling, 2010; Ghosh et al., 2011; McCracken et al., 1997).

A range of purposes of capping nascent RNA have been identified over the past two decades. The cap-binding complex (CBC), a heterodimer consisting of CBC80 and CBC20, binds to the m7G cap during the act of transcription and binds the U4, U5,

and U6 snRNPs involved in splicing (Pabis et al., 2013), which will be discussed below. In fact, the m7G cap is required to enhance splicing in both simpler eukaryotes such as yeast, and higher eukaryotes such as *Xenopus* or humans (Fresco and Buratowski, 1996; Inoue et al., 1989; Konarska et al., 1984; Ohno et al., 1987; Ramanathan et al., 2016).

In addition to aiding splicing, the cap is relevant in promoting nuclear export in higher eukaryotes: CBC interacts with a range of nuclear export proteins including REF (RNA export factor) which is part of TREX, the transcription export complex, leading to the 5'end of the mRNA leaving the nucleus first (Carmody and Wente, 2009; Daneholt, 2001; Izaurralde and Adam, 1998; Nojima et al., 2007). The TREX complex is loaded onto the entire body of the transcript, which is directed by the CBC as well as the exon-junction complex (Viphakone et al., 2019), which is involved in splicing, as discussed below. After nuclear export of mRNA, the cap-binding complex (CBC) is still bound to the m7G cap and facilitates the initiation of translation: It recruits a range of translation initiation factors, namely CTIF (CBP80/20-dependent translation initiation factor), eIF3G, eIF4III, and the methionyl-tRNA (Chiu et al., 2004; Choe et al., 2012, 2014; Fortes et al., 2000; Maquat et al., 2010).

Eukaryotic genes, from yeast to humans, can contain non-coding sequences called introns that need to be removed from pre-mRNAs in order to merge the protein-coding parts of the mRNAs, exons, together, in a process called splicing (reviewed in Wilkinson et al. (2020)). In *Saccharomyces cerevisiae*, only 400 out of their 5000 protein-coding genes have introns (Qin et al., 2016; Spingola et al., 1999). On the other hand, the vast majority of human genes undergo splicing, with 8 introns on average,

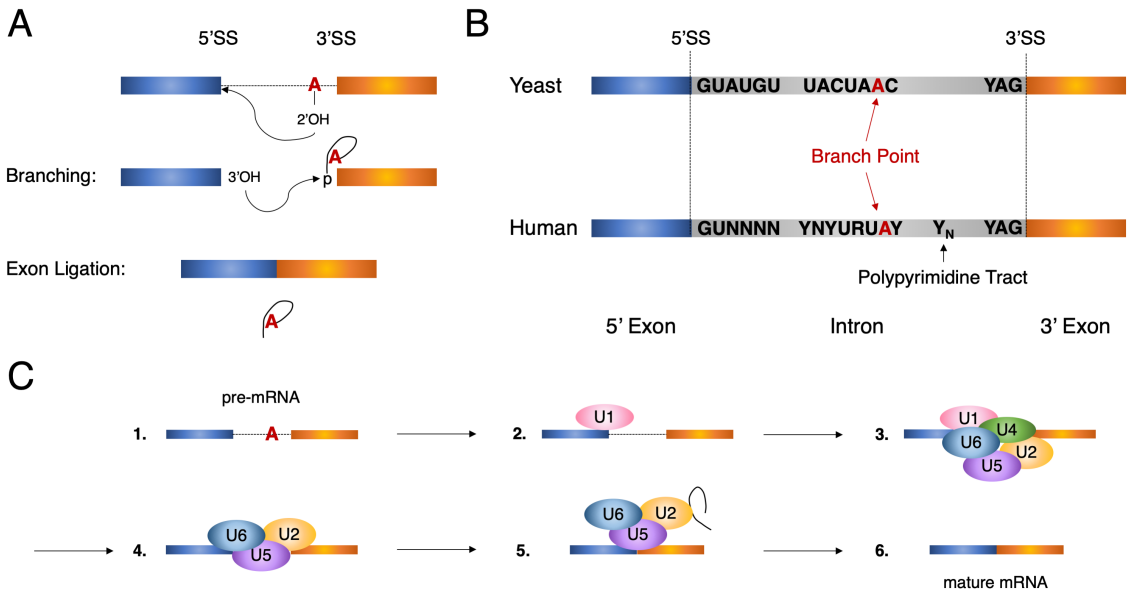


Figure 1.8: Splicing in Eukaryotes

A. Simplified chemical mechanism of branching, lariat formation, and exon ligation during splicing. p represents a phosphate group. B. Sequence motifs in yeast and human introns relevant for splicing. Y represents a pyrimidine. R represents a purine. N can be any nucleotide. C. Spliceosome assembly and disassembly during the splicing process. Figure partially based on Nguyen et al. (2016) and Wilkinson et al. (2020). 5' exons are shown in blue, 3' exons are shown in orange, and introns are shown in dashed lines/grey. Branch-point adenine nucleotides are highlighted in red.

and 95% of human genes being subject to alternative splicing (Hong et al., 2006; Pan et al., 2008; Wang et al., 2008). The sequence features of introns relevant for the splicing mechanism, as well as an overview of the chemical process of branching required for splicing, are depicted in 1.8 A and B. Splicing has been found to proceed in two steps of phosphoryl transfers termed branching and exon ligation: In the first step, the branch point's adenosine 2' hydroxyl group performs a nucleophilic attack on the phosphate group at the 5' splice site, generating a lariat intron/exon intermediate and a cleaved 5' exon (Domdey et al., 1984; Padgett et al., 1984). Second, the exons are ligated by the 3' hydroxyl group of the 5' exon attacking the 3' splice site phosphate, which removes the lariat intron (Rodriguez et al., 1984; Ruskin et al., 1984).

The splicing process is catalysed by the spliceosome, which consists of more than

100 proteins as well as 5 small nuclear RNAs (snRNAs) termed U1, U2, U4, U5, and U6 (Kastner et al., 2019). The first four snRNAs (U1, U2, U4, U5) are transcribed by RNA polymerase II and have a trimethyl-guanosine cap, but U6 is transcribed by RNA polymerase III and acquires a monomethyl-guanosine cap (Reddy and Busch, 1988; Wissink et al., 2019).

During the splicing process, a pre-spliceosome is formed that consists of the U1 snRNP which recognises the 5' splice site, and U2, which recognises the branch point, before the presplicesosome then binds to the U4/U6.U5 tri-snRNP which then constitutes the full spliceosome (Kastner et al., 2019; Plaschka et al., 2019; Yan et al., 2019). The 5' splice site is released from the U1 snRNP by the helicase Prp28 in order to transfer it to the U6 snRNA (Laggerbauer et al., 1998). Subsequently, the U4 snRNA is separated from the U6 snRNA by the RNA helicase Brr2, which enables the U6 snRNA to bind the U2 snRNA, resulting in the formation of the active site that contains the two catalytic metal ions (Fica et al., 2013; Hang et al., 2015; Laggerbauer et al., 1998; Raghunathan and Guthrie, 1998; Steitz and Steitz, 1993). This also allows the branch point adenosine to reach the active site, allowing for the branching reaction to take place that results in the released 5' exon and the intermediary lariat-intron (Galej et al., 2016; Wan et al., 2016). Now, the branch point adenosine leaves the active site to allow for the exon-ligation reaction and the resulting mRNA is released from the active site (Galej et al., 2016; Hang et al., 2015; Yan et al., 2019).

Like capping, splicing is considered to be a co-transcriptional rather than post-transcriptional process (reviewed in Herzel et al. (2017)). This is frequently reflected by links between splicing and the differential phosphorylation states of the Rpb1 CTD

of RNA Pol II, as exemplified by spliced yeast genes where the relative levels of serine 5 phosphorylation decreases over introns whereas those CTD PTMs associated with the later stage of the transcription cycle start increasing at the 3' splice site (Harlen et al., 2016; Herzel et al., 2017; Milligan et al., 2016). Crucially, phosphorylation of serine 5 has been shown to be involved in recruitment of the U1 snRNP, and conditional mutation of serine 5 to alanine in the first 8 repeats of the CTD (S5A(1-8)) can lead to intron retention, whilst threonine 4 phosphorylation marks the release of the spliceosome, but doesn't cause its release (Harlen et al., 2016).

As discussed above, a majority of human genes are spliced alternatively, which is considered to be one of the mechanisms adding to the complexity of humans despite the relatively low number of genes compared to various other organisms (Wang et al., 2008). Alternative splicing is the process of differential exon ligation which means one gene can have a range of different mRNAs and therefore protein isoforms (reviewed in Lee and Rio (2015) and Naftelberg et al. (2015)). Mechanistically, alternative splicing can be regulated in a number of ways, including through protein-protein interactions, through RNA-RNA interactions, through chromatin, and more (Naftelberg et al., 2015).

One prominent example of a class of proteins relevant in the regulation of alternative splicing are hnRNPs (heterogenous ribonucleoprotein particles), of which the hnRNPA1 protein can alter splicing behaviour in various ways, including through exon exclusion by sterically prohibitive binding or promoting loop formation between intronic silencing elements located on both sides of the respective exon (Cartegni et al., 2002; Jean-Philippe et al., 2013). One example of how chromatin can affect splicing includes H3K36me3 and another hnRNP, PTB (polypyrimidine tract binding protein):

MRG15 (MORF-related gene 15), a chromatin-binding protein with a chromodomain, is recruited to H3K36me3, and MRG15 can then recruit the splicing factor PTB, which modulates alternative splicing based on the position of the binding site and its distance to the affected exon (Llorian et al., 2010; Luco et al., 2010; Venables et al., 2008).

Polyadenylation and Termination

In order to complete the transcription cycle of RNA Pol II, the polymerase needs to be able to dissociate from chromatin, which requires cleavage of the transcript (reviewed in Porrua and Libri (2015)). As shown in Figure 1.9, termination requires the key protein complexes CPSF (cleavage and polyadenylation specificity factor) and CSTF (cleavage stimulation factor), which, as their name suggest, mediate the cleavage of RNA as well as another type of modification on the RNA: polyadenylation (reviewed in Kuehner et al. (2011)).

For most genes, the termination of transcription is linked to polyadenylation (Porrua and Libri, 2015). This is initiated by RNA Pol II recognising the poly-A site (PAS), which has the consensus sequence AAUAAA in humans (Birse et al., 1998; Connelly and Manley, 1988; Logan et al., 1987). 18 to 30 nucleotides downstream of this signal, cleavage is carried out by CPSF (Porrua and Libri, 2015). Polyadenylation is carried out by the enzyme poly(A)-polymerase (PAP), which has a very low basal activity and only binds to RNA weakly, but specifically and efficiently adds the poly(A) tail on the 3' end of the nascent transcript in the presence of CPSF (Kaufmann et al., 2004; Keller et al., 1991; Wahlet, 1991).

The CTD of RNA Polymerase II plays a crucial role in recruiting the various factors involved in cleavage and polyadenylation (Hsin and Manley, 2012). CSTF-50, a component of CSTF, and CPSF160 (or Yhh1 in yeast), a subunit of CPSF, bind to the RNA Pol II CTD which is enhanced by the presence of serine2 phosphorylation (Ahn et al., 2004; Dichtl et al., 2002; Fong and Bentley, 2001). In addition, PCF11, a cleavage and polyadenylation factor, binds to phosphorylated serine 2 specifically (Wang et al., 2019). On a genome-wide scale, ChIP-Seq peaks of 3'end processing factors are generally found downstream of phospho-serine2 peaks, which would provide further support for the idea that the CTD along with its phosphorylation states are involved in the recruitment of the cleavage and polyadenylation complexes (Hsin and Manley, 2012; Kim et al., 2009).

The way in which termination occurs is still not completely understood, with two main models proposing different mechanisms: The torpedo model suggest that cleavage by CPSF enables the exoribonuclease XRN2 to reach the nascent RNA and therefore catch up with and evict RNA Pol II, which is supported by studies that identified kinetic competition due to the effects of XRN2 and elongation rates on RNA Pol II (Fong et al., 2015). On the other hand, the allosteric model proposes that it is the process of polyadenylation that causes a conformational change of RNA Pol II at the 3'end of genes and therefore terminates transcription, which is supported by the finding that cleavage is not absolutely required for poly(A)-signal dependent termination (Zhang et al., 2015). A "sitting duck torpedo" model has also been suggested based on the observation that RNA Pol II is decelerated through Spt5 dephosphorylation which makes it an easier target for XRN2 (Cortazar et al., 2019).

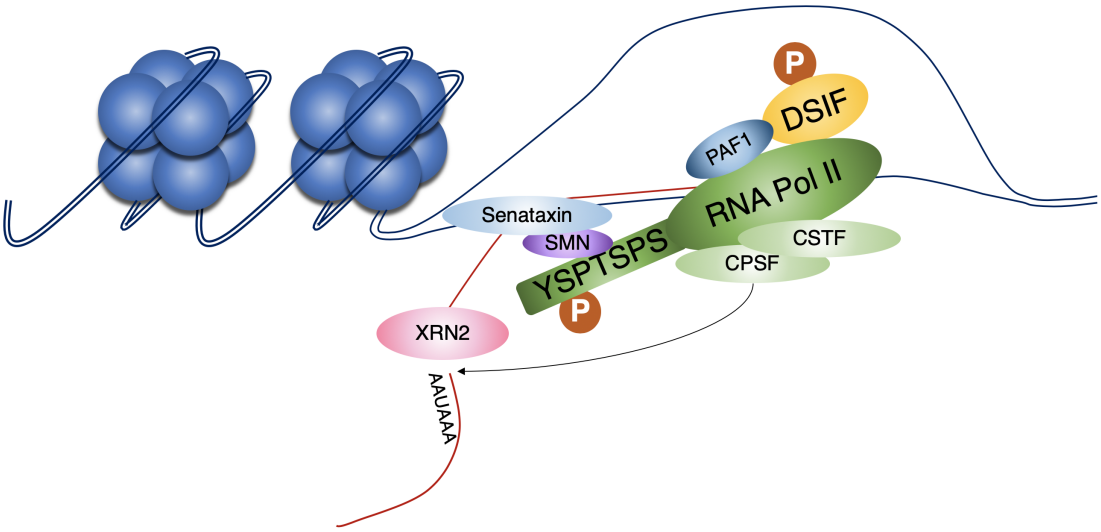


Figure 1.9: Termination of Transcription

A model for the termination of transcription through cleavage and polyadenylation (not shown). Cleavage is mediated through CPSF (cleavage and polyadenlyation specificity factor) and CSTF (cleavage stimulation factor). RNA is shown in red. SMN: survival of motor neuron protein. XRN2: 5' to 3' exoribonuclease 2. Figure partially based on Porrua and Libri (2015).

During the termination process, RNA Pol II undergoes another regulated pausing step (Core et al., 2008; Gromak et al., 2006), which might be an explanation of why XRN2 can catch up with RNA Pol II (Porrua and Libri, 2015). At the termination site, RNA-DNA hybrids are frequently found, which are generated when the nascent RNA anneals to the template DNA, leading to a structure referred to as R-loops (reviewed in Santos-Pereira and Aguilera (2015)). Recently, R-loops have been suggested to trigger antisense transcription in general, and more specifically at the 3' end of genes where antisense transcripts promote the deposition of the repressive histone modification H3K9me3 through the RNAi pathway (Nojima et al., 2018; Skourti-Stathaki et al., 2011). In turn, H3K9me3 can be bound by HP1 (heterochromatin protein 1) that could promote pausing over the termination region (Porrua and Libri, 2015; Skourti-Stathaki et al., 2014). Senataxin (SETX), which resolves and unwinds R-loops, could therefore allow XRN2 to access the nascent transcript after cleavage and therefore

be a critical part of the termination procedure (Banerjee et al., 2009; Suraweera et al., 2009; Wagschal et al., 2012).

1.1.2 Transcriptional Pausing

The Genomic Context and Biological Role of Pausing

Promoter-Proximal pausing is a widely studied phenomenon of the transcription cycle across a range of eukaryotic organism. However, the reason why RNA Polymerase II pauses is not entirely clear. Mechanistically, pausing could be considered a way of preventing new transcription initiation in order to reduce transcriptional noise (Gaertner and Zeitlinger, 2014). Stalled polymerase also gives the cell a time window to integrate regulatory inputs into the expression of a target gene before productive elongation (Henriques et al., 2013), and would avoid polymerase crowding, providing a greater time window for co- and post-transcriptional processing (Marbach-Bar et al., 2013).

Originally, based on *Drosophila* studies on heat-shock genes such as *hsp70*, pausing was considered to provide a way of rapid induction, considering that RNA Pol II is already positioned at the gene to be transcribed and only needs to be released into the gene body without requiring PIC assembly or initiation steps (Boehm et al., 2003; Nechaev et al., 2010). However, this view has been challenged by a variety of counter examples (Smith and Shilatifard, 2013): The human homologue to *hsp70* has a paused PolII that takes several hours to be released (Lin et al., 2010). Crucially, genes that are rapidly induced do not require a paused RNA Pol II, and more generally, genes that exhibit pausing can be induced at variable speed (Lin et al.,

2010). This would suggest that something other than rapid induction should explain the biological purpose of pausing.

The link between pausing or pause-release and the DNA damage response (DDR) as well as genome integrity via proteins such as TRIM28 and RECQL5, respectively, is well established (Bunch et al., 2014, 2015; Saponaro et al., 2014). Another crucial nexus between pausing and DNA repair is mediated by PAF1 together with R-Loops and BRCA2, a protein involved in DDR (Yoshida and Miki, 2004): Reduction of PAF1 levels leads to lower levels of R-loops at the pause site, and reduction of PAF1 and BRCA2 leads to a similar level of reduction, suggesting epistasis (Shivji et al., 2018). However, PAF1 overexpression in BRCA2-deficient cells restores R-loop levels, suggesting that the accumulation of R-loops at pause sites due to BRCA2 loss is orchestrated through reduced PAF1 recruitment at the site of paused RNA Pol II (Shivji et al., 2018).

A range of studies and genome-wide analyses in *Drosophila melanogaster* has helped our understanding of what the biological purpose of pausing could be in the context of development. As summarised in Figure 1.10 A, genes exhibiting promoter-proximal pausing have a distinct promoter architecture, which includes GAGA and Inr elements, and are depleted for TATA-boxes and DRE elements (Gaertner and Zeitlinger, 2014; Hendrix et al., 2008; Kwak et al., 2013; Lee et al., 2008; Smale and Kadonaga, 2003). The GAGA element serves as a binding site for the GAGA factor (GAF), which has been shown to promote open chromatin conformations to allow for the access of RNA Pol II to stalled promoters (Hendrix et al., 2008; Wilkins and Lis, 1997). Inr (initiator sequence) was originally identified as a promoter element that

acts independently from, but functionally similar to, the TATA-Box (Javahery et al., 1994). Promoters exhibiting dispersed transcriptional behaviour are enriched for the Ohler motifs (Ohler et al., 2002). The function of these elements remains elusive. Furthermore, dispersed initiation genes have more clearly delineated nucleosome-depleted regions (NDRs) as well as higher levels of the histone variant H2A.Z (Rach et al., 2011).

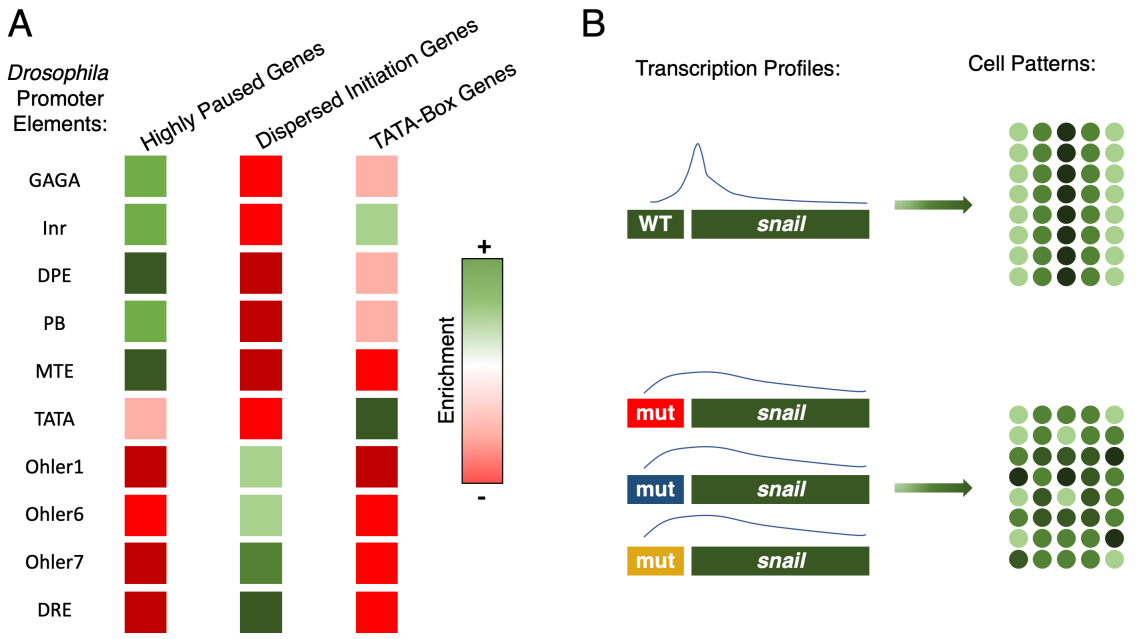


Figure 1.10: Promoter Architecture and RNA Pol II Pausing in *Drosophila*.

A. Enrichment of promoter elements for genes exhibting pausing, dispersed initiation, or TATA-box genes (Gaertner and Zeitlinger, 2014; Hendrix et al., 2008). B. Summary of experiments showing that swapping the promoter of *snail* to a promoter from non-pausing genes leads to invagination defects during *Drosophila* development (Lagha et al., 2013)

In mammalian cells, the link between promoter architecture, PIC assembly, and promoter-proximal pausing is evident, too. The general transcription factor known to initiate PIC assembly, TFIID, for instance, is required for promoter-proximal pausing in human cells, whilst NELF or DSIF enhance pausing, but are not required (Fant et al., 2020). Furthermore, TFIID recruits the super elongation complex (SEC) through a poly-serine domain-dependent interaction (Yadav et al., 2019). The observation that

TATA-boxes are depleted amongst genes that exhibit pausing is also true for mammalian cells, alongside the observation of higher promoter GC skew (Day et al., 2016). Similar to *Drosophila*, H2A.Z deposition in human cells was found to antagonize pausing, in a proposed mechanism by which H2A.Z decreases the height of nucleosomal barriers to RNA Pol II during early transcription (Weber et al., 2014).

The impact of the promoter sequence and PIC assembly on transcriptional pausing behaviour is evident in *Drosophila* development (Levine, 2011). Using the example of *snail* gene expression during *Drosophila* development, Lagha et al. (2013) have shed a light on the crucial role promoter-proximal pausing can play to coordinate developmental cues: As outlined in Figure 1.10 B, *snail* is transcribed with a high degree of promoter-proximal pausing, and is responsible for the coordination of mesoderm cell invagination during gastrulation (Leptin, 2005). When the strongly paused promoter of *snail* is swapped with promoters from genes that exhibit only moderate, or no pausing at all, the onset of *snail* expression is much more random. This leads to a random "salt-and-pepper" expression pattern of *snail* across the mesoderm, rather than a controlled pattern required for invagination (Lagha et al., 2013).

On a larger scale, Zeitlinger et al. (2007) found that amongst those *Drosophila* genes that display the most stalled transcriptional profile, developmental control genes are heavily enriched. Similarly, in Zebrafish, pausing has been found to regulate hematopoietic stem cell (HSC) development: Abolishing the promoter-proximal pause leads to premature (over)-expression of TGF-β which leads to inhibition of HSC specification (Yang et al., 2016).

The idea of co-ordinated onset of transcription due to promoter-proximal pausing

may have the specific purpose to reduce stochastic or 'bursty' transcription initiation (Levine, 2011). This has been proposed after a series of quantitative *in situ* hybridization experiments on several developmental genes in *Drosophila* embryos (Boettiger and Levine, 2009), which demonstrated that key developmental genes show distinctive promoter-proximal peaks of RNA Pol II distributions and are induced highly synchronously, compared to genes that have much flatter RNA Pol II profiles (Boettiger and Levine, 2009).

Taken together, these key findings point to a potential key role of pausing - to allow for synchronicity of induction of genes, especially for metazoans during key stages of development. Pausing could therefore provide a mechanism for multicellular organisms to coordinate regulatory signals by having RNA Pol II of the target gene(s) stalled across cells, until a regulatory signal allows for a timely, precise, and coordinated onset of transcription elongation and therefore synchronised gene expression.

Pausing versus Early Termination

The increased signal of polymerase II levels (or nascent RNA) in metagenes just downstream of the transcription start site is commonly assigned as a promoter proximal pausing signal, but one has to also allow for another potential explanation of this signal: early or premature transcription termination (PTT, reviewed in Kamieniarz-Gdula et al. (2019)). In this context, a range of protein factors that are involved in canonical termination of protein-coding genes are also found at the 5' end of genes, which could explain mechanistically how short nascent transcripts arise due to termination, instead of pausing of RNA Pol II.

PCF11, a cleavage and polyadenylation (CPA) factor, for example, has been shown to not only regulate termination and processing at the 3' end of genes, but also promotes termination and polyadenylation across vertebrates much closer to the 5' end of genes (Kamieniarz-Gdula et al., 2019; Wang et al., 2019). Crucially, amongst those genes that are intragenic PCF11-targets, half of them do not undergo intragenic polyadenylation which might suggest that a significant proportion of RNAs synthesised through PTT are unstable (Kamieniarz-Gdula et al., 2019; Wang et al., 2019).

Another non-canonical factor involved in premature transcription termination is the exosome: In *Saccharomyces cerevisiae*, the RNA degradation activity of the exosome is coupled to premature termination, and in human cells, the DIS3 subunit of the exosome prevents accumulation of truncated transcripts, which are likely the product of premature termination (Schmid and Jensen, 2018; Szczepińska et al., 2015). In addition to this, the cap-binding complex (CBC), which is known to interact with the exosome, can promote premature termination within the first exon of a subset of protein coding genes (Iasillo et al., 2017; Kamieniarz-Gdula et al., 2019).

On the other hand, there are a range of mechanisms in which premature termination of transcription can be suppressed, too. As discussed above, TFIIS can rescue RNA Pol II from nucleosomal 'roadblocks' and stimulates the endonuclease activity of RNA Pol II, resulting in a new 3' end of the nascent transcript - and this stimulation of cleavage prevents early termination that might otherwise occur (Sheridan et al., 2019). Interestingly, the RNA Pol II CTD heptad repeat and its differential phosphorylation states can exert regulatory effects on premature termination too: CDK12, which phosphorylates serine 2 of the CTD, suppresses premature termination since CDK12

depletion leads to genome-wide increases in intragenic polyadenylation events (Dubbury et al., 2018). Additionally, SCAF4 and SCAF8, which are involved in termination and elongation, respectively, bind the phosphorylated CTD of RNA Pol II at both phosphorylated serine 2 and phosphorylated serine 5, acting as anti-terminators by binding to nascent transcripts upstream of early polyadenylation sites and preventing early termination events (Gregersen et al., 2019).

One crucial protein complex relevant in this context is the integrator complex, which was originally discovered as a multiprotein complex that interacts with RNA Pol II and promotes termination of snRNA genes through two of its subunits, IntS9 and IntS11, which are homologs of the CPA factors CPSF100 and CPSF74, respectively (Baillat et al., 2005; Dominski et al., 2005; Kamieniarz-Gdula and Proudfoot, 2019; Rienzo and Casamassimi, 2016). Crucially, integrator also has a role in regulating protein-coding genes, and specifically in pause-release near the TSS, which is dependent on the endonucleolytic activity of IntS11 (Skaar et al., 2015; Stadelmayer et al., 2014). More recently, integrator has been shown to make use of its endonuclease activity to degrade nascent RNA associated with the 5' end of human and *Drosophila* genes and the promoter-proximal pause, resulting in termination of transcription at the pause site, especially at those genes that have unstable RNA Pol II association with promoter DNA (with RNA Pol II stability determined by the half life of RNA Pol II levels at the promoter after triptolide treatment) (Elrod et al., 2019).

This could also point towards a way of uniting the two opposing viewpoints of whether true pausing or early termination is the leading mechanism that drives the observed Pol II-peaks in metagenes of nascent RNA sequencing methods: Some

genes, potentially those with stably associated RNA Pol II at their promoters, undergo pausing regulated by NELF, DSIF and other factors, whereas termination of the short nascent transcripts at the pause site occurs at those genes that have a much less stable RNA Pol II complex associated with their promoter DNA. However, more analysis and additional experiments would be required to understand more about this and to what extent pausing or early termination are mutually exclusive mechanisms.

1.2 Next-Generation Sequencing Methods in Nascent Transcriptomics

In order to detect nascent transcripts as well as the underlying regulatory steps of the transcription cycle, such as pausing, a range of next-generation sequencing methods have been developed. While RNA-Seq measures steady-state RNA levels, nascent sequencing methods (Figure 1.11, Wissink et al. (2019)), aim to measure the act of transcription, which allows for a much better detection of less stable or transient RNA molecules such as enhancer RNAs (Andersson et al., 2014) and antisense RNAs (Brown et al., 2018; Katayama et al., 2005).

1.2.1 NET-Seq

Sequencing of nascent RNA associated with RNA Pol II, referred to as NET-Seq (native elongating transcript sequencing), was originally developed in *Saccharomyces cerevisiae*, relying on precipitation of a FLAG-tagged Rpb1 subunit of RNA Pol II (Churchman and Weissman, 2011). Subsequently, the method was developed for

Technique	GRO-Seq	PRO-Seq	NET-Seq	TT-Seq	START-Seq	CAGE-Seq
In Vivo ?	No	No	Yes	Yes	Yes	Yes
Nascent RNA Selection	BrdU-Antibody	Biotin-NTPs / Streptavidin	Pol II –Antibody	4sU-Specific Biotin / Streptavidin	None Nuclear RNA selection followed by size-selection.	5'cap of mature RNAs
Fragmentation	Alkaline Hydrolysis	Alkaline Hydrolysis	MNase Digestion	Sonication	Not necessary	Random Priming
Resolution	Lower	Single bp	Single bp	Lower	Single bp	Single bp
Benefits / Usage	Nascent Transcription	Enhancer Detection; High-resolution profile of nascent (run-on) transcription	Detect nascent RNA specific to Pol II CTD modifications; Protection of nascent RNA within the Polymerase; Detects promoter-proximal pause	Accurate detection of poly-(A)-site; Determination of Elongation Rates; Very Sensitive; Straightforward protocol	Can receive high resolution information of pausing sites and transcription start sites	High resolution information about transcription start sites
Considerations	Lower Resolution; *in vitro*;	*In vitro*; Hard to carry out, complicated protocol	High Background; Hard to carry out, complicated protocol;	Lower Resolution; Cannot detect promoter-proximal pausing	Not truly nascent RNA; Very limited coverage over gene bodies	Not truly nascent RNA; Detection bias towards more stable RNAs

Figure 1.11: Comparison of Sequencing Methods used for Genome-Wide Transcriptomics

A comparison between GRO-Seq, PRO-Seq, NET-Seq, TT-Seq, START-Seq, and CAGE-Seq is shown, summarising the respective benefits and weaknesses as well as some of the technical features of each technique.

mammalian cells (Mayer and Stirling Churchman, 2016; Mayer et al., 2015; Nojima et al., 2015, 2016). It's noteworthy that the NET-Seq presented in Mayer et al. (2015) is a chromosomal RNA-Seq approach, where some of the signal is from partially processed transcripts or fully processed but nuclear-retained transcripts, whereas the NET-Seq developed by Nojima et al. (2015), referred to as mNET-Seq, uses Rpb1-specific antibodies, as shown in Figure 1.12. Crucially, mNET-Seq can be used to characterise regulatory or RNA-processing-related features of transcription by immunoprecipitating specific RNA Pol II CTD post-translational modifications (Nojima et al., 2018; Schlackow et al., 2017). The latter is achieved by utilising antibodies specific to the various CTD heptad phosphorylation states, such as serine 2 phosphorylation, which is associated with transcription elongation, as discussed above.

NET-Seq profiles reflect the distribution of RNA Pol II associated with RNA that

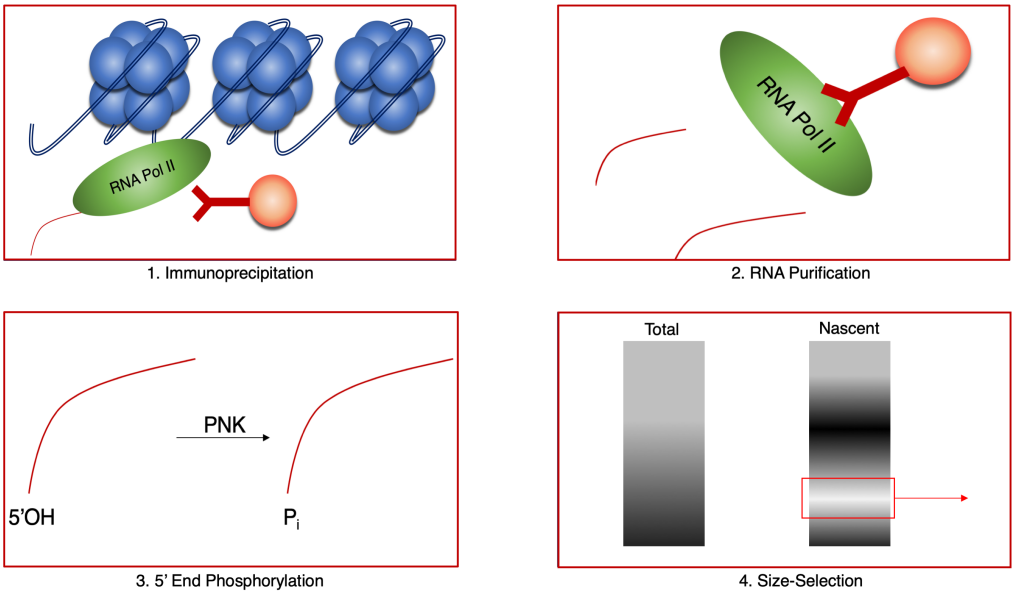


Figure 1.12: mNET-Seq Workflow

The mNET-Seq workflow is shown: This includes immunoprecipitation of RNA Pol II with an anti-RPB1 antibody. Nascent RNA associated with the polymerase is then purified, fragmented, and phosphorylated at its 5' end using PNK (Polynucleotide Kinase) before size selection. The method was developed by Nojima et al. (2015).

ideally is nascent, or in the process of still being synthesised. However, non-specific nascent transcripts, such as RNA Pol I- or Pol III-synthesised RNAs, as well as non-nascent RNAs need to be removed, but some small nuclear RNAs still remain stably bound to RNA Pol II and may therefore be part of the resulting sequencing library, which is an important consideration of the NET-Seq approach (Nojima et al., 2018; Schlackow et al., 2017; Wissink et al., 2019). However, this can be addressed with updated protocols that includes the addition of empigen to disassociate RNA complexes from RNA Pol II (Nojima et al., 2018).

1.2.2 4sU-Seq and TT-Seq

Alternative approaches for nascent RNA sequencing methods are based on addition of cell-permeable nucleoside analogues such as 5-ethenyl-uridine or 4-thiouridine

(Akbalik et al., 2017; Cleary et al., 2005; Fuchs et al., 2014). As an extension of 4sU-Seq, TT-Seq adds a sonication step that removes unlabelled parts of the RNA and therefore measures transient transcription with a reduced 5' bias (Schwalb B et al., 2016), with the workflow shown in Figure 1.13.

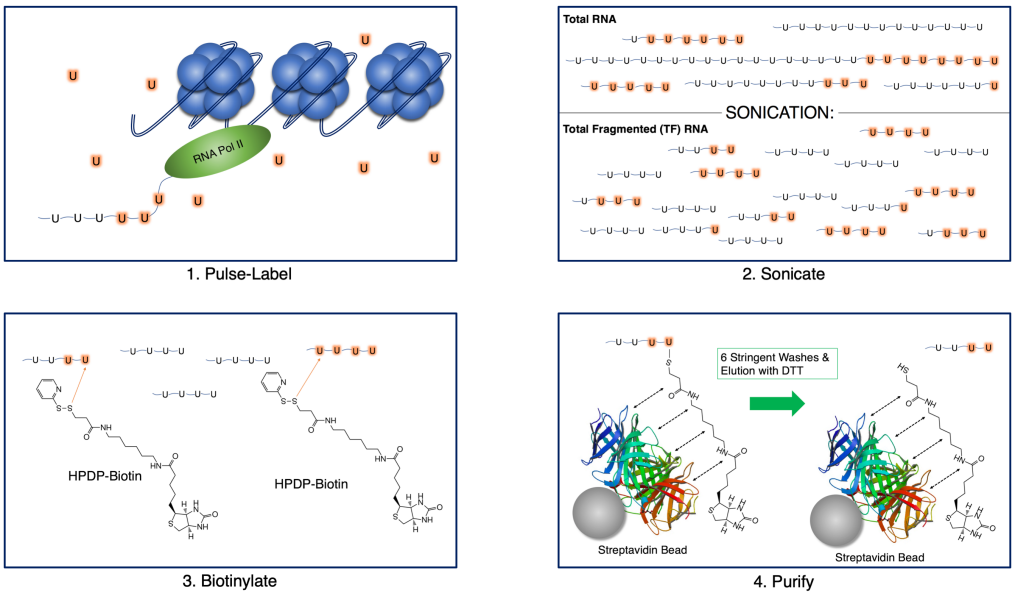


Figure 1.13: TT-Seq Workflow

The TT-Seq workflow is shown: This includes thiolabelling (1.) *in vivo* followed by RNA purification and sonication (2.) for one cycle. The sonicated RNA is then biotinylated with HPDP-biotin (3.), which specifically forms a covalent bond to thio-labelled RNA. This RNA is then purified using streptavidin columns (4.). Thiolabelled nucleotides are indicated in orange. Strand specific libraries are prepared with rRNA depletion and random primers before sequencing. This workflow follows the protocol established by Schwalb B et al. (2016).

With the aim of measuring the act of transcription as accurately and comprehensively as possible, there is a range of advantages and disadvantages of the above mentioned techniques that can be addressed: TT-Seq is highly sensitive, can be used for estimating elongation and half-life dynamics (Schwalb B et al., 2016), and metabolic labelling methods can be used in living organisms (Sidaway-Lee et al., 2014; Stoeckius et al., 2014). However, these methods lack high resolution and cannot detect the promoter-proximal pause, despite recent improvements on the resolu-

tion of TT-Seq by chemical fragmentation (Gregersen et al., 2020).

Using thiolabelling, 4sU-based sequencing approaches have also been used for single-cell RNA-Sequencing experiments: NASC-Seq (new transcriptome alkylation-dependent single-cell RNA sequencing) utilises 4sU-labelling followed by alkylation and immobilisation of RNAs using oligo-dT beads, which is then followed by the SMART-Seq2 protocol for single-cell isolation, and reverse transcription that triggers T-to-C conversion because of the alkylated thio-uridines, aiding in the identification of newly synthesised transcripts in single cells (Hendriks et al., 2019; Picelli et al., 2013). Whilst this approach does have the advantage of providing single-cell information, which most other nascent transcriptomic methods do not, it still only helps in characterising newly transcribed RNAs, rather than providing insights into the process of nascent transcription.

1.2.3 GRO-Seq and PRO-Seq

Another approach to measuring nascent transcription are run-on methods performed on isolated nuclei or permeabilized cells. This includes GRO-Seq (global run-on sequencing, (Core et al., 2008)), which is based on nascent RNA labelling with 5-bromouridine 5'-triphosphate followed by immunoprecipitation with an antibody that is specific to the bromine-containing nucleotide analogue. Similarly, PRO-Seq (Kwak et al., 2013) is a run-on technique based on biotin-NTPs that terminate transcription, followed by streptavidin-based isolation, yielding a transcriptional profile at single nucleotide-resolution.

PRO-Seq has single-nucleotide resolution because of the nature of the biotin-

NTPs that terminate transcription during the run on, which means that this technique is highly suitable for analyses of the promoter-proximal pause or early termination events in both *Drosophila* as well as human cells (Elrod et al., 2019). PRO-Seq is considered to be a difficult, multi-day procedure which is why attempts have been made to simplify the protocol by carrying out some of the enzymatic steps whilst the RNA remains bound to the streptavidin-beads (Judd et al., 2020). Another drawback is the fact that both PRO-Seq as well as GRO-Seq are *in vitro* techniques, as they are not carried out on living cells, and cells need to be permeabilised by treatment with sarkosyl, which inhibitis new transcription initiation events (Szentirmay and Sawadogo, 1994).

1.2.4 CAGE

Cap-analysis of gene expression (CAGE, Kodzius et al. (2006); Shiraki et al. (2003)) and its extension, CAGE-Seq (Takahashi et al., 2011), are not truly nascent transcriptomic methods as they are based on usage of steady-state RNA and do not require a selection of nascent transcripts in any way, but crucially, are capable of mapping transcription start sites with high accuracy: This is achieved by reverse transcription of RNAs using random primers, followed by biotinylation of the 5'end before RNase digestion of single-stranded RNA. cDNA with completed 5' ends can then be selected with streptavidin before linker ligation and sequencing. This yields a single-nucleotide resolution of transcription start sites and has been widely used in landmark projects for genome annotations (Andersson et al., 2014). One key drawback of this method is that due to the lack of nascent RNA-selection, CAGE is biased towards the most

stable RNAs and can therefore miss a large range of transcripts that are less stable.

1.2.5 START-Seq

START-Seq was originally developed in *Drosophila* by performing enrichment of nuclear RNAs based on the presence of a 5'cap and, most importantly, size of less than 100 bp (Nechaev et al., 2010). The advantage of START-Seq profiles is that, thanks to the small size of the selected RNA fragments, one can get single-nucleotide resolution information of both the transcription process up until the end of the short fragment (which could be the promoter proximal pause), and TSSs. In humans, START-Seq was used to extensively annotate different classes of antisense transcription start sites, such us upstream or downstream asTSSs (Lavender et al., 2016). However, it is worth pointing out that START-Seq is based on nuclear RNA, not truly nascent RNA. This is crucial in the context of considering short RNAs that are synthesised up until the promoter-proximal pause: A subset of these transcripts are likely very unstable, an observation that is based on studies involving the integrator complex and PRO-Seq experiments (Elrod et al., 2019). This would mean that using nuclear, and not truly nascent short RNAs, could therefore lead to the exclusion of the more unstable short transcripts and therefore make this technique much less comprehensive than it intends to be.

1.2.6 Other Approaches in Nascent Transcriptomics

One major drawback of all methods mentioned above is the lack of single-cell resolution, since these methods require large numbers of cells and therefore generate

readouts that correspond to averaged signals across a population of cells. Image-based methods, on the other hand, can provide single-cell information although they might lack genome-wide information. In addition, they can provide additional spatial and 3-dimensional information about nascent transcription.

One common approach is the usage of fluorescently labelled RNAs that can hybridise with transcripts in the cell using RNA Fluorescence *in situ* hybridization (FISH, Bauman et al. (1980); Gall and Pardue (1969); Singer and Ward (1982), reviewed in Lim (2018); Wissink et al. (2019)). An alternative option is to engineer the transcript so that it forms hairpin structures and make use of hairpin-binding proteins fused to GFP, such as MS2, a bacterial coat protein (Bertrand et al., 1998; Yunger et al., 2010).

A key way of ensuring that the output signal that is measured is as close to nascent transcription as possible, would be to make sure that the probes hybridise to intron sequences as these represent parts of the transcripts that are only present in the very short-lived pre-mRNAs which are excised co-transcriptionally, as discussed above (Raj et al., 2008; Tantale et al., 2016).

Recent advances in these approaches have yielded much higher throughput that allows for the monitoring of many more transcription processes than previously possible: iceFISH (intron chromosomal expression FISH), for example, uses different combinations of spectrally distinguishable fluorescent labels that allows for simultaneous hybridisation and detection of several intronic transcripts at once (Levesque and Raj, 2013). Intron seqFISH (intron sequential FISH) has been a major advancement in this field by allowing for the detection of thousands of intronic transcription signals through barcoded hybridisation in a sequential manner (Shah et al., 2018).

While the ability to look at (close to) nascent transcription in single cells for thousands of transcripts is a key advance, it is worth pointing out that these methods are not capable at distinguishing different stages of the transcription cycle or locate the exact site of active transcription carried out by RNA Pol II (Wissink et al., 2019).

1.2.7 The Need for Further Method Development

With the aim of understanding as much as possible about the transcription cycle and its key regulatory steps, measuring the act of transcription as accurately and comprehensively as possible is crucial. There are a range of advantages and disadvantages of the above mentioned techniques that can be addressed: TT-Seq is highly sensitive, can be used for estimating elongation and half-life dynamics (Schwalb B et al., 2016), and metabolic labelling methods can be used in living organisms (Sidaway-Lee et al., 2014; Stoeckius et al., 2014). However, these methods lack a high resolution and cannot detect the promoter-proximal pause, despite recent improvements on the resolution of TT-Seq by chemical fragmentation (Gregersen et al., 2020).

As discussed above, the various next-generation sequencing approaches have a range of advantages and disadvantages to consider, with the main points of concern being resolution, practicability and accessibility of the method, sensitivity, and background. With NET-Seq and PRO-Seq showing high resolution but seeming to be more difficult to carry out, and TT-Seq lacking the resolution but showing high sensitivity and an easier protocol, this study aims to find an approach to combine these considerations into one method. Therefore, this study presents single-nucleotide resolution 4sU-sequencing, or SNU-Seq.

1.3 Aims and Research Outlines

A first step in this study will be to establish TT-Seq in cell lines that will be used to develop SNU-Seq, namely HeLa and HEK293 cells. This is because a lot more nascent transcriptomic and other next-generation sequencing data is available in these cell lines, rather than K562, which is the cell line in which TT-Seq was developed in originally (Schwalb B et al., 2016). Getting the TT-Seq protocol to work in these cell lines will then allow for comparison with NET-Seq or GRO-Seq and identify direct differences in the transcriptional profiles.

Furthermore, TT-Seq with different labelling times in HEK293 was performed, not only to find the best labelling time for this cell line, but also to study further nascent RNA processing dynamics. Being able to carry out TT-Seq in HEK293 cells will then allow for the establishment of the new method presented in this study, SNU-Seq (single-nucleotide resolution 4sU-Seq).

Similar to TT-Seq, SNU-Seq relies on 4sU-pulse labelling of cells but omits sonication and includes addition of an artificial poly-(A)-tail. This allows for library preparation with oligo-d(T) primers, retaining information about the last incorporated nucleotide and therefore generating a transcriptional profile at near basepair resolution. This approach therefore retains the accessible procedure of TT-Seq, whilst at the same time aiming to retrieve the resolution observed in NET-Seq or PRO-Seq, as represented in the Venn diagrams shown in Figure 1.14 A and B.

Moreover, size-selection of the purified nascent RNA was performed to enrich for shorter RNAs (50-100 bp) with the aim to sequence the fraction of labelled RNA that predominantly reflect transcripts generated up until the RNA Pol II pause site. This

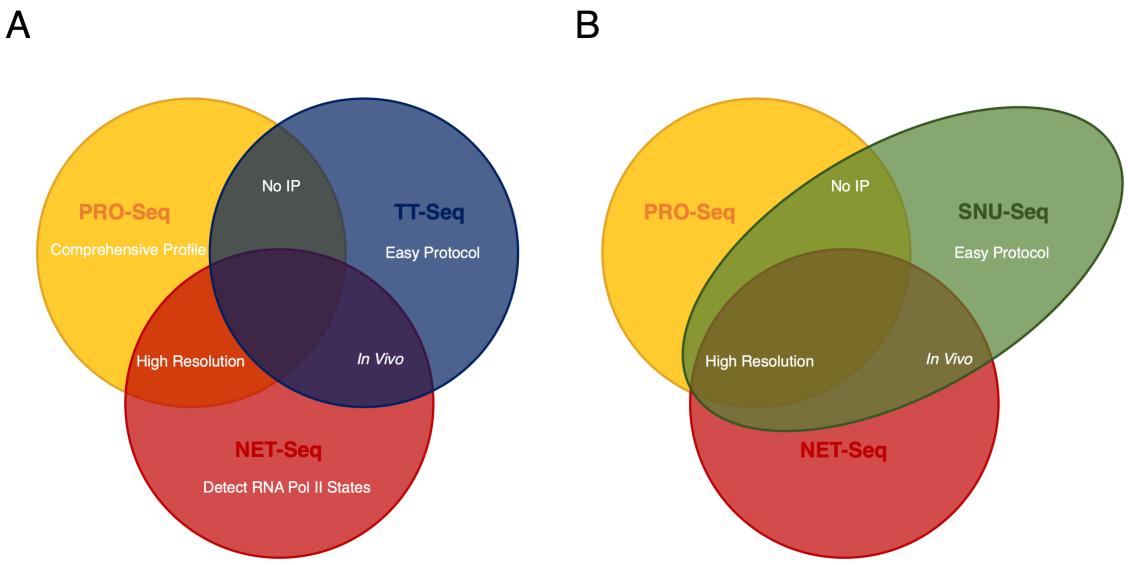


Figure 1.14: Venn Diagram of Nascent RNA Sequencing Methods and their Benefitis

A. Venn diagram comparison of PRO-Seq, NET-Seq, and TT-Seq. B. Venn diagram comparison of PRO-Seq, NET-Seq, and SNU-Seq.

approach can show that SNU-Seq can replicate pausing profiles at single base-pair resolution as observed in NET-Seq. Analyses of gene ontology and TF enrichment reveal the types of genes more likely to exhibit pausing behaviour and demonstrate a biological role for this step of transcription.

Finally, a workflow is proposed on how to use SNU-seq with size selection for transcription start-site annotation using the 5'end information of the small, nascent RNAs purified. These TSSs are then characterised as either already annotated or novel. They can further be classified in their genomic context, their relation to other TSSs, and their performance of accurately marking the start of active transcription.

Chapter 2

TT-Seq Generates Nascent Human Transcriptome Profiles

2.1 TT-Seq Measures Nascent Transcription with High Sensitivity

As a first approach towards advancing nascent RNA-sequencing methodology, this study aimed to perform TT-Seq in adherent cells which would then be the subject for advancement of the technique towards developing SNU-Seq. This is because the original TT-Seq protocol was developed in the non-adherent K562 cell line (Schwalb B et al., 2016). For this, the workflow for TT-Seq, developed by Schwalb *et al*. (2016) was adopted and slightly adjusted with further details explained in the methods section.

Crucially, the usage of thio-labelled spike-in RNAs allows for calibration, if necessary, for genome-wide comparisons. This was achieved by *in vitro* transcription of 3 yeast genes of approximately 1000 bp in length with 4sUTP added to the NTP mix. This is described in more detail in the methods section.

As an additional control, non-thio-labeled RNA from HEK and HeLa cells was sonicated and run through the same workflow of biotinylation and streptavidin columns. However, after running the streptavidin columns, there was no detectable RNA (<0.1 ng) that could have been carried forward to library preparation. This supports the notion that metabolic labelling with 4sU followed by biotin-streptavidin-based enrichment is highly specific and creates a relatively low background. The high specificity of metabolic labelling approaches is complemented by a high sensitivity. In order to show this, a comparative experiment was carried out by performing TT-Seq and TF-Seq (total fragmented RNA-Seq), which follows the TT-Seq protocol but skips the

biotinylation/streptavidin purification, to allow for direct comparison with TT-Seq. As shown in Figure 2.1, the normalised coverage of TT-Seq is significantly greater than that of fragmented RNA-Seq, not only in breadth, but also in depth, genome-wide.

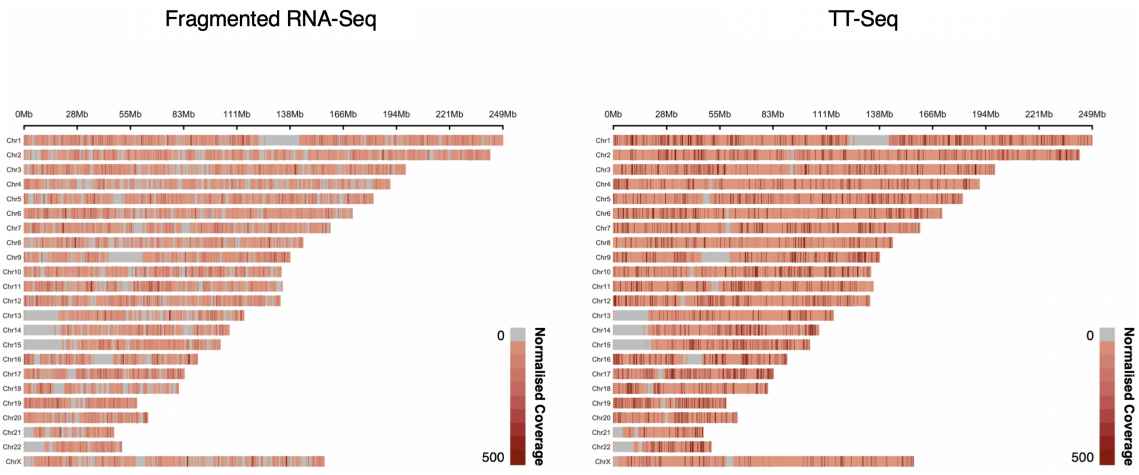


Figure 2.1: Genome-Wide Normalised Coverage of Fragmented RNA-Seq and TT-Seq

The coverage plots show the normalised read coverage, genome wide, of total fragmented RNA-Seq and TT-Seq performed in HeLa WT cells. For total fragmented RNA-Seq, the TT-Seq protocol was followed but the biotin/streptavidin purification was skipped, to allow for direct comparison with TT-Seq.

In addition to characterising the genome-wide coverage, a next step was to compare the TT-Seq metagene in HeLa and HEK293 cells with other existing nascent RNA sequencing methods such as GRO-Seq (Laitem et al., 2015) or NET-Seq (Nojima et al., 2015), with their metagenes shown in Figure 2.2 A and B, respectively. In this context it is worth pointing out that for HeLa cells, the chosen 4sU-labelling time was 5 minutes, the same time as the original publication used (Schwalb B et al., 2016). However, for HEK293 cells, a 5 minute labelling time did not yield sufficient nascent RNA that could be carried forward to library preparation after the streptavidin column purification (data not shown).

One speculative reason for this is that given that HeLa cells are cancer cells, their metabolism might be optimised towards faster nucleotide metabolism and therefore

faster conversion of 4sU into the thiolabelled nucleotide that is incorporated during transcription. On the other hand HEK293 cells, which are embryonic kidney cells and don't have the same cancer-like properties, might have slower nucleotide metabolism than HeLa cells and therefore require a longer labelling time in order to generate sufficient 4sUTP from the 4sU that is added during the pulse labelling time.

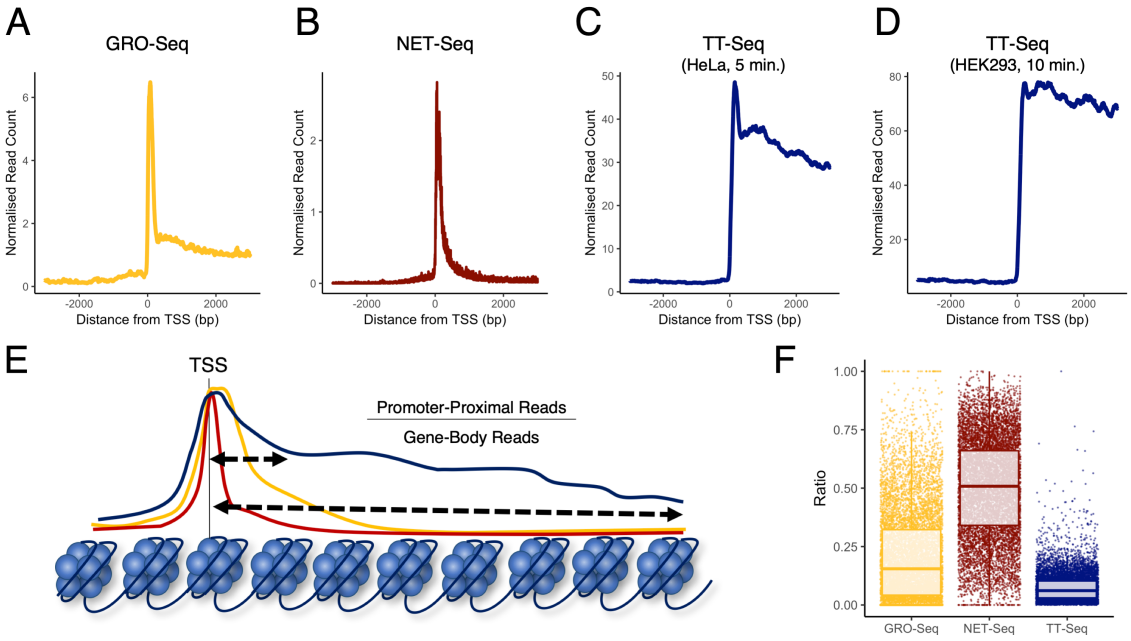


Figure 2.2: Metagenes and Positional Read Ratios

A. GRO-Seq metagene in HeLa cells. The sequencing dataset was generated in Laitem et al. (2015) and re-aligned to hg38. 9336 TSSs annotated previously (Brown et al., 2018) were used. B. NET-Seq metagene in HeLa cells. The sequencing dataset was generated in Nojima et al. (2015) and re-aligned to hg38. C. TT-Seq metagene in HeLa cells (from this study). D. TT-Seq metagene in HEK293 cells (from this study). E. Calculation of positional read ratios: The summed reads over the first 200 nucleotides for each gene are divided by the summed reads over the first 1000 nucleotides. F. Comparison of positional read ratios for GRO-Seq, NET-Seq, and TT-Seq (HeLa). P-values (Mann-Whitney test, corrected using the Bonferroni method) are smaller than 1 x 10^{-20}.

As shown in Figure 2.2 A - D, GRO-Seq and NET-Seq metagene profiles have much sharper peaks near the 5' end of genes, whereas TT-Seq has continuously high levels over the gene body with only a slight decrease after several hundreds of nucleotides. In this context it is worth pointing out that TT-Seq has a much lower resolution than NET-Seq so would not be able to capture the promoter-proximal pause at

the same precision that NET-Seq could. In order to quantify how much these methods capture nascent transcription over the gene body versus the promoter-proximal region, a positional read ratio was defined (see Figure 2.2 E), which calculates the ratios of reads near the TSS (first 200 nucleotides) over the reads over the gene body (first 1000 nucleotides) for each gene. As shown in Figure 2.2, this ratio is particularly high for NET-Seq, which is expected, and much lower for GRO-Seq (Mann-Whitney, Bonferroni-corrected p-value < 1×10^{-20}), and again much lower for TT-Seq (Mann-Whitney, Bonferroni-corrected p-value < 1×10^{-20}).

This speaks to the different nature of the various methods. With NET-Seq being based on the immunoprecipitation of RNA Pol II, it actually has higher signals at positions where RNA polymerase is slower, backtracked, or stalled, whereas for TT-Seq, the higher the 4sU incorporation, the higher the signal, reflecting the activity of the elongating RNA polymerase. This means that NET-Seq has a tendency to predominantly capture pausing events, whilst TT-Seq has much higher, and even, coverage over the entire gene body.

This is also reflected when looking at transcriptional profiles at individual loci on the IGV genome browser. Figure 2.3 shows the TT-Seq profiles for HeLa (5 minute labelling) and HEK293 cells (10 minute labelling) around the *ANRIL* locus, a long non-coding antisense transcript that overlaps the cell-cycle inhibitor coding genes *CDKN2A* and *CDKN2B*. For comparison, the corresponding NET-Seq and GRO-Seq profiles are shown (Laitem et al., 2015; Nojima et al., 2007)). In contrast to TT-Seq, transcription of *ANRIL* as measured by NET-Seq, is much lower than the neighbouring *CDKN2A* gene, with the TSS of *ANRIL* being located near the first intron-exon

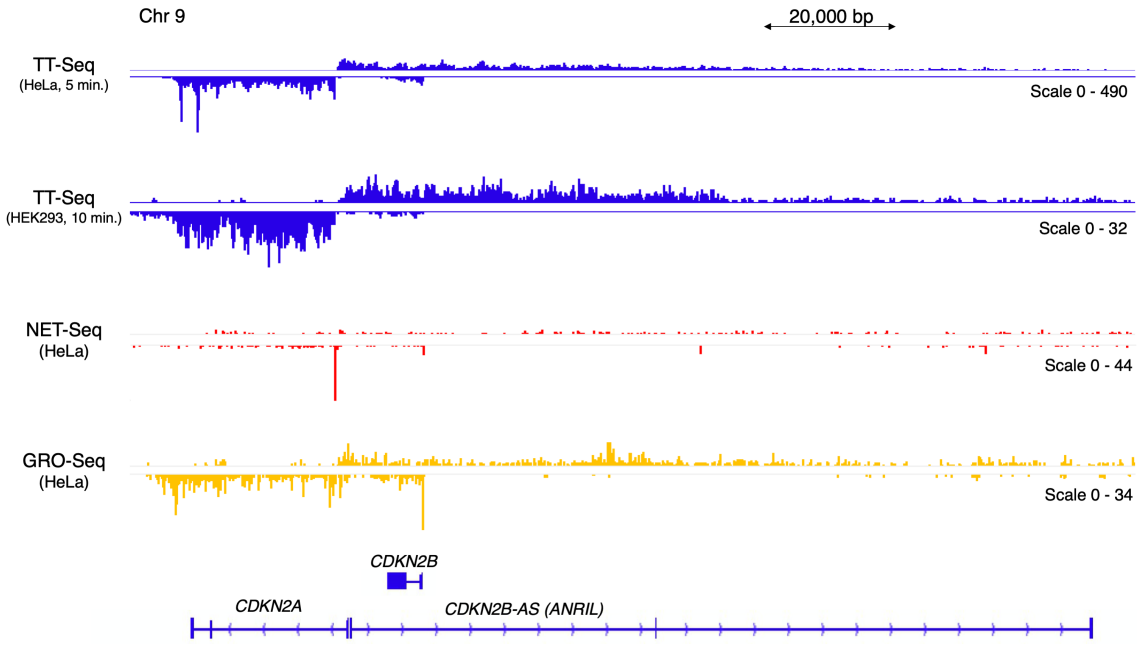


Figure 2.3: IGV Genome Browser Screenshot at the *ANRIL* Locus

TT-Seq profiles from HeLa cells (5 minute labelling) and HEK293 cells (10 minute labelling) are shown, in comparison to NET-Seq (HeLa, Nojima et al. (2007)), and GRO-Seq (HeLa, Laitem et al. (2015)) around the *ANRIL* locus on chromosome 9.

boundary of *CDKN2A* (Brown et al., 2018). Interestingly, *ANRIL* has been suggested to contribue towards *CDKN2A* and *CDKN2B* repression through recruitment of polycomb complexes (Yap et al., 2010). Crucially, TT-Seq levels in both HeLa and HEK293 are much higher for the *ANRIL* / *CDKN2B*-antisense transcript than *CDKN2B* itself, demonstrating the capacity of TT-Seq to capture nascent non-coding transcription, especially when compared to NET-Seq. This is discussed in more detail, and with a genome-wide comparison, in section 2.3.

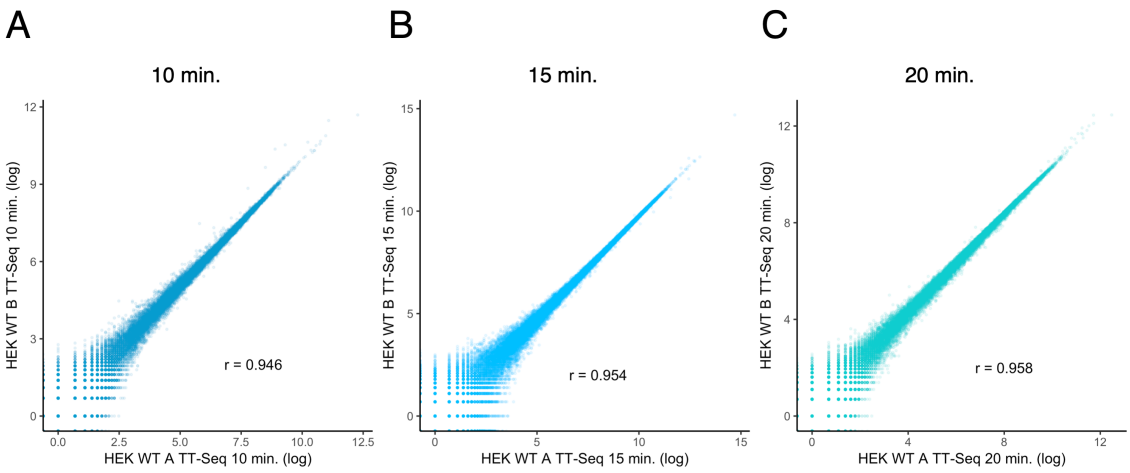
2.2 Nascent RNA Processing Dynamics and TT-Seq

One additional feature that can be observed in Figure 2.3 is that for *CDKN2A*, the last two exons seem to have much higher TT-Seq levels than the surrounding intronic regions in the HeLa profile. TT-Seq has a tendency to cover exonic reads more fre-

quently than intronic reads for some genes. One explanation for this is that, due to the co-transcriptional nature of splicing, as soon as just one thiolabelled nucleotide is contained in an exon that has been spliced into the mature mRNA, the entire corresponding fragment (after sonication) will contribue to the TT-Seq signals. If the processing has occurred early in the labelling time, this will enhance the effect even more.

In order to study this relationship further, an experiment was set up where HEK293 WT cells were labelled with 4sU for different pulse labelling periods: 10 minutes, 15 minutes, and 20 minutes. The TT-Seq protocol was applied subsequently for two biological repeats for each labelling time point. Usage of thio-labelled spike-in RNAs allows for calibration of the experiments, which is described in more detail in the methods section. The technical sequencing statistics of this set of experiments is summarised in Figure 2.4. The scatter plots of log-transformed read counts per gene between the two repeats for 10 minutes, 15 minutes, and 20 minutes are shown in Figure. 2.4 A, B, and C respectively.

The corresponding Spearman correlation coefficients are 0.946, 0.954, and 0.958 for the 10 minutes, 15 minutes, and 20 minutes labelling times, respectively. This points to a high level of reproducibility. Furthermore, this sequencing experiment shows high alignment percentages of between 78 % and 87 %, which is much higher than the typical NET-Seq or GRO-Seq alignment percentages (Laitem et al., 2015; Nojima et al., 2015). It is noteworthy that for this study, NET-Seq and GRO-Seq data were realigned to hg38 (the original reads were aligned to hg19), which yielded an alignment percentage of approximately 34 % for NET-Seq, and 56 % for GRO-Seq.



D

% Alignment to hg38	10 Min	15 Min	20 Min
Repeat A	84.3 %	82.5 %	85.0 %
Repeat B	78.0 %	81.4 %	86.6 %

Figure 2.4: TT-Seq Scatter Plots and Alignment Percentages

A. Scatter plot of log-transformed read counts for two repeats of 10-minute labelled TT-Seq experiment in HEK293 cells. B. Scatter plot of log-transformed read counts for two repeats of 15-minute labelled TT-Seq experiment in HEK293 cells. C. Scatter plot of log-transformed read counts for two repeats of 20-minute labelled TT-Seq experiment in HEK293 cells. D. Table displaying alignment percentages of TT-Seq experiments.

The high alignment percentages of TT-Seq also points to the relatively low background in contrast to NET-Seq, where high background signals are a concern (Wissink et al., 2019).

Having characterised the sequencing output and read count correlations between repeats, the next step was to look at the transcriptional profile of these different time-labelling samples and compare the coverage of exons and introns. Initially, this can be observed at individual loci, such as the *FBL* gene, which is shown in an IGV genome browser screenshot in Figure 2.5.

There is a strong difference in overall coverage of the reads that map to the first and penultimate introns in particular: Whilst the 10-minute labelled sample shows moderately high levels of transcription in the first intron, the penultimate intron in particular has a coverage at a similar level to the surrounding exons. This is strikingly

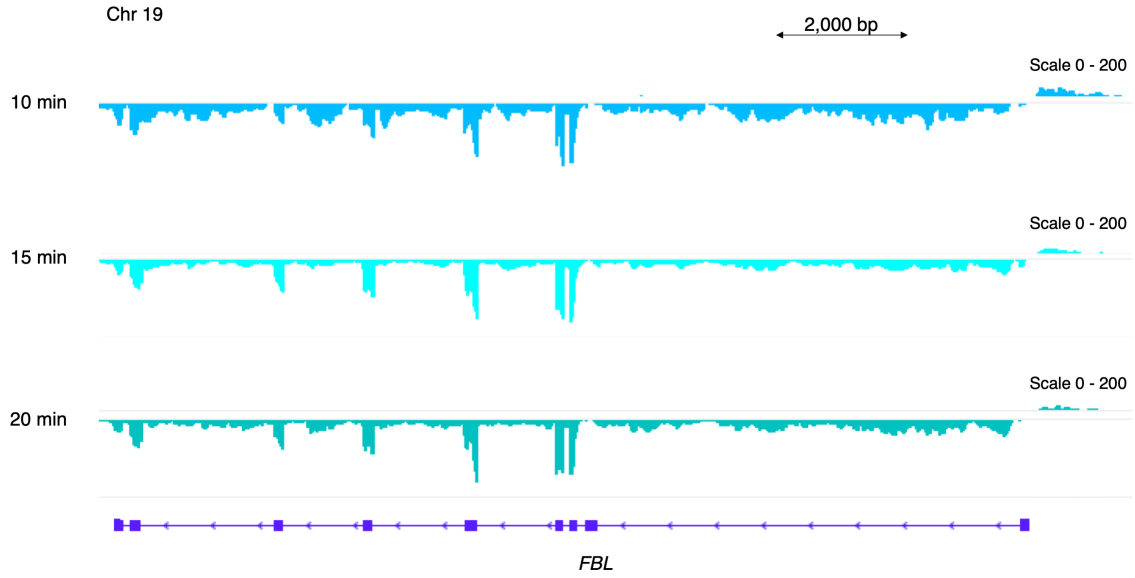


Figure 2.5: TT-Seq IGV Genome Browser Screenshot at the *FBL* Locus

TT-Seq profiles from HEK293 cells (10, 15, and 20 minute labelling times) at the *FBL* locus are shown.

different for the 15-minute and 20-minute labelled sample, where the discrepancy between the intronic and exonic reads is getting much greater.

In effect, the longer the labelling time, the more the TT-Seq profile resembles a fragmented RNA-Seq profile that only covers the exonic reads because of the nature of RNA-Seq, which is biased towards stable, processed transcripts as opposed to nascent, un-processed transcripts. The explanation for this is based on the co-transcriptionality of splicing (reviewed in Herzel et al. (2017)): As soon as just one labelled nucleotide is incorporated in an exonic part of the nascent RNA that is in the process of undergoing splicing, the entire (sonicated) RNA fragment, which may span several exons, counts towards the ultimate TT-Seq read. Given that the sonicated fragments are on average 1000-2000 bp long (Schwalb B et al., 2016), it is entirely possible that multiple exons are covered in a subset of the sonicated fragments.

In this context, it is important to point out that, whilst this feature of TT-Seq means

that splicing dynamics can be assessed given the changing intron coverage depending on the labelling time, it also demonstrates a weakness of TT-Seq: When monitoring nascent transcription, there is no reason to think that there is inherently higher levels of transcription occuring over exons as compared to introns, which is why TT-Seq may not give a real reflection of the transcription process. A truly nascent transcription profile would deliver a much more even coverage over the entire gene body, both including exons and introns, and not vary between genes depending on the speed of splicing.

Based on the TT-Seq profiles, one could, for example, conclude that *FBL* is spliced more quickly than *CDKN2A*, as shown in Figures 2.5 and 2.3, respectively. A nascent RNA sequencing method that truly reflects the act of nascent transcription would not have its profiles determined by the speed of splicing, however, and would show a more even or flat coverage for both genes, even if *FBL* is spliced more rapidly. This is why a novel nascent RNA sequencing method was developed, which makes up the majority of the work presented in this study, and is covered in chapters 2 - 5.

Having observed a noticeable difference in the intron-to-exon coverage from individual loci when using different labelling times in HEK293 cells, a next step was to perform a genome-wide analysis on this TT-Seq dataset. As shown in Figure 2.6 A, the metagenes for 20-minute 4sU-labelled HEK293 show the largest drop off of normalised reads at the first exon-to-intron boundary, followed by 15 minutes, and then 10 minutes.

It is worth pointing out that these metagenes were generated so that the coverage within each window (-100 to + 100 around each exon-intron boundary), is normalised

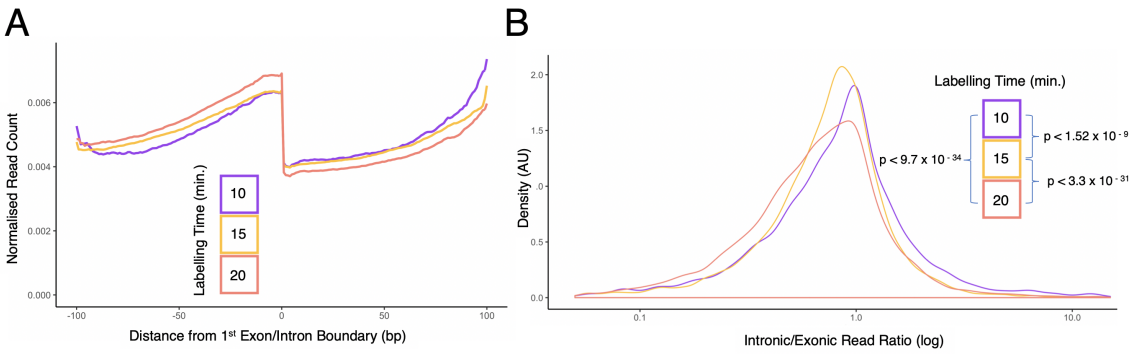


Figure 2.6: Genome-Wide RNA Processing Assessed by TT-Seq

A. TT-Seq metagenes for 10-minute, 15-minute, and 20-minute 4sU-labelled HEK293 cells. Metagenes are normalised so that each gene contributes the same towards the metagene in the given window (-100 to + 100 bp around each exon-intron boundary). 9336 exon-intron boundaries were utilised, which were annotated previously (Brown et al., 2018). B. Distribution of Intron/Exon read ratio (log scale) determined by the summed reads in 25-bp windows upstream and downstream of the first exon-intron boundary of 9336 genes, as annotated previously (Brown et al., 2018), for 10, 15, and 20 minute 4sU-labelled HEK293 cells. p-values were calculated using a Wilcoxon-rank sum test, and corrected using the Bonferroni method.

to a sum of 1. This means that each gene contributes to the metagene equally, and that all differences across and between the profiles are down to the within-gene TT-Seq-based change of the transcriptional coverage. As expected, and as indicated by the *FBL* example, the discrepancy of coverage at the exon-intron boundary is the largest for 20 minutes 4sU-labelled HEK293 cells.

In addition to a metagene-based visualisation demonstrating this behaviour, a numerical approach was considered too: An intron/exon read ratio was calculated for each gene, by dividing the summed TT-Seq reads of the 25 basepairs of the first intron by the summed TT-Seq reads of the last 25 basepairs of the first exon of each gene. These were log transformed and visualised in distribution plots in Figure 2.6 B. As can be seen here, the ratio is significantly higher for 10 minutes compared to 15 minutes labelled HEK293 cells (corrected p-value < 1.52×10^{-9}), and 20 minutes labelled HEK293 cells (corrected p-value < 9.7×10^{-34}).

One interesting feature of the first exon-intron boundary in the human genome is

the frequent presence of downstream antisense transcription start sites in very close proximity (Brown et al., 2018). Figure 2.7 shows the distance of downstream antisense transcription start sites from the first exon-intron boundary in a scatter plot (A) or a density plot (B). This shows that overall, a specific set of antisense transcirption start sites centers around the first exon-intron boundary, genome-wide. The annotations used for this analysis were originally generated in the hg19 version of the human genome, but for this analysis everything was transferred into the hg38 version, using the UCSC liftover tool. The median distance between these downstream antisense transcription start sites and their corresponding sense start sites is 166 basepairs, whereas the median distance between these antisense TSSs and the corresponding first exon-intron boundary is only 7 basepairs.

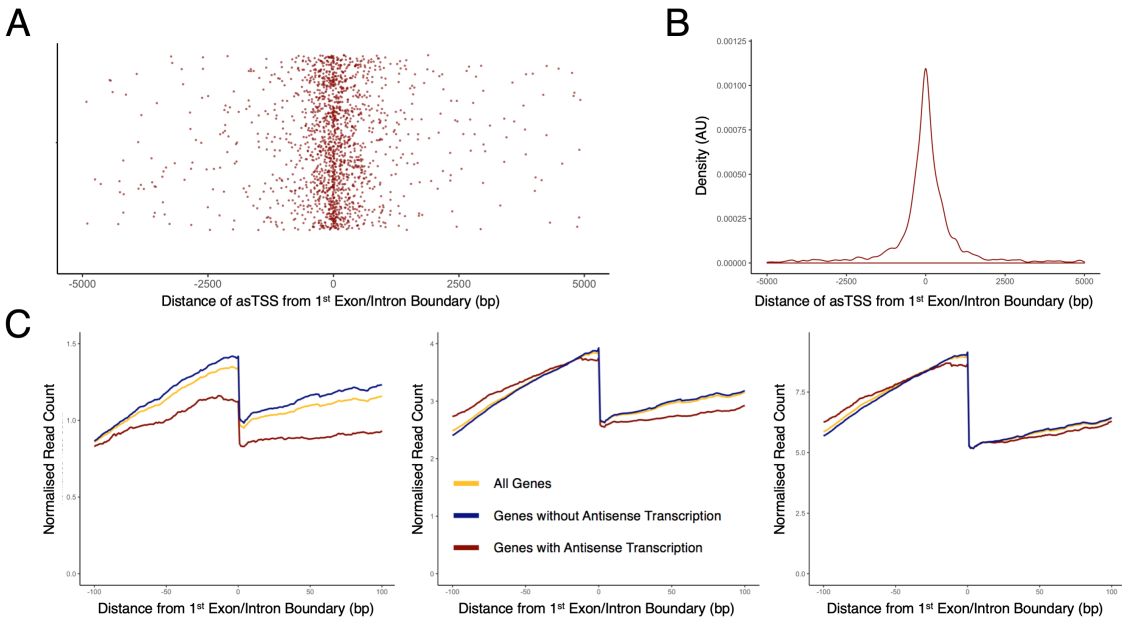


Figure 2.7: Genome-Wide RNA Processing Assessed by TT-Seq

A. Scatter plot of antisense TSSs around the first exon-intron boundary (n = 2546). B. Density plot of antisense TSSs around the first exon-intron boundary (n = 2546). C. TT-Seq metagene plots of 10 minute (left), 15 minute (middle), or 20 minute labelled HEK293 cells. Shown are all genes from the annotation set generated in Brown et al. (2018) in yellow, (n = 9336), genes without downstream antisense TSSs (n = 6790) in blue, and genes with downstream antisense TSSs (n = 2546) in red.

Crucially, genes that have downstream antisense transcription start sites also have

a higher intron/exon read ratio than genes without downstream antisense TSSs ($p < 3.74 \times 10^{-15}$). This difference is shown more clearly by looking at the TT-Seq metagenes shown in Figure 2.7 C, with the 10 minute (left), 15 minute (middle), or 20 minute labelled HEK293 cells. The 10-minute labelled TT-Seq metagene shows the strongest discrepancy in the relative intron-to-exon coverage between genes with and without antisense transcription, whereas this difference disappears with longer labelling times. This is expected since longer labelling times lead to TT-Seq profiles resembling fragmented RNA-Seq profiles more and more, and therefore predominantly cover exonic reads. The strikingly different TT-Seq profile of genes with antisense start sites for the lower labelling time would strongly suggest that having an antisense start site near the first exon-intron boundary coincides with a slower rate of splicing. However, it is important to bear in mind that these metagene profile changes along with higher intron/exon read ratios do not prove causality, just an association.

Overall, TT-Seq provides a way to detect and analyse the nascent RNA processing dynamics during transcription. The different coverage of intronic versus exonic regions is both a feature that is useful for this purpose, but also an inherent drawback of this technique, because it varies depending on the labelling time, and a truly nascent transcriptional profile should be more even over the gene body, not significantly higher over exonic regions.

2.3 4sU-based Metabolic Labelling is Highly Suitable for Measuring Non-coding Transcription

As discussed above, and described in the original publication of TT-Seq for K562 cells, 4sU-labelling-based methods are suitable for measuring nascent transcription of non-coding RNAs (Schwalb B et al., 2016). Using the downstream antisense transcription start sites located near the first exon-intron boundary (Brown et al., 2018), as well as antisense TSSs assigned in the ENCODE database, a comparison to sense TSSs transcription is shown with TT-Seq metagenes in Figure 2.8 A. Note that this is using HeLa TT-Seq data generated in this study rather than HEK293 cells. As can be seen from the metagenes, the average antisense transcription levels from TSSs annotated in Brown et al. (2018) have similar levels to sense transcription, whereas transcription emanating from antisense TSSs from the ENCODE database have much lower average levels.

In order to see if the similar order of magnitude of transcription levels between sense and antisense TSSs holds true, genome-wide heatmaps were generated and are shown in Figure 2.8 B. This shows that the observed similar average levels of transcription between sense and antisense start sites is not due to subsets of genes or outliers, but because of a genome-wide behaviour. This is crucial because it shows that levels of transcription can be high for non-coding transcripts, even if they may not be stable enough to appear to a significant degree in a standard RNA library pool.

The next aim was to test the unique suitability of TT-Seq to capture nascent transcription of these non-coding transcripts exemplified by the set of antisense transcripts

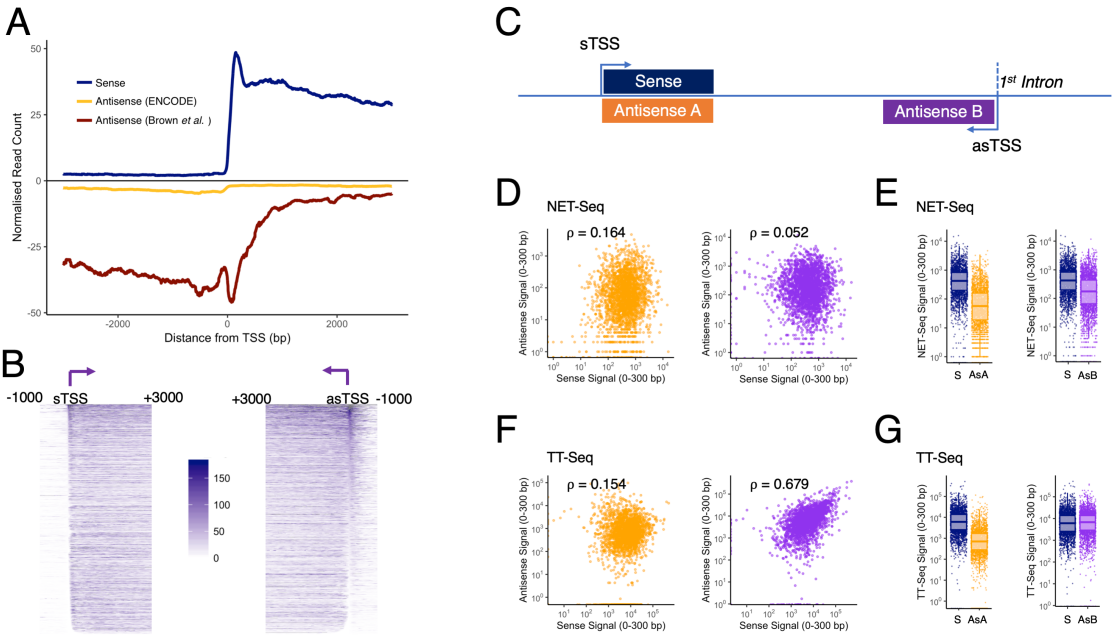


Figure 2.8: Genome-wide Assessment of Antisense Transcription by TT-Seq and NET-Seq

A. TT-Seq metagenes (HeLa WT cells, 5-minute labelling) representing transcription from sense (n = 9336), ENCODE antisense TSSs (n = 5701) and antisense TSSs annotated by Brown et al. (2018) (n = 2546). B. TT-Seq heatmaps (HeLa WT cells, 5-minute labelling) around sense and antisense TSSs (Brown et al., 2018). C Layout of transcriptomics comparison between NET-Seq and TT-Seq for sense and antisense transcription. Read sums in 300-basepair long windows were determined downstream of the sense TSS on the forward and reverse strand (labelled 'Sense' or 'Antisense A', respectively), and a 300-basepair long window on the reverse strand following the downstream antisense TSSs determined by (Brown et al., 2018) located near the first exon-intron boundary (labelled 'Antisense B'). D NET-Seq correlation analysis. Scatter plots are shown between 'Sense' and 'Antisense A' (left) or 'Sense' and 'Antisense B' (right). Spearman correlation coefficients (rho) are indicated. E. Boxplots of NET-Seq levels are shown comparing 'Sense' with 'Antisense A' (left) or with 'Antisense B'. For both comparisons, Wilcoxon-rank sum tests were performed. Bonferroni-corrected p-values < 1.6×10^{-16}. F. TT-Seq correlation analysis. Scatter plots are shown between 'Sense' and 'Antisense A' (left) or 'Sense' and 'Antisense B' (right). Spearman correlation coefficients (rho) are indicated. G. Boxplots of TT-Seq levels are shown comparing 'Sense' with 'Antisense A' (left) or with 'Antisense B'. For both comparisons, Wilcoxon-rank sum tests were performed. Bonferroni-corrected p-value < 1.6×10^{-16} for the comparison with 'Antisense A' and 0.374 for the comparison with 'Antisense B'.

annotated by Brown et al. (2018) as compared to NET-Seq, which was performed as follows: The summed sequencing reads on the sense strand (first 300 basepairs) was compared to the summed reads in the same window but on the opposite strand ('Antisense A'), or the summed reads in a 300-basepair-window on the opposite strand following the downstream antisense TSSs annotated by Brown et al. (2018), as shown in Figure 2.8 C. Note that for this analysis, HeLa sequencing data for NET-Seq (Nojima et al., 2015) and TT-Seq (5 min. labelling, this study) was used to avoid comparison

between different cell lines. In a way, one could consider the 'Antisense A' window as a negative control that one would expect to have lower levels of nascent transcription. A comparison with GRO-Seq has not been done, because of 35 % of the 'Antisense B' windows having zero reads.

Figure 2.8 D and E show that the NET-Seq signal in the 'Antisense A' or 'Antisense B' regions do not correlate with the sense transcription signal (Spearman rho equals 0.164 and 0.052, respectively). Further to the lack of correlation, the median NET-Seq signal of both of the antisense regions is almost an entire order of magnitue lower than the corresponding sense signal (Bonferroni-corrected p-values < 1.6×10^{-16}).

However, when doing the same comparison using TT-Seq data, a significant difference can be observed. The lack of any correlation between sense and antisense transcription in the 'Antisense A' region remains (Spearman rho equals 0.154), but there is a noticeable correlation between sense transcription and the antisense transcription that arises from the downstream asTSSs annotated by Brown et al. (2018) when measured by TT-Seq (Spearman rho = 0.679). This is depicted in Figure 2.8 F. Most strikingly, however, is the fact that there is no significant difference between the median levels of sense transcription and the antisense transcription emanating from the downstream asTSSs ('Antisense B'), as shown in Figure 2.8 G. Whilst transcription measured by TT-Seq in the 'Antisense A' region remains significantly lower compared to sense transcription (Bonferroni-corrected p-value < 1.6×10^{-16}), the median transcription levels from the 'Antisense B' region is not different to sense transcription (Bonferroni-corrected p-avlue = 0.374), with median summed read counts of 6024 and 6616 for the sense transcription and transcription from the 'Antisense B' window,

respectively.

This strong difference in how NET-Seq and TT-Seq are able to measure nascent non-coding transcription, as exemplified by antisense transcription from downstream antisense TSSs, shows that there are a few technical aspects to consider when performing genome-wide measurements of the nascent transcriptome: It appears that based on this work, as well as the original study on TT-Seq (Schwalb B et al., 2016), metabolic labelling using 4-thio-uridine followed by biotin-streptavidin enrichment is a highly sensitive approach that may well capture some of the more unstable nascent transcripts. NET-Seq, on the other hand, has an immunoprecipitation step that may lead to higher background levels which may reduce coverage of these non-coding nascent transcripts (Wissink et al., 2019).

An alternative explanation for this discrepancy regarding the coverage of these antisense transcripts would be the possibility that these are not actually transcribed by RNA Pol II but by a different polymerase. NET-Seq, because of the usage of antibodies specific to RNA Pol II, would not be able to capture transcripts that are produced from other polymerase (unless it is part of the background), whereas the 4sU-labelling approach of TT-Seq does not discriminate between different polymerases.

In order to test this hypothesis, TT-Seq was performed with and without alpha-amanitin treatment in Hela cells. Transcription of Pol II target genes is 100 times more sensitive to alpha-amanitin than transcription from Pol III target genes, while Pol I target genes remain unaffected (Bensaude, 2011). Therefore, a concentration was chosen that has been shown to inhibit Pol II transcription but not Pol III transcription (Bensaude, 2011). After treatment, TT-Seq was performed with thiolabelled spike-in

to allow for genome-wide comparison. Because of the usage of spike-in, a global comparison of genome-wide TT-Seq levels can be made, the results of which are shown in metagenes in Figure 2.9 for sense TSSs (A) and antisense TSSs (B).

A surprising observation was made when looking at the spike-in calibrated metagenes for TT-Seq around sense TSSs: It seems that treatment with alpha-amanitin may still allow RNA Pol II to initiate transcription but inhibits elongation much more strongly, given the sharp peak of the metagene curve that corresponds to the sample that underwent alpha-amanitin treatment compared to the untreated control. This would raise the possibility that the initiating version of RNA Pol II is more resistant to alpha-amanitin than the elongating version.

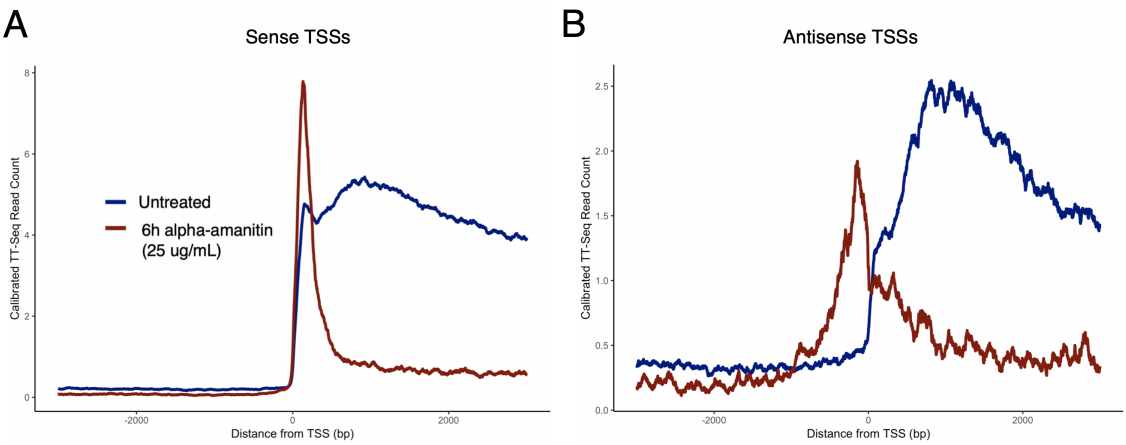


Figure 2.9: TT-Seq of HeLa Cells with or without alpha-amanitin Treatment

TT-Seq of HeLa Cells with or without alpha-amanitin Treatment around sense TSSs as annotated by Brown et al. (2018). TT-Seq of HeLa Cells with or withou alpha-amanitin Treatment around antisense TSSs as annotated by Brown et al. (2018).

On the other hand, the metagenes around antisense TSSs (annotated by Brown et al. (2018)), are less conclusive. The curve representing the alpha-amanitin treated sample has a sharper peak, again, but the position of the peak shifts slightly further upstream, whilst coverage further downstream is much lower compared to the curve

of the untreated control. Therefore, a statistical comparison analogous to what has been done in Figure 2.8 was not possible. These observations would point to the conclusion that the hypothesis of potential Pol III transcription of these antisense TSSs is likely wrong - otherwise the metagenes of the treated and untreated samples around antisense TSSs would look more alike. A less likely, but still possible explanation was that the concentration of alpha-amanitin used to inhibit Pol II but not Pol III, as adopted from Bensaude (2011), still inhibited significant proportions of Pol III complexes. This possibility, albeit unlikely, cannot be fully dismissed.

However, this experiment has still given useful insights: First, the observation that TT-Seq still shows elevated levels at the 5' end but does not give a signal over the gene body after alpha-amanitin treatment is an interesting observation that will be used further in a later chapter of this study, concerned with genome-wide analysis and computational approaches to investigate promoter-proximal pausing. Second, this experiment provided an additional control, in that the alpha-amanitin treated sample had 90 % less nascent RNA than the untreated control (5 ng vs 50 ng). If the alpha-amanitin treated sample had a similar amount of nascent RNA after the streptavidin column purification, that would be an indicator for contamination issues with non-nascent RNAs, which does not seem to be the case. 4sU-labelling followed by biotinylation and streptavidin-column purification therefore seems to provide a viable way of obtaining nascent transcripts for next-generation sequencing.

To summarise, 4sU-based labelling provides a sensible way to measure non-coding transcription. However, TT-Seq, as discussed has proven disadvantages, most importantly the lack of resolution and the uneven coverage over exons and intron. The

next step in this study is therefore to develop a method that retains the sensitivity of TT-Seq but drastically increases resolution.

Chapter 3

Development of SNU-Seq

3.1 Motivation for SNU-Seq

As discussed above, the various methods aiming to measure nascent transcription each have a set of benefits and considerations or disadvantages. In addition to the efficiency of the techniques, the expense of time or resources as well as the difficulty of carrying these out are important considerations too. NET-Seq and PRO-Seq are not considered to be very straightforward methods and require a lot of time and resources. Furthermore, NET-Seq has alignment percentages of often 50 % or less, potentially due to high background levels caused by ribosomal or other RNAs binding to RNA Pol II (Nojima et al., 2015; Wissink et al., 2019). On the other hand, NET-Seq and PRO-Seq have single basepair resolution, which TT-Seq, a method that is much easier to carry out and shows high efficiency, does not have.

In order to bridge this divide between these nascent RNA sequencing methodologies, this study aims to find a way to combine the high resolution of NET-Seq or PRO-Seq with the simpler TT-Seq workflow. In effect, this can be achieved by developing a method that turns TT-Seq into a high-resolution method, which was accomplished with SNU-Seq (single-nucleotide resolution 4sU-Seq), as described and characterised below.

3.2 SNU-Seq Workflow

Similar to 4sU-Seq methods, SNU-Seq is based on 4sU-labelling of cells for 10 minutes followed by purification of nascent RNA using biotin-labelling and streptavidin pull-down. However, in order to generate single-nucleotide resolution, an artificial

poly-A-tail is added before library preparation using *E. coli* poly-(A)-polymerase. The artificial poly-(A)-tail allows for library preparation with oligo-dT primers. This means that positional information about the last incorporated nucleotide during the pulse labelling window is retained. In contrast to TT-Seq (Schwalb *et al.*, 2016), SNU-Seq does not incorporate a sonication step after total RNA purification. This is shown in Figure 3.1. A 10-minute labelling time was chosen because of the same reasons 4sU-labelling of 10 minutes was performed for TT-Seq in HEK293 cells: As discussed in the previous chapter, HEK293 cells seem to need a longer labelling time than HeLa cells for example, which may be due to a slower nucleotide metabolism as 4-thiouridine needs to be converted into 4s-UTP in order to be able to be incorporated into nascent RNA as a nucleotide.

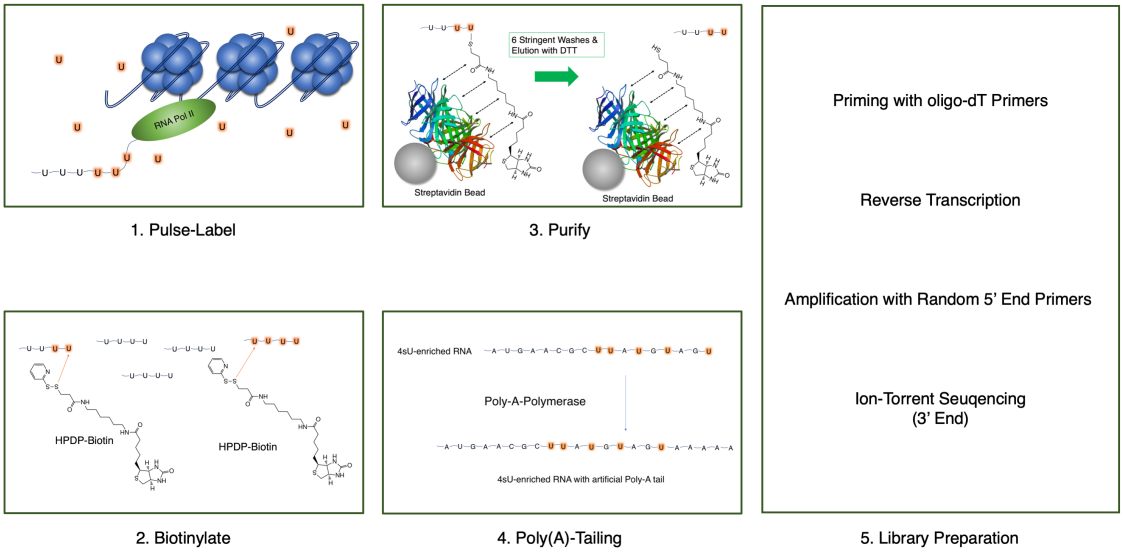


Figure 3.1: SNU-Seq Workflow

The SNU-Seq workflow is shown: This includes thiolabelling *in vivo* followed by RNA purification without sonication The RNA is then biotinylated with HPDP-biotin, which specifically forms a covalent bond to thio-labelled RNA. This RNA is then purified using strepdavidin columns. An artificial poly-A tail is added before performing strand-specific library preparation using oligo-dT primers. Library preparation steps include reverse transcription and amplification with random primers. Libraries are sequenced from the 3' end on an ion torrent sequencer. Thiolabelled nucleotides are indicated in orange.

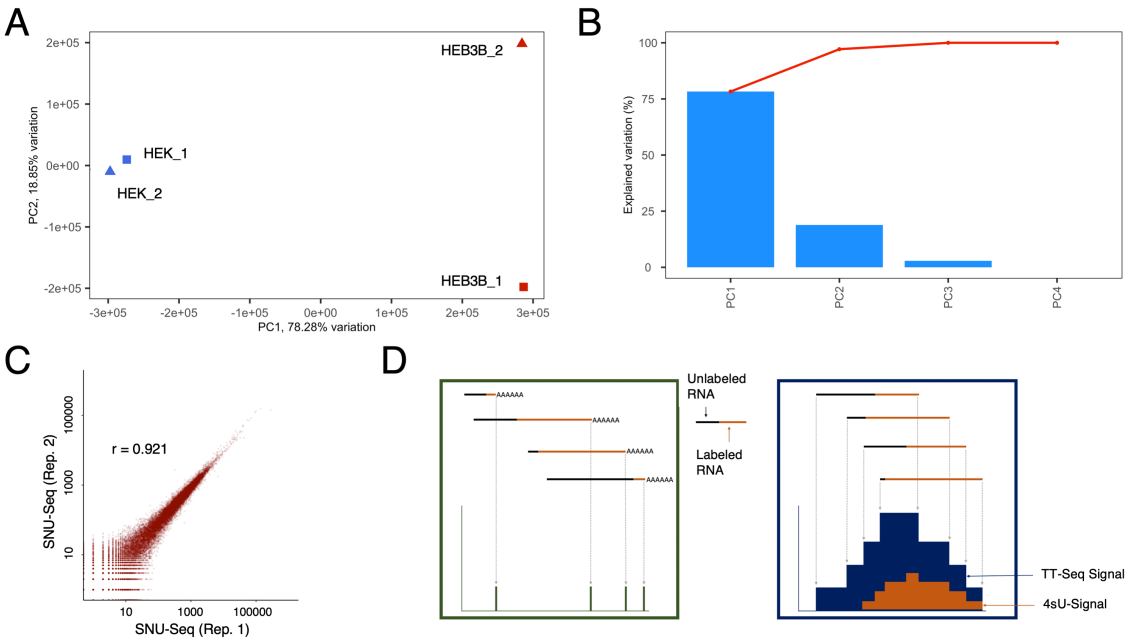


Figure 3.2: SNU-Seq Technical Details

A. Principal component analysis of SNU-Seq in HEK293 and HEP3B cells with two repeats each. The first (x-axis) and second (y-axis) principal components are plotted against each other in a biplot. B. Scree plot of principal component analysis of SNU-Seq in HEK293 and HEP3B cells with two repeats each, showing the individual variance contribution (bar graph) as well as the cumulative variance (red line). C. Scatter plot of SNU-Seq repeats read counts (log scale, HEK293 cells). The pearson correlation coefficient is shown (r = 0.921). D. Comparison between SNU-Seq (left, green) and TT-Seq (right, blue) with respect to how their signals are generated.

To demonstrate the ability to perform SNU-Seq across different mammalian cell lines, the method was performed in HEK293 as well as HEP3B cells with two biological repeats each. The corresponding principal component analysis along with their scree plot (individual and cumulative variance) for the two repeats in each cell line is shown in Figure 3.2 A and B. The first two principal components, as indicated by the scree plot, make up more than 95 % of the variation, so the biplot of first and second component is sufficient as a way to examine the closeness of the two biological repeats in the different cell lines.

It is noteworthy that, while the HEB3B repeats are close to each other in the first principal component of the PCA biplot, they are a lot further apart in the second component. However, as can be seen in the scree plot, the vast majority of the variation

of read counts is based in the first component (more than 75 %).

The alignment percentage of approximately 65% is high compared to other nascent RNA-sequencing methods and biological repeats correlate well. The repeatability of SNU-Seq is also confirmed by correlating the read counts between two repeats, yielding a correlation coefficient of 0.921, with the corresponding scatter plot shown in Figure 3.2 C.

For SNU-Seq, the eventual signal is derived predominantly from the act of nucleotide incorporation. Unlike TT-Seq, however, SNU-Seq not only has near single-basepair resolution but also completely eliminates any 5' bias that TT-Seq still has, despite fragmentation. This is displayed in Figure 3.2 D, showing that for TT-Seq, parts of the eventual signal are not actually 4sU-labelled, whereas for SNU-Seq, every signal has to come from the last incorporated nucleotide in a part of the RNA that contains the thiolabel.

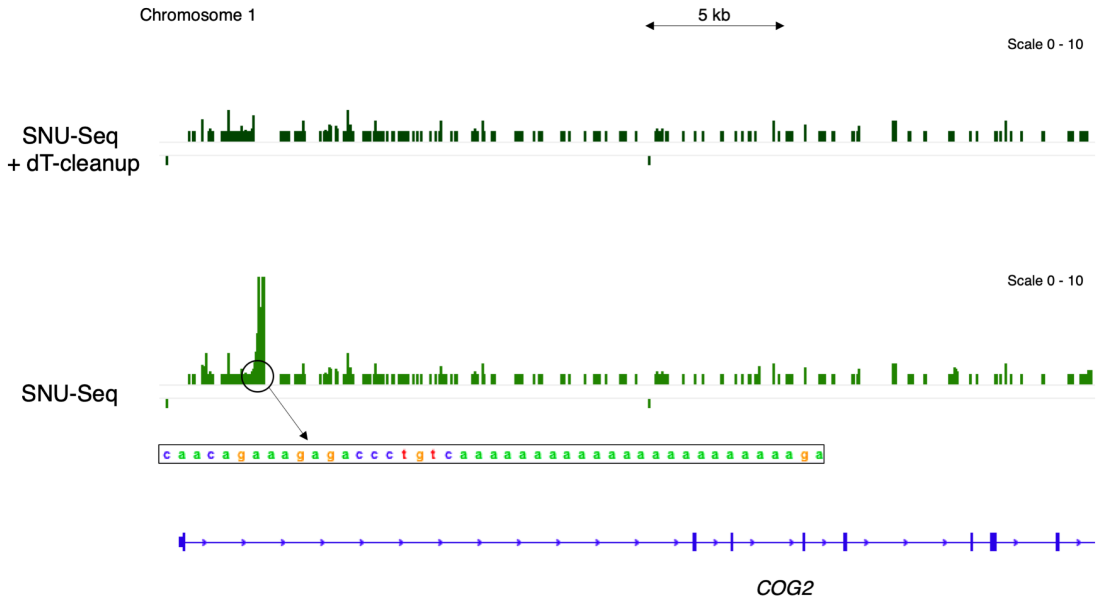


Figure 3.3: Oligo-dT Cleanup of SNU-Seq

IGV screenshots of SNU-Seq with and without dT-cleanup are shown around the *COG2* locus. The DNA sequence of the forward strand surrounding the peak in SNU-Seq signal is indicated.

One crucial aspect about SNU-Seq is the possible, albeit rare, occurrence of poly(A) stretches within genes. During nascent transcription, RNAs containing these sequences may be transcribed and therefore be part of the pool of nascent RNAs. Because the library preparation is based on oligo-dT primers, internal priming may occur, generating signals that do not reflect the true 3'end of nascent transcripts.

In order to address this, a bioinformatic poly(A) clean-up step was introduced utilising peak-calling and sequence-based filtering on the forward and reverse strand followed by signal removal, which is explained in more detail in the methods section. The result of this is shown in Figure 3.3, exemplified on the *COG2* locus. Going forward, SNU-Seq with oligo-dT cleanup will simply be referred to as SNU-Seq.

Having successfully generated and sequenced SNU-Seq libraries, a next step was to compare the SNU-Seq profiles to that of other methods, namely NET-Seq (Nojima et al. (2015), HeLa), TT-Seq (this study, HEK293), and PRO-Seq (Woo et al. (2018), HEK293). This comparison is displayed in Figure 3.4 for a larger scale, and in Figure 3.5 for a smaller scale.

Different scales were chosen in order to be able to distinguish different profile shapes across the various genes. A normalisation based on number of reads across these four methods was considered but ultimately rejected, because NET-Seq only measures transcripts from RNA Pol II (whereas the other methods measure all transcripts), plus TT-Seq has an rRNA removal step, leading to the different methods representing different pools of nascent transcripts which would make a calibration based on raw read number a problematic approach.

The higher PRO-Seq density can therefore be explained by the high read number

(204 million reads, Woo et al. (2018); compared to 105 million reads for SNU-Seq, 165 million reads for TT-Seq, and approximately 182 million reads for NET-Seq (Nojima et al., 2007)).

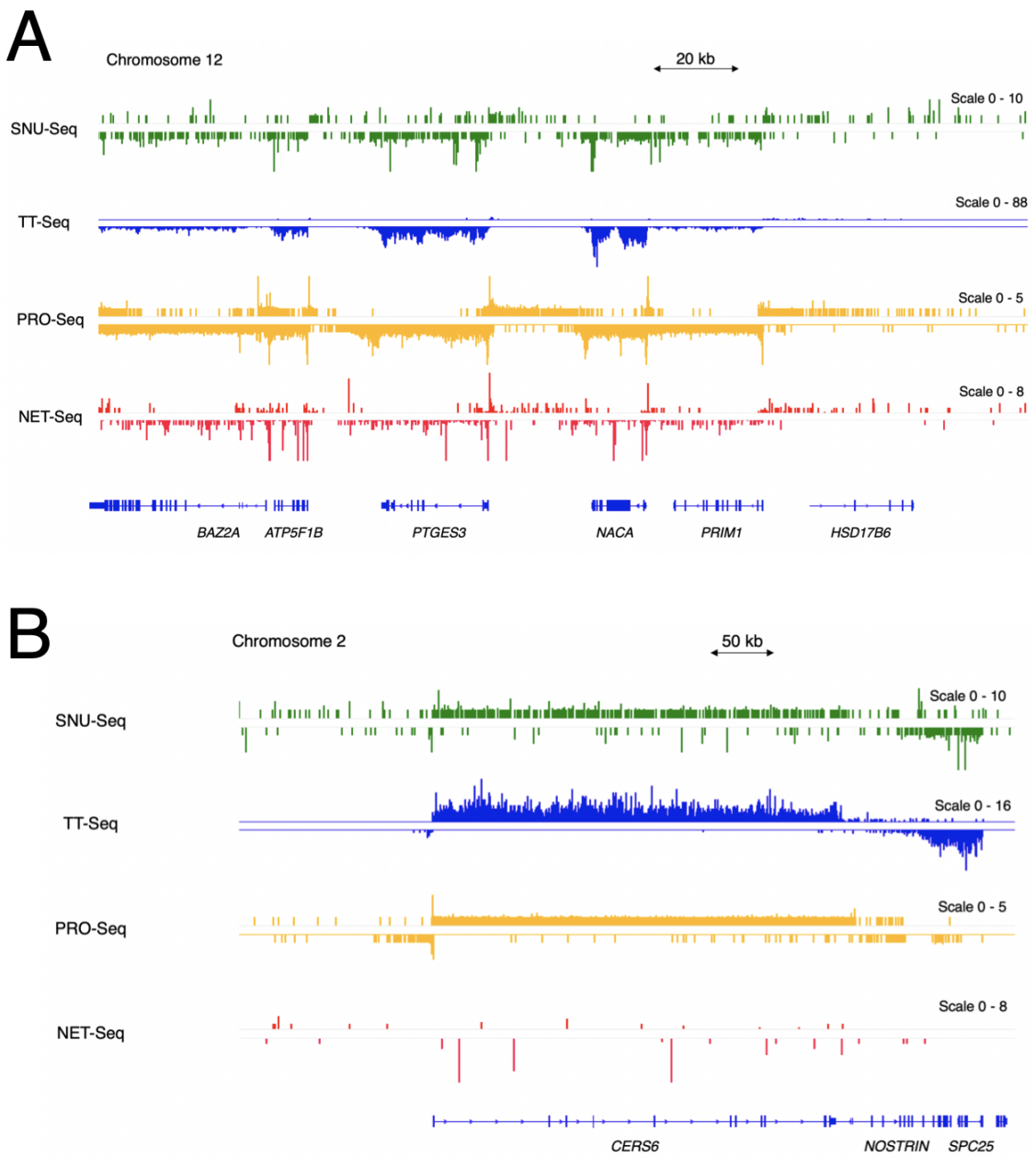


Figure 3.4: Transcription Profiles of SNU-Seq, TT-Seq, PRO-Seq, and NET-Seq (Large Scale)

A. IGV screenshot displaying a comparison of SNU-Seq (HEK293, this study), NET-Seq (Nojima et al. (2015), HeLa), TT-Seq (this study, HEK293), and PRO-Seq (Woo et al. (2018), HEK293) on a region of chromosome 12. B. IGV screenshot displaying a comparison of SNU-Seq (HEK293, this study), NET-Seq (Nojima et al. (2015), HeLa), TT-Seq (this study, HEK293), and PRO-Seq (Woo et al. (2018), HEK293) on a region of chromosome 2.

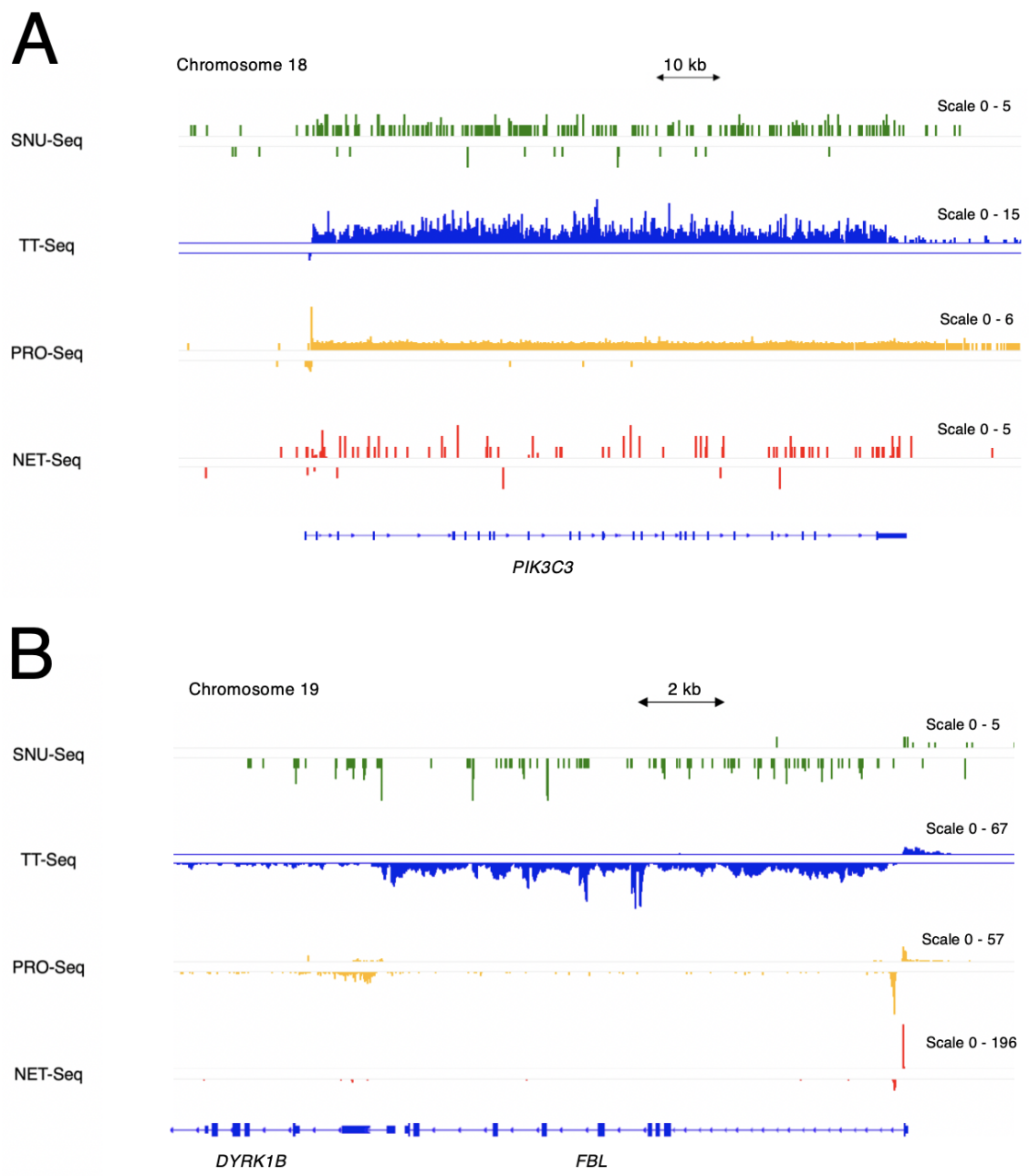


Figure 3.5: Transcription Profiles of SNU-Seq, TT-Seq, PRO-Seq, and NET-Seq (Small Scale)

A. IGV screenshot displaying a comparison of SNU-Seq (HEK293, this study), NET-Seq (Nojima et al. (2015), HeLa), TT-Seq (this study, HEK293), and PRO-Seq (Woo et al. (2018), HEK293) around the *PIK3C3* locus. B. IGV screenshot displaying a comparison of SNU-Seq (HEK293, this study), NET-Seq (Nojima et al. (2015), HeLa), TT-Seq (this study, HEK293), and PRO-Seq (Woo et al. (2018), HEK293) around the *FBL* locus.

As can be seen from these IGV screenshots, NET-Seq profiles generally show distinct, sparse, and sharp peaks, often at the 5' end of genes that may represent the promoter proximal pause. This could be explained by the fact that the pull-down

during the immunoprecipitation is more likely to occur when RNA polymerase is stalled or transcribes slowly. However, additional peaks occur within the gene body as well, which could be the result of additional pausing or backtracking events occurring during transcription. PRO-Seq, on the other hand, does have promoter-proximal peaks too, but also shows a dense and slightly more even coverage over most gene bodies as well.

TT-Seq, as discussed above, does not have the resolution to display the promoter proximal pause but shows dense coverage over gene bodies. SNU-Seq, shows some peaks at the 5' end or the 3' end of genes, which could represent the pause as well as polyadenylation processes, respectively. It also shows a very uniform, and even coverage over gene bodies, like PRO-Seq and unlike NET-Seq. The flat profile with most of the signals having a value of 1 could reflect a probability density distribution over the gene body, where each 'hit' represents one nucleotide incorporation from the nascent transcription process during the pulse labelling. NET-Seq misses some of the non-coding transcription events, such as the antisense transcription emanating from the *CERS6* promoter, as shown in Figure 3.4 B. Additionally, PRO-seq reads for transcription over *SPC25* (Figure 3.4 B) are markedly different from those obtained using TT-seq or SNU-seq when compared to the reads over *CERS6*, suggesting that there are additional, more subtle gene-specific differences in these techniques that remain to be uncovered.

Overall, it seems like there is greater variability in the TT-Seq signal than the SNU-Seq or PRO-Seq signal when comparing the overall levels of nascent transcription between different genes, such as in Figure 3.4 A. This is potentially a consequence

of the different natures of the methods, with TT-Seq having a bias towards exonic coverage so may not exclusively reflect nascent transcription, which, based on the PRO-Seq or SNU-Seq profiles, seems to be more even and similar between the genes depicted in the section of chromosome 12 in Figure 3.4 A.

Similar conclusions can be made when observing the IGV screenshots on a smaller scale (3.5). Interestingly, the *FBL* locus, which was discussed in the previous chapter as an example for TT-Seq bias towards exonic reads in some instances. Crucially, the SNU-Seq signal at the *FBL* does not seem to vary depending whether the read originated from an intron or an exon.

In terms of drawing conclusions regarding benefits and disadvantages of each techniques, one can see that PRO-Seq and SNU-Seq profiles are more alike to each other than the other two methods (NET-Seq and TT-Seq). However, PRO-Seq, as discussed above, is expensive and hard to carry out experimentally, whilst also not reflecting *in vivo* transcription because of the requirement to carry out a run-on in isolated nuclei or in permeabilised cells. These disadvantages may not obviously manifest themselves in the resulting transcriptional profiles, but are still important considerations.

3.3 SNU-Seq at the 5' and 3' Ends of Genes

In addition to comparing the profiles across individual genes or loci, a metagene analysis on the SNU-Seq data in HEK293 cells was performed, with the results displayed in Figure 3.6 A and B. For the generation of metagenes, the average read count at each basepair position between 1000 bp upstream and 2000 bp downstream of each

TSS (annotations from Brown et al. (2018)) was calculated after removing the top and bottom 5 % of reads at each position and normalising each gene entry to have a read sum equal to 1. This ensures that each gene contributes equally to the metagene. A more detailed explanation of the bioinformatic analysis of SNU-Seq data and the downstream computational procedures is provided in the methods section.

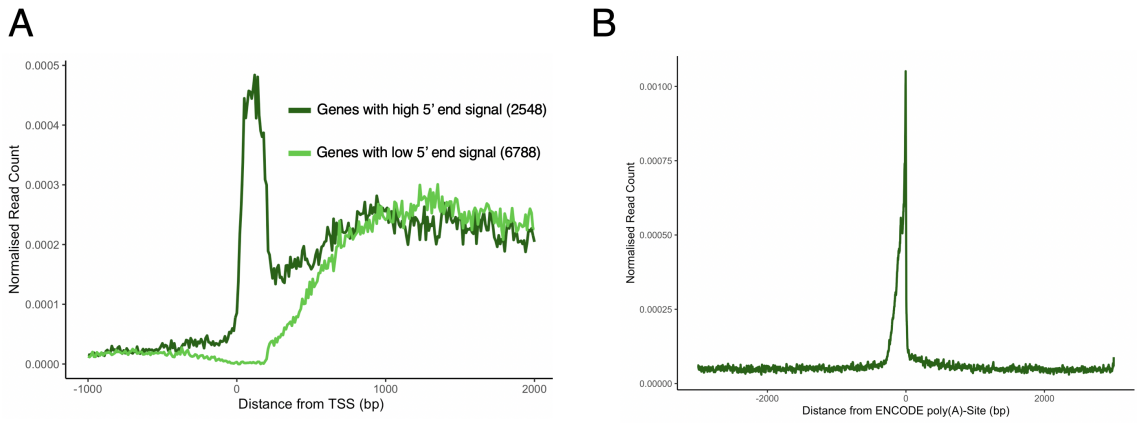


Figure 3.6: Metagene Analysis of SNU-Seq Approach

A. A metagene plot of SNU-Seq reads around transcription start sites (annotations from Brown et al. (2018)) are shown. The genes were split up based on whether there was a considerable number of reads in the first 150 basepairs (> 5 % of reads in the window) or not. B. SNU-Seq metagene around ENCODE poly(A)-sites (n = 9336). If multiple poly(A)-sites are known, the most proximal option was chosen.

As shown in Figure 3.6 A, SNU-Seq profiles have two distinct shapes. The clear bifurcation of SNU-Seq signals shows either genes with very high or very low read numbers in the very 5' region of genes. This could potentially represent a promoter-proximal pausing signal or lack thereof, respectively. Crucially, however, the average SNU-Seq signal within the gene body is comparable between the two sets of genes, independent of the SNU-Seq signal in the 5' end of either group.

This would strongly suggest that the relative levels of promoter-proximal pausing is not an indicator for the average nascent transcription levels in the gene body, or nascent elongating transcription in general. In fact, there is no correlation between

the SNU-Seq levels within the first 150 bp and within the gene body (500 - 700 bp downstream of the TSS), as the Spearman correlation coefficient between these levels across 9336 genes is only 0.063. One explanation for the low read count at the 5' end of a significant proportion of genes is not clear but could represent a technical problem with SNU-Seq where during library preparation the RT step does not reach the 5' region or could arise because the usage of random primer at the 5' end may not allow for amplification of shorter fragments without sequences represented in the primer pool. This is further discussed below in a discussion about SNU-Seq, and addressed by advancing the SNU-Seq methodology in chapter 4. Nevertheless, SNU-seq reveals a signal expected of a promoter proximal pause at some genes and this is investigated further in chapter 5.

Due to the fact that the library preparation for SNU-Seq is based on oligo-dT primers, SNU-Seq is able to detect poly(A)-sites of nascent transcripts. Figure 3.6 B shows the metagene of normalised SNU-Seq reads around ENCODE poly(A) sites. It appears that the singular sharp peak observed in this metagene would suggest that the precise position of previously annotated poly(A)-sites in the ENCODE database can be captured with SNU-Seq with high accuracy. This is because in the SNU-Seq libraries used for this analysis, no poly(A)-tail removal has been done before the library preparation, so fully synthesised transcripts with a 4sU label that are polyadenylated were part of the library and therefore highly likely to be efficiently amplified with oligo-dT primers.

In addition to singular poly(A)-sites, some genes undergo alternative polyadenylation events, one of which is *SSR3*, as described previously (Fischl et al., 2019). Figure

3.7 shows the SNU-Seq and TT-Seq profiles around the *SSR3* locus on chromosome 3, demonstrating that SNU-Seq can clearly pick up the two alternative poly(A)-sites. Being able to detect two distinct poly(A)-sites is much more difficult to do based on the TT-Seq, PRO-Seq, or NET-Seq profiles.

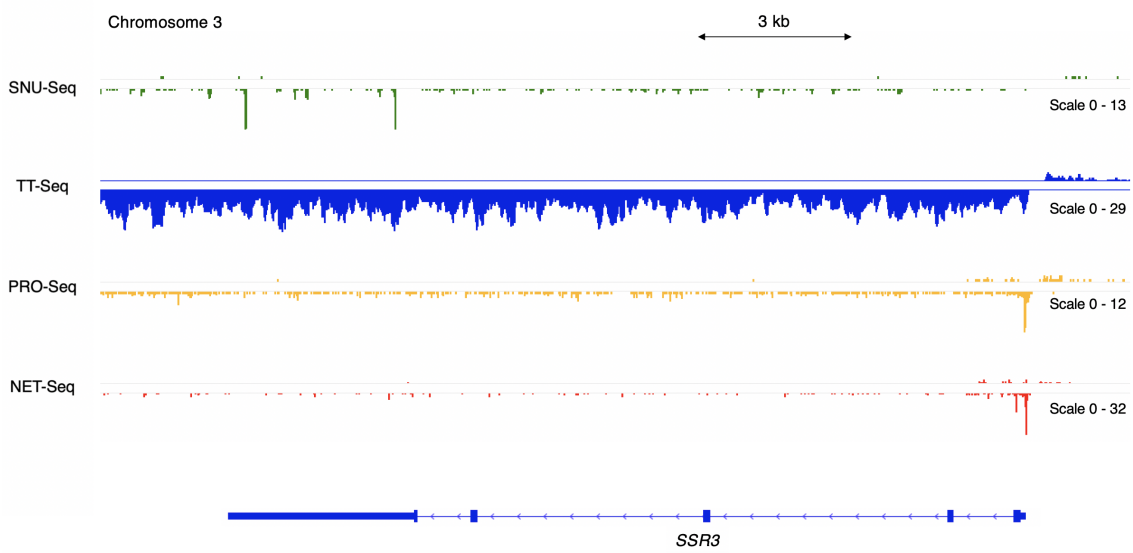


Figure 3.7: SNU-Seq, TT-Seq, PRO-Seq, and NET-Seq profiles at the *SSR3* locus.

The TT-Seq (HEK293, navy) and SNU-Seq (HEK293, green), PRO-Seq (yellow, HEK293, Woo et al. (2018)), and NET-Seq (red, HeLa, Nojima et al. (2015)) profiles at the *SSR3* locus are shown. Scales are indicated for comparison.

The comparison between these sequencing methods also reveals another distinction, especially between SNU-Seq and TT-Seq: Because of the random sonication step introduced during the TT-Seq protocol, full-length transcripts that contain any 4sU-label will contribute towards signals that are reflected onto a large part of the gene body. In contrast, signals emanating from full-length transcript with incorporated 4sU-labels from SNU-Seq will only be reflected at the corresponding peaks of the poly(A) site (or, if relevant, the multiple alternative poly(A) sites).

3.4 SNU-Seq and Gene Expression

SNU-Seq is capable of generating a transcriptional profile that reflects nascent transcription; the very beginning of gene expression. One important validation that was performed on the SNU-Seq data was to calculate the decay and synthesis rate based on a gene expression and transcriptional model developed by Villamil et al. (2019). This has previously been done with TT-Seq in order to identify mathematical constants reflecting the gene expression process (Schwalb B et al., 2016). A correlation between the TT-Seq constants and the corresponding constants calculated for SNU-Seq would therefore indicate to what extent SNU-Seq can be an alternative method of determining the mathematics of transcription

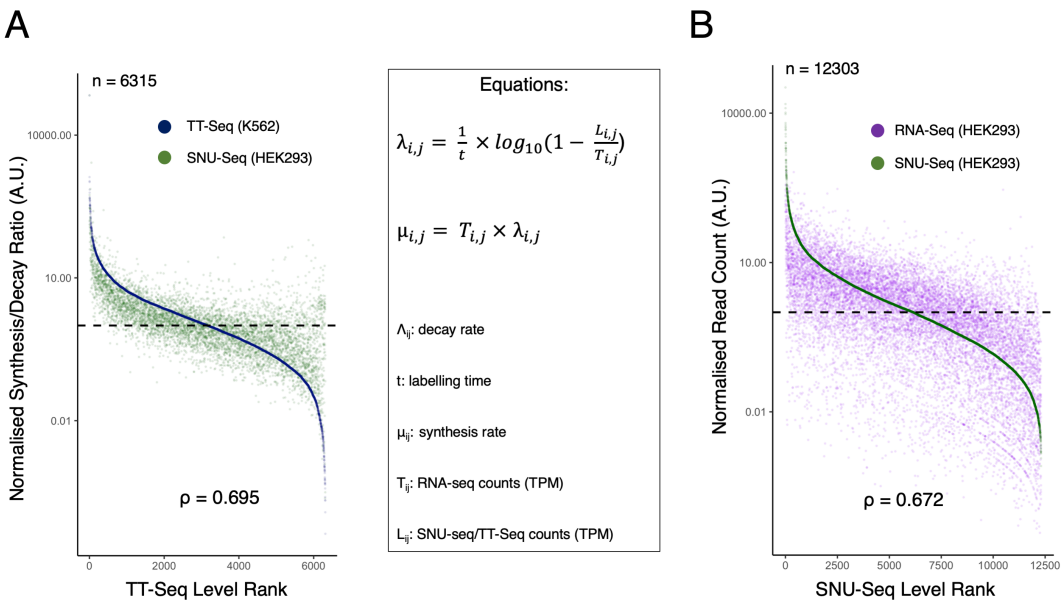


Figure 3.8: Correlative Comparison of SNU-Seq with TT-Seq and RNA-Seq

A. Ranked correlation plot of the synthesis-to-decay ratios between SNU-Seq and TT-Seq. The equations used are indicated in the box on the right. Equations were applied by Aksel Saukko-Paavola, with the original model developed by Villamil et al. (2019). SNU-Seq data are in HEK293 cells (from this study), and, for comparison, to see if SNU-Seq can be used for this mathematical approach, the original TT-Seq dataset in K562 cells was used (Schwalb B et al., 2016). The Spearman correlation coefficient (rho) is indicated. B. Correlation between SNU-Seq (this study) and RNA-Seq in HEK293 cells (Banks et al., 2014). The Spearman correlation coefficient (rho) is indicated.

Figure 3.8 A shows the rank correlation plot of the normalised synthesis-to-decay ratios of TT-Seq and SNU-Seq. For reference, the original TT-Seq dataset in K562 cells was used. The equations used to calculate this ratio are indicated on the left and the Spearman correlation of 0.695 indicates a good correlation between the two methods, despite the different cell lines. This would support the idea that SNU-Seq is in fact a viable method for mathematically describing the gene expression process.

In addition to the mathematical modelling of gene expression, a correlation analysis between SNU-Seq (HEK293, this study) and RNA-Seq (HEK293, Banks et al. (2014)) was performed. RNA-Seq in effect measures steady-state levels of RNA, resulting from the synthesis of RNA minus RNA degradation. A nascent transcriptomics method should therefore (slightly) correlate with steady state RNA levels (as measured by RNA-Seq), but not too perfectly because of transcripts that are synthesised rapidly and degraded slowly, or those that are synthesised slowly but degraded rapidly. Figure 3.8 B shows a rank correlation plot between SNU-Seq and RNA-Seq counts for more than 12000 genes. The Spearman (rho) correlation coefficient of 0.672 reflects the moderate level of correlation that would be expected between nascent transcription and steady-state gene expression.

3.5 Benefits and Considerations of SNU-Seq

3.5.1 Resolution

SNU-Seq has, because of the addition of the artificial poly(A)-tail followed by the oligo-dT library preparation, a near-single-nucleotide resolution. This is the main reason

for developing SNU-Seq in the first place: To develop a 4sU-based labelling method that, unlike TT-Seq, has a high resolution. It is important to point out that intragenic sequences with poly(A)-stretches will not be generating single-nucleotide resolution with SNU-Seq because after library-preparation with oligo-dT primers, one cannot tell which of the adenine residues of the poly(A)-stretch was the last to be incorporated during the metabolic labelling time. This issue, however, is resolved with the usage of immediate adapter ligation of short nascent RNAs on both the 3' and 5' ends for size-selected SNU-Seq, which is presented in the next chapter. Other than the exception of intragenic poly(A)-stretches, SNU-Seq has a near single-nucleotide resolution.

3.5.2 5' End Biases

Because of the sonication step for random fragmentation, TT-Seq has a 5' bias: After 4sU-enrichment, some fragments will be entirely thiolabelled, whereas other fragments will only be partially thiolabelled. The part of those fragments that have no thiolabel is always at the 5' end. This means that TT-Seq will create libraries that may, to some extent, generate signals that do not actually reflect the incorporation of the thiolabel, but just happened to be at the 5' end of fragments that do also contain a thiolabel further downstream. TT-Seq therefore has a 5' bias in its signal coverage. In contrast, SNU-Seq does not have this bias, because the signals exclusively originate from the most 3' nucleotide incorporated during the thiolabelling pulse. In addition to the higher resolution, this is one of the major advantages of SNU-Seq over TT-Seq.

On the other hand, SNU-Seq may have a different type of disadvantage when it comes to biases at the ends of the nascent transcripts: Because of the usage of

random priming at the 5' end during the library preparation for amplification, some very short nascent RNAs may not be fully reflected in the SNU-Seq library. Again, this is addressed by the usage of immediate adaptor ligation at the 3' and 5' ends of short nascent RNAs for size-selected SNU-Seq, which is discussed in the next chapter.

3.5.3 Practicability and Cost

SNU-Seq has a straightforward protocol. Similar to TT-Seq, the thiolabelling can be carried out on cells in tissue culture dishes without any additional procedures or prior steps. This is in large contrast to GRO-Seq and PRO-Seq which require difficult nuclei isolation steps followed by a run-on reaction with biotin-NTPs on ice. GRO-Seq and PRO-Seq are therefore *in vitro* methods and may not entirely reflect the transcriptional process of living cells, especially considering the necessary addition of sarkosyl, which inhibits (re-) inititation of transcription (Szentirmay and Sawadogo, 1994). In contrast, SNU-Seq has the benefit of *in vivo* labelling. Furthermore, the polyadenylation step allows for an easy library preparation using well-established oligo-dT-based library preparation kits. This also makes SNU-Seq a lot more affordable than other methods such as PRO-Seq or even TT-Seq, which usually require more complicated library preparation methods.

Furthermore, the alignment percentages is an important factor when considering how (cost)-efficient a sequencing method is. Because of RNA contaminants binding to the polymerase during the immunoprecipitation step of NET-Seq, and other reasons, NET-Seq can have relatively low alignment percentages that effectively make

this method more costly compared to SNU-Seq, which has much higher alignment percentages of >65 %.

Chapter 4

SNU-Seq and Size-Selection

4.1 Motivation for Size-Selection

One major aim for developing next-generation sequencing methods that measure nascent transcription is to understand key steps of the transcription cycle. Two crucial steps in this cycle are initiation and promoter-proximal pausing followed by pause-release. With the SNU-Seq metagenes showing a peak for some genes, but not all, near the 5' end of genes before levelling off to a flat, steady level in the gene body, one reason for advancing SNU-Seq was to see whether it was possible to 'zoom in' on this region. Performing a size-selection on the nascent RNA could potentially allow for the nascent transcripts that have been synthesised up to the promoter proximal pause to be sequenced.

In addition, this approach would also eliminate two potential, minor, issues associated with the standard SNU-Seq approach; one, size-selection makes the input fragments for the library preparation small enough to allow for immediate adaptor ligation not only at the 3' end but also at the 5' end, which would avoid the need for random priming at the 5' end. Two, whilst usage of oligo-dT primers for library preparation is well-established and easy to carry out, it can lead to amplification of intragenic poly(A)-stretches when measuring nascent transcription. This will be avoided when using adaptors at the 3' and 5' ends.

Additionally, this approach will lead to single-nucleotide resolution at both the 3' end and the 5' end. This does not only allow for measuring the last incorporated nucleotide during the act of (early) transcription, but would also facilitate the identification of transcription start sites from nascent RNAs. The latter is described in chapter 6.

4.2 Size-Selected SNU-Seq Workflow

Having identified pausing behaviour of genes from SNU-Seq at high precision, a further aim was to enrich the SNU-Seq libraries for the 5'end of genes. Therefore, an advancement of SNU-Seq was developed for this study, in order to get deep nascent transcriptional sequencing data of the 5' end of human genes. This was accomplished by developing size-selected SNU-Seq. The workflow for a size-selected SNU-Seq procedure is shown in Figure 4.1.

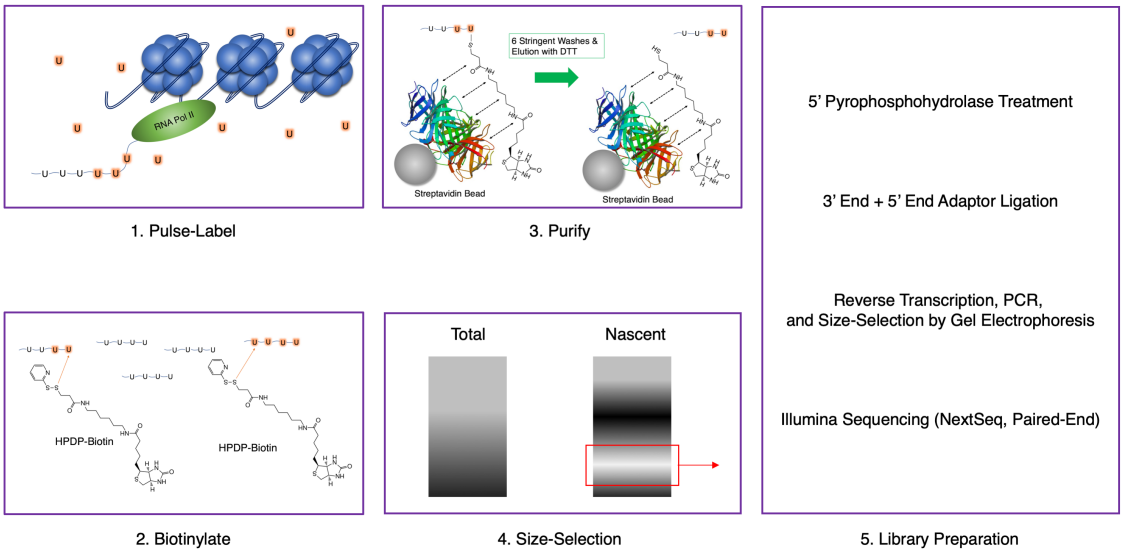


Figure 4.1: Size-Selected SNU-Seq Workflow

The SNU-Seq with size-selection is shown: This includes thiolabelling, biotinylation, and streptavidin purification similar to SNU-Seq, but instead of polyadenylation, includes a gel-based size selection step. After purifying RNAs of a size range between 50 and 100 nucleotides, adapters were added immediately for library preparation, retaining single nucleotide resolution at both the 5' and the 3' ends. No polyadenylation was performed. Further library preparation steps include reverse transcription, PCR amplification, size-selection by gel electrophoresis. Libraries were sequenced on an Illumina NextSeq platform with 75 cycles. Thiolabelled nucleotides are indicated in orange.

When run on a gel, nascent, thio-labelled RNA shows a clear band at approximately 50-100 bp and a smear around 600 bp and above. This is shown in Figure 4.2 A. It could be hypothesized that the clear band of small nascent RNAs could represent

those RNAs that were synthesized up until the promoter-proximal pausing step of the transcription cycle and the remaining smear for larger RNAs corresponding to nascent transcripts generated within the body of the gene. However, some of the signal from this band might also represent nascent transcripts undergoing random cleavage and degradation.

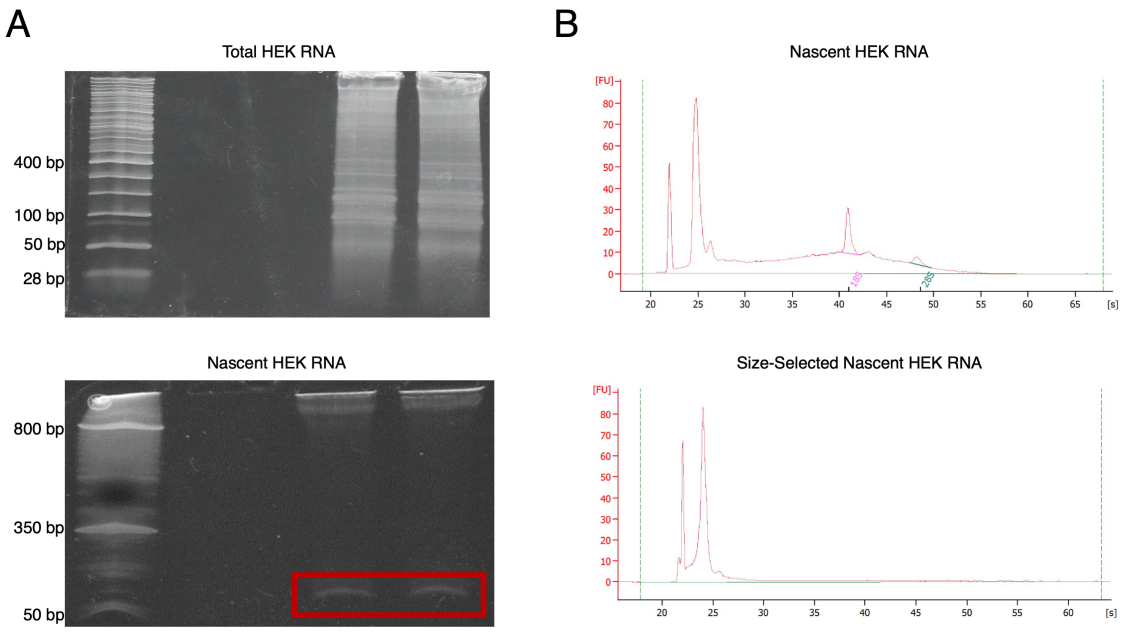


Figure 4.2: Size-Selection of Nascent RNA

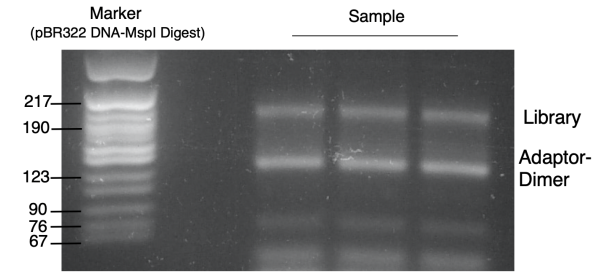
A. RNA-Gels run with total and nascent (thiolabeled) HEK293 RNA. B. RNA pico-chip Bioanalyzer traces of nascent HEK293 RNA before and after gel-based size selection.

Figure 4.2 B shows the Bioanalyzer traces of nascent RNA before (top) and after size-selection of the small bands (bottom). The lower trace shows no background or contamination of larger transcripts, which therefore were deemed pure enough to carry forward in the size-selected SNU-Seq procedure. Treatment with RppH (to remove the 5' cap) after gel-based size selection and immediate adaptor ligation then allows for sequencing of the short nascent RNA fraction exclusively.

The library preparation with immediate ligation of the 5' and 3' end adaptors (of 54 and 75 bp length, respectively) can cause a series of different DNA fragments, some

of which are not intended to be part of the ultimate sequencing library: These would include adapter-dimers or inserts with only one of the adaptors ligated.

A



B

Sequencing Stats	Number	Percentage
Number of Total Reads	333,002,721	
Reads that align 0 times	44,549,628	13.3 %
Aligned Reads	288,453,093	86.7 %

Figure 4.3: Size-selected SNU-Seq Library Preparation and Sequencing Stats

A. DNA Gel of size-selected SNU-Seq DNA products after adaptor ligation and PCR amplification. The top bands represent the desired library product whereas the second highest bands are a result of adaptor dimerisation. B. Size-selected SNU-Seq sequencing depth and alignment percentage.

A DNA gel of these library preparation products is shown in Figure 4.3 A, of which the top bands were gel-purified, quality-controlled on a Bioanalyzer and then sequenced. The libraries were sequenced on an Illumina platform to allow for very deep sequencing runs. The resulting sequencing stats are displayed in Figure 4.3 B, showing an alignment percentage of more than 86 %. This is significantly higher than what is usually observed with other nascent transcriptomics methods such as NET-Seq. For the analysis of the last incorporated nucleotide of the short nascent RNA, the most 3' signal was collected bioinformatically after alignment. A more detailed explanation of the wet-lab and bioinformatic procedures is described in the methods section.

4.3 Detection of Pausing by Size-Selected SNU-Seq

Having suspected that the clear, sharp band of nascent RNAs with a size of approximately 80 bp could (partially) represent the transcripts synthesised up to the promoter proximal pause, a first step was to observe the profile of size-selected SNU-Seq on IGV and compare it to other 4sU-based methods, namely SNU-Seq and TT-Seq.

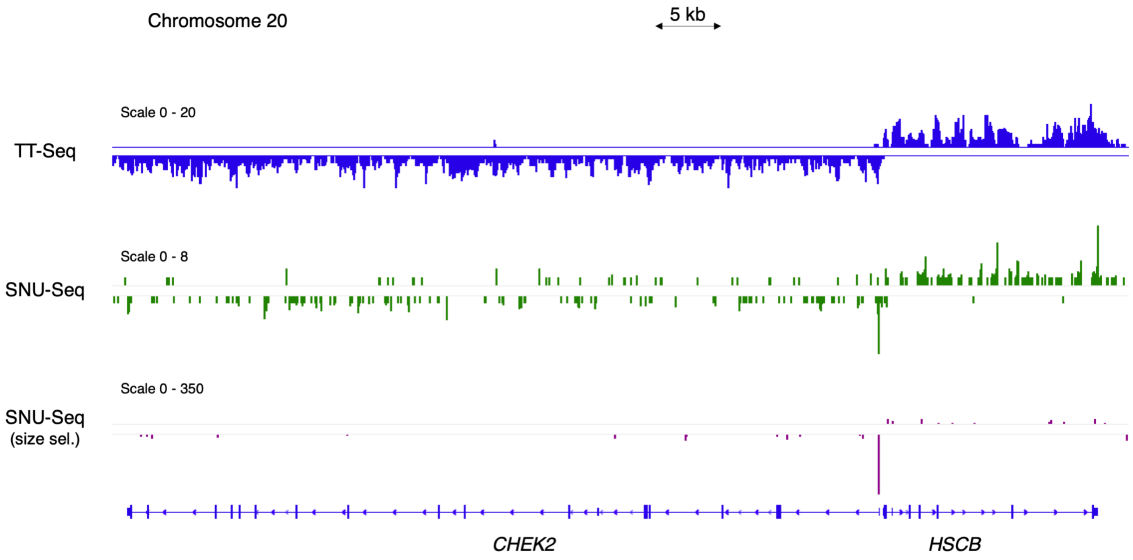


Figure 4.4: IGV Screenshot of size-selected SNU-Seq, TT-Seq, and SNU-Seq around the *CHEK2* locus.

IGV screenshot displaying TT-Seq (HEK293, this study), SNU-Seq (HEK293, this study), and size-selected SNU-Seq (HEK293, this study) profile around the *CHEK2* locus on chromosome 20.

Figure 4.4 shows an IGV screenshot comparing the TT-Seq (HEK293, this study), SNU-Seq (HEK293, this study), and size-selected SNU-Seq (HEK293, this study) profile around the *CHEK2* locus on chromosome 20. As can be seen from these transcription profiles of *CHEK2*, which is transcribed on the reverse strand, SNU-Seq shows a strong peak at the very 5' end of the gene followed by a more even coverage over the gene body. In contrast, TT-Seq has no such peak. Crucially, the size-selected SNU-Seq track shows a strong singular peak at the 5' end of *CHEK2* at

the same position as the SNU-Seq peak, but no other significant signals across the gene body or elsewhere.

This exemplifies and demonstrates how size-selection of short, nascent RNAs followed by deep sequencing may provide a viable approach to detecting and measuring the early stages of transcription, namely initiation and promoter-proximal pausing. This locus obviously only provides a singular example, which is why, as a next step, a genome-wide metagene analysis was performed to identify the predominant location of these short nascent RNAs within the genome.

For this, as shown in Figure 4.5 A, a calibrated metagene of the size-selected SNU-Seq 3' end signal was created. As can be seen in this metagene, the average signal from size-selected SNU-Seq reads peaks around 60-80 bp downstream of the transcription start site. Crucially, no significant averaged levels can be observed further downstream into the gene body or within the first 40-50 bp of the gene. This would confirm the suspicion that the short nascent RNAs that were observed on the RNA-gel with a clear band at approximately 80 bp do in fact, to a degree, represent the nascent transcripts synthesised within the 5' end of genes.

As an additional piece of analysis, instead of considering all reads for metagenes, the location of the maximum size-selected SNU-Seq read for each gene was mapped. This is shown in a 2-dimensional density plot in Figure 4.5 B. Importantly, the maximum reads are again centred around a distance of approximately 80 bp downstream of the TSS. Interestingly, the density plot also shows a minor, secondary, broader peak centering at approximately 300 bp downstream of the TSS. The meaning of this peak is harder to explain, but one approach in which this could be studied would be

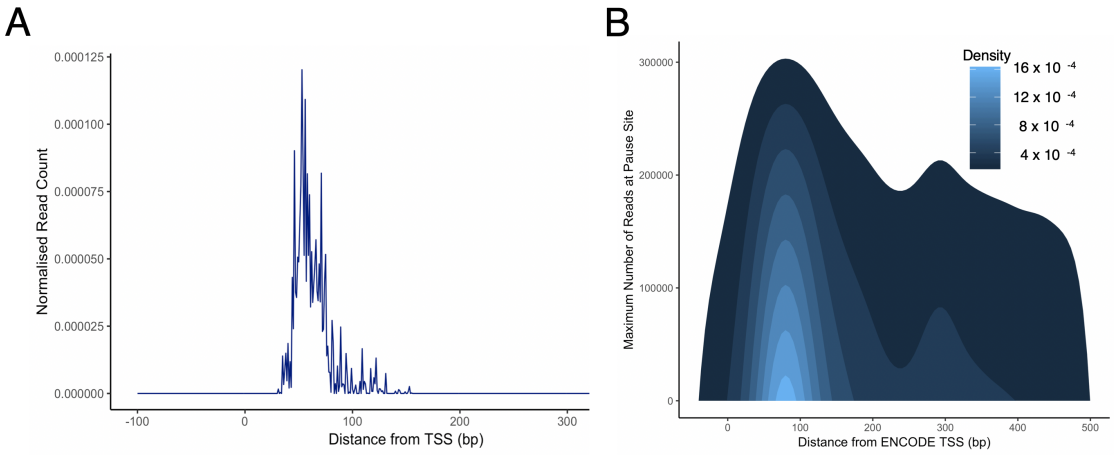


Figure 4.5: Metagene Analysis of Size-Selected SNU-Seq

A. Normalised metagene of size-selected SNU-Seq reads around human TSSs. TSS annotations were derived from Brown et al. (2018) (n = 9336). B. 2-dimensional density plot showing the distribution of the maximum read number of each gene against its location relative to the gene's ENCODE transcription start site. Density refers to the number of occurrences per pixel (n = 19874).

to cluster genes into different groups based on their size-selected SNU-Seq profiles to identify differential profiles and figure out if these groups have different attributes or features. This will be done as part of a clustering approach in the next chapter.

To summarise, the 3' end signal of size-selected SNU-Seq libraries show a localisation relative to the transcription start site that would suggest that the short nascent fragments could correspond to the transcripts synthesised up to the point of promoter-proximal pausing. So far, the localisation alone is no proof of this, and could just be an artefact of the technique itself. Therefore, the next chapter is concerned with the size-selected SNU-Seq profiles in different gene clusters and comparing them to those of NET-Seq, which is known to capture promoter-proximal pausing well, as well as a range of other sequencing methods, such as ChIP-Seq of the pausing factor NELF.

4.4 Benefits and Considerations of Size-Selected SNU-Seq

With the development of size-selected SNU-Seq, a range of benefits and considerations are addressed below to assess its performance.

4.4.1 Resolution

Because of the immediate adaptor ligation during the library preparation, and the avoidance of any fragmentation during the entire protocol, size-selected SNU-Seq has full single-nucleotide resolution. This makes size-selected SNU-Seq a suitable extension of SNU-Seq. However, unlike SNU-Seq, size-selected SNU-Seq does not lose resolution because of potential intragenic poly(A)-stretches. Albeit extremely rare, these can cause SNU-Seq to lose several basepairs of resolution because of the polyadenylation step. The adaptors used during the library preparation process for size-selected SNU-Seq on the other hand have a defined non-repetitive sequence that would exclude the risk of false and unwanted priming sites.

The high resolution of the 3'end of short nascent RNAs, combined with the knowledge that these locate to a large extent at the very 5' end of genes, can therefore provide a useful basis to create better insights into the early stages of transcription. This is done in the following chapter, where size-selected SNU-Seq profiles are compared to NET-Seq and other methods, as well as a study into the biological context of promoter-proximal pausing that is measured by this method. Crucially, size-selected SNU-Seq does not only have single-nucleotide resolution at the 3' end, but also at the

5' end. This is useful not directly for studying nascent transcription profiles, but for the annotation of transcription start sites. This analysis is presented in chapter 6.

One potential downside of size-selected SNU-Seq, however, is that despite the genome-wide signal centering near the TSSs, there is still a proportion of reads located elsewhere across the genome. This could either be considered non-specific background, or the product of RNA cleavage events during or shortly after transcription that lead to a smaller size of these transcripts and may therefore end up in the purified small nascent RNA fraction utilised for library preparation.

4.4.2 Practicability and Cost

With the added resolution on both ends of the transcripts for size-selected SNU-Seq over SNU-Seq alone comes additional labour necessary to carry out this method. This includes an RppH step to remove the 5' cap. Additionally, the size-selection itself can be considered a difficult endeavour in which RNA degradation needs to be avoided at all cost and requires assessment of the fragment size with an RNA bioanalyzer.

Otherwise, the main features of size-selected SNU-Seq remain comparable to the SNU-Seq protocol. One crucial consideration for nascent transcriptomic sequencing methods is the alignment percentage. This percentage is particularly high for size-selected SNU-Seq (more than 80 %). This makes size-selected SNU-Seq a cost-efficient method, especially compared to other widely used single-nucleotide resolution methods in nascent transcriptomics, such as NET-Seq, which is known to have issues with high background levels (as reviewed in Wissink et al. (2019)).

Chapter 5

Genome-Wide Analysis of the Promoter-Proximal Pause

5.1 K-Means Clustering Reveals Differential Pausing Behavior

In order to identify differential pausing behaviour of different groups of genes, a k-means clustering approach was used to group genes into different classes. k-means clustering is a non-supervised machine learning algorithm (Lloyd, 1957) that is explained visually in Figure 5.1.

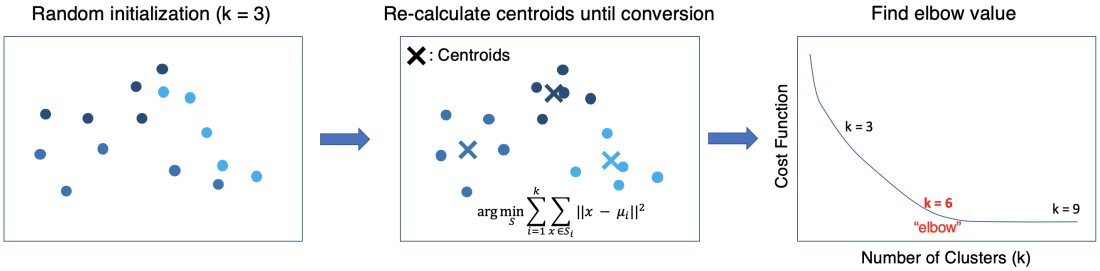


Figure 5.1: k-Means Clustering Approach

The k-means clustering algorithm, in a simplified version, is shown. The data to be used is initialised into clusters before calculating centroids. The optimisation problem is to reduce the distance between the data points and the centroids of their corresponding cluster by re-calculating the centroids and re-assigning the clusters until convergence. This is performed by calculating Euclidian distance. In this example, only two dimensions are shown, but for sequencing data, each basepair position is considered one dimension. The number of clusters is determined heuristically by finding the elbow value, that corresponds to the postion after which increasing k does not further reduce the cost function.

In the specific context of classifying genes based on their transcription profiles, NET-Seq was used to determine clusters as described in Brown et al. (2018). In their study, Brown et al. performed k-means clusters on a set of approximately 3000 human genes, which represent the data points, and using NET-Seq reads (Nojima et al., 2015) from the first 1000 bp downstream of the transcription start site (which represent the number of dimensions, or features).

As shown in Figure 5.2, the NET-Seq profiles differ significantly between the various clusters. Cluster 1 shows a flat profile across the 1000 basepair window down-

stream of the TSSs. Clusters 2, 3, and 4 all show significant peaks just downstream of the TSS that could correspond to promoter-proximal pausing signals. Of the three, cluster 4 has the most narrow peak whilst cluster 2 has the earliest peak. Cluster 5 on the other hand, has a much later, and broader NET-Seq peak that centers around 300-400 basepairs downstream of the TSS. Cluster 6 has a slightly narrower and earlier peak than cluster 5 but is still much broader and further downstream than the peaks from clusters 2-4.

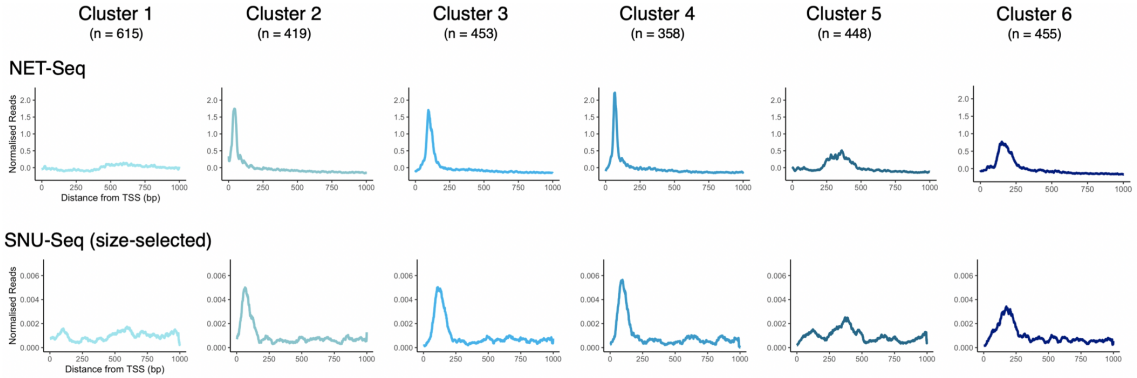


Figure 5.2: k-Means Clusters of Human NET-Seq and Size-selected SNU-Seq

For each cluster, the NET-Seq (Nojima et al., 2015) and size-selected SNU-Seq metagenes are shown. The clusters were generated using NET-Seq and the same cluster indices for each gene were applied to sort the size-selected SNU-Seq data into the respective clusters. Metagenes are normalised to make every gene contribute equally. The mNET-Seq data is from HeLa whereas size-selected SNU-Seq was performed in HEK293 cells.

When using the cluster indices determined from NET-Seq and grouping the size-selected SNU-Seq into the same clusters, the resulting metagenes show a remarkably similar pattern: Clusters 2-4 especially, show similarly sharp peaks with, again, cluster 4 showing the most narrow and cluster 2 showing the earliest peak amongst the three. Cluster 1 has a noisier, but still flat profile, while cluster 5 shows, similar to the mNET-Seq profile, mostly a broader and later peak centering around 300 bp downstream of the TSS. However, the size-selected SNU-Seq metagene for cluster 5 also shows an additional, smaller peak closer to the TSS. Cluster 6 again has a

profile that matches the mNET-Seq profile, having a slightly broader peak located further downstream compared to cluster 2-4. Some of the low background signals in the size-selected SNU-Seq metagenes observed within gene bodies may be the result of some cleavage or degradation events, as discussed in the previous chapter.

NET-Seq and the newly developed size-selected SNU-Seq seem to be very capable of detecting the pause, and based on this clustering approach, show a capability to show different pausing behaviour of different genes. On the other hand TT-Seq, which has a much lower resolution, is not able to do so, as shown in Figure 5.3. Standard SNU-Seq without size-selection mostly covers the nascent transcription process within the gene body, as discussed above, but, as shown in Figure 5.3, clusters 2, 3, and 4 have minor peaks in their SNU-Seq profiles. This is due to the higher resolution of SNU-Seq compared to TT-Seq and supports the idea that SNU-Seq yields a comprehensive profile of the nascent transcription process.

TT-Seq profiles of alpha-amanitin treated cells look different to the standard TT-Seq profiles and mostly reflect transcription of the very 5' end of the gene. When grouping the alpha-amanitin TT-Seq data into the same clusters, the show a remarkably similar pattern to the NET-Seq and size-selected SNU-Seq profiles, with clusters 2-4 having sharp peaks near the TSS, and cluster 1 having no peak. Cluster 5 seems to have a major bipartite peak much further downstream and a smaller peak closer to the TSS, generating a profile that seems to resemble the size-selected SNU-Seq profile more than the NET-Seq profile. Like the NET-Seq and size-selected SNU-Seq profiles, the alpha-amanitin TT-Seq profile for cluster 6 has a broader and slightly more downstream peak compared to clusters 2-4.

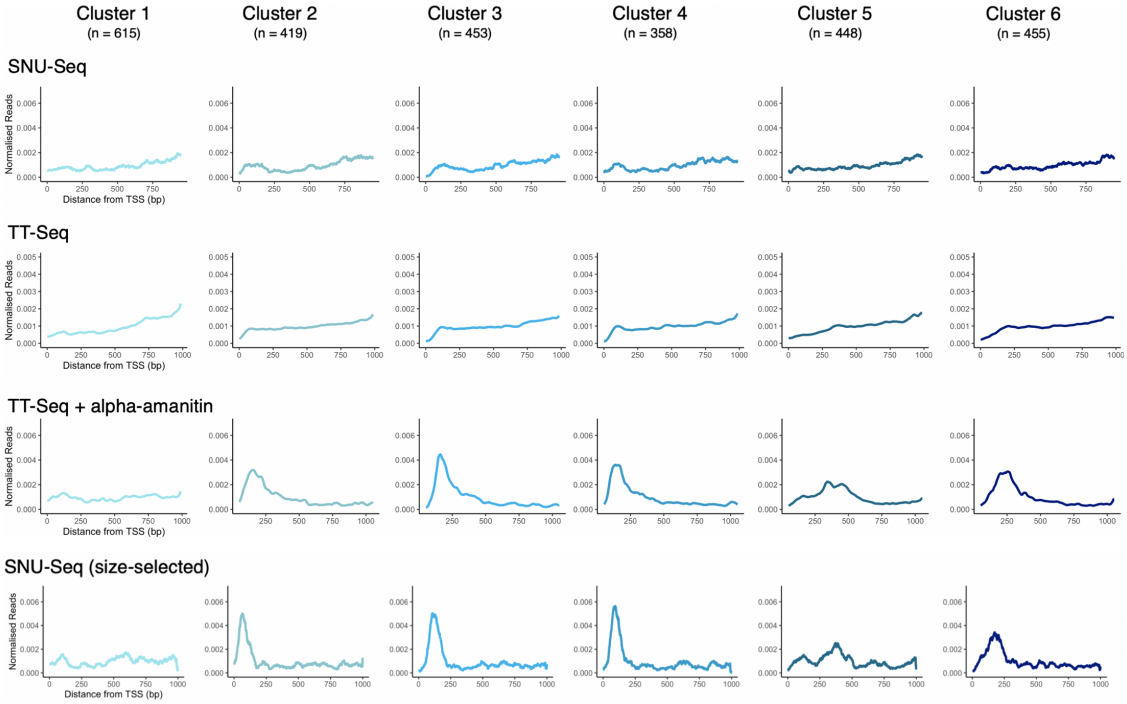


Figure 5.3: k-Means Cluster Profiles of Human SNU-Seq, TT-Seq and TT-Seq after alpha-amanitin treatment

For each cluster (k = 1 to k = 6), the SNU-Seq, TT-Seq, and TT-Seq after alpha-amanitin treatment are shown. The clusters were originally generated using NET-Seq and the same cluster indices for each gene were applied to sort the SNU-Seq and TT-Seq data into the respective clusters. Metagenes are normalised to make every gene contribute equally. SNU-Seq and TT-Seq was performed in HEK293 cells. For comparison, the size-selected SNU-Seq metagenes in the respective clusters are shown in the bottom panel.

Having generated these clusters and finding remarkable agreements between NET-Seq and SNU-Seq (with and without size-selection) in terms of their metagene profiles, a next step was to find out to what extent these profiles correspond to true differential pausing behaviour. In order to test this, publicly available NELF, INTS3, and MED26 ChIP-Seq data was used to cluster the data into the same 6 clusters (Liu et al., 2014; Nojima et al., 2018; Takahashi et al., 2020). The reason for using ChIP-Seq data for these three factors is that, as described above, NELF is universally considered to be a pausing factor that needs to be phosphorylated by p-TEFb and removed from the pause site in order to allow release of Pol II into the gene body (Adelman and Henriques, 2018; Cheng and Price, 2007; Guo et al., 2000; Marshall

et al., 1996; Vos et al., 2018a,b; Wu et al., 2003; Yamada et al., 2006; Yamaguchi et al., 1999). DSIF was not considered because of its dual action as both a positive and a negative elongation factor (Adelman and Henriques, 2018; Cheng and Price, 2007; Wada et al., 1998; Wu et al., 2003). As discussed above, integrator function has been linked to the pausing signal (or the alternative of premature termination), which is why a ChIP-Seq for the integrator subunit INTS3 was included (Elrod et al., 2019). Finally, mediator was included because of its role in transcription at the 5' end of genes, such as PIC assembly and promoter escape (Esnault et al., 2008; Guidi et al., 2004; Wong et al., 2014).

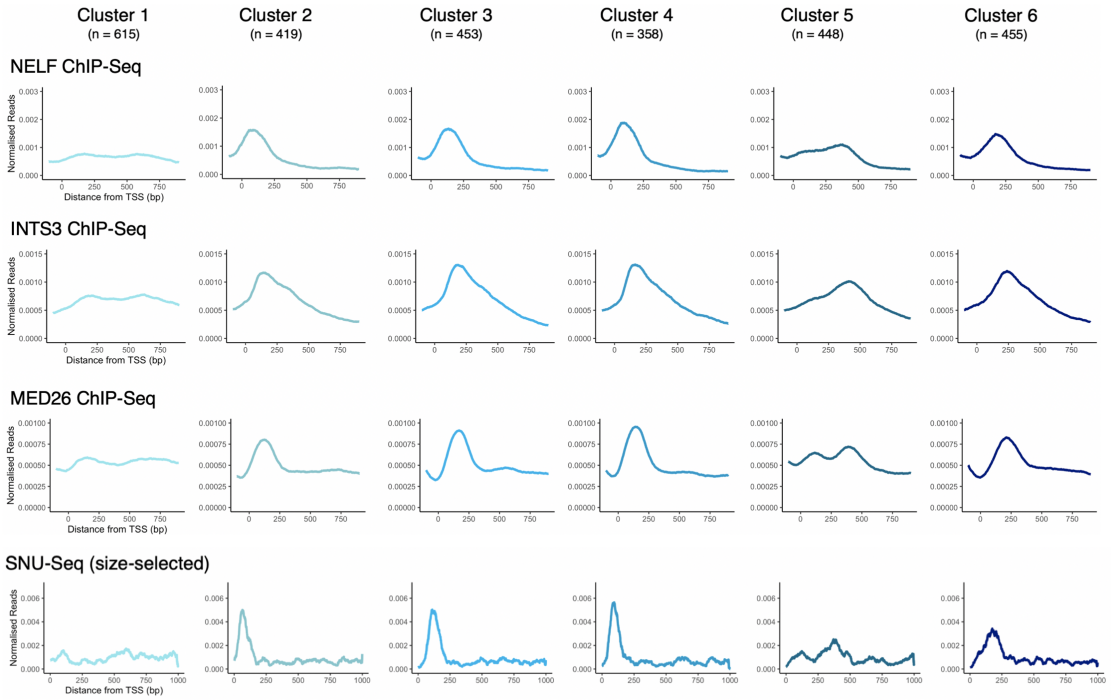


Figure 5.4: k-Means Cluster Profiles of NELF, INTS3, and MED26 ChIP-Seq

For each cluster (k = 1 to k = 6), the NELF, INTS3, and MED26 ChIP-Seq metagenes are shown (Liu et al., 2014; Nojima et al., 2018; Takahashi et al., 2020). The clusters were originally generated using NET-Seq and the same cluster indices for each gene were applied to sort the ChIP-Seq data into the respective clusters. Metagenes are normalised to make every gene contribute equally. ChIP-Seq was performed in HeLa cells. For comparison, the size-selected SNU-Seq metagenes in the respective clusters are shown in the bottom panel.

As can be seen in Figure 5.4, the profiles of factors involved in 5'end transcription

and pausing behaviour show metagene profiles within the 6 clusters as one would expect. For all three factors, the most pronounced and most 5' peaks are found in clusters 2, 3, and 4 whereas cluster 1 shows very flat profiles. Cluster 5 has a later peak, although only the mediator Chip-Seq profile (MED26) shows the dual peaks that were also observed in the size-selected SNU-Seq and the TT-Seq with alpha-amanitin treatment. Cluster 6 has again a slightly further downstream and broader peak in its profile for all three factors compared to cluster 2, 3, and 4.

This could be considered a validation of the clustering approach, and that the differential profiles found reflect a true difference in transcriptional behaviour that is supported by the corresponding presence and occupancy profiles of relevant protein factors involved in the pausing process and transcription at the 5' end of genes. In order to test this further, and with statistical significance, the ChIP-Seq levels for each protein factor were compared on a gene-by-gene basis and grouped into the 6 clusters. For this, the summed normalised reads of the first 300 basepairs were compared. In addition, SNU-Seq levels (raw reads corrected for gene length within the first 1000 basepairs) were compared on a gene-by-gene basis too. This was done in order to determine whether the reduced levels of NELF, INTS3, or MED26 is an explanation for the differential pausing profiles, or whether this is merely down to absolute levels of nascent transcription.

As shown in Figure 5.5 A, boxplot diagrams for these gene-by-gene comparisons grouped within the 6 clusters show a stark difference between NELF levels when comparing cluster 1 to cluster 2 to 4 in particular. This difference is slightly less pronounced for INTS3 and MED26 levels, but particularly small when looking at the

levels of nascent transcription as measured by SNU-Seq (without size-selection).

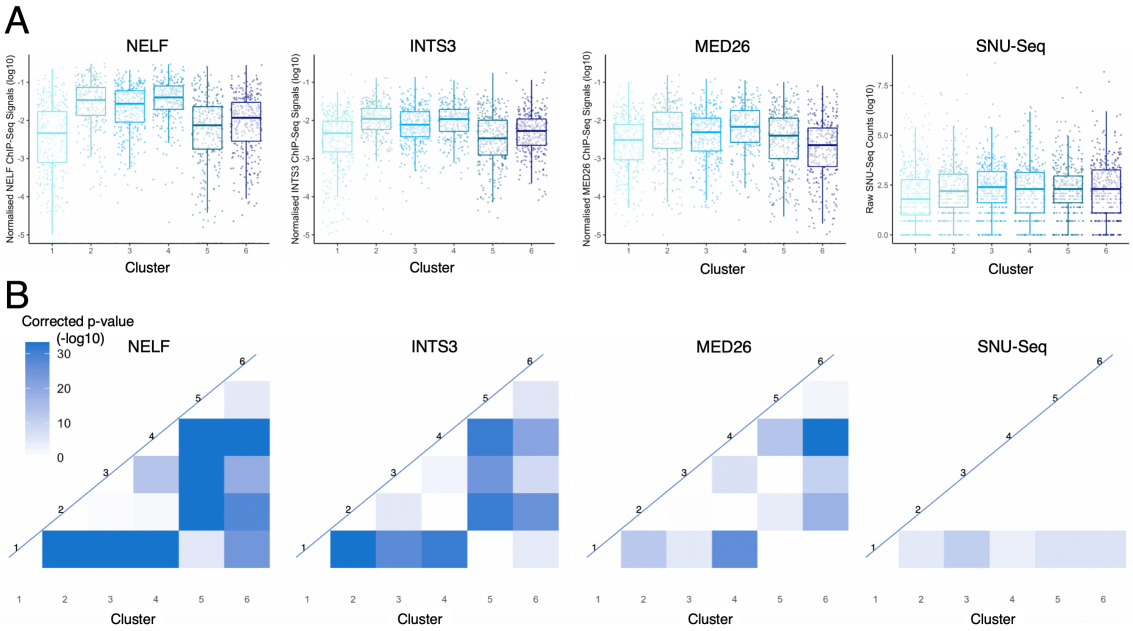


Figure 5.5: Statistical Analysis of NELF, Intgretor, and Mediator Occupancy within Gene Clusters

A. Boxplot Diagrams for NELF, INTS3, MED26, and SNU-Seq levels at the 5' end of genes for each cluster (300 bp window for ChIP-Seq and 1000 bp window for SNU-Seq). ChIP-Seq data was normalised to make each gene contribute equally. For SNU-Seq, the raw counts were used. All data was log-transformed. B. Triangular heatmaps of between-cluster significance for NELF, INTS3, MED26, and SNU-Seq levels. P-values were calculated using the Wilcox-rank sum test, and corrected using the Bonferroni method.

Figure 5.5 B shows the corresponding triangular heatmaps of the between-cluster corrected p-values. As can be seen here, NELF shows the strongest statistically significant difference between the clusters, with particular distinctions visible between cluster 2 to 4 and the other clusters. This distinction gets weaker for the integrator and mediator subunit ChIP-Seq levels, but is nonetheless again more significant when comparing cluster 2 to 4 and the other clusters. Levels of nascent transcription are much less capable of providing statistically significant distinctions between the clusters, and only shows significant differences when comparing cluster 1 to the other clusters, but not amongst the latter.

With NELF being considered the most prominent factor involved in promoting

promoter-proximal pausing of RNA Pol II (Adelman and Henriques, 2018; Cheng and Price, 2007; Guo et al., 2000; Marshall et al., 1996; Vos et al., 2018a,b; Wu et al., 2003; Yamada et al., 2006; Yamaguchi et al., 1999), this analysis can be considered a validation of a true biological meaning of these clusters: NELF, more so than the integrator and particularly more so than the mediator complex, shows highly significant differential occupancy levels between those clusters that show very 'pausy' profiles (2, 3, and 4) and those that do not (cluster 1) or a delayed/broader profile (cluster 5 and 6).

5.2 Assessing The Predictability of Pausing Clusters from ChIP-Seq Data

So far, the analysis on the k-means clusters found that NELF, followed by the integrator and then mediator complexes, are enriched at promoter-proximal pausing sites. These clusters were determined using NET-Seq and verified with size-selected SNU-Seq. A consequential next step of this analysis would then be to assess whether NELF, integrator, or mediator levels alone are predictive of which cluster a gene may belong to.

In order to perform this analysis, the same ChIP-Seq data for NELF, integrator (INTS3), and mediator (MED26) was used as above (Liu et al., 2014; Nojima et al., 2018; Takahashi et al., 2020), and assigned into 7 50-basepair long windows, with one upstream and 6 downstream of the TSS of each gene, as displayed in Figure 5.6 A.

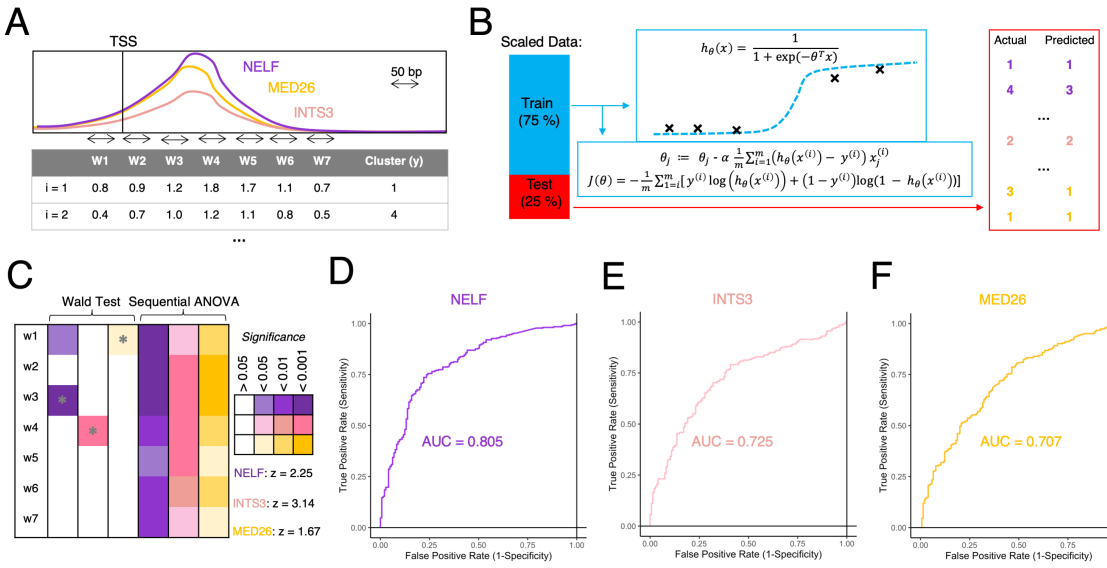


Figure 5.6: Classification Approach of NELF, Integrator, and Mediator Levels into NET-Seq based Clusters

A. ChIP-Seq data preparation for machine learning approach. 50-bp long summed windows are generated with one window upstream and 6 windows downstream of the TSS of each gene. B. Logistic regression approach depicting the data partitioning into training and testing sets and the sigmoid hypothesis function for classification. The gradient descent and cost function equations are presented in the lower blue box. m denotes the total number of training examples (total number of genes, i.e. 1845). J denotes the cost function, i denotes the training example (i.e. the gene), and j the feature (i.e. the window). The vector containing the 7 parameters is termed θ. $x^{(i)}$ is a vector containing the summed ChIP-Seq level for the i-th gene in the 7 windows. $y^{(i)}$ denotes the value to be predicted, i.e. the cluster. h is the hypothesis function. C. P-values associated with the probability of parameters being significantly different from zero are displayed for NELF (purple), INTS3 (pink), and MED26 (gold), as determined by the Wald test, which tests the hypothesis whether the variable assigned to the window can be removed without affecting the prediction model. P-values in the left half of the table are determined by ANOVA which sequentially compares the model containing only the previous variables to the model containing the respective variable. Z-values indicate the magnitude of how this parameter influences cluster classification towards clusters 2-4. A negative number would indicated that higher levels in this window would make it more likely to classify the gene as belonging to cluster 1. The window to which the maximum z-values belong are indicated with a grey star. D. Sensitivity vs FPR curves and the area under the curve for NELF. E. Sensitivity vs FPR curves and the area under the curve for INTS3. D. Sensitivity vs FPR curves and the area under the curve for MED26.

It is important to point out that instead of differentiating between 6 clusters, only two classes were considered for this analysis: Whether a gene belongs to cluster 1, which does not exhibit any pausing behaviour, or to cluster 2, 3, or 4, which have a sharp pausing profile but are very similar and may be too difficult to distinguish with this approach. Clusters 5 and 6 were ignored since they exhibit more nuanced profiles

of two pausing peaks or a broader profile, respectively. The data was also scaled within each window to make each column have the same mean of 1. As described in Figure 5.6 B, the data was then split randomly into a 'training' and a 'testing' set, which make up 75 % and 25 %, respectively.

The training set was used to estimate the parameters governing the hypothesis function, which is the sigmoid function that allows for a bifurcation of the data points and therefore yields a classification into two classes. The algorithm used to optimise the parameters is gradient descent that works by finding the minimum of the cost function in an iterative manner. A more detailed explanation of the algorithm and its implementation is provided in the methods section.

Using the training data, parameters for each window were calculated so that these fit the hypothesis function in order to predict the correct cluster assignment with the lowest possible value for the cost function. The significance level of each parameter for NELF, INTS3, and MED26 is shown in Figure 5.6 C. The parameter that is significant for determining the cluster assignment for NELF is associated with window 6. For INTS3 it is window 8, and for MED26 it is window 7.

The z-values of 2.25, 3.14, and 1.67, for NELF, INTS3, and MED26, respectively, show that within the specified windows, higher ChIP-Seq levels make it more likely for a gene to be classified as belonging to cluster 2 to 4. (A negative z-value would increase the likelihood of classifying the gene as being part of cluster 1). It is particularly striking that for NELF and INTS3, the most significant regions that are predictive of cluster assignments are windows 3-4, corresponding to 50-150 bp downstream of the TSS, which is approximately where one would expect the promoter proximal pause to

occur.

As discussed in the previous section, NELF has the strongest enrichment in cluster 2-4 out of the three factors looked at. Correspondingly, NELF has the strongest p-values in the logistic regression parameter determination (5.6 C). This would indicate that NELF ChIP-Seq levels are better suited for predicting the cluster assignment of a gene than integrator or mediator ChIP-Seq levels. This is also confirmed by Figure 5.6 D-F, which shows the receiver operating characteristic (ROC) curves plotting the specificity against the false positive rate.

A numerical characteristic value, the area under the curve (AUC), based on this graph is indicative of the overall accuracy of the algorithm. The AUC values are 0.805, 0.725, and 0.707 for NELF, INTS3, and MED26, respectively. A value of 0.5 would indicate an algorithm no better than pure chance of correctly predicting the cluster of a gene. Therefore, NELF has much better predictive power than the other two factors in terms of being able to assign NET-Seq based clusters on ChIP-Seq levels of protein factors alone.

The fact that mediator ChIP-Seq levels have the least predictive power for assigning a gene as belonging to a pausing or non-pausing cluster, might be due to the fact that mediator is considered to be universally involved in transcription, and mediator deletions are lethal across eukaryotes (Holstege et al., 1996; Ito et al., 2000, 2002; Malik and Roeder, 2010; Soutourina et al., 2011). Integrator, on the other hand, is involved in early termination that may result in profiles resembling promoter-proximal pausing profiles, but only for a subset of genes with less stable Pol II in the promoter regions, which may be the intermediate predictive power of integrator (Elrod et al.,

2019). Crucially, however, the significant window for cluster assignment prediction for INTS3 is further downstream of the TSS than NELF. This may, again, be explained by the role of integrator in early termination, where the cleavage subunits of integrator act on the 3' end of nascent transcripts which would explain the significant window for INTS3 being approximately 50 bp further downstream than the significant window for NELF.

Overall, however, NELF shows the strongest predictive power out of the three which may demonstrate that the profiles observed are largely caused by true pausing and early termination playing a more limited role. However, this conclusion has to be taken with caution because it assumes that NELF is not involved in early termination processes, for which there is no indication so far but could of course always be shown in future studies.

5.3 Identification of Alternative TSS Usage from Cluster Analysis

As can be observed in the metagene profiles shown in Figures 5.2, 5.3, and 5.4, cluster 5 seems to behave particularly different compared to the other clusters, by having a minor peak closer to the TSS and a major, broader, bipartite peak further into the gene body, centering around approximately 300 basepairs downstream of the TSS. One potential explanation for this behaviour could be the usage of alternative start sites.

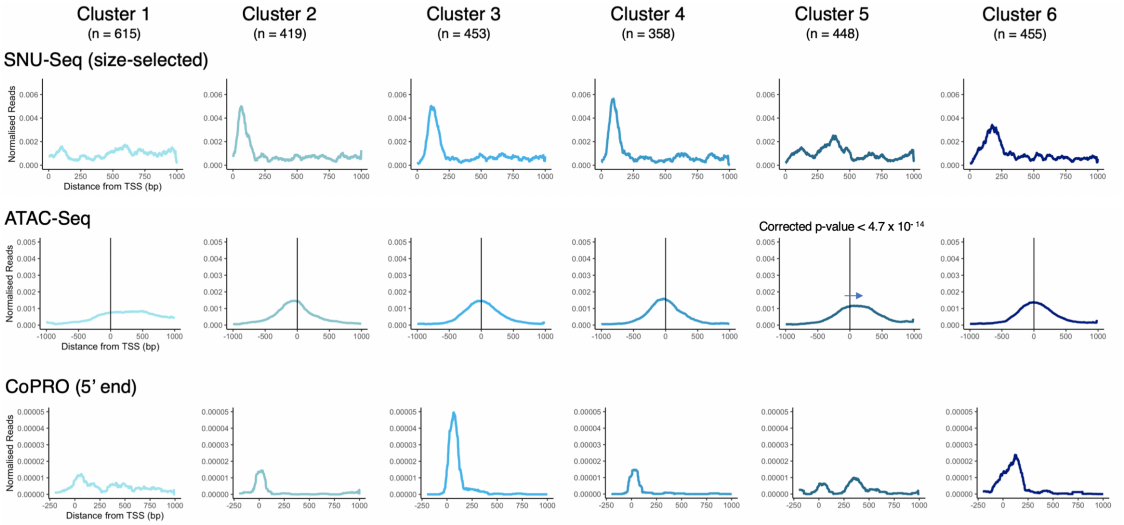


Figure 5.7: k-Means Cluster Profiles of Size-selected SNU-Seq, ATAC-Seq, and CoPRO

For each cluster (k = 1 to k = 6), the size-selected SNU-Seq, ATAC-Seq, and CoPRO metagenes are shown (Tome et al., 2018; Weltner et al., 2018). The clusters were originally generated using NET-Seq and the same cluster indices for each gene were applied to sort the SNU-Seq, ATAC-Seq, or CoPRO data into the respective clusters. Metagenes are normalised to make every gene contribute equally. Size-selected SNU-Seq and ATAC-Seq were performed in HEK cells, whereas CoPRO was performed in K562 cells.

In order to test this, a first step was to look at the peaks of ATAC-Seq profiles, which is shown in Figure 5.7. ATAC-Seq (Assay of Transposase Accessible Chromatin sequencing) measures chromatin accessibility (Yan et al., 2020). All ATAC-Seq peaks are centered around the TSS with a slight shift to the left reflecting the nucleosome-depleted region surrounding, and upstream of, the TSS. Cluster 1 has a much lower peak, which could be explained by the slightly lower levels of nascent transcription as demonstrated by SNU-Seq, TT-Seq and NET-Seq profiles.

Crucially, all ATAC-seq peaks are centered at or near the transcription start sites for clusters 2, 3, 4, and 6 (indicated by vertical black line). However, cluster 5 has a minor shift in the ATAC-peak position slightly further downstream, which is statistically significant when compared to all the other maximum peak positions (Wilcoxon rank-sum test, $p < 4.7 \times 10^{-14}$). This could potentially indicate a slight shift in the nu-

cleosome depleted region from which transcription begins, rather than an additional, or later pause site. To further confirm this, CoPRO data was utilized, which is a PRO-Seq approach with 5'cap enrichment (Tome et al., 2018), of which the 5' end signal is shown in the bottom row of Figure 5.7. There are two CoPRO peaks in cluster 5, which provides confirmatory evidence for the idea that the later and broader pausing peaks are in fact due to alternative transcription start site usage rather than a 'late pause'.

5.4 Gene Ontology Analysis

To characterize the regulatory context of promoter-proximal pausing, a calibrated pausing signal was calculated by dividing the summed size-selected SNU-Seq signals for each gene (0 – 150 bp downstream of the TSS) by the sum of TT-Seq reads for each gene (0 – 500 bp downstream of the TSS). This was the basis for performing ontology analysis on the top and bottom decile of genes ranked by their relative pausing signal. To determine gene ontologies, GORILLA (Eden et al., 2009) was used, utilising the 'biological process' option.

The result of this analysis is presented in Figure 5.8. Those genes that exhibit no promoter-proximal pausing behaviour tend to be involved in biological processes that could generally be summarised as 'house-keeping' functions. This includes translation ('Cytoplasmic Translation, 'Translation Initiation', 'Positive Regulation of Translational Elongation'), and protein trafficking ('Protein Targeting', Intracellular Protein Transport').

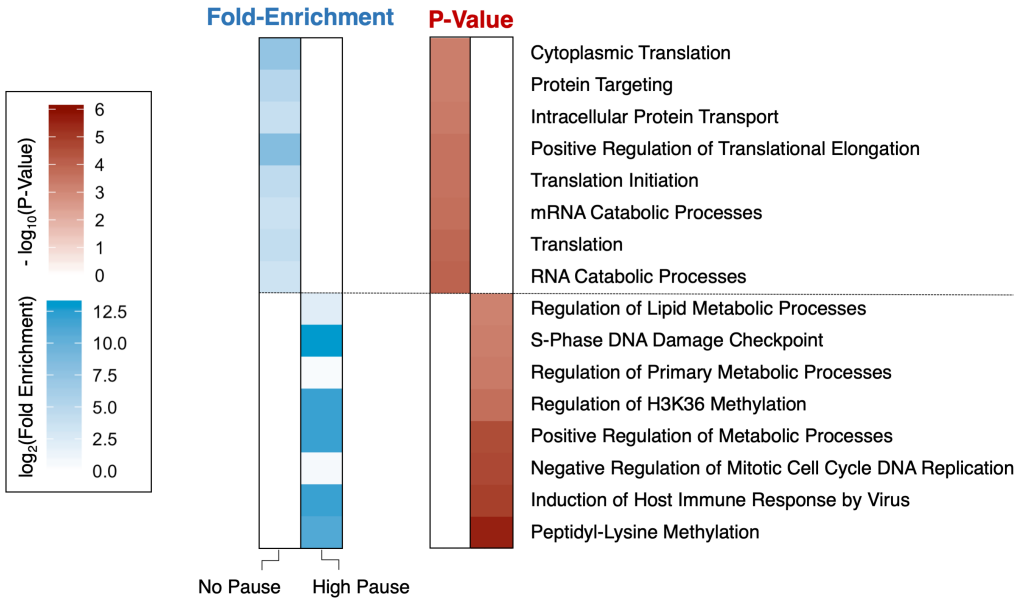


Figure 5.8: Gene Ontology Analysis of Genes with and without Promoter-Proximal Pausing

Fold-enrichment and p-values of gene ontologies as determined by GORILLA are shown for genes that exhibit high levels of calibrated pausing (top decile) or no pausing (bottom decile). log2-transformed values for the fold-enrichment and negative log10-transformed values for p-values are shown. The top and bottom decile each contained 934 genes (total number of genes was 9336). Genes with zero TT-Seq or zero SNU-Seq reads were ignored.

On the other hand, the Figure also shows that genes exhibiting high levels of pausing could generally be classified into 'regulatory' and 'synchronicity' functional groups, with some of the key ontologies involving cell cycle and DNA-repair ('S-Phase DNA Damage Checkpoint', 'Negative Regulation of Mitotic Cell Cycle DNA Replication'), or metabolism ('Regulation of Lipid Metabolic Processes', 'Positive Regulation of Primary Metabolic Processes', 'Positive Regulation of Metabolic Processes').

The ontologies found in this analysis largely agree with previous analyses on promoter-proximal pausing based on Pol II ChIP-Seq (Day *et al.*, 2016). One can conclude that because of the single-nucleotide resolution as well as the high sensitivity of size-selected SNU-Seq, this approach managed to generate highly statistically significant results on the gene ontology of promoter-proximal pausing compared to the

ChIP-Seq approach which generates low-resolution Pol II positioning profiles rather than precise information about the act of transcription.

Crucially, the analysis presented in Figure 5.8 shows that there is no overlap whatsoever in the biological function of gene ontologies between genes that display high levels of pausing behaviour and genes that do not. Further to this, there is not only no overlap; for all gene ontologies listed, those that show enrichment for genes with no pause, the enrichment for genes with high pausing is consistently zero, with the p-values consistently near one, and vice versa. This clear-cut distinction could point to the biological relevance of undergoing promoter-proximal pausing during the early stages of the transcription cycle.

In accordance with the observations discussed in the introduction and the potential importance of promoter-proximal pausing to enable the highly coordinated onset of gene expression of key regulatory genes, this gene ontology analysis points towards similar conclusions. The ontologies of the top decile are associated with biological processes that could be considered highly regulatable and synchronised procedures too, most importantly cell cycle regulation and DNA damage control. Likewise, the genes with the lowest levels of relative pausing encode proteins involved in those cellular processes that may require less synchronous or co-ordinated regulatability.

However, this analysis did not yield direct developmental ontologies *per se*. In order to provide an additional viewpoint at the regulatory framework of promoter-proximal pausing behaviour, the next step was to therefore also look at transcription factor enrichment.

5.5 Transcription Factor Enrichment Analysis

To further characterise the biological context of promoter-proximal pausing, the same relative pausing score was used to identify groups of transcription factors that are enriched at genes that exhibit high levels of pausing. For this, the OPOSSUM tool was used and single-site analysis was performed for each gene that had above zero TT-Seq and SNU-Seq reads (Kwon et al., 2012).

The resulting TFs that are enriched in the top decile of genes with respect to their relative pausing score, were given Fisher scores that correspond to the significance of their enrichment. Figure 5.9 shows those transcription factors with a Fisher score above 4. The transcription factor family is indicated by the bar colour. This enrichment analysis yielded a total of 60 transcription factors, from a wide range of 17 transcription factor families incluing GATA, Homeo box, Forkhead, Leucine Zipper, and others.

The presence of TBP in this list of enriched TFs is of some interest, especially since the TATA-box is considered to be absent in most *Drosophila* genes that exhibit strong levels of pausing - however, the TATA-binding protein itself, is present at promoters of pausing genes (Gaertner and Zeitlinger, 2014; Hendrix et al., 2008). Crucially, the most noticeable feature about the outcome of this transcription factor analysis is the high prevalence of factors involved in development: Out of the 60 enriched transcription factor in those genes that exhibit high relative levels of promoter-proximal pausing, a total of 48 transcription factors (80 %) are involved in developmental processes (GeneCards database). Amongst these, 15 are involved in regulation of early embryogenesis, morphogenesis, or sex determination, whereas the remaining 33 are involved in the regulation of tissue-specific development.

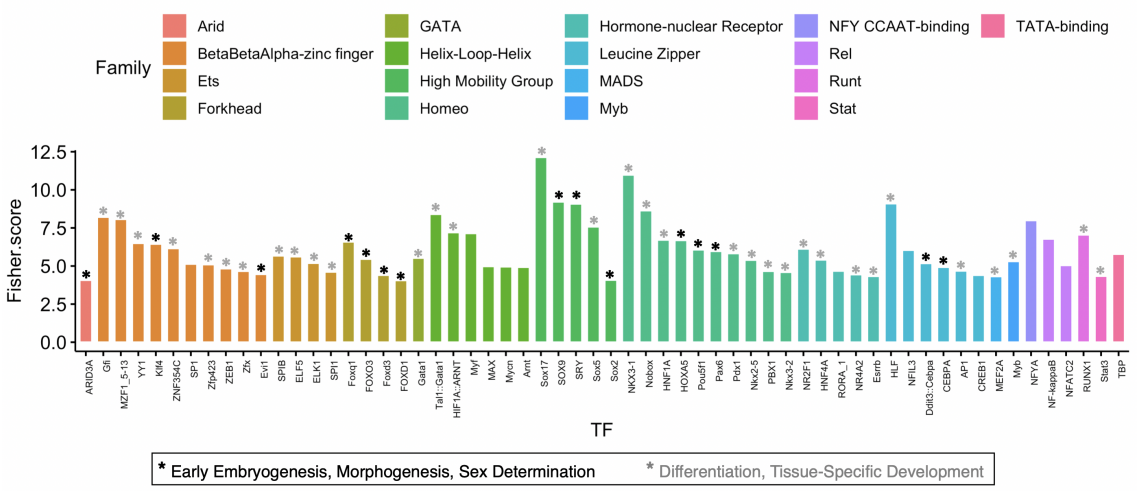


Figure 5.9: Transcription Factor Enrichment in Genes with High Levels of Pausing

Transcription factor enriched in genes within the top decile (n = 934) of the relative pausing score are shown. A Fisher score threshold of 4 was applied. The transcription factor family is indicated by the bar colour. A black or grey star indicates wether the transcription factor is involved in early embryogenesis or tissue-specific development, respectively. ARID3A: AT-rich interaction domain 3A. Gfi: Growth-factor independent repressor. MZF1: Myleoid zinc finger. YY1: ying and yang protein 1. KLF4: Kruppel-like factor 4. ZNF354: Zinc finger protein 354. Sp1: Specificity protein 1. ZFP423: Zinc finger ring protein 423. ZEB1: Zinc finger E-box binding homeobox 1. Zfx: Zinc finger x-linked. Evi1: Ectopic viral integration site 1. SPIB: Spi-B transcription factor. ELF5: E74-like ETS transcription factor 5. ELK1: ETS transcription factor ELK1. SPI1: Spi-1 proto-oncogene. Foxq1: Forkhead box Q1. FOXO3: Forkhead box O3. Foxd3: Forkhead box D3. Foxd1: Forkhead box D1. GATA1: GATA binding protein 1. TAL1: TAL BHLH transcription factor 1 / erythroid differentiation factor. Myf: Myogenic factor. MAX: MYC associated factor X. MYC: Myelocytomatosis viral oncogene homolog. MYCN: Myelocytomatosis viral oncogene neuroblastoma. ARNT: Aryl hydrocarbon receptor nuclear translocator. SOX17: SRY-box transcription factor 17. SOX9: SRY-box transcription factor 9. SRY: SRY: Sex-determining regino Y. SOX5: SRY-box transcription factor 5. SOX2: SRY-box transcription factor 2. NKX3-1: NK3 homeobox 1. NOBOX: Newborn ovary homeobox-encoding gene. HNF1A: Hepatic nuclear factor 1 homeobox A. HOXA5: Homeobox A5. POU5F1: POU class 5 homeobox 1. Pax6: Paired box 6. Pdx1: Pancreatic and duodenal homeobox 1. NKX2-1: NK2 homeobox 1. NR2F1: Nuclear receptor subfamily 2 group F member 1. HNF4A: Hepatocyte nuclear factor 4 alpha. RORA1: RAR related orphan receptor A 1. NR4A2: Nuclear receptor subfamily 4 grou A member 2. ESRRB: Estrogen related receptor beta. HLF: Hepatic leukemia factor. NFIL3: Nuclear factor, interleukin 3 regulated. CEBPA: CCAAT enhancer binding protein alpha. DDIT3: DNA damage inducible transcript 3. AP1: Activator protein 1. CREB1: cAMP responsive element binding protein 1. MEF2A: Myocyte enhancer factor 2A. MYB: Myeloblastosis viral oncogene homolog. NFYA: Nuclear transcription factor Y subunit alpha. NF-kB: Nuclear factor kappa-light-chain enhancer of activated B cells. NFATC2: Nuclear factor of activated T-cells, cytoplasmic 2. RUNX1: Runt-related transcription factor 1. STAT3: Signal transducer and activator of transcription 3. TBP: TATA-binding protein.

This strikingly high occurrence of developmental transcription factors binding to the genes that exhibit the highest levels of promoter-proximal pausing could be considered a genome-wide, and more global, support for the idea that pausing happens in order to coordinate the onset of transcription (elongation) across cells during devel-

opment, which originated in *Drosophila* (Gaertner and Zeitlinger, 2014; Hendrix et al., 2008; Lagha et al., 2013; Zeitlinger et al., 2007).

The reverse analysis was performed, too, in order to identify transcription factors that are enriched at those genes with the lowest levels of relative pausing (bottom decile). Interestingly, only one transcription factor, PLAGL1, came out of this analysis. This TF is a zinc finger protein known to suppress cell proliferation (Kas et al., 1998). This observation may be a reflection of the idea that genes that undergo promoter proximal pausing during their transcription cycle share common features and may be involved in key synchronous, regulatory features of metazoans - but that those genes that do not pause may not form a group of genes with similar features the way 'pausy' genes do.

In summary, the global, human, transcription factor enrichment at genes showing the highest relative degrees of promoter-proximal pausing seem to support the biological reasoning originally developed for single genes in *Drosophila*: That pausing occurs deliberately in order to allow for the coordination of the onset of transcription elongation across cells in metazoans for processes that require high synchronicity such as development (Gaertner and Zeitlinger, 2014; Lagha et al., 2013).

Chapter 6

A Pipeline for TSS Annotations using SNU-Seq

6.1 Establishing A Pipeline For SNU-Seq-based TSS Annotations

Having used the 3'end signal of the size-selected SNU-Seq dataset, a final step of this study was to develop a pipeline that allows for the annotation of transcription start sites based on the 5' signal. This is in effect using the same principle that has been used for TSS-annotations based on START-Seq (Nechaev *et al.*, 2010; Scruggs *et al.*, 2015; Lavender *et al.*, 2016). However, with the 4sU-labelling and -enrichment steps before size-selection, truly nascent RNA was used for this TSS annotation approach instead of the chromatin-associated or nuclear RNAs used in the traditional START-Seq protocols.

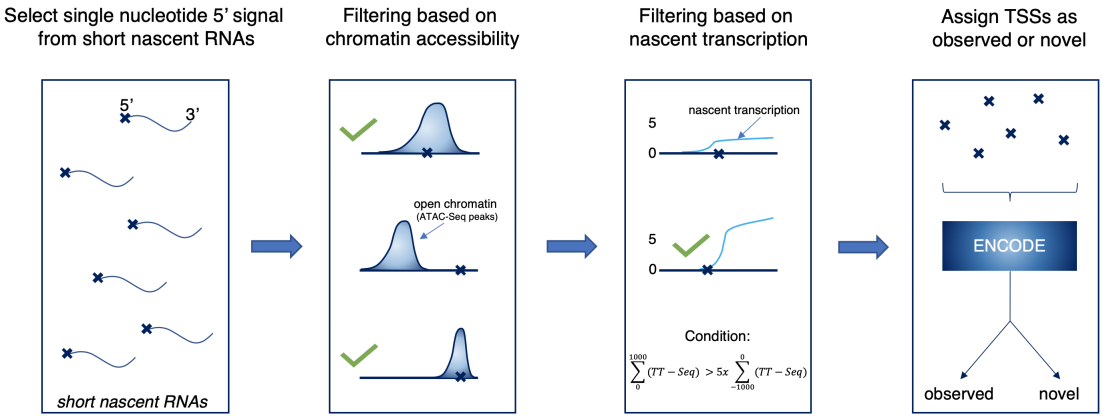


Figure 6.1: SNU-Seq based TSS Annotation Workflow

A simplified overview of the TSS annotation workflow is shown. For this, 5' end signals of size-selected SNU-Seq was filtered through ATAC-Seq peaks and, after clustering into TSS-clusters (not shown), filtered based on a minimum 5-fold increase in the TT-Seq signal immediately downstream of the TSS candidate. The verified TSSs were then classified as either previously annotated (obsTSSs) or novel / un-annotated TSSs (uTSSs).

A simplified overview of the workflow used to generate TSS annotations from size-selected SNU-Seq data is outlined in Figure 6.1. This includes usage of the 5'end signal, ATAC-Seq peak overlapping, TT-Seq signal verification, and TSS classification.

Initially, using the 5' end signal of the size-selected SNU-Seq bedgraph files, the signals are filtered through ATAC-Seq peaks to ensure only signals located in accessible chromatin are considered, using previously published ATAC-Seq data from HEK293 cells (Weltner et al., 2018). Since accessible chromatin is predominantly a feature of the nucleosome-depleted region that contains the transcription start site (reviewed in Klemm et al. (2019)), this step may help filtering out potential signals from new 5'ends that may be generated within the gene bodies during nascent transcription because of RNA cleavage events.

The filtered coordinates are then clustered into TSS clusters using paraclu, a clustering mechanism that has been applied to CAGE-Seq based TSS annotations too (Takahashi et al., 2011). As shown in Figure 6.2 A, the length of the cluster is predominantly very small which means that when taking the average position of each cluster forward, there is no considerable loss of resolution. The paraclu-based clusters were required to have a total read count of either singular or multiple reads of at least 30.

As a check on where the clustered, and ATAC-Seq filtered 5' end signals are located, a metagene of the full-length clusters was generated and shown in Figure 6.2. This shows that the signals are truly aligned at the ENCODE start sites and the cluster signals trail off with increasing distance downstream of the ENCODE TSSs.

The next step was to calculate the cluster centres. As described above, most of the paraclu clusters determined are very short, so using the cluster centres would mostly retain the high resolution of this approach. The cluster centres are then verified as true, active TSSs based on their TT-Seq profiles: If the summed TT-Seq signal up

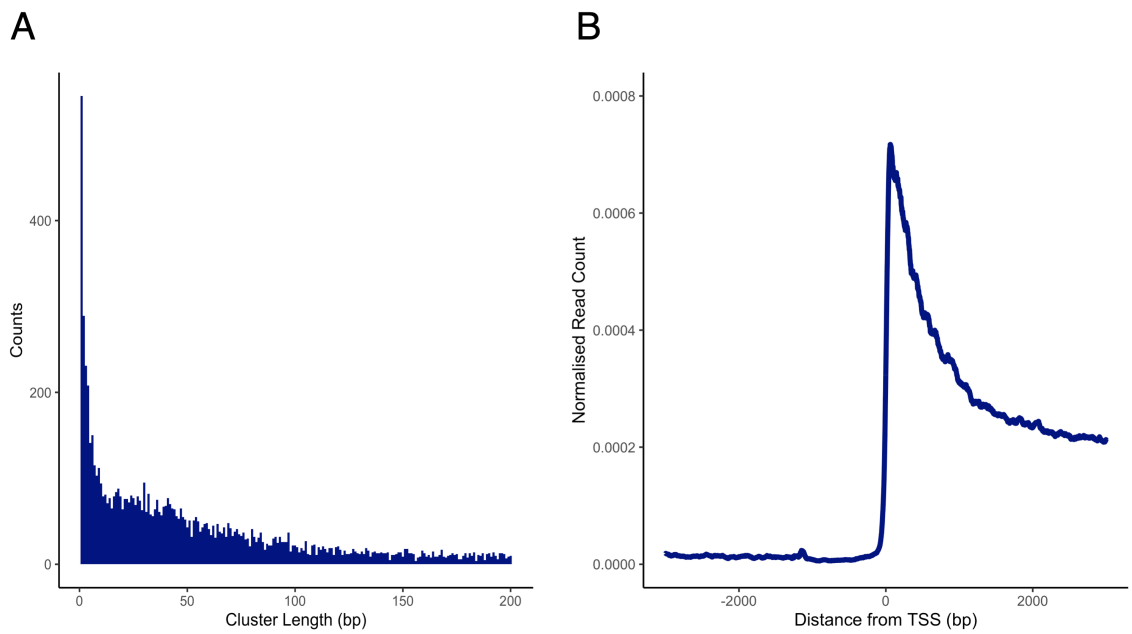


Figure 6.2: Clustering of SNU-Seq based TSS candidates

A. A histogram of the cluster lengths is shown. B. A metagene of the full-length clusters around ENCODE transcription start sites is shown. The signal is normalised so that each locus contributes equally to the metagene, by making the sum of reads for each cluster equal to 1.

to 1000 basepairs downstream of the cluster-centre position is at least 5 times greater than the summed TT-Seq signal within 1000 basepairs upstream of the cluster-centre position, then the candidate will be considered as a verified TSS. This was the case for 4383 candidates, which were then considered to be verified and transcriptionally active transcription start sites.

The verified TSSs are mapped in a Manhattan plot shown in Figure 6.3. This figure shows the location of each TSS according to their chromosome and strand coordinates, as well as the corresponding read number of tags and the TSS density (in a binsize of 10 million basepairs).

As shown in the Manhattan plot in Figure 6.3, the density of transcription start sites determined by size-selected SNU-Seq is mostly very even. In this context it is important to point out that it was not possible to determine transcription start sites on

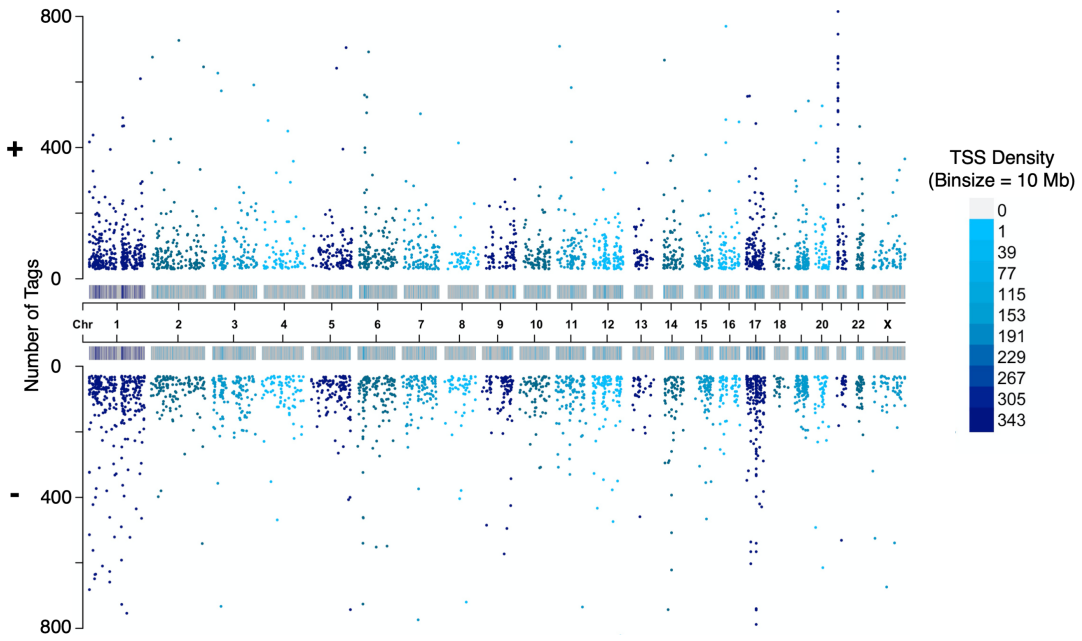


Figure 6.3: Manhattan Plot of SNU-Seq-based TSS annotations

The plot shows the location of each verified TSS annotation on the respective strand (+ or -) and chromosome. The y-axis corresponds to the number of reads that were associated with the original 5' end signal. The bars show the density of TSSs within each chromosome on each strand.

the Y chromosome because the underlying sequencing libraries were prepared from HEK293 cells which are female and therefore lack a Y chromosome.

To visualise the fact that the SNU-Seq-based TSSs correspond to the onset of active transcription, a heatmap of TT-Seq reads is shown in Figure 6.4 A. The nascent transcription emanating from these TSSs is predominantly evenly spread over several hundreds or thousands of basepairs downstream, and in general is of considerable length, which would support the notion that the TSS annotations verified correspond to full-length transcripts and are not just transcriptional noise.

6.2 Characterisation of Novel TSSs

In addition to annotating start sites that are already present in the ENCODE database (obsTSSs), the SNU-Seq based approach uncovered more than 1400 novel, and pre-

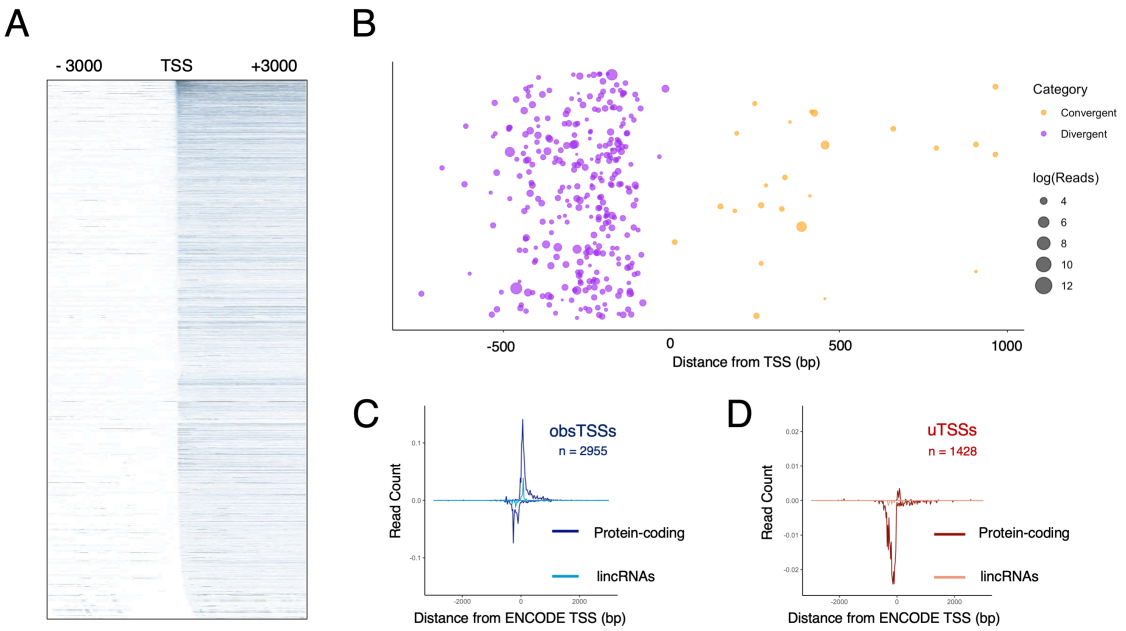


Figure 6.4: Characterisation of SNU-Seq-based TSS Annotations

A. A heatmap of TT-Seq around SNU-Seq based TSS annotations (signal range 0 - 500 reads), ordered by number of reads from 1 basepair downstream of the TSS in a descending order and displayed in a 6000-basepair window around the TSS. B. Categories of SNU-Seq-based TSSs that have not been annotated previously are classified as either convergent or divergent if they are in the vicinity of already annotated TSSs and on the opposite strand. The size of the data points corresponds to the log-transformed read count. C. Metagene of already annotated TSSs (obsTSSs, n = 2955) around protein-coding or lincRNA ENCODE TSSs. D. Metagene of novel/un-annotated TSSs (uTSSs, n = 1428) around protein-coding or lincRNA ENCODE TSSs.

viously un-annotated TSSs (uTSSs). This was determined by TSScall, a python code originally developed for START-Seq (Lavender et al., 2016). Further characterisation of these novel TSSs is provided in Figure 6.4 B-D. When looking at the novel TSSs in the context of where they are located with respect to the already annotated TSSs, those that are in the vicinity of obsTSSs can only be novel if they are on the opposite strand. A classification of these antisense uTSSs was performed and found that the majority of the antisense uTSSs form a divergent TSS pair with the obsTSSs, as shown in Figure 6.4 B.

In order to compare the positional information between those TSSs that were already annotated (obsTSSs) and the new TSSs (uTSSs), the TSS positions and SNU-

Seq read counts were mapped around ENCODE protein-coding and lincRNA TSSs, as shown in Figure 6.4 C and D. For obsTSSs, a familiar divergent TSS pattern can be seen, especially around protein-coding TSSs, with a sharp peak of the obsTSSs mapping exactly at the ENCODE TSSs. A slight trail of obsTSSs are slightly further downstream of the protein-coding ENCODE TSSs.

In agreement with the conclusions of Figure 6.4 B, the metagene of uTSSs around ENCODE protein-coding and lincRNA TSSs shows that the majority of obsTSSs, both in their position as well as their magnitude of reads, are positioned as upstream, divergent, antisense TSSs. Overall, this observation agrees with the idea that 4sU-based labelling for nascent transcriptomics is particularly suitable for detection of non-coding (antisense) transcription, as discussed in chapter 2. These might be missed by other methods that are used to annotate TSSs and have contributed to TSSs database.

One commonly used method for TSS annotation is CAGE, which relies on the selection of the 5'cap and does not have a nascent RNA-selection, which means it might miss the more unstable RNAs or those that have not been capped (Kodzius et al., 2006; Shiraki et al., 2003). This would explain the bias of the newly found SNU-Seq-based uTSSs towards upstream antisense TSSs, which may be less stable and harder to detect without nascent RNA selection.

Another aspect to bear in mind about these transcription start site annotations is, as can be seen in Figure 6.4 C, the trail of some SNU-Seq base annotations being located downstream of the ENCODE annotations. Using CoPRO data in the different pausing cluster has already shown that there is a disagreement about the exact 5' location of a subset of genes, with ENCODE commonly having the more 5' position

and other approached (CoPRO, size-selected SNU-Seq) a more 3' position.

A further piece of analysis that was done in this context is to see where the 3' ends of the short nascent RNAs are located with respect to either the ENCODE TSSs or the SNU-Seq based TSS annotations. This comparison is shown in Figure 6.5 for ENCODE annotations (A) and the SNU-Seq based annotations (B).

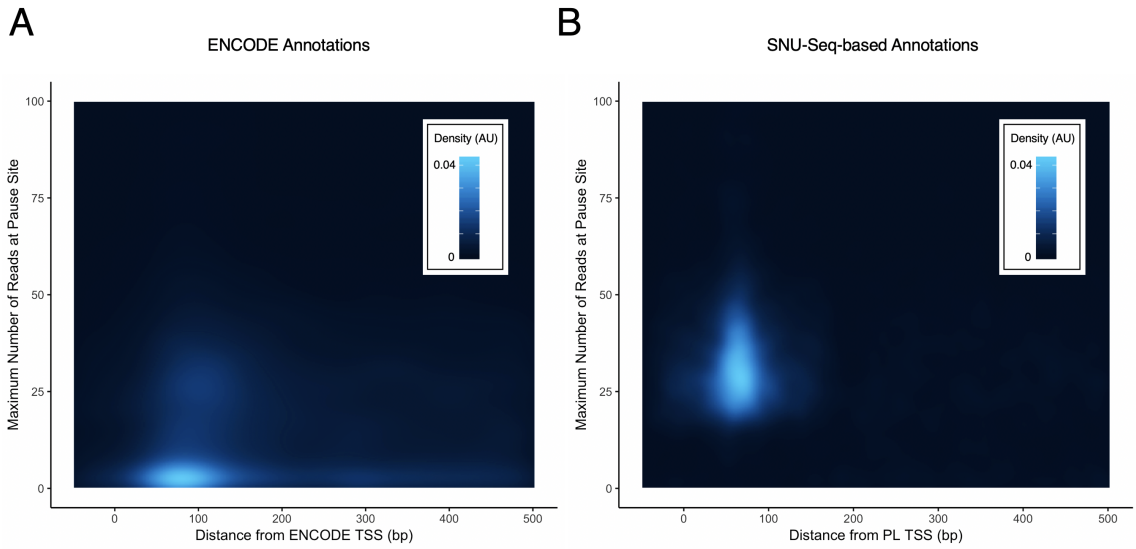


Figure 6.5: Density Plots of 3'ends of Short Nascent RNAs around Different TSS Annotations

A. A 2-dimensional density plot depicting the density of 3'ends of short nascent RNAs (from size-selected SNU-Seq). The maximum read value and its corresponding position was chosen, with the density indicated by the colour. The distance from the ENCODE TSS and maximum read value are indicated on the x-axis and y-axis, respectively. The density (arbitrary units) corresponds to the number of TSSs per pixel (n = 19874) B. A 2-dimensional density plot depicting the density of 3'ends of short nascent RNAs (from size-selected SNU-Seq). The maximum read value and its corresponding position was chosen, with the density indicated by the colour. The distance from the SNU-Seq based TSSs and maximum read value are indicated on the x-axis and y-axis, respectively. The density (arbitrary units) corresponds to the number of TSSs per pixel (n = 2955).

Strikingly, the pause signals, when mapped around the ENCODE TSSs, are almost distributed horizontally, centered between 50 and 150 bp downstream of the TSS, but with lower densities positioned at a much wider distance range. In contrast, when mapping the 3'end of the short nascent RNAs around the SNU-Seq based annotations, they are distributed almost vertically, with a sharp positional distribution between 50 and 100 bp downstream of the TSSs. Only negligible numbers of 3'ends

are located outside this range.

This would also point to potential issues with the ENCODE database of TSS annotations, since some of these TSSs may not have been verified through nascent transcription filters, for example. They might therefore not reflect the true starting point of nascent transcription, but only reflect the most 5' end signal of (alternative) transcripts that were picked as the TSS, even though the true TSS from which some of the nascent transcription emanates is actually located further downstream.

In conclusion, size-selected SNU-Seq can be a viable basis for transcription start site annotation, especially because, in contrast to other methods such as CAGE, the transcripts used for the annotations are nascent. The way in which TSSs were annotated for this study relied on a set of very strict filters, such as ATAC-Seq peaks, and a 5-fold increase in TT-Seq signal, which lead to a relatively small set of TSSs. However, because of these filters, one can assume that these TSSs belong to transcription units that are transcriptionally active and found in open and accessible chromatin.

As noticed with TT-Seq for measuring antisense transcription, a 4sU-based approach seems particularly suitable when interested in detecting non-coding and antisense transcription. This holds true for 4sU-based TSS annotations through size-selected SNU-Seq, too, given the high proportion of unkown uTSSs that are antisense to the TSSs of already annotated protein coding genes.

Chapter 7

Discussion

7.1 Summary

As outlined in chapter 1, the key aim of this study was to develop new nascent transcriptomic next-generation sequencing methods in order to gain a better understanding of the transcription process in human cells.

One outcome of this study is the usage of the already established TT-Seq method (developed by Schwalb B et al. (2016)) in new, and different cell lines (HEK293 HeLa, instead of K562 which are non-adherent), and for measuring antisense transcription. In contrast to other methods such as NET-Seq, TT-Seq captures transcription emanating from downstream antisense transcription start sites (Brown et al., 2018) at a similar average coverage than transcription from sense start sites. Although TT-Seq has a high sensitivity for non-coding transcription, one downside of this method is the lack of high resolution.

Therefore, this study aimed to develop a method that, like TT-Seq, is based on 4-sU labelling, but by skipping the sonication step and using artificial polyadenlyation, can retain single-nucleotide resolution at the 3' end of the nascent transcript. The resulting signals show a flat, even profile over the gene body irrespective of intronic or exonic regions. Because of the nature of this method, SNU-Seq can also detect poly(A)-sites of nascent transcripts.

SNU-Seq has a high resolution at the 3' end of the nascent transcripts, but the resolution can be lower in the rare case of intragenic poly(A)-tracks. By performing a size-selection of nascent transcripts with a size of approximately 60-80 basepairs, a library preparation procedure that is based on immediate adaptor ligation at both ends can therefore yield a sequencing signal of short nascent transcripts with single

nucleotide resolution at the 3' and 5' ends. This method was termed size-selected SNU-Seq and can capture promoter proximal pausing signal from the 3' end of these transcripts.

The profiles of size-selected SNU-Seq in defined clusters of genes match the corresponding NET-Seq clusters. The differential pausing behaviour as measured by these profiles is also reflected in differential NELF occupancy in these clusters. In fact, NELF ChIP-Seq levels are predictive of which clusters a gene may belong to.

Finally, the short nascent transcript sequencing signals were used to develop a SNU-Seq-based pipeline for the annotation of actively transcribed human transcription start-sites, which are verified through TT-Seq. Amongst these TSSs are both already annotated ones as well as previously unknown TSSs.

7.2 Why 4sU-Labelling?

Transient transcriptome sequencing (TT-Seq), developed by Schwalb B et al. (2016), like 4sU-Seq, utilises a nucleotide precursor analogue, 4-thiouridine for pulse labelling of nascent transcripts. 4sU can diffuse into cells where it is converted into 4-thio-UTP, which can then be incorporated into nascent RNA during the transcription process. A key feature of this approach is how chemically similar 4-thio-UTP is to cellular UTP. The only difference is one oxygen being replaced by sulfur, an atom of similar chemical properties and size. It is therefore safe to assume that 4-thio-UTP does not heavily interfere with the nascent transcription process. This is in stark contrast to other methods that require metabolic or other types of pulse labelling.

GRO-Seq (Nechaev et al., 2010), for example, uses a bromine-based nucleotide

analogue. Bromine is not only more different to oxygen, chemically and in size, than sulfur. The run-on approach with bromo-uridine also requires a subsequent antibody-based immunoprecipitation step. PRO-Seq (Kwak et al., 2013), on the other hand, uses nucleotides that have a covalent chemical bond to biotin to allow for subsequent strepdavidin purification. Because of the size of these nucleotides, the transcription process is immediately stopped upon incorporation. In addition, GRO-Seq and PRO-Seq require the isolation of nuclei or cell permeabilisation in addition to the usage of sarkozyl, which inhibits transcription re-initiation (reviewed in Wissink et al. (2019)).

NET-Seq (Nojima et al., 2015), on the other hand, is based on the precipitation of RNA polymerase II followed by the purification of the nascent RNA within the polymerase. This approach does not require a labelling step but there is a different reason why NET-Seq profiles may not directly reflect the true nascent transcription process: Using an antibody pull-down step will in fact bias the method towards RNA Pol II states that correspond to either pausing or slower transcription rates. In a way, NET-Seq signals are lower for faster transcription, and higher for slower transcription rates, which of course is counter-intuitive to how one would imagine a nascent transcription method should represent the transcription process.

One could therefore consider the thio-uridine approach for run-on/metabolic labelling of the nascent transcription process to be highly suitable for, and the least intrusive to, the ongoing nascent transcription process, but it also provides a way not to interfere in any way with the growing cell as it does not require any chemical or biochemical interference of the cells prior to or during the pulse-labelling.

7.3 Detection of Key Steps of the Eukaryotic Transcription Cycle by SNU-Seq and Size-selected SNU-Seq

Having developed SNU-Seq and size-selected SNU-Seq for measuring nascent transcription at high resolution, the two methods can detect various steps of the eukaryotic transcription cycle and map the corresponding annotation sites with high precision. This is summarised in Figure 7.1, which compounds the SNU-Seq and size-selected SNU-Seq profiles along a generalised gene body and which steps of the transcription cycles are measured by each method.

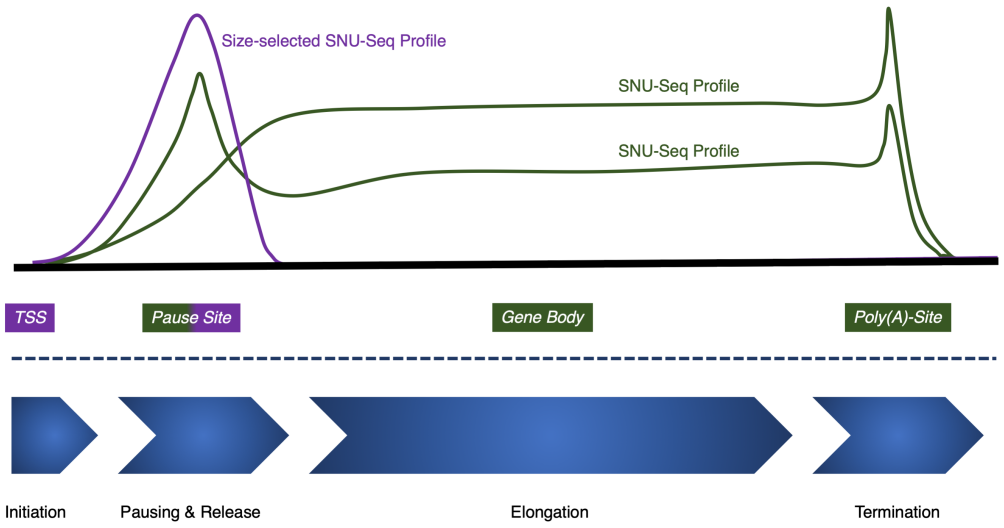


Figure 7.1: SNU-Seq Methods and Detection of Eukaryotic Transcription Cycle Steps

A summary of the SNU-Seq and size-selected SNU-Seq profiles along with the corresponding sites and steps of the transcription cycle they measure. Note that the distances between the sites are not to scale.

Furthermore, as discussed in Chapter 3, SNU-Seq (without size-selection) may have a minor bias towards the 3' end and may, in some rare cases, miss nascent transcriptional signals at the very 5' end of genes. This is because of the usage of

random primers during the amplification of the cDNAs during the library preparation. In some rare cases, the nascent transcript is too short to present enough combinations of nucleotides to form a sequence that could be bound by one of the random primers. Alternatively, the reverse transcriptase may not always reach the 5' end of long transcripts. This would make SNU-Seq on its own more suitable for the detection of overall nascent transcription levels over the gene body and, if required, poly(A)-site detection, but there is a risk that for some loci, SNU-Seq may not capture the ongoing nascent transcription process up until the pause or within the very 5' end of a small subset of genes. This is why SNU-Seq and size-selected SNU-Seq in combination deliver a complimentary, and comprehensive, representation of the eukaryotic transcription cycle. SNU-Seq with size selection is particularly suited for the detection of the TSS and the pause site, whereas SNU-Seq (without size-selection) is well-suited for measuring the ongoing nascent transcription process in the gene body as well as the detection of poly(A)-sites, and even alternative poly(A)-site usage that may be missed from standard RNA-Seq / 3'-end sequencing approaches because of the sensitivity of this method that is based on the nascent RNA selection.

As discussed in Chapter 5, SNU-Seq with size-selection captures small transcripts that are synthesised up until what would be considered the promoter-proximal pausing site. A range of recent studies in the field of eukaryotic transcription has tried to discuss and address the question of whether what is commonly referred to as pausing is in reality an early termination event (reviewed in Kamieniarz-Gdula and Proudfoot (2019)). The size-selected SNU-Seq profiles, as well as NET-Seq or PRO-Seq, on their own do not necessarily provide a direct answer of whether some of the elevated

signals near the 5' end of genes correspond to pausing or early termination.

This study, overall, is not aiming to take sides on this issue. However it is worth pointing out a range of observations from the bioinformatic analysis performed in this study that may provide at least a few insights: First, Elrod et al. (2019) have shown that the integrator complex is a key factor in promoting early termination events, leading to elevated PRO-Seq signals within the first 80-100 bp of *Drosophila* or human genes. However, their study pointed out that this is only relevant at genes with unstable RNA Pol II complexes in the 5' end of genes, meaning that this mechanism may only play a role in a subset of genes. In fact, when looking at the statistical and machine-learning approaches performed in this study with respect to the k-means clusters of differential pausing behaviours, it is NELF, and not integrator, that is both more enriched in, and predictive for, genes that exhibit elevated NET-Seq and size-selected SNU-Seq signals in their 5'.

This is by no means a rejection of the idea that genes can undergo early termination events during the early stages of the transcription cycle, but this study would potentially indicate that (NELF-mediated) pausing is more common genome-wide. This may perhaps provide a way of unifying the opposing ideas, in that pausing and early termination are not mutually exclusive overall, but that whether one or the other occurs depends on the subset of genes that is examined.

7.4 Future and Additional Work

Based on the observation and results offered in this study covering the development of SNU-Seq and size-seleted SNU-Seq, there are a range of experiments that can be

done on top of the already executed work. In this context, it is worth pointing out that due to the global COVID-19 pandemic, these experiments were not possible to carry out as part of this study and will therefore have to be part of future projects.

One additional set of experiments that could be carried out has to do with addressing some of the intrinsic features and potential issues with the standard SNU-Seq protocol. Whilst this method has high resolution, a straightforward protocol, and could generally be considered an efficient method, the poly(A)-tailing steps means that in rare cases, the method shows artificially high signals for naturally occurring, intragenic poly(A)-sequences. This issue has been addressed bioinformatically, as described in chapter 2, and in the methods section.

Additionally, however, one could make use of oligo-dT beads or poly(A)+ columns to remove those nascent transcripts that have any naturally occurring poly(A)-sequences. The advantage of this procedure is that it would ensure that SNU-Seq has single nucleotide resolution for 100 % of the reads – whereas before, the resolution was lower in these very rare cases. The downside of this, however, could be that usage of such additional selection steps would increase the number of times an RNA purification would have to be performed, which can reduce the overall yield each time.

An alternative approach to achieving high resolution form 4sU-labelled RNAs would be to use an alternative tail or adapter at the 3' end. Library preparation with oligo-dT primers is well established and a handy procedure. However, usage of a specific adapter at the 3' end would allow for accurate, nucleotide-resolution detection of the 3' end of nascent transcripts even from those RNA with intergenic poly(A) tracks.

With the development of size-selected SNU-Seq, on the other hand, a sensitive

and precise approach was set up to identify promoter-proximal pausing sites and, using k-means clustering (chapter 5), differential pausing profiles in different clusters of genes were identified, and these clusters match the corresponding NET-Seq profiles. In order to interrogate to what extent these cluster profiles are a true reflection of pausing behaviour, or whether they are in some way just an artefact of size-selected SNU-Seq and / or NET-Seq, ChIP-Seq data of the well-established pausing factor NELF was utilised. The cluster analysis of NELF, combined with statistical enrichment analysis, shows that the clusters displaying distinctive pausing peaks have significantly higher levels of NELF in conjunction with matching NELF-profiles. An additional verification of the biological relevance of these clusters was that, in a reverse approach, NET-Seq levels in defined windows around the TSS can actually predict whether a gene belongs into a 'pausy' cluster or not.

This would support the idea that size-selected SNU-Seq does in fact capture promoter-proximal pausing of the nascent transcription process. However, when it comes to the biological verification, one was required to utilise transcription factor ChIP-Seq data of proteins that are already known to be relevant in pausing, and, crucially, not in other steps of the transcription cycle. Therefore, one experiment that could be of considerable interest to understand more about the regulation of promoter-proximal pausing, is to perform an *in vitro* RNA immunoprecipitation: One could add either the bulk nascent short RNAs (with their 4sU-label), or a single, *in vitro* transcribed short nascent RNA to a nuclear cell extract. Then, after crosslinking, the proteins bound to the short RNAs with their 4sU label could be selected through HPDP-biotinylation followed by streptavidin purification. Finally proteins can then be

identified through mass spectrometry. Although this experiment may require a large amount of short nascent RNAs, it could be of considerable interest to identify the proteome that is associated with such short nascent RNAs which are synthesised up until the point of promoter-proximal pausing.

In addition to these experiments, it is also worth considering what combinational approaches that combine different sequencing methods may achieve. For example, applying 4sU-based sequencing methods (high sensitivity, ease of use) together with NET-Seq (high resolution, RNA processing information) on the same sample could provide a much more comprehensive, and quantifiable read-out of the nascent transcription process.

Finally, on remaining limitation that is common to all currently used nascent transcritpomic sequencing methods is the fact that it is performed on a batch of cells and that the resulting profiles present an average of the population.

Chapter 8

Methods and Materials

8.1 Methods

8.1.1 Cell Culture

HeLa and HEK293 cells were cultured in DMEM supplemented with 10 % (v/v) FBS and 1 % (v/v) Penicillin-Streptomycin. The incubator was set at 37 °C at 5 % CO2. Cells were passaged 24 hours before harvesting for next-generation sequencing experiments at which point they were grown to 80 % confluency. Cells were counted on a Nexcelom Biosciences Auto 2000 Cell Counter.

8.1.2 TT-Seq

TT-Seq was performed as described in the original publication (Schwalb B et al., 2016), with few modifications to the protocol as described below:

Thiolabeled spike-ins were generated by amplifying 3 1000-bp long DNA fragments from genomic yeast DNA with a 42 % GC content, covering the genes *bat2*, *hxt1*, and *gal1* using forward primers containing the T7 promoter sequence. The PCR product was purified and then transcribed *in vitro* following the MEGAscript T7 transcription kit instructions, using a thio-UTP to UTP ratio of 1:5.

For alpha-amanitin treatment, 60 million HeLa cells were incubated with or without 25 µg / mL alpha-amanitin for 6 hours at 37 °C. This concentration was determined based on the aim to inhibit RNA Pol II, but not RNA Pol III (Bensaude, 2011).

For TT-Seq labelling, 60 million cells per sample were incubated with 500 µM 4sU for 10 minutes (HEK293) or 5 minutes (HeLa) at 37 °C. Cells were washed rapidly in PBS before adding 0.7 mL QIAZOL per 10 million cells. The cells were scraped

off into 2mL Eppendorffs, vortexed for 10 s and placed on ice for 10 minutes. If required for comparative calibration, 2.4 ng of thio-labelled RNA spike-ins per 1 million cells were added at this stage. Total RNA was purified by adding 320 µL chloroform, vortexing for 10 s, incubating at room temperature for 3 minutes, and centrifugation at 4 °C for 15 minutes at 12000 rpm. The upper phase was transferred to a new Eppendorf with 800 uL isopropanol. The sample was left at room temperature for 5 minutes before centrifugation at 4 °C for 15 minutes at 15000 rpm. After removing the supernatant, the pellet was washed 3 times in 320 uL 75% ethanol (prepared freshly), with centrifugation at 4 °C for 15 minutes at 15000 rpm. The pellet was then air-dried for 10 minutes before dissolving in 160 µL RNAse-free water. At this stage samples were frozen at -80 °C.

RNA amounts were checked with Nanodrop. 300 µg of total RNA in 400 µL water were chilled on ice for 10 minutes. Samples were sonicated in a Bioruptor at the 'HIGH' setting for 1 cycle at 30s on / 30 s off. Samples were split equally into two Eppendorfs, incubated at 65 °C for 10 minutes, and then on ice for 5 minutes. After quick-spinning, 300 µL RNase-free water, 100 µL 10x Biotinylation Buffer (100 mM Tris pH 7.5, 1m mM EDTA pH 8.0), 200 µL DMF, and 200 mL EZ-link HPDP Biotin in DMF (1mg / mL, stored in the dark at 4 °C). Incubate in the dark at room temperature for 2 hours with shaking at 800 rpm. RNA was purified using Phase-Lock tubes and resuspended in 100 µL RNase-free water. Samples were pooled back together.

For thio-labelled RNA separation, a wash buffer (100 mM Tris pH 7.5, 10 mM EDTA pH 8.0, 1 M NaCl, 0.1 % Tween-20) was prepared freshly – one kept at room temperature (RT-WB), and one at 65 °C (65-WB). 100 µL uMACS streptavidin beads

were added to 200 µL RNA-Biotin and incubated in a thermomixer at 4 °C for 15 minutes at 800 rpm. During this time, the uMACS magnetic stand and rack were prepared alongside a drainer. uMACS columns were placed into the magnetic rack and equilibrated with 900 µL RT-WB for 10 minutes.

The biotin-RNA-bead mix was added to the uMACS column. The flow-through was collected and re-loaded to the columns. The second flow-through was discared. Columns were washed 3 times with 900 µL of 65-WB and then 3 times with 900 µL RT-WB. Elution of labelled RNA was carried out with 100 µL 100 mM DTT. After 5 minutes, elution was performed again with an additional 100 uL 100 mM DTT. Samples were then immediately carried through for RNA purification using the miRNeasy kit. RNA was eluted with 15 µL RNase-free water. Quality and quantity of thiolabeled RNA was checked using the Agilent Bioanalyzer with an RNA pico chip.

Sequencing libraries were prepared using the NEBNext rRNA Depletion kit for Human Cells followed by the NEBNext II Ultra Directional RNA Library prep kit, following both kit instructions. Library quality was assessed using the Agilent Bioanalyzer with a DNA-High Sensitivity chip. Pooled libraries were sequenced on the Illumina NextSeq 500 platform using the High-Output Illumina Nextseq 500 kit (150 cycles, paire-end).

8.1.3 TT-Seq - Sequencing Analysis

Paired-end sequencing reads were quality-checked with fastqc and then trimmed using Trim-Galore to remove sequences with a quality below a Q score of 20. Trimmed reads were then aligned to the human genome (hg38 assembly) with hisat2 (Kim et al., 2015) using the –no-mixed and the -no-discordant flags. Aligned files in the

sam format were then filtered using samtools (Li et al., 2009) with the flags -q 40, -f 99, -F3852, -bS.

Calibration of samples was achieved by calculating scaling factors of spike-in counts between samples based on counts tables generated by the bioconductor package featureCounts (Liao et al., 2013). This was the case for the time-labelling series experiment (10, 15, 20 minute 4sU-labelling in HEK293 cells), and the alpha-amanitin treatment experiment (TT-Seq with or without alpha-amanitin treatment in HEK293 cells with a labelling time of 10 minutes). Bedgraph and bigwig files were generated from bam files using bedtools (Quinlan and Hall, 2010), and wigTobigWig (UCSC Genome Browser), respectively.

8.1.4 SNU-Seq without size-selection

Cells were thio-labelled and harvested as described in the TT-Seq section. RNA was purified with QIAZOL followed by isopropanol precipitation as described above. Following purification, the RNA was not sonicated but carried through to biotinylation and streptavidin purification. RNAs were polyadenylated using the NEB Poly A Polymerase kit following the kit instructions. The reaction with 150 ng thiolabeled RNA was left for 45 minutes at 37 °C before isopropanol precipitation. The pellet was resuspended in 11 µL RNase-free water. Qualities and amounts were checked on the Qubit fluoremeter and Agilent bioanalyzer. Libraries were prepared following the Quant-Seq Lexogen 3' mRNA kit instructions with 13 PCR cycles. Following the manufacturer's instructions, the chips loaded by the Ion Chef were then sequenced on an Ion Proton Sequencing platform (single-end). SNU-Seq in HEB3B cells was

performed by Anna Lamstaes.

8.1.5 SNU-Seq without size-selection – Sequencing Analysis

After quality control with fastqc, single-end fastq files were trimmed with Trimmomatic (Bolger et al., 2014) to remove reads with a quality score below 20 in a sliding window of 5 bp. The poly(A) signals were removed with the clip option. Sequences were then aligned with hisat2 (Kim et al., 2015) with the same settings as for TT-Seq. Sorted bam files were generated using samtools (Li et al., 2009). Calibration of samples (if necessary) was achieved by calculating scaling factors of spike-in counts between samples based on counts tables generated by the bioconductor package featureCounts. Bedgraph and bigwig files were generated from bam files using bedtools (Quinlan and Hall, 2010), and wigTobigWig (UCSC Genome Browser), respectively. 3'end single basepair resolution was achieved by using the bedtools -bg -3 option.

In order to remove potential signals arising from internal priming due to intragenic poly-A stretches, peak-calling on the SNU-Seq bedgraph files (separately for each strand) was performed using macs2 (Zhang et al., 2008) with the narrowpeaks and no-model options. DNA sequences of these peaks were retrieved using the bedtools getfasta command before generating a bed file containing all peak coordinates with sequences containing at least 8 consecutive As (or Ts on the reverse strand). Finally, the 'bedtools intersect -wa' option was used to filter out any SNU-Seq signals that lie within these peaks.

8.1.6 SNU-Seq with size-selection

Nascent, thio-labelled RNA was generated as described for SNU-Seq without size selection. The nascent RNA was then run on a 3.5 % TBE-Urea gel (8M urea). For this, 250 ng of nascent RNA was mixed with 2x loading dye and boiled for 2 minutes at 80 °C. The gel was kept on ice and pre-run for 10 minutes at 80 V, using ice-cold ultra-pure TBE (1x) as the running buffer. Wells of the gel were resuspended with TBE before loading samples. Samples were then run at 80 V for 90 minutes. The gel was transferred onto a square petri dish and incubated in 1x ultra-pure TBE with SYBR Gold (1/10,000 v/v) for 5 minutes. The gel was rinsed twice in 1x ultra-pure TBE before imaging.

For purification of the small nascent RNA fraction, the gel was incubated at -80 °C for 10 minutes. The small band of 70-100 bp size was cut using a razor blade. RNA was extracted from the gel slice using dialysis: The gel slice, along with 0.8 mL 10 mM Tris-HCl (pH 7), was placed into SnakeSkin dialysis membrane sealed by two Eppendorf caps. The dialysis was run for 35 minutes at 80 V. The RNA was then purified with isopropanol precipitation and eluted in 11 µL RNase-free water. Quality and size distribution of the size-selected RNA was checked on the Agilent Bioanalyzer using an RNA pico chip.

The size selected nascent RNA was treated with 5' Pyrophosphohydrolase (NEB) in ThermoPol® reaction buffer following the manufacturer's instructions. The reaction was stopped with 1 µL 500 mM EDTA and heat-inactivated by incubation at 65 °C for 5 minutes. RNA was again purified with isopropanol precipitation and resuspended in 6 µL RNase-free water. The NEBNext Small RNA Library kit for Illumina following

the kit instructions. Library quality was assessed with the Agilent Bioanalyzer using a DNA High Sensitivity chip. Sequencing was performed on an Illumina NextSeq 500 platform.

8.1.7 SNU-Seq with size-selection – Sequencing Analysis

Fastq files were quality-trimmed with Trim-Galore (Babraham Bioinformatics) and aligned with bowtie2 (Langmead and Salzberg, 2012). Sorted bam files were generated from aligned sam files using samtools (Quinlan and Hall, 2010) with a filtering step added to only retain reads that have a mapping quality score above 30 (-q 30). As above, 3' end single basepair resolution was achieved by using the bedtools -bg -3 option when generating bedgraph files.

8.1.8 TSS annotations

To generate TSS annotations from the size-selected SNU-Seq samples, 5' end coverage was generated using the -bg -5 option (instead of -3) when generating bedgraphs from bam files in bedtools (Quinlan and Hall, 2010). To generate 5'end reads that only occur in nucleosome-depleted regions, ATAC-Seq data in HEK293 cells (Weltner et al. (2018), E-MTAB-6195) were used. The ATAC-Seq fastq files were trimmed with Trim-Galore, aligned with bowtie2, and converted to bam and bed files using samtools and bedtools, respectively.

Subsequently, macs2 (Zhang et al., 2008) was used to call peaks (no model assumed). The intersect option in bedtools was then used to retain only those 5'end SNU-Seq (size-selected) signals that are located within ATAC-Seq peaks. The result-

ing bedgraph files were then used as an input to identify TSS cluster centres using Paraclu (Takahashi et al., 2011) which identifies clusters in a sliding window, and a cluster value threshold of 30 was applied.

Cluster-centres were calculated using the mean position of the cluster, and TSS candidates were verified in Matlab as follows: Only those TSS candidates were assigned as true TSSs where the summed TT-seq signal (using HEK WT 10 minute-labelled TT-Seq) in the 100 bp downstream of the candidate TSS was at least 5 times greater than in the 100 bp upstream of the candidate TSS. TSScall, a python code (Lavender et al., 2016), was then used to identify annotated and un-annotated (novel) TSSs using the comprehensive Gencode (v29) annotation file which yielded 2955 previously annotated TSSs and 1428 novel transcription start sites. The read threshold for this was set to 5. The Linux commands for this pipeline are attached in the appendix (section 10.2).

8.1.9 Bioinformatics and Data Analysis

For metagenes, TSS annotations that were generated previously (Brown et al., 2018) were used, or, if indicated otherwise, ENCODE TSS and poly(A)-site annotations (v29). In MATLAB, the read counts for each gene in the specified window were normalised to a total sum of 1, in order to make each gene contribute equally to the metagene plots. At each position, the top and bottom 0.5 % of entries were removed, and the final signals were then averaged over 10bp-moving windows. An exemplary code (for NELF ChIP-Seq metagenes) is shown in the Appendix (section 10.1).

To be able to compare the different sequencing methods, NET-Seq (Nojima et al.,

2015) and GRO-Seq (Laitem et al., 2015) reads were re-aligned to the latest human genome version (hg38) as reads from the original publications were aligned to hg19. The remaining processing steps were performed as described in the respective original publications (Laitem et al., 2015; Nojima et al., 2015).

Positional read ratios were calculated by division of reads near the TSS (first 200 nucleotides) over the reads over the gene body (first 1000 nucleotides) for each gene (n = 9336, annotations by Brown et al. (2018))/ Intron/Exon read ratio (log scale) were determined by the summed reads in 25-bp windows upstream and downstream of the first exon-intron boundary of 9336 genes, as annotated previously (Brown et al., 2018), for 10, 15, and 20 minute 4sU-labelled HEK293 cells.

p-values were determined using the non-parametric Wilcoxon rank sum test, and the Bonferroni correction for multiple testing was applied when required. To determine the correlation between repeats, the Pearson correlation coefficient was calculated based on the counts tables. In all other cases, the Spearman ranked correlation (rho) was calculated. Princple Component Analysis (PCA) and data visualization were performed in R using the PCAtools and ggplot2 packages, respectively.

For gene ontology analysis, a list of gene ranked by the relative pause signal was used to generate GO terms for 'biological process' using GORILLA (Eden et al., 2009). For TF enrichment, the SSA (Single-site analysis) tool provided by Opossum-3 (Kwon et al., 2012) was used.

8.1.10 Cluster Analysis

Clustering was performed on calibrated profiles of NET-Seq reads. Only genes with a length of at least 1000 bp were considered for this analysis, in addition to the requirement of having NET-Seq reads above the average background across the genome (t-test, $p > 0.01$). MATLAB was then used to perform k-means clustering for values of k between 1 and 20. k=6 was eventually chosen by the elbow method, as the value that encompasses the variance amongst all genes without any redundancy. This means that addition of any more clusters would have split one group into two subgroups with (almost) identical profiles. The original generation of these clusters was performed in collaboration with Tom Brown.

Normalised size-selected SNU-Seq, TT-Seq (with or without alpha-amanitin treatment), ATAC-Seq, ChIP-Seq, and CoPRO (5') bigwig files were then transformed into dataframes with signals around TSSs within the same cluster annotations in MATLAB. Metagenes for these sequencing methods were generated as above, however the removal of the top and bottom 0.5 % of signals was omitted apart from the CoPRO metagenes.

For logistic regression of ChIP-Seq data, 25-basepair summed windows for each dataset were generated and scaled across columns to make each window mean equal 1. Cluster assignments were simplified to have genes either assigned into cluster 1 or clusters 2-4 (clusters 5 and 6 were ignored). The scaled data for each factor was split into training (75 %) and testing (25 %) sets. A general linear model was built in R using the glm function with the binomial parameter family and the sigmoid function as the hypothesis function, making use of the training data as the input.

The summary function of the glm R tool for the model provided the z-value for each parameter for each window as well as their corresponding significance levels. Two sets of p-values were determined to assess the significance of the parameters. One set of p-values corresponds to the probability of the parameters being significantly different from zero, which were determined by the Wald test. This tests the hypothesis whether the variable assigned to a given window can be removed without affecting the prediction model. The second set of p-values were determined by ANOVA through sequentially comparing the model containing only the previous variables to the model containing the respective variable. These tests were performed through the 'summary' and 'anova' functions in the glm tool for R.

The overall performance was calculated using the testing set by comparing the actual versus predicted cluster assignment. This allowed for the subsequent determination of the true positive rates, false positive rates, false negative rate, and true negative rates, and by extension the receiver operating characteristic curve (ROC curve) with the corresponding area under the curve (AUC).

The resources of the publicly available datasets used in this study are listed in the materials section below.

8.1.11 Mathematical Modelling and Determination of Transcription Constants

Synthesis and decay rates were determined based on a previously published modelling approached (Villamil et al., 2019). HEK293 RNA-seq counts were used from Banks et al. (2014). The following equations were used:

$$\lambda(i,j) = \frac{1}{t} \times log(\frac{1 - L(i,j)}{T(i,j)})$$

$$\mu(i,j) = T(i,j) \times \lambda(i,j)$$

where λ(i,j) is the decay rate, t is the labelling time, μ(i,j) is the synthesis rate, T(i,j) is the RNA-seq counts (length-normalised tags per million, TPM) and L(i,j) is the SNU-seq counts (TPM). Scaling factors will become cancelled out in the calculation of the Synthesis Rate / Decay Rate ratio. Application of these equations to the SNU-Seq and TT-Seq data was done by Aksel Saukko-Paavola. The downstream analysis for determining Spearman correlation coefficients and visualisation with ggplot was performed by the author of this study.

8.2 Resources and Materials

8.2.1 Sequencing Data

Dataset	Cell Line	Source
SNU-Seq	HEK293	This Study
SNU-Seq	HEP3B	Anna Lamstaes
SNU-Seq with Size Selection	HEK293	This Study
TT-Seq	HEK293	This Study
TT-Seq	HeLa	This Study
TT-Seq (+/- alpha-amanitin)	HeLa	This Study
TF-Seq	HeLa	This Study
TT-Seq	K562	GSE75792
NET-Seq	HeLa	GSE60358
GRO-Seq	HeLa	E-MTAB-3360
CoPRO	K562	GSE116472
ATAC-Seq	HEK293	E-MTAB-6195
RNA-Seq	HEK293	E-GEOD-57027
ChIP-Seq (NELF)	HeLa	GSM1280296
ChIP-Seq (INTS3)	HeLa	GSM2976713
ChIP-Seq (MED26)	HeLa	GSE121024
PRO-Seq	HEK293	GSE103719

8.2.2 Kits and Reagents

Kit / Reagent Name	Provider	Code
NEBNext rRNA Depletion Kit (Human/Mouse/Rat)	NEB	E6350S
NEBNext Multiplex Oligos (Set 1)	NEB	E7335L
NEBNext Ultra II Directional RNA Library Prep	NEB	E7765S
NEBNext Small RNA Library Prep Kit	NEB	E7300S
QuantSeq 3'end Kit for Ion Torrent	Lexogen	012.24A
uMACS Streptavidin Kit	Miltenyi Biotech	130074101
NextSeq 500/550 High Output v2.5 kit (150 cycles)	Illumina	20024907
4-thiouridine	Carbosynth	NT06186
EZ-link HPDP-Biotin Thermo	Fisher	21341
SnakeSkin Dialysis Tubing, 16 mm	Thermo Fisher	88242
Poly(A) Polymerase Kit	NEB	M0276S
RNA 5' Pyrophosphohydrolase (RppH)	NEB	M0356S
ThermoPol® Reaction Buffer Pack	NEB	B9004S
miRNeasy Micro Kit	Qiagen	217084
5-PRIME Phase Lock Gel Heavy	VWR	733-2478

Chapter 9

Appendix

9.1 Metagene Code (MATLAB)

```
% Make metagene of NELF ChIP-Seq.

    % Upload bigwig converted to txt file:

    NELF = readtable('hg38_NELF_ChIP_Seq.txt');

    NELF_col1 = table2array(NELF(:,1));

    NELF_col1 = str2double(NELF_col1);

    NELF_col2 = table2array(NELF(:,2));

    NELF_col3 = table2array(NELF(:,3));

    NELF_col4 = table2array(NELF(:,4));

    NELF_array = [NELF_col1 NELF_col2 NELF_col3 NELF_col4];


    % Upload TSSs and strand vector from ENCODE database

    sense_TSSs = TSSs;

    sense_directions = strand_vector;

    position_counter = -1000:1:1000;

    all_positions_s = repmat(position_counter, 58344, 1);

    s_positions = all_positions_s + sense_TSSs;


    % Create empty arrays. chr23 = X; chr24 = Y.

    NELF_chr1 = zeros(300000000,1);

    NELF_chr10 = zeros(300000000,1);

    NELF_chr11 = zeros(300000000,1);
```

```
NELF_chr12 = zeros(300000000,1);

NELF_chr13 = zeros(300000000,1);

NELF_chr14 = zeros(300000000,1);

NELF_chr15 = zeros(300000000,1);

NELF_chr16 = zeros(300000000,1);

NELF_chr17 = zeros(300000000,1);

NELF_chr18 = zeros(300000000,1);

NELF_chr19 = zeros(300000000,1);

NELF_chr2 = zeros(300000000,1);

NELF_chr20 = zeros(300000000,1);

NELF_chr21 = zeros(300000000,1);

NELF_chr22 = zeros(300000000,1);

NELF_chr3 = zeros(300000000,1);

NELF_chr4 = zeros(300000000,1);

NELF_chr5 = zeros(300000000,1);

NELF_chr6 = zeros(300000000,1);

NELF_chr7 = zeros(300000000,1);

NELF_chr8 = zeros(300000000,1);

NELF_chr9 = zeros(300000000,1);

NELF_chr23 = zeros(300000000,1);

NELF_chr24 = zeros(300000000,1);


idxchr1 = find(NELF_array(:,1)==1);

idxchr2 = find(NELF_array(:,1) == 2);

idxchr3 = find(NELF_array(:,1) == 3);
```

```
idxchr4 = find(NELF_array(:,1) == 4);

idxchr5 = find(NELF_array(:,1) == 5);

idxchr6 = find(NELF_array(:,1) == 6);

idxchr7 = find(NELF_array(:,1) == 7);

idxchr8 = find(NELF_array(:,1) == 8);

idxchr9 = find(NELF_array(:,1) == 9);

idxchr10 = find(NELF_array(:,1) == 10);

idxchr11 = find(NELF_array(:,1) == 11);

idxchr12 = find(NELF_array(:,1) == 12);

idxchr13 = find(NELF_array(:,1) == 13);

idxchr14 = find(NELF_array(:,1) == 14);

idxchr15 = find(NELF_array(:,1) == 15);

idxchr16 = find(NELF_array(:,1) == 16);

idxchr17 = find(NELF_array(:,1) == 17);

idxchr18 = find(NELF_array(:,1) == 18);

idxchr19 = find(NELF_array(:,1) == 19);

idxchr20 = find(NELF_array(:,1) == 20);

idxchr21 = find(NELF_array(:,1) == 21);

idxchr22 = find(NELF_array(:,1) == 22);

idxchr23 = find(NELF_array(:,1) == 23);

idxchr24 = find(NELF_array(:,1) == 24);


% For each chromosome, take the reads from ranges (e.g. 50-53) into

% single-bp positions (50, 51, 52, 53).

for i = 1:length(idxchr1)
```

```
NELF_chr1(NELF_array(idxchr1(i),2):NELF_array(idxchr1(i),3)) =
NELF_array(idxchr1(i),4);
end
NELF_chr1 = sparse(NELF_chr1);

for i = 1:length(idxchr10)
NELF_chr10(NELF_array(idxchr10(i),2):NELF_array(idxchr10(i),3)) =
NELF_array(idxchr10(i),4);
end
NELF_chr10 = sparse(NELF_chr10);

for i = 1:length(idxchr11)
NELF_chr11(NELF_array(idxchr11(i),2):NELF_array(idxchr11(i),3)) =
NELF_array(idxchr11(i),4);
end
NELF_chr11 = sparse(NELF_chr11);

for i = 1:length(idxchr12)
NELF_chr12(NELF_array(idxchr12(i),2):NELF_array(idxchr12(i),3)) =
NELF_array(idxchr12(i),4);
end
NELF_chr12 = sparse(NELF_chr12);

for i = 1:length(idxchr13)
NELF_chr13(NELF_array(idxchr13(i),2):NELF_array(idxchr13(i),3)) =
```

```
NELF_array(idxchr13(i),4);

end

NELF_chr13 = sparse(NELF_chr13);


for i = 1:length(idxchr14)

NELF_chr14(NELF_array(idxchr14(i),2):NELF_array(idxchr14(i),3)) =

NELF_array(idxchr14(i),4);

end

NELF_chr14 = sparse(NELF_chr14);


for i = 1:length(idxchr15)

NELF_chr15(NELF_array(idxchr15(i),2):NELF_array(idxchr15(i),3)) =

NELF_array(idxchr15(i),4);

end

NELF_chr15 = sparse(NELF_chr15);


for i = 1:length(idxchr16)

NELF_chr16(NELF_array(idxchr16(i),2):NELF_array(idxchr16(i),3)) =

NELF_array(idxchr16(i),4);

end

NELF_chr16 = sparse(NELF_chr16);


for i = 1:length(idxchr17)

NELF_chr17(NELF_array(idxchr17(i),2):NELF_array(idxchr17(i),3)) =

NELF_array(idxchr17(i),4);
```

```
end

NELF_chr17 = sparse(NELF_chr17);


for i = 1:length(idxchr18)

NELF_chr18(NELF_array(idxchr18(i),2):NELF_array(idxchr18(i),3)) =

NELF_array(idxchr18(i),4);

end

NELF_chr18 = sparse(NELF_chr18);


for i = 1:length(idxchr19)

NELF_chr19(NELF_array(idxchr19(i),2):NELF_array(idxchr19(i),3)) =

NELF_array(idxchr19(i),4);

end

NELF_chr19 = sparse(NELF_chr19);


for i = 1:length(idxchr2)

NELF_chr2(NELF_array(idxchr2(i),2):NELF_array(idxchr2(i),3)) =

NELF_array(idxchr2(i),4);

end

NELF_chr2 = sparse(NELF_chr2);


for i = 1:length(idxchr20)

NELF_chr20(NELF_array(idxchr20(i),2):NELF_array(idxchr20(i),3)) =

NELF_array(idxchr20(i),4);

end
```

```
NELF_chr20 = sparse(NELF_chr20);


for i = 1:length(idxchr21)
NELF_chr21(NELF_array(idxchr21(i),2):NELF_array(idxchr21(i),3)) =
NELF_array(idxchr21(i),4);
end
NELF_chr21 = sparse(NELF_chr21);


for i = 1:length(idxchr22)
NELF_chr22(NELF_array(idxchr22(i),2):NELF_array(idxchr22(i),3)) =
NELF_array(idxchr22(i),4);
end
NELF_chr22 = sparse(NELF_chr22);


for i = 1:length(idxchr3)
NELF_chr3(NELF_array(idxchr3(i),2):NELF_array(idxchr3(i),3)) =
NELF_array(idxchr3(i),4);
end
NELF_chr3 = sparse(NELF_chr3);


for i = 1:length(idxchr4)
NELF_chr4(NELF_array(idxchr4(i),2):NELF_array(idxchr4(i),3)) =
NELF_array(idxchr4(i),4);
end
NELF_chr4 = sparse(NELF_chr4);
```

```
for i = 1:length(idxchr5)
NELF_chr5(NELF_array(idxchr5(i),2):NELF_array(idxchr5(i),3)) =
NELF_array(idxchr5(i),4);
end
NELF_chr5 = sparse(NELF_chr5);

for i = 1:length(idxchr6)
NELF_chr6(NELF_array(idxchr6(i),2):NELF_array(idxchr6(i),3)) =
NELF_array(idxchr6(i),4);
end
NELF_chr6 = sparse(NELF_chr6);

for i = 1:length(idxchr7)
NELF_chr7(NELF_array(idxchr7(i),2):NELF_array(idxchr7(i),3)) =
NELF_array(idxchr7(i),4);
end
NELF_chr7 = sparse(NELF_chr7);

for i = 1:length(idxchr8)
NELF_chr8(NELF_array(idxchr8(i),2):NELF_array(idxchr8(i),3)) =
NELF_array(idxchr8(i),4);
end
NELF_chr8 = sparse(NELF_chr8);

for i = 1:length(idxchr9)
```

```
NELF_chr9(NELF_array(idxchr9(i),2):NELF_array(idxchr9(i),3)) =
NELF_array(idxchr9(i),4);
end
NELF_chr9 = sparse(NELF_chr9);

for i = 1:length(idxchr23)
NELF_chr23(NELF_array(idxchr23(i),2):NELF_array(idxchr23(i),3)) =
NELF_array(idxchr23(i),4);
end
NELF_chr23 = sparse(NELF_chr23);

for i = 1:length(idxchr24)
NELF_chr24(NELF_array(idxchr24(i),2):NELF_array(idxchr24(i),3)) =
NELF_array(idxchr24(i),4);
end
NELF_chr24 = sparse(NELF_chr24);

% Create vectors for each chromosome from chromosome_array (ENCODE)
chromosome_nos = chromosome_array;
idx_genes_chr1 = find(chromosome_nos ==1);
idx_genes_chr2 = find(chromosome_nos ==2);
idx_genes_chr3 = find(chromosome_nos ==3);
idx_genes_chr4 = find(chromosome_nos ==4);
idx_genes_chr5 = find(chromosome_nos ==5);
idx_genes_chr6 = find(chromosome_nos ==6);
```

```
idx_genes_chr7 = find(chromosome_nos ==7);

idx_genes_chr8 = find(chromosome_nos ==8);

idx_genes_chr9 = find(chromosome_nos ==9);

idx_genes_chr10 = find(chromosome_nos ==10);

idx_genes_chr11 = find(chromosome_nos ==11);

idx_genes_chr12 = find(chromosome_nos ==12);

idx_genes_chr13 = find(chromosome_nos ==13);

idx_genes_chr14 = find(chromosome_nos ==14);

idx_genes_chr15 = find(chromosome_nos ==15);

idx_genes_chr16 = find(chromosome_nos ==16);

idx_genes_chr17 = find(chromosome_nos ==17);

idx_genes_chr18 = find(chromosome_nos ==18);

idx_genes_chr19 = find(chromosome_nos ==19);

idx_genes_chr20 = find(chromosome_nos ==20);

idx_genes_chr21 = find(chromosome_nos ==21);

idx_genes_chr22 = find(chromosome_nos ==22);

idx_genes_chr23 = find(chromosome_nos ==23);

idx_genes_chr24 = find(chromosome_nos ==24);


NELF_F_chr1 = zeros(length(idx_genes_chr1),2001);

NELF_F_chr2 = zeros(length(idx_genes_chr2),2001);

NELF_F_chr3 = zeros(length(idx_genes_chr3),2001);

NELF_F_chr4 = zeros(length(idx_genes_chr4),2001);

NELF_F_chr5 = zeros(length(idx_genes_chr5),2001);

NELF_F_chr6 = zeros(length(idx_genes_chr6),2001);
```

```
NELF_F_chr7 = zeros(length(idx_genes_chr7),2001);

NELF_F_chr8 = zeros(length(idx_genes_chr8),2001);

NELF_F_chr9 = zeros(length(idx_genes_chr9),2001);

NELF_F_chr10 = zeros(length(idx_genes_chr10),2001);

NELF_F_chr11 = zeros(length(idx_genes_chr11),2001);

NELF_F_chr12 = zeros(length(idx_genes_chr12),2001);

NELF_F_chr13 = zeros(length(idx_genes_chr13),2001);

NELF_F_chr14 = zeros(length(idx_genes_chr14),2001);

NELF_F_chr15 = zeros(length(idx_genes_chr15),2001);

NELF_F_chr16 = zeros(length(idx_genes_chr16),2001);

NELF_F_chr17 = zeros(length(idx_genes_chr17),2001);

NELF_F_chr18 = zeros(length(idx_genes_chr18),2001);

NELF_F_chr19 = zeros(length(idx_genes_chr19),2001);

NELF_F_chr20 = zeros(length(idx_genes_chr20),2001);

NELF_F_chr21 = zeros(length(idx_genes_chr21),2001);

NELF_F_chr22 = zeros(length(idx_genes_chr22),2001);

NELF_F_chr23 = zeros(length(idx_genes_chr23),2001);

NELF_F_chr24 = zeros(length(idx_genes_chr24),2001);


% Fill positions -1000 to 1000 around TSSs

for i = 1:length(idx_genes_chr1)

for j = 1:2001

NELF_F_chr1(i,j) = NELF_chr1([s_positions(idx_genes_chr1(i),j)]);

end

end
```

```
for i = 1:length(idx_genes_chr2)

for j = 1:2001

NELF_F_chr2(i,j) = NELF_chr2([s_positions(idx_genes_chr2(i),j)]);

end

end


for i = 1:length(idx_genes_chr3)

for j = 1:2001

NELF_F_chr3(i,j) = NELF_chr3([s_positions(idx_genes_chr3(i),j)]);

end

end


for i = 1:length(idx_genes_chr4)

for j = 1:2001

NELF_F_chr4(i,j) = NELF_chr4([s_positions(idx_genes_chr4(i),j)]);

end

end


for i = 1:length(idx_genes_chr5)

for j = 1:2001

NELF_F_chr5(i,j) = NELF_chr5([s_positions(idx_genes_chr5(i),j)]);

end

end


for i = 1:length(idx_genes_chr6)
```

```
for j = 1:2001
NELF_F_chr6(i,j) = NELF_chr6([s_positions(idx_genes_chr6(i),j)]);
end
end

for i = 1:length(idx_genes_chr7)
for j = 1:2001
NELF_F_chr7(i,j) = NELF_chr7([s_positions(idx_genes_chr7(i),j)]);
end
end

for i = 1:length(idx_genes_chr8)
for j = 1:2001
NELF_F_chr8(i,j) = NELF_chr8([s_positions(idx_genes_chr8(i),j)]);
end
end

for i = 1:length(idx_genes_chr9)
for j = 1:2001
NELF_F_chr9(i,j) = NELF_chr9([s_positions(idx_genes_chr9(i),j)]);
end
end

for i = 1:length(idx_genes_chr10)
for j = 1:2001
```

```
NELF_F_chr10(i,j) = NELF_chr10([s_positions(idx_genes_chr10(i),j)]);
end
end

for i = 1:length(idx_genes_chr11)
for j = 1:2001
NELF_F_chr11(i,j) = NELF_chr11([s_positions(idx_genes_chr11(i),j)]);
end
end

for i = 1:length(idx_genes_chr12)
for j = 1:2001
NELF_F_chr12(i,j) = NELF_chr12([s_positions(idx_genes_chr12(i),j)]);
end
end

for i = 1:length(idx_genes_chr13)
for j = 1:2001
NELF_F_chr13(i,j) = NELF_chr13([s_positions(idx_genes_chr13(i),j)]);
end
end

for i = 1:length(idx_genes_chr14)
for j = 1:2001
NELF_F_chr14(i,j) = NELF_chr14([s_positions(idx_genes_chr14(i),j)]);
```

```
end
end

for i = 1:length(idx_genes_chr15)
for j = 1:2001
NELF_F_chr15(i,j) = NELF_chr15([s_positions(idx_genes_chr15(i),j)]);
end
end

for i = 1:length(idx_genes_chr16)
for j = 1:2001
NELF_F_chr16(i,j) = NELF_chr16([s_positions(idx_genes_chr16(i),j)]);
end
end

for i = 1:length(idx_genes_chr17)
for j = 1:2001
NELF_F_chr17(i,j) = NELF_chr17([s_positions(idx_genes_chr17(i),j)]);
end
end

for i = 1:length(idx_genes_chr19)
for j = 1:2001
NELF_F_chr19(i,j) = NELF_chr19([s_positions(idx_genes_chr19(i),j)]);
end
```

```
end

for i = 1:length(idx_genes_chr20)
for j = 1:2001
NELF_F_chr20(i,j) = NELF_chr20([s_positions(idx_genes_chr20(i),j)]);
end
end

for i = 1:length(idx_genes_chr21)
for j = 1:2001
NELF_F_chr21(i,j) = NELF_chr21([s_positions(idx_genes_chr21(i),j)]);
end
end

for i = 1:length(idx_genes_chr22)
for j = 1:2001
NELF_F_chr22(i,j) = NELF_chr22([s_positions(idx_genes_chr22(i),j)]);
end
end

for i = 1:length(idx_genes_chr23)
for j = 1:2001
NELF_F_chr23(i,j) = NELF_chr23([s_positions(idx_genes_chr23(i),j)]);
end
end
```

```
for i = 1:length(idx_genes_chr18)

for j = 1:2001

NELF_F_chr18(i,j) = NELF_chr18([s_positions(idx_genes_chr18(i),j)]);

end

end


for i = 1:length(idx_genes_chr24)

for j = 1:2001

NELF_F_chr24(idx_genes_chr24(i),j) = NELF_chr24([s_positions(idx_genes_chr24(i),j)]);

end

end


% Combine reads from the chromosomes

NELF_s_TSSs_all_chrs = [NELF_F_chr1;

NELF_F_chr2; NELF_F_chr3;

NELF_F_chr4; NELF_F_chr5;

NELF_F_chr6; NELF_F_chr7;

NELF_F_chr8; NELF_F_chr9;

NELF_F_chr10; NELF_F_chr11;

NELF_F_chr12; NELF_F_chr13;

NELF_F_chr14; NELF_F_chr15;

NELF_F_chr16; NELF_F_chr17;

NELF_F_chr18; NELF_F_chr19;

NELF_F_chr20; NELF_F_chr21;

NELF_F_chr22; NELF_F_chr23; NELF_F_chr24];
```

```
% Sort metagene in the right strand orientation:

NELF_s_TSS_all_genes_new_annot_hg38 = zeros(58344, 2001);


for i = 1:58344

if (strand_vector(i) == 0)

NELF_s_TSS_all_genes_new_annot_hg38(i,:) = fliplr(NELF_s_TSSs_all_chrs(i,:));

else

NELF_s_TSS_all_genes_new_annot_hg38(i,:) = NELF_s_TSSs_all_chrs(i,:);

end

end


% Remove top and bottom 0.5

nelf = NELF_s_TSS_all_genes_new_annot_hg38;

nelf_removed_05 = zeros(57761,2001);

for j = 1:2001

sorted_nelf(:,j) = sortrows(nelf(:,1), 'descend');

nelf_removed_05(:,i) = sorted_nelf(292:58052,j);

end


% Normalise NELF ChIP-Seq:

normalised_nelf = zeros(57761,2001);

for i = 1:57761

sum_nelf = sum(nelf_removed_05(i, 1:2001));

normalised_nelf(i,:) = nelf_removed_05(i,1:2001)./sum_nelf;

end
```

```
% Write output as csv files:

csvwrite("NELF_metagene.csv", NELF_s_TSS_all_genes_new_annot_hg38);

csvwrite("NELF_metagene_normalised.csv", nelf_normalised_removed_05);
```

9.2 TSS Annotation Pipeline (Linux)

```
# HEK TSS Annotation Bioinformatics Commands:

    conda install fastqc

    fastqc HEK-FII_S1_L001_R1_001.fastq.gz

    fastqc HEK-FII_S1_L002_R1_001.fastq.gz

    fastqc HEK-FII_S1_L003_R1_001.fastq.gz

    fastqc HEK-FII_S1_L004_R1_001.fastq.gz


    # Run trim-galore

    conda install trim-galore

    trim_galore -o trimmed_fastqs/ –fastqc HEK-FII_S1_L001_R1_001.fastq.gz

    trim_galore -o trimmed_fastqs/ –fastqc HEK-FII_S1_L002_R1_001.fastq.gz

    trim_galore -o trimmed_fastqs/ –fastqc HEK-FII_S1_L003_R1_001.fastq.gz

    trim_galore -o trimmed_fastqs/ –fastqc HEK-FII_S1_L004_R1_001.fastq.gz


    # Merge Lanes and gzip:

    zcat HEK-FII_S1_L001_R1_001_trimmed.fq.gz

    HEK-FII_S1_L002_R1_001_trimmed.fq.gz

    HEK-FII_S1_L003_R1_001_trimmed.fq.gz

    HEK-FII_S1_L004_R1_001_trimmed.fq.gz

    >HEK_FII_trimmed_merged.fq


    gzip HEK_FII_trimmed_merged.fq
```

```
# Perform Alignment with bowtie2.

bowtie2-build -f hg38NoAltNoHap.fa hg38_NoAlt_NoHap >genome_build_report.txt


# Align with bowtie2

bowtie2 -q -k 1 -p 1 -x hg38_NoAlt_NoHap -U HEK_FII_trimmed_merged.fq.gz -S HEK_FII.sam


# Filter with samtools using a mapping quality of at least 30.

samtools view -Sb -q 30 HEK_FII.sam > HEK_FII_filtered.bam


# Sort bam file:

samtools sort HEK_FII_filtered.bam HEK_FII_filtered_sorted.bam


# Convert to strand-specific bedgraph files:

bedtools genomecov -ibam HEK_FII_filtered_sorted.bam -bg -5 -strand +

> HEK_FII_5_F_filtered.bedgraph


bedtools genomecov -ibam HEK_FII_filtered_sorted.bam -bg -5 -strand -

> HEK_FII_5_R_filtered.bedgraph


# Now, download an ATAC-Seq file (HEK WT) from Weltner et al (2018).

trim_galore -q 20 —illumina —fastqc —trim-n ATAC_seq_HEK_weltner_1.fastq.gz

trim_galore -q 20 —illumina —fastqc —trim-n ATAC_seq_HEK_weltner_2.fastq.gz

trim_galore -q 20 —illumina —fastqc —trim-n ATAC_seq_HEK_weltner_3.fastq.gz

trim_galore -q 20 —illumina —fastqc —trim-n ATAC_seq_HEK_weltner_4.fastq.gz
```

```
zcat ATAC_seq_HEK_weltner_1.fastq_trimmed.gz
ATAC_seq_HEK_weltner_2_trimmed.fastq.gz
ATAC_seq_HEK_weltner_3_trimmed.fastq.gz
ATAC_seq_HEK_weltner_4_trimmed.fastq.gz >
ATAC_seq_HEK_merged.fq

gzip ATAC_seq_HEK_merged.fq

bowtie2 -q -k 1 -p 1 -x hg38_NoAlt_NoHap -U ATAC_seq_HEK_merged.fq.gz
-S ATAC_seq_HEK_weltner.sam

samtools view -Sb -q 30 ATAC_seq_HEK_weltner.sam >ATAC_seq_HEK_weltner.bam
samtools sort ATAC_seq_HEK_weltner.bam ATAC_seq_HEK_weltner_sorted
bedtools bamtobed -i ATAC_seq_HEK_weltner_sorted.bam
> ATAC_seq_HEK_weltner.bed

# Use macs2 for peak-calling:
macs2 callpeak -t ATAC_seq_HEK_weltner.bed -n ATAC_seq_HEK_weltner_peaks
—-nomodel —shift -100 —extsize 200 —broad

# Use bedtools intersect to retain entries that overlap with ATAC-seq peaks:
bedtools intersect -abam HEK_FII_filtered_sorted.bam
-b ATAC_seq_HEK_weltner_peaks_peaks.broadPeak
> HEK_FII_filtered_ATAC_overlap.bam
```

```
# Convert the ATAC-Seq filtered bam file into strand specific bedrgraph files:

bedtools genomecov -ibam HEK_FII_filtered_ATAC_overlap.bam -bg -5 -strand +

> HEK_FII_ATAC_5_F_filtered.bedgraph


bedtools genomecov -ibam HEK_FII_filtered_ATAC_overlap.bam -bg -5 -strand -

> HEK_FII_ATAC_5_R_filtered.bedgraph


#Run TSScall (Lavender et al., 2016):

python TSScall.py -—annotation_file gencode.v29.annotation.gtf

-—detail_file tss_annot_HEK_FII_filtered.details.txt

HEK_FII_5_F_filtered.bedgraph

HEK_FII_5_R_filtered.bedgraph hg38_chrom_sizes.txt

tss_annot_HEK_FII_filtered.bed


python TSScall.py -—annotation_file gencode.v29.annotation.gtf

-—detail_file tss_annot_HEK_FII_ATAC_filtered.details.txt

HEK_FII_ATAC_5_F_filtered.bedgraph

HEK_FII_ATAC_5_R_filtered.bedgraph hg38_chrom_sizes.txt

tss_annot_HEK_FII_ATAC_filtered.bed


# Following Takahashi et al (2012), use paraclu for TSS cluster identification:

# Modify bedgraph files so that the second column only contains strand information in R.


# Now run paraclu, setting the total cluster value threshold to 30:

paraclu 30 HEK_FII_5_F_filtered_paraclu.bedgraph
```

> HEK_FII_5_F_filtered_tsscluster.bedgraph

paraclu 30 HEK_FII_5_R_filtered_paraclu.bedgraph

> HEK_FII_5_R_filtered_tsscluster.bedgraph

paraclu-cut.sh HEK_FII_5_F_filtered_tsscluster.bedgraph

> HEK_FII_5_F_filtered_tsscluster_cut.bedgraph

paraclu-cut.sh HEK_FII_5_R_filtered_tsscluster.bedgraph

> HEK_FII_5_R_filtered_tsscluster_cut.bedgraph

paraclu 30 HEK_FII_5_ATAC_F_filtered_paraclu.bedgraph

> HEK_FII_5_ATAC_F_filtered_tsscluster.bedgraph

paraclu 30 HEK_FII_5_ATAC_R_filtered_paraclu.bedgraph

> HEK_FII_5_ATAC_R_filtered_tsscluster.bedgraph

paraclu-cut.sh HEK_FII_5_ATAC_F_filtered_tsscluster.bedgraph

> HEK_FII_5_ATAC_F_filtered_tsscluster_cut.bedgraph

paraclu-cut.sh HEK_FII_5_ATAC_R_filtered_tsscluster.bedgraph

> HEK_FII_5_ATAC_R_filtered_tsscluster_cut.bedgraph

Transform into a bedgraph-type file.

Remaining format: chr; start; end; sum_of_all_tags_in_cluster

Awk '$2=$5=$7=$8=""; print' HEK_FII_5_F_filtered_tsscluster_cut.bedgraph

> HEK_FII_F_filtered.tsscluster_readable.bedgraph

awk '$2=$5=$7=$8=""; print' HEK_FII_5_R_filtered_tsscluster_cut.bedgraph

> HEK_FII_R_filtered.tsscluster_readable.bedgraph

In MATLAB, use clusters to

- create metagenes of clusters around gencode TSSs to verify

- use TT-seq (10 min, replicates A and B combined) for metagenes around clustercentres

- retain TSSs where summed TT-Seq signal (TSS to TSS+1000) > 5 x (TSS-1000 to TSS).

#Add 'chr' at the start of each row/first column to prepare files for TSScall:

sed -i "" 's/^/chr/g' TSS_clusters_centres_TT_seq_verified_F.txt

> TSS_clusters_centres_TT_seq_verified_chr_F.txt

sed -i "" 's/^/chr/g' TSS_clusters_centres_TT_seq_verified_R.txt

> TSS_clusters_centres_TT_seq_verified_chr_R.txt

#Run TSScall to identify novel/annotated TSSs:

python TSScall.py -—annotation_file gencode.v29.chr_patch_hapl_scaff.annotation.gtf

–detail_file tss_annot_HEK_FII_filtered_clustercentres.details.txt

TSS_clusters_centres_TT_seq_verified_F.txt

TSS_clusters_centres_TT_seq_verified_R.txt hg38_chrom_sizes.txt

tss_annot_HEK_FII_filtered_cluster_centres.bed

Bibliography

K. Adelman and T. Henriques. Transcriptional speed bumps revealed in high resolution. 560(7720):560–561, 2018.

K. Adelman and J. T. Lis. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. 13(10):720–731, 2012.

S. H. Ahn, M. Kim, and S. Buratowski. Phosphorylation of Serine 2 within the RNA Polymerase II C-Terminal Domain Couples Transcription and 3 End Processing. *Mol. Cell*, 13(1):67–76, 2004.

G. Akbalik, K. Langebeck-Jensen, G. Tushev, S. Sambandan, J. Rinne, I. Epstein, I. Cajigas, I. Vlatkovic, and E. M. Schuman. Visualization of newly synthesized neuronal RNA in vitro and in vivo using click-chemistry. *RNA Biol.*, 14(1):20–28, 2017.

M. S. Akhtar, M. Heidemann, J. R. Tietjen, D. W. Zhang, R. D. Chapman, D. Eick, and A. Z. Ansari. TFIIH Kinase Places Bivalent Marks on the Carboxy-Terminal Domain of RNA Polymerase II. *Mol. Cell*, 34(3):387–393, 2009.

R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li,

L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jørgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller, A. R. Forrest, P. Carninci, M. Rehli, and A. Sandelin. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, 2014.

K.-J. Armache, S. Mitterweger, A. Meinhart, and P. Cramer. Structures of Complete RNA Polymerase II and Its Subcomplex, Rpb4/7. *J. Biol. Chem.*, 280:7131–7134, 2004.

C. Bachmann, H. Nguyen, J. Rosenbusch, L. Pham, T. Rabe, M. Patwa, G. Sokpor, R. H. Seong, R. Ashery-Padan, A. Mansouri, A. Stoykova, J. F. Staiger, and T. Tuoc. mSWI/SNF (BAF) Complexes Are Indispensable for the Neurogenesis and Development of Embryonic Olfactory Epithelium. *PLoS Genet.*, 12(9), 2016.

D. Baillat, M. A. Hakimi, A. M. Näär, A. Shilatifard, N. Cooch, and R. Shiekhattar. Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell*, 123(2):265–276, 2005.

A. Banerjee, M. C. Sammarco, S. Ditch, J. Wang, and E. Grabczyk. A Novel Tandem Reporter Quantifies RNA Polymerase II Termination in Mammalian Cells. *PLoS One*, 4(7):e6193, 2009.

C. A. Banks, Z. T. Lee, G. Boanca, M. Lakshminarasimhan, B. D. Groppe, Z. Wen, G. L. Hattem, C. W. Seidel, L. Florens, and M. P. Washburn. Controlling for gene

expression changes in transcription factor protein networks. *Mol. Cell. Proteomics*, 13(6):1510–1522, 2014.

T. Baptista, S. Grünberg, N. Minoungou, M. J. Koster, H. T. Timmers, S. Hahn, D. Devys, and L. Tora. SAGA Is a General Cofactor for RNA Polymerase II Transcription. *Mol. Cell*, 68(1):130–143.e5, oct 2017.

L. Baranello, D. Wojtowicz, K. Cui, B. N. Devaiah, H. J. Chung, K. Y. Chan-Salis, R. Guha, K. Wilson, X. Zhang, H. Zhang, J. Piotrowski, C. J. Thomas, D. S. Singer, B. F. Pugh, Y. Pommier, T. M. Przytycka, F. Kouzine, B. A. Lewis, K. Zhao, and D. Levens. RNA Polymerase II Regulates Topoisomerase 1 Activity to Favor Efficient Transcription. *Cell*, 165(2):357–371, 2016.

A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4):823–837, 2007.

K. Batta, Z. Zhang, K. Yen, D. B. Goffman, and B. Franklin Pugh. Genome-wide function of H2B ubiquitylation in promoter and genic regions. *Genes Dev.*, 25(21): 2254–2265, 2011.

J. G. Bauman, J. Wiegant, P. Borst, and P. van Duijn. A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA. *Exp. Cell Res.*, 128(2):485–490, 1980.

O. Bensaude. Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? *Transcription*, 2(3):103–108, 2011.

C. Bernecky, J. M. Plitzko, and P. Cramer. Structure of a transcribing RNA polymerase II-DSIF complex reveals a multidentate DNA-RNA clamp. *Nat. Struct. Mol. Biol.*, 24 (10):809–815, 2017.

E. Bertrand, P. Chartrand, M. Schaefer, S. M. Shenoy, R. H. Singer, and R. M. Long. Localization of ASH1 mRNA particles in living yeast. *Mol. Cell*, 2(4):437–445, 1998.

A. P. Bird. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.*, 8(7):1499–1504, 1980.

C. E. Birse, L. Minvielle-Sebastia, B. A. Lee, W. Keller, and N. J. Proudfoot. Coupling termination of transcription to messenger RNA maturation in yeast. *Science (80-.).*, 280(5361):298–301, 1998.

N. P. Blackledge, N. R. Rose, and R. J. Klose. Targeting Polycomb systems to regulate gene expression: modifications to a complex story. *Nat. Rev. Mol. Cell Biol.*, 16(11): 643–649, 2015.

A. K. Boehm, A. Saunders, J. Werner, and J. T. Lis. Transcription Factor and Polymerase Recruitment, Modification, and Movement on dhsp70 In Vivo in the Minutes following Heat Shock. *Mol. Cell. Biol.*, 23(21):7628–7637, 2003.

A. N. Boettiger and M. Levine. Synchronous and stochastic patterns of gene activation in the drosophila embryo. *Science (80-.).*, 325(5939):471–473, 2009.

O. Bogdanović and G. J. C. Veenstra. DNA methylation and methyl-CpG binding proteins: Developmental requirements and function. 118(5):549–565, 2009.

A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

V. A. Bondarenko, L. M. Steele, A. Újvári, D. A. Gaykalova, O. I. Kulaeva, Y. S. Polikanov, D. S. Luse, and V. M. Studitsky. Nucleosomes Can Form a Polar Barrier to Transcript Elongation by RNA Polymerase II. *Mol. Cell*, 24(3):469–479, 2006.

J. Bonnet, C. Y. Wang, T. Baptista, S. D. Vincent, W. C. Hsiao, M. Stierle, C. F. Kao, L. Tora, and D. Devys. The SAGA coactivator complex acts on the whole transcribed genome and is required for RNA polymerase II transcription. *Genes Dev.*, 28(18): 1999–2012, 2014.

V. Brès, S. M. Yoh, and K. A. Jones. The multi-tasking P-TEFb complex. 20(3):334–340, 2008.

T. Brown, F. S. Howe, S. C. Murray, M. Wouters, P. Lorenz, E. Seward, S. Rata, A. Angel, and J. Mellor. Antisense transcription-dependent chromatin signature modulates sense transcript dynamics. *Mol. Syst. Biol.*, 14(2):e8007, 2018.

H. Bunch, X. Zheng, A. Burkholder, S. T. Dillon, S. Motola, G. Birrane, C. C. Ebmeier, S. Levine, D. Fargo, G. Hu, D. J. Taatjes, and S. K. Calderwood. TRIM28 regulates RNA polymerase II promoter-proximal pausing and pause release. *Nat. Struct. Mol. Biol.*, 21(10):876–883, 2014.

H. Bunch, B. P. Lawney, Y. F. Lin, A. Asaithamby, A. Murshid, Y. E. Wang, B. P. Chen, and S. K. Calderwood. Transcriptional elongation requires DNA break-induced signalling. *Nat. Commun.*, 6, 2015.

S. Buratowski, S. Hahn, L. Guarente, and P. A. Sharp. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell*, 56(4):549–561, 1989.

P. Čabart, A. Újvári, M. Pal, and D. S. Luse. Transcription factor TFIIF is not required for initiation by RNA polymerase II, but it is essential to stabilize transcription factor TFIIB in early elongation complexes. *Proc. Natl. Acad. Sci. U. S. A.*, 108(38):15786–15791, 2011.

S. R. Carmody and S. R. Wente. mRNA nuclear export at a glance. *J. Cell Sci.*, 122 (12):1933–1937, 2009.

L. Cartegni, S. L. Chew, and A. R. Krainer. Listening to silence and understanding nonsense: Exonic mutations that affect splicing. 3(4):285–298, 2002.

E. Cavellán, P. Asp, P. Percipalle, and A. K. Ö. Farrants. The WSTF-SNF2h chromatin remodeling complex interacts with several nuclear proteins in transcription. *J. Biol. Chem.*, 281(24):16264–16271, 2006.

R. D. Chapman, M. Heidemann, C. Hintermair, and D. Eick. Molecular evolution of the RNA polymerase II CTD. 24(6):289–296, 2008.

F. X. Chen, P. Xie, C. K. Collings, K. Cao, Y. Aoi, S. A. Marshall, E. J. Rendleman, M. Ugarenko, P. A. Ozark, A. Zhang, R. Shiekhattar, E. R. Smith, M. Q. Zhang, and A. Shilatifard. PAF1 regulation of promoter-proximal pause release via enhancer activation. *Science (80-.).*, 357(6357):1294–1298, 2017.

F. X. Chen, E. R. Smith, and A. Shilatifard. Born to run: control of transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.*, 19(7):464–478, 2018.

B. Cheng and D. H. Price. Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. *J. Biol. Chem.*, 282(30):21901–21912, 2007.

S. Y. Chiu, F. Lejeune, A. C. Ranganathan, and L. E. Maquat. The pioneer translation initiation complex is functionally distinct from but structurally overlaps with the steady-state translation initiation complex. *Genes Dev.*, 18(7):745–754, 2004.

E. J. Cho, T. Takagi, C. R. Moore, and S. Buratowski. mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. *Genes Dev.*, 11(24):3319–3326, 1997.

H. Cho, T. K. Kim, H. Mancebo, W. S. Lane, O. Flores, and D. Reinberg. A protein phosphatase functions to recycle RNA polymerase II. *Genes Dev.*, 13(12):1540–52, 1999.

J. Choe, N. Oh, S. Park, Y. K. Lee, O. K. Song, N. Locker, S. G. Chi, and Y. K. Kim. Translation initiation on mRNAs bound by nuclear cap-binding protein complex CBP80/20 requires interaction between CBP80/20-dependent translation initiation factor and eukaryotic translation initiation factor 3g. *J. Biol. Chem.*, 287(22):18500–18509, 2012.

J. Choe, I. Ryu, O. H. Park, J. Park, H. Cho, J. S. Yoo, S. W. Chi, M. K. Kim, H. K. Song, and Y. K. Kim. EIF4AIII enhances translation of nuclear cap-binding complex-bound mRNAs by promoting disruption of secondary structures in UTR. *Proc. Natl. Acad. Sci. U. S. A.*, 111(43):E4577–E4586, 2014.

L. S. Churchman and J. S. Weissman. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature*, 469:368–375, 2011.

M. A. Cianfrocco, G. A. Kassavetis, P. Grob, J. Fang, T. Juven-Gershon, J. T. Kadonaga, and E. Nogales. Human TFIID binds to core promoter DNA in a reorganized structural state. *Cell*, 152(1-2):120–131, 2013.

C. R. Clapier and B. R. Cairns. The Biology of Chromatin Remodeling Complexes. *Annu. Rev. Biochem.*, 78(1):273–304, 2009.

M. D. Cleary, C. D. Meiering, E. Jan, R. Guymon, and J. C. Boothroyd. Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nat. Biotechnol.*, 23(2):232–237, 2005.

T. Clouaire, S. Webb, P. Skene, R. Illingworth, A. Kerr, R. Andrews, J. H. Lee, D. Skalnik, and A. Bird. Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes Dev.*, 26(15):1714–1728, 2012.

R. C. Conaway and J. Weliky Conaway. An RNA polymerase II transcription factor has an associated DNA-dependent ATPase (dATPase) activity strongly stimulated by the TATA region promoters. *Proc. Natl. Acad. Sci. U. S. A.*, 86(19):7356–7360, 1989.

S. Connelly and J. L. Manley. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev.*, 2(4):440–452, 1988.

L. Core and K. Adelman. Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. *Genes Dev.*, 33(15-16), 2019.

L. J. Core, J. J. Waterfall, and J. T. Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–8, 2008.

S. Corless and N. Gilbert. Effects of DNA supercoiling on chromatin architecture. 8 (Suppl 1):51–64, 2016.

M. A. Cortazar, R. M. Sheridan, B. Erickson, N. Fong, K. Glover-Cutter, K. Brannan, and D. L. Bentley. Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a "Sitting Duck Torpedo" Mechanism. *Mol. Cell*, 76(6), 2019.

V. H. Cowling. Regulation of mRNA cap methylation. 425(2):295–302, 2010.

B. Daneholt. Assembly and transport of a premessenger RNP particle. *Proc. Natl. Acad. Sci. U. S. A.*, 98(13):7012–7017, 2001.

D. S. Day, B. Zhang, S. M. Stevens, F. Ferrari, E. N. Larschan, P. J. Park, and W. T. Pu. Comprehensive analysis of promoter-proximal RNA polymerase II pausing across mammalian cell types. *Genome Biol.*, 17(1):120, 2016.

A. M. Deaton and A. Bird. CpG islands and the regulation of transcription. *Genes Dev.*, 25(10):1010–1022, 2011.

M. L. Dechassa, A. Sabri, S. Pondugula, S. R. Kassabov, N. Chatterjee, M. P. Kladde,

and B. Bartholomew. SWI/SNF Has Intrinsic Nucleosome Disassembly Activity that Is Dependent on Adjacent Nucleosomes. *Mol. Cell*, 38(4):590–602, 2010.

B. Dichtl, D. Blank, M. Sadowski, W. Hubner, S. Weiser, and W. Keller. Yhh1p/Cft1p directly links poly(A) site recognition and RNA polymerase II transcription termination. *EMBO*, 21(15):4125–4135, 2002.

S. K. Doamekpor, A. M. Sanchez, B. Schwer, S. Shuman, and C. D. Lima. How an mRNA capping enzyme reads distinct RNA polymerase II and Spt5 CTD phosphorylation codes. *Genes Dev.*, 28(12):1323–1336, 2014.

H. Domdey, B. Apostol, R. J. Lin, A. Newman, E. Brody, and J. Abelson. Lariat structures are in vivo intermediates in yeast pre-mRNA splicing. *Cell*, 39(3 PART 2): 611–621, 1984.

Z. Dominski, X.-c. Yang, M. Purdy, E. J. Wagner, and W. F. Marzluff. A CPSF-73 Homologue Is Required for Cell Cycle Progression but Not Cell Growth and Interacts with a Protein Having Features of CPSF-100. *Mol. Cell. Biol.*, 25(4):1489–1500, 2005.

S. J. Dubbury, P. L. Boutz, and P. A. Sharp. CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. *Nature*, 564(7734):141–145, 2018.

E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10:48, 2009.

S. Egloff, J. Zaborowska, C. Laitem, T. Kiss, and S. Murphy. Ser7 phosphorylation

of the CTD recruits the RPAP2 ser5 phosphatase to snRNA genes. *Mol. Cell*, 45: 111–122, 2012. ISSN 10972765. doi: 10.1016/j.molcel.2011.11.006.

N. D. Elrod, T. Henriques, K. L. Huang, D. C. Tatomer, J. E. Wilusz, E. J. Wagner, and K. Adelman. The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes. *Mol. Cell*, 76(5):738–752.e7, 2019.

M. A. Erb, T. G. Scott, B. E. Li, H. Xie, J. Paulk, H. S. Seo, A. Souza, J. M. Roberts, S. Dastjerdi, D. L. Buckley, N. E. Sanjana, O. Shalem, B. Nabet, R. Zeid, N. K. Offei-Addo, S. Dhe-Paganon, F. Zhang, S. H. Orkin, G. E. Winter, and J. E. Bradner. Transcription control by the ENL YEATS domain in acute leukaemia. *Nature*, 543 (7644):270–274, 2017.

C. Esnault, Y. Ghavi-Helm, S. Brun, J. Soutourina, N. Van Berkum, C. Boschiero, F. Holstege, and M. Werner. Mediator-Dependent Recruitment of TFIIH Modules in Preinitiation Complex. *Mol. Cell*, 31(3):337–346, 2008.

T. Eychenne, E. Novikova, M. B. Barrault, O. Alibert, C. Boschiero, N. Peixeiro, D. Cornu, V. Redeker, L. Kuras, P. Nicolas, M. Werner, and J. Soutourina. Functional interplay between Mediator and TFIIB in preinitiation complex assembly in relation to promoter architecture. *Genes Dev.*, 30(18):2119–2132, 2016.

C. B. Fant, C. B. Levandowski, K. Gupta, Z. L. Maas, J. Moir, J. D. Rubin, A. Sawyer, M. N. Esbin, J. K. Rimel, O. Luyties, M. T. Marr, I. Berger, R. D. Dowell, and D. J. Taatjes. TFIID Enables RNA Polymerase II Promoter-Proximal Pausing. *Mol. Cell*, 2020.

S. M. Fica, N. Tuttle, T. Novak, N. S. Li, J. Lu, P. Koodathingal, Q. Dai, J. P. Staley, and J. A. Piccirilli. RNA catalyses nuclear pre-mRNA splicing. *Nature*, 503(7475): 229–234, 2013.

H. Fischl, J. Neve, Z. Wang, R. Patel, A. Louey, B. Tian, and A. Furger. hnRNPC regulates cancer-specific alternative cleavage and polyadenylation profiles. *Nucleic Acids Res.*, 47(14):7580–7591, 2019.

J. Fishburn and S. Hahn. Architecture of the Yeast RNA Polymerase II Open Complex and Regulation of Activity by TFIIF. *Mol. Cell. Biol.*, 32(1):12–25, 2012.

O. Flores, H. Lu, and D. Reinberg. Factors involved in specific transcription by mammalian RNA polymerase II. Identification and characterization of factor IIH. *J. Biol. Chem.*, 267(4):2786–2793, 1992.

N. Fong and D. L. Bentley. Capping, splicing, and 3 processing are independently stimulated by RNA polymerase II: Different functions for different segments of the CTD. *Genes Dev.*, 15(14):1783–1795, 2001.

N. Fong, K. Brannan, B. Erickson, H. Kim, M. A. Cortazar, R. M. Sheridan, T. Nguyen, S. Karp, and D. L. Bentley. Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Mol. Cell*, 60(2):256–267, 2015.

P. Fortes, T. Inada, T. Preiss, M. W. Hentze, I. W. Mattaj, and A. B. Sachs. The yeast nuclear cap binding complex can interact with translation factor eIF4G and mediate translation initiation. *Mol. Cell*, 6(1):191–196, 2000.

P. S. Francis Tsai. Structural basis of preinitiation complex assembly on human Pol II promoters. *EMBO*, 19(1):25–36, 2000.

L. D. Fresco and S. Buratowski. Conditional mutants of the yeast mRNA capping enzyme show that the cap enhances, but is not required for, mRNA splicing. *RNA*, 2(6):584–596, 1996.

G. Fuchs, D. Hollander, Y. Voichek, G. Ast, and M. Oren. Cotranscriptional histone H2B monoubiquitylation is tightly coupled with RNA polymerase II elongation rate. *Genome Res.*, 24(10):1572–1583, 2014.

T. Furukawa and N. Tanese. Assembly of partial TFIID complexes in mammalian cells reveals distinct activities associated with individual TATA box-binding protein-associated factors. *J. Biol. Chem.*, 275(38):29847–29856, 2000.

B. Gaertner and J. Zeitlinger. RNA polymerase II pausing during development. *Development*, 141(6):1179–83, 2014.

W. P. Galej, M. E. Wilkinson, S. M. Fica, C. Oubridge, A. J. Newman, and K. Nagai. Cryo-EM structure of the spliceosome immediately after branching. *Nature*, 537 (7619):197–201, 2016.

J. G. Gall and M. L. Pardue. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proc. Natl. Acad. Sci. U. S. A.*, 63(2):378–383, 1969.

C. Ganem, F. Devaux, C. Torchet, C. Jacq, S. Quevillon-Cheruel, G. Labesse, C. Facca, and G. Faye. Ssu72 is a phosphatase essential for transcription termination of snoRNAs and specific mRNAs in yeast. *EMBO*, 22(7):1588–1598, 2003.

C. B. Gerhold and S. M. Gasser. INO80 and SWR complexes: Relating structure to function in chromatin remodeling. 24(11):619–631, 2014.

A. Ghosh, S. Shuman, and C. D. Lima. Structural Insights to How Mammalian Capping Enzyme Reads the CTD Code. *Mol. Cell*, 43(2):299–310, 2011.

D. S. Gilmour and J. T. Lis. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. *Mol. Cell. Biol.*, 6 (11):3984–3989, 1986.

L. H. Gregersen, R. Mitter, A. P. Ugalde, T. Nojima, N. J. Proudfoot, R. Agami, A. Stewart, and J. Q. Svejstrup. SCAF4 and SCAF8, mRNA Anti-Terminator Proteins. *Cell*, 177(7):1797–1813.e18, 2019.

L. H. Gregersen, R. Mitter, and J. Q. Svejstrup. Using TTchem-seq for profiling nascent transcription and measuring transcript elongation. *Nat. Protoc.*, 15:604–672, 2020.

D. Grohmann, J. Nagy, A. Chakraborty, D. Klose, D. Fielden, R. H. Ebright, J. Michaelis, and F. Werner. The Initiation Factor TFE and the Elongation Factor Spt4/5 Compete for the RNAP Clamp during Transcription Initiation and Elongation. *Mol. Cell*, 43(2):263–274, 2011.

N. Gromak, S. West, and N. J. Proudfoot. Pause Sites Promote Transcriptional Termination of Mammalian RNA Polymerase II. *Mol. Cell. Biol.*, 26(10):3986–3996, 2006.

M. G. Guenther, S. S. Levine, L. A. Boyer, R. Jaenisch, and R. A. Young. A Chromatin

Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell*, 130: 77–88, 2007.

B. W. Guidi, G. Bjornsdottir, D. C. Hopkins, L. Lacomis, H. Erdjument-Bromage, P. Tempst, and L. C. Myers. Mutual targeting of Mediator and the TFIIH kinase Kin28. *J. Biol. Chem.*, 279(28):29114–29120, 2004.

S. Guo, Y. Yamaguchi, S. Schilbach, T. Wada, J. Lee, A. Goddard, D. French, H. Handa, and A. Rosenthal. A regulator of transcriptional elongation controls vertebrate neuronal development. *Nature*, 408(6810):366–369, 2000.

J. Hang, R. Wan, C. Yan, and Y. Shi. Structural basis of pre-mRNA splicing. *Science (80-.).*, 349(6253):1191–1198, 2015.

D. C. Hargreaves and G. R. Crabtree. ATP-dependent chromatin remodeling: Genetics, genomics and mechanisms. *Cell Res.*, 21(3):396–420, 2011.

K. M. Harlen and L. S. Churchman. The code and beyond: Transcription regulation by the RNA polymerase II carboxy-terminal domain. 18(4):263–273, 2017.

K. M. Harlen, K. L. Trotta, E. E. Smith, M. M. Mosaheb, S. M. Fuchs, and L. S. Churchman. Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. *Cell Rep.*, 15(10):2147–2158, 2016.

N. He, M. Liu, J. Hsu, Y. Xue, S. Chou, A. Burlingame, N. J. Krogan, T. Alber, and Q. Zhou. HIV-1 Tat and Host AFF4 Recruit Two Transcription Elongation Factors into a Bifunctional Complex for Coordinated Activation of HIV-1 Transcription. *Mol. Cell*, 38(3):428–438, 2010.

G. J. Hendriks, L. A. Jung, A. J. Larsson, M. Lidschreiber, O. Andersson Forsman, K. Lidschreiber, P. Cramer, and R. Sandberg. NASC-seq monitors RNA synthesis in single cells. *Nat. Commun.*, 10(1):1–9, 2019.

D. A. Hendrix, J. W. Hong, J. Zeitlinger, D. S. Rokhsar, and M. S. Levine. Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. *Proc. Natl. Acad. Sci. U. S. A.*, 105(22):7762–7767, 2008.

C. J. Hengartner, V. E. Myer, S. M. Liao, C. J. Wilson, S. S. Koh, and R. A. Young. Temporal regulation of RNA polymerase II by Srb10 and Kin28 cyclin-dependent kinases. *Mol. Cell*, 2(1):43–53, 1998.

T. Henriques, D. A. Gilchrist, S. Nechaev, M. Bern, G. W. Muse, A. Burkholder, D. C. Fargo, and K. Adelman. Stable pausing by rna polymerase II provides an opportunity to target and integrate regulatory signals. *Mol. Cell*, 52(4):517–528, 2013.

L. Herzel, D. S. Ottoz, T. Alpert, and K. M. Neugebauer. Splicing and transcription touch base: Co-transcriptional spliceosome assembly and function. 18(10):637–650, 2017.

A. R. Hieb, W. A. Halsey, M. D. Betterton, T. T. Perkins, J. F. Kugel, and J. A. Goodrich. TFIIA Changes the Conformation of the DNA in TBP/TATA Complexes and Increases their Kinetic Stability. *J. Mol. Biol.*, 372(3):619–632, 2007.

F. Hirose, N. Ohshima, E.-J. Kwon, H. Yoshida, and M. Yamaguchi. Drosophila Mi-2 Negatively Regulates dDREF by Inhibiting Its DNA-Binding Activity. *Mol. Cell. Biol.*, 22(14):5182–5193, 2002.

F. C. Holstege, P. C. van der Vliet, and H. T. Timmers. Opening of an RNA polymerase II promoter occurs in two distinct steps and requires the basal transcription factors IIE and IIH. *EMBO J.*, 1996.

M. Hondele, T. Stuwe, M. Hassler, F. Halbach, A. Bowman, E. T. Zhang, B. Nijmeijer, C. Kotthoff, V. Rybin, S. Amlacher, E. Hurt, and A. G. Ladurner. Structural basis of histone H2A-H2B recognition by the essential chaperone FACT. *Nature*, 499 (7456):111–114, 2013.

X. Hong, D. G. Scofield, and M. Lynch. Intron Size, Abundance, and Distribution within Untranslated Regions of Genes. *Molecual Biol. Evol.*, 23(12):2392–2404, 2006.

L. Hou, Y. Wang, Y. Liu, N. Zhang, I. Shamovsky, E. Nudler, B. Tian, and B. D. Dynlacht. Paf1C regulates RNA polymerase II progression by modulating elongation rate. *Proc. Natl. Acad. Sci. U. S. A.*, 116(29):14583–14592, 2019.

F. S. Howe, H. Fischl, S. C. Murray, and J. Mellor. Is H3K4me3 instructive for transcription activation? *BioEssays*, 39(1):e201600095, 2017.

J. P. Hsin and J. L. Manley. The RNA polymerase II CTD coordinates transcription and RNA processing. 26(19):2119–2137, 2012.

J. P. Hsin, A. Sheth, and J. L. Manley. RNAP II CTD phosphorylated on threonine-4 is required for histone mRNA 3 end processing. *Science (80-.).*, 334(6056):683–686, 2011.

C. Iasillo, M. Schmid, Y. Yahia, M. A. Maqbool, N. Descostes, E. Karadoulama,

E. Bertrand, J.-C. Andrau, and T. Heick Jensen. ARS2 is a general suppressor of pervasive transcription. *Nucleic Acids Res.*, 45:10229–10241, 2017.

K. Inoue, M. Ohno, H. Sakamoto, and Y. Shimura. Effect of the cap structure on pre-mRNA splicing in Xenopus oocyte nuclei. *Genes Dev.*, 3:1472–1479, 1989.

M. Ito, C. X. Yuan, H. J. Okano, R. B. Darnell, and R. G. Roeder. Involvement of the TRAP220 component of the TRAP/SMCC coactivator complex in embryonic development and thyroid hormone action. *Mol. Cell*, 5(4):683–93, 2000.

M. Ito, H. J. Okano, R. B. Darnell, and R. G. Roeder. The TRAP100 component of the TRAP/Mediator complex is essential in broad transcriptional events and development. *EMBO J.*, 21(13):3464–75, 2002.

T. Ito, M. Bulger, M. J. Pazin, R. Kobayashi, and J. T. Kadonaga. ACF, an ISWI-containing and ATP-utilizing chromatin assembly and remodeling factor. *Cell*, 90 (1):145–155, 1997.

E. Izaurralde and S. Adam. Transport of macromolecules between the nucleus and the cytoplasm. *RNA*, 4:351–364, 1998.

M. G. Izban and D. S. Luse. The RNA polymerase II ternary complex cleaves the nascent transcript in a 3' to 5' direction in the presence of elongation factor SII. *Genes Dev.*, 6:1342–1356, 1992.

K. James, P. Gamba, S. J. Cockell, and N. Zenkin. Misincorporation by RNA polymerase is a major source of transcription pausing in vivo. *Nucleic Acids Res.*, 45 (3):1105–1113, 2016.

R. Javahery, A. Khachi, K. Lo, B. Zenzie-Gregory, and S. T. Smale. DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol. Cell. Biol.*, 14(1):116–127, 1994.

J. Jean-Philippe, S. Paz, and M. Caputi. hnRNP A1: The Swiss Army Knife of Gene Expression. *Int. J. Mol. Sci.*, 14(9):18999–19024, 2013.

M. Jishage, X. Yu, Y. Shi, S. J. Ganesan, W. Y. Chen, A. Sali, B. T. Chait, F. J. Asturias, and R. G. Roeder. Architecture of Pol II(G) and molecular mechanism of transcription regulation by Gdown1. *Nat. Struct. Mol. Biol.*, 25(9):859–867, sep 2018.

J. Judd, L. A. Wojenski, L. M. Wainman, N. D. Tippens, A. Dziubek, G. J. Villafano, E. M. Wissink, P. Versluis, L. Bagepalli, S. R. Shah, D. B. Mahat, J. M. Tome, C. G. Danko, J. T. Lis, and L. J. Core. A rapid, sensitive, scalable method for Precision Run-On sequencing (PRO-seq) . *bioRxiv*, 5, 2020.

T. Kafri, M. Ariel, M. Brandeis, R. Shemer, L. Urven, J. McCarrey, H. Cedar, and A. Razin. Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. *Genes Dev.*, 6(5):705–714, 1992.

M. H. Kagey, J. J. Newman, S. Bilodeau, Y. Zhan, D. A. Orlando, N. L. Van Berkum, C. C. Ebmeier, J. Goossens, P. B. Rahl, S. S. Levine, D. J. Taatjes, J. Dekker, and R. A. Young. Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435, 2010.

K. Kamieniarz-Gdula and N. J. Proudfoot. Transcriptional Control by Premature Termination: A Forgotten Mechanism. 35(8):553–564, 2019.

K. Kamieniarz-Gdula, M. R. Gdula, K. Panser, T. Nojima, J. Monks, J. R. Wiśniewski, J. Riepsaame, N. Brockdorff, A. Pauli, and N. J. Proudfoot. Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination. *Mol. Cell*, 74(1):158–172.e9, 2019.

T. Kanno, Y. Kanno, G. Leroy, E. Campos, H. W. Sun, S. R. Brooks, G. Vahedi, T. D. Heightman, B. A. Garcia, D. Reinberg, U. Siebenlist, J. J. O'Shea, and K. Ozato. BRD4 assists elongation of both coding and enhancer RNAs by interacting with acetylated histones. *Nat. Struct. Mol. Biol.*, 21(12):1047–1057, 2014.

K. Karmodiya, A. R. Krebs, M. Oulad-Abdelghani, H. Kimura, and L. Tora. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics*, 13(1):424, 2012.

K. Kas, M. L. Voz, K. Hensen, E. Meyen, and W. J. Van De Ven. Transcriptional activation capacity of the novel PLAG family of zinc finger proteins. *J. Biol. Chem.*, 273(36):23026–23032, 1998.

B. Kastner, C. L. Will, H. Stark, and R. Lührmann. Structural insights into nuclear pre-mRNA splicing in higher eukaryotes. *Cold Spring Harb. Perspect. Biol.*, 11(11): a032417, 2019.

S. Katayama, Y. Tomaru, T. Kasukawa, K. Waki, M. Nakanishi, M. Nakamura, H. Nishida, C. C. Yap, M. Suzuki, J. Kawai, H. Suzuki, P. Carninci, Y. Hayashizaki, C. Wells, M. Frith, T. Ravasi, K. C. Pang, J. Hallinan, J. Mattick, D. A. Hume, L. Lipovich, S. Batalov, P. G. Engstrom, Y. Mizuno, M. A. Faghihi, A. Sandelin, A. M.

Chalk, S. Mottagui-Tabar, Z. Liang, B. Lenhard, and C. Wahlestedt. Antisense transcription in the mammalian transcriptome. *Science (80-.).*, 309(5740):1564–1566, 2005.

I. Kaufmann, G. Martin, A. Friedlein, H. Langen, and W. Keller. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J.*, 23(3):616–626, 2004.

W. Keller, S. Bienroth, K. Lang, and G. Christofori. Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3 processing signal AAUAAA. *EMBO J.*, 10(13):4241–4249, 1991.

D. J. Kemble, L. L. McCullough, F. G. Whitby, T. Formosa, and C. P. Hill. FACT Disrupts Nucleosome Structure by Binding H2A-H2B with Conserved Peptide Motifs. *Mol. Cell*, 60(2):294–306, 2015.

D. Kim, B. Langmead, and S. L. Salzberg. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods*, 12(4):357–360, 2015.

M. Kim, H. Suh, E. J. Cho, and S. Buratowski. Phosphorylation of the yeast Rpb1 C-terminal domain at serines 2,5, and 7. *J. Biol. Chem.*, 284(39):26421–26426, 2009.

I. F. King, C. N. Yandava, A. M. Mabb, J. S. Hsiao, H. S. Huang, B. L. Pearson, J. M. Calabrese, J. Starmer, J. S. Parker, T. Magnuson, S. J. Chamberlain, B. D. Philpot, and M. J. Zylka. Topoisomerases facilitate transcription of long genes linked to autism. *Nature*, 501(7465):58–62, 2013.

M. L. Kireeva, W. Walter, V. Tchernajenko, V. Bondarenko, M. Kashlev, and V. M. Studitsky. Nucleosome remodeling induced by RNA polymerase II: Loss of the H2A/H2B dimer during transcription. *Mol. Cell*, 9(3):541–552, 2002.

B. J. Klein, D. Bose, K. J. Baker, Z. M. Yusoff, X. Zhang, and K. S. Murakami. RNA polymerase and transcription elongation factor Spt4/5 complex structure. *Proc. Natl. Acad. Sci. U. S. A.*, 108(2):546–550, 2011.

S. L. Klemm, Z. Shipony, and W. J. Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.*, 20(4):207–220, 2019.

R. J. Klose and A. P. Bird. Genomic DNA methylation: The mark and its mediators. 31(2):89–97, 2006.

R. Kodzius, M. Kojima, H. Nishiyori, M. Nakamura, S. Fukuda, M. Tagami, D. Sasaki, K. Imamura, C. Kai, M. Harbers, Y. Hayashizaki, and P. Carninci. CAGE: cap analysis of gene expression. *Nat. Methods*, 3(3):211–222, 2006.

P. Komarnitsky, E. J. Cho, and S. Buratowski. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev.*, 14(19):2452–2460, 2000.

M. M. Konarska, R. A. Padgett, and P. A. Sharp. Recognition of cap structure in splicing in vitro of mRNA precursors. *Cell*, 38(3):731–736, 1984.

J. N. Kuehner, E. L. Pearson, and C. Moore. Unravelling the means to an end: RNA polymerase II transcription termination. 12(5):283–294, 2011.

A. K. Kutach and J. T. Kadonaga. The Downstream Promoter Element DPE Appears To Be as Widely Used as the TATA Box in Drosophila Core Promoters. *Mol. Cell. Biol.*, 20(13):4754–4764, 2000.

H. Kwak, N. J. Fuda, L. J. Core, and J. T. Lis. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science (80-.).*, 339(6122):950–953, 2013.

A. T. Kwon, D. J. Arenillas, R. W. Hunt, and W. W. Wasserman. Opossum-3: Advanced analysis of regulatory motif over-representation across genes or chip-seq datasets. *G3 Genes, Genomes, Genet.*, 2(9):987–1002, 2012.

B. Laggerbauer, T. Achsel, and R. Lührmann. The human U5-200kD DEXH-box protein unwinds U4/U6 RNA duplices in vitro. *Proc. Natl. Acad. Sci. U. S. A.*, 95(8): 4188–4192, 1998.

M. Lagha, J. P. Bothma, E. Esposito, S. Ng, L. Stefanik, C. Tsui, J. Johnston, K. Chen, D. S. Gilmour, J. Zeitlinger, and M. S. Levine. XPaused Pol II coordinates tissue morphogenesis in the drosophila embryo. *Cell*, 153(5):976, 2013.

C. Laitem, J. Zaborowska, N. F. Isa, J. Kufs, M. Dienstbier, and S. Murphy. CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II–transcribed genes. *Nat. Struct. Mol. Biol.*, 22(5):396–403, 2015.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, 2012.

S. M. Lauberth, T. Nakayama, X. Wu, A. L. Ferris, Z. Tang, S. H. Hughes, and R. G.

Roeder. H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell*, 152(5):1021–1036, 2013.

C. A. Lavender, K. R. Cannady, J. A. Hoffman, K. W. Trotter, D. A. Gilchrist, B. D. Bennett, A. B. Burkholder, C. J. Burd, D. C. Fargo, and T. K. Archer. Downstream Antisense Transcription Predicts Genomic Features That Define the Specific Chromatin Environment at Mammalian Promoters. *PLOS Genet.*, 12(8):e1006224, 2016.

X. Le Guezennec, M. Vermeulen, A. B. Brinkman, W. A. M. Hoeijmakers, A. Cohen, E. Lasonder, and H. G. Stunnenberg. MBD2/NuRD and MBD3/NuRD, Two Distinct Complexes with Different Biochemical and Functional Properties. *Mol. Cell. Biol.*, 26(3):843–851, 2006.

C. Lee, X. Li, A. Hechmer, M. Eisen, M. D. Biggin, B. J. Venters, C. Jiang, J. Li, B. F. Pugh, and D. S. Gilmour. NELF and GAGA Factor Are Linked to Promoter-Proximal Pausing at Many Genes in Drosophila. *Mol. Cell. Biol.*, 28(10):3290–3300, 2008.

J. S. Lee, E. Smith, and A. Shilatifard. The Language of Histone Crosstalk. 142(5): 682–685, 2010.

J. S. Lee, A. S. Garrett, K. Yen, Y. H. Takahashi, D. Hu, J. Jackson, C. Seidel, B. Franklin Pugh, and A. Shilatifard. Codependency of H2B monoubiquitination and nucleosome reassembly on Chdl. *Genes Dev.*, 26(9):914–919, 2012.

W. Lee, D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes, and C. Nislow. A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, 39(10): 1235–1244, 2007.

Y. Lee and D. C. Rio. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annu. Rev. Biochem.*, 84(1):291–323, 2015.

M. Leptin. Gastrulation movements: The logic and the nuts and bolts. 8(3):305–320, 2005.

M. J. Levesque and A. Raj. Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. *Nat. Methods*, 10(3):246–248, 2013.

M. Levine. Paused RNA polymerase II as a developmental checkpoint. 145(4):502–511, 2011.

H. Li, S. Ilin, W. Wang, E. M. Duncan, J. Wysocka, C. D. Allis, and D. J. Patel. Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature*, 442(7098):91–95, 2006.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

J. Li and D. S. Gilmour. Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor. *EMBO J.*, 32(13):1829–1841, 2013.

Y. Liao, G. K. Smyth, and W. Shi. featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *biorxiv*, 4, 2013.

B. Lim. Imaging transcriptional dynamics. 52:49–55, 2018.

C. Y. Lim, B. Santoso, T. Boulay, E. Dong, U. Ohler, and J. T. Kadonaga. The MTE, a new core promoter element for transcription by RNA poymerase II. *Genes Dev.*, 18 (13):1606–1617, 2004.

C. Lin, E. R. Smith, H. Takahashi, K. C. Lai, S. Martin-Brown, L. Florens, M. P. Washburn, J. W. Conaway, R. C. Conaway, and A. Shilatifard. AFF4, a Component of the ELL/P-TEFb Elongation Complex and a Shared Subunit of MLL Chimeras, Can Link Transcription Elongation to Leukemia. *Mol. Cell*, 37(3):429–437, 2010.

P. S. Lin, M. F. Dubois, and M. E. Dahmus. TFIIF-associating carboxyl-terminal domain phosphatase dephosphorylates phosphoserines 2 and 5 of RNA polymerase II. *J. Biol. Chem.*, 277(48):45949–45956, 2002.

P. Liu, Y. Xiang, K. Fujinaga, K. Bartholomeeusen, K. A. Nilson, D. H. Price, and B. M. Peterlin. Release of positive transcription elongation factor b (P-TEFb) from 7SK small nuclear ribonucleoprotein (snRNP) activates hexamethylene bisacetamide-inducible protein (HEXIM1) transcription. *J. Biol. Chem.*, 289(14):9918–9925, 2014.

X. Liu, C. K. Lee, J. A. Granek, N. D. Clarke, and J. D. Lieb. Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome Res.*, 16(12):1517–1528, 2006.

Y. Liu, K. Zhou, N. Zhang, H. Wei, Y. Z. Tan, Z. Zhang, B. Carragher, C. S. Potter, S. D'Arcy, and K. Luger. FACT caught in the act of manipulating the nucleosome. *Nature*, 577(7790):426–431, 2020.

M. Llorian, S. Schwartz, T. A. Clark, D. Hollander, L. Y. Tan, R. Spellman, A. Gordon, A. C. Schweitzer, P. De La Grange, G. Ast, and C. W. Smith. Position-dependent

alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nat. Struct. Mol. Biol.*, 17(9):1114–1123, 2010.

S. Lloyd. Least Squares Quantization in PCM. *Technical Report*, 1957.

J. Logan, E. Falck-Pedersen, J. E. Darnell, and T. Shenk. A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. *Proc. Natl. Acad. Sci. U. S. A.*, 84(23):8306–8310, 1987.

J. Lovén, H. A. Hoke, C. Y. Lin, A. Lau, D. A. Orlando, C. R. Vakoc, J. E. Bradner, T. I. Lee, and R. A. Young. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, 153(2):320–334, 2013.

R. F. Luco, Q. Pan, K. Tominaga, B. J. Blencowe, O. M. Pereira-Smith, and T. Misteli. Regulation of alternative splicing by histone modifications. *Science (80-.).*, 327 (5968):996–1000, 2010.

Z. Luo, C. Lin, E. Guest, A. S. Garrett, N. Mohaghegh, S. Swanson, S. Marshall, L. Florens, M. P. Washburn, and A. Shilatifard. The super elongation complex family of RNA polymerase II elongation factors: gene target specificity and transcriptional output. *Mol. Cell. Biol.*, 32(13):2608–17, 2012a.

Z. Luo, C. Lin, and A. Shilatifard. The super elongation complex (SEC) family in transcriptional control. 13(9):543–547, 2012b.

D. S. Luse. The RNA polymerase II preinitiation complex: Through what pathway is the complex assembled? 5(NOV), 2013.

M. D. Lynch, A. J. H. Smith, M. De Gobbi, M. Flenley, J. R. Hughes, D. Vernimmen, H. Ayyub, J. A. Sharpe, J. A. Sloane-Stanley, L. Sutherland, S. Meek, T. Burdon, R. J. Gibbons, D. Garrick, and D. R. Higgs. An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. *EMBO J.*, 31(2):317–329, 2012.

S. Malik and R. G. Roeder. The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. 11(11):761–772, 2010.

L. E. Maquat, J. Hwang, H. Sato, and Y. Tang. CBP80-promoted mRNP rearrangements during the pioneer round of translation, nonsense-mediated mRNA decay, and thereafter. *Cold Spring Harb. Symp. Quant. Biol.*, 75:127–134, 2010.

N. Marbach-Bar, A. Ben-Noon, S. Ashkenazi, A. T. B. Harush, T. Avnit-Sagi, M. D. Walker, and R. Dikstein. Disparity between microRNA levels and promoter strength is associated with initiation rate and Pol II pausing. *Nat. Commun.*, 4(1):1–12, 2013.

J. Marhold, A. Brehm, and K. Kramer. The Drosophila methyl-DNA binding protein MBD2/3 interacts with the NuRD complex via p55 and MI-2. *BMC Mol. Biol.*, 5:20, 2004.

N. F. Marshall, J. Peng, Z. Xie, and D. H. Price. Control of RNA polymerase II elongation potential by a novel carboxyl- terminal domain kinase. *J. Biol. Chem.*, 271(43): 27176–27183, 1996.

T. N. Mavrich, I. P. Ioshikhes, B. J. Venters, C. Jiang, L. P. Tomsho, J. Qi, S. C. Schuster, I. Albert, and B. F. Pugh. A barrier nucleosome model for statistical positioning

of nucleosomes throughout the yeast genome. *Genome Res.*, 18(7):1073–1083, 2008a.

T. N. Mavrich, C. Jiang, I. P. Ioshikhes, X. Li, B. J. Venters, S. J. Zanton, L. P. Tomsho, J. Qi, R. L. Glaser, S. C. Schuster, D. S. Gilmour, I. Albert, and B. F. Pugh. Nucleosome organization in the Drosophila genome. *Nature*, 453(7193):358–362, 2008b.

M. E. Maxon, J. A. Goodrich, and R. Tjian. Transcription factor IIE binds preferentially to RNA polymerase IIa and recruits TFIIH: A model for promoter clearance. *Genes Dev.*, 8(5):515–524, 1994.

A. Mayer and L. Stirling Churchman. Genome-wide profiling of RNA polymerase transcription at nucleotide resolution in human cells with native elongating transcript sequencing. *Nat. Protoc.*, 11, 2016.

A. Mayer, J. di Iulio, S. Maleri, U. Eser, J. Vierstra, A. Reynolds, R. Sandstrom, J. A. Stamatoyannopoulos, and L. S. Churchman. Native Elongating Transcript Sequencing Reveals Human Transcriptional Activity at Nucleotide Resolution. *Cell*, 161(3): 541, 2015.

S. McCracken, N. Fong, E. Rosonina, K. Yankulov, G. Brothers, D. Siderovski, A. Hessel, S. Foster, S. Shuman, and D. L. Bentley. 5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Genes Dev.*, 11(24):3306–3318, 1997.

S. L. McDaniel and B. D. Strahl. Shaping the cellular landscape with Set2/SETD2 methylation. 74(18):3317–3334, 2017.

A. Meinhart, T. Kamenski, S. Hoeppner, S. Baumli, and P. Cramer. A structural perspective of CTD function. *Genes Dev.*, 19:1401–1415, 2005.

E. M. Mendenhall, R. P. Koche, T. Truong, V. W. Zhou, B. Issac, A. S. Chi, M. Ku, and B. E. Bernstein. GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells. *PLoS Genet.*, 6(12):e1001244, 2010.

T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Mexissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560, 2007.

L. Milligan, V. A. Huynh-Thu, C. Delan-Forino, A. Tuck, E. Petfalski, R. Lombraña, G. Sanguinetti, G. Kudla, and D. Tollervey. Strand-specific, high-resolution mapping of modified RNA polymerase II. *Mol. Syst. Biol.*, 12(6):874, 2016.

A. Missra and D. S. Gilmour. Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex. *Proc. Natl. Acad. Sci. U. S. A.*, 107(25):11301–11306, 2010.

M. Mohan, H. M. Herz, Y. H. Takahashi, C. Lin, K. C. Lai, Y. Zhang, M. P. Washburn, L. Florens, and A. Shilatifard. Linking H3K79 trimethylation to Wnt signaling through a novel Dot1-containing complex (DotCom). *Genes Dev.*, 24(6):574–589, 2010.

F. Mohn, M. Weber, M. Rebhan, T. C. Roloff, J. Richter, M. B. Stadler, M. Bibel, and D. Schübeler. Lineage-Specific Polycomb Targets and De Novo DNA Methylation

Define Restriction and Potential of Neuronal Progenitors. *Mol. Cell*, 30(6):755–766, 2008.

A. Morillon, N. Karabetsou, A. Nair, and J. Mellor. Dynamic lysine methylation on histone H3 defines the regulatory phase of gene transcription. *Mol. Cell*, 18(6): 723–734, 2005.

N. Morohashi, K. Nakajima, D. Kurihara, Y. Mukai, A. P. Mitchell, and M. Shimizu. A nucleosome positioned by α2/Mcm1 prevents Hap1 activator binding in vivo. *Biochem. Biophys. Res. Commun.*, 364(3):583–588, 2007.

S. Moteki and D. Price. Functional coupling of capping and transcription of mRNA. *Mol. Cell*, 10(3):599–609, 2002.

F. Müller, A. Zaucker, and L. Tora. Developmental regulation of transcription initiation: More than just changing the actors. 20(5):533–540, 2010.

S. C. Murray, P. Lorenz, F. S. Howe, M. Wouters, T. Brown, S. Xi, H. Fischl, W. Khushaim, J. R. Rayappu, A. Angel, and J. Mellor. H3K4me3 is neither instructive for, nor informed by, transcription. *biorvix*, 2019.

V. E. Myer and R. A. Young. RNA Polymerase II Holoenzymes and Subcomplexes. *J. Biol. Chem.*, 273(43):27757–27760, 1998.

A. M. Näär, B. D. Lemon, and R. Tjian. Transcriptional Coactivator Complexes. *Annu. Rev. Biochem.*, 70(1):475–501, 2001.

S. Naftelberg, I. E. Schor, G. Ast, and A. R. Kornblihtt. Regulation of Alternative

Splicing Through Coupling with Transcription and Chromatin Structure. *Annu. Rev. Biochem.*, 84(1):165–198, 2015.

R. Narayanan and T. C. Tuoc. Roles of chromatin remodeling BAF complex in neural differentiation and reprogramming. 356(3):575–584, 2014.

S. Nechaev, D. C. Fargo, G. dos Santos, L. Liu, Y. Gao, and K. Adelman. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science*, 327(5963):335–8, 2010.

F. Neri, S. Rapelli, A. Krepelova, D. Incarnato, C. Parlato, G. Basile, M. Maldotti, F. Anselmi, and S. Oliviero. Intragenic DNA methylation prevents spurious transcription initiation. *Nature*, 543(7643):72–77, 2017.

L. V. Ngoc, C. J. Cassidy, C. Y. Huang, S. H. Duttke, and J. T. Kadonaga. The human initiator is a distinct and abundant element that is precisely positioned in focused core promoters. *Genes Dev.*, 31(1):6–11, 2017.

T. H. D. Nguyen, W. P. Galej, S. M. Fica, P. C. Lin, A. J. Newman, and K. Nagai. CryoEM structures of two spliceosomal complexes: Starter and dessert at the spliceosome feast. 36:48–57, 2016.

Z. Ni, A. Saunders, N. J. Fuda, J. Yao, J.-R. Suarez, W. W. Webb, and J. T. Lis. P-TEFb Is Critical for the Maturation of RNA Polymerase II into Productive Elongation In Vivo. *Mol. Cell. Biol.*, 28(3):1161–1170, 2008.

T. Nojima, T. Hirose, H. Kimura, and M. Hagiwara. The interaction between cap-

binding complex and RNA export factor is required for intronless mRNA export. *J. Biol. Chem.*, 282(21):15645–15651, 2007.

T. Nojima, M. Carmo-Fonseca, N. J. Proudfoot, T. S. Gomes, A. R. Fialho Grosso, H. Kimura, M. J. Dye, and S. Dhir. Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell*, 161:526–540, 2015.

T. Nojima, M. Carmo-Fonseca, and N. J. Proudfoot. Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. *Nat. Protoc.*, 11(3):413–428, 2016.

T. Nojima, M. Tellier, J. Foxwell, C. Ribeiro de Almeida, S. M. Tan-Wong, S. Dhir, G. Dujardin, A. Dhir, S. Murphy, and N. J. Proudfoot. Deregulated Expression of Mammalian lncRNA through Loss of SPT6 Induces R-Loop Formation, Replication Stress, and Cellular Senescence. *Mol. Cell*, 72(6):970–984.e7, 2018.

E. Nudler. RNA polymerase backtracking in gene regulation and genome instability. 149(7):1438–1445, 2012.

Y. Ohkuma, S. Hashimoto, C. K. Wang, M. Horikoshi, and R. G. Roeder. Analysis of the role of TFIIE in basal transcription and TFIIH-mediated carboxy-terminal domain phosphorylation through structure-function studies of TFIIE-alpha. *Mol. Cell. Biol.*, 15(9):4856–4866, 1995.

U. Ohler, G. chun Liao, H. Niemann, and G. M. Rubin. Computational analysis of core promoters in the Drosophila genome. *Genome Biol.*, 3(12):research0087.1, 2002.

M. Ohno, H. Sakamoto, and Y. Shimura. Preferential excision of the 5' proximal intron

from mRNA precursors with two introns as mediated by the cap structure. *Proc. Natl. Acad. Sci. U. S. A.*, 84(15):5187–5191, 1987.

V. Oksenych and F. Coin. The long unwinding road: XPB and XPD helicases in damaged DNA opening. 9(1):90–96, 2010.

G. Orphanides, G. LeRoy, C. H. Chang, D. S. Luse, and D. Reinberg. FACT, a factor that facilitates transcript elongation through nucleosomes. *Cell*, 92(1):105–116, 1998.

F. Ozsolak, J. S. Song, X. S. Liu, and D. E. Fisher. High-throughput mapping of the chromatin structure of human promoters. *Nat. Biotechnol.*, 25(2):244–248, 2007.

M. Pabis, N. Neufeld, M. C. Steiner, T. Bojic, Y. Shav-Tal, and K. M. Neugebauer. The nuclear cap-binding complex interacts with the U4/U6·U5 tri-snRNP and promotes spliceosome assembly in mammalian cells. *RNA*, 19(8):1054–1063, 2013.

R. A. Padgett, M. M. Konarska, P. J. Grabowski, S. F. Hardy, and P. A. Sharp. Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors. *Science (80-.).*, 225(4665):898–903, 1984.

Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40(12):1413–1415, 2008.

R. Pavri, B. Zhu, G. Li, P. Trojer, S. Mandal, A. Shilatifard, and D. Reinberg. Histone H2B Monoubiquitination Functions Cooperatively with FACT to Regulate Elongation by RNA Polymerase II. *Cell*, 125(4):703–717, 2006.

B. Payer and J. T. Lee. X Chromosome Dosage Compensation: How Mammals Keep the Balance. *Annu. Rev. Genet.*, 42(1):733–772, 2008.

B. M. Peterlin and D. H. Price. Controlling the Elongation Phase of Transcription with P-TEFb. 23(3):297–305, 2006.

S. J. Petesch and J. T. Lis. Overcoming the nucleosome barrier during transcript elongation. 28(6):285–294, 2012.

S. Picelli, Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, and R. Sandberg. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, 10(11):1096–1100, 2013.

C. Plaschka, A. J. Newman, and K. Nagai. Structural basis of nuclear pre-mRNA splicing: Lessons from Yeast. *Cold Spring Harb. Perspect. Biol.*, 11(5):a032391, 2019.

D. K. Pokholok, C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, E. Herbolsheimer, J. Zeitlinger, F. Lewitter, D. K. Gifford, and R. A. Young. Genome-wide Map of Nucleosome Acetylation and Methylation in Yeast. *Cell*, 122:517–527, 2005.

O. Porrua and D. Libri. Transcription termination and the control of the transcriptome: Why, where and how to stop. 16(3):190–202, 2015.

Z. C. Poss, C. C. Ebmeier, and D. J. Taatjes. The Mediator complex and transcription regulation. 48(6):575–608, 2013.

J. A. Pruneski, S. J. Hainer, K. O. Petrov, and J. A. Martens. The Paf1 Complex Represses SER3 Transcription in Saccharomyces cerevisiae by Facilitating Intergenic Transcription-Dependent Nucleosome Occupancy of the SER3 Promoter . *Eukaryot. Cell*, 10(10):1283–1294, 2011.

D. Qin, L. Huang, A. Wlodaver, J. Andrade, and J. P. Staley. Sequencing of lariat termini in S. cerevisiae reveals 5' splice sites, branch points, and novel splicing events. *RNA*, 22(2):237–253, 2016.

A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–2, 2010.

E. A. Rach, D. R. Winter, A. M. Benjamin, D. L. Corcoran, T. Ni, J. Zhu, and U. Ohler. Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet.*, 7(1), 2011.

P. L. Raghunathan and C. Guthrie. RNA unwinding in U4/U6 snRNPs requires ATP hydrolysis and the DEIH-box splicing factor Brr2. *Curr. Biol.*, 8(15):847–855, 1998.

A. Raj, P. van den Bogaard, S. A. Rifkin, A. van Oudenaarden, and S. Tyagi. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods*, 5 (10):877–879, 2008.

A. Ramanathan, G. B. Robb, and S.-H. Chan. mRNA capping: biological functions and applications. *Nucleic Acids Res.*, 44(16):7511–7526, 2016.

R. Reddy and H. Busch. Small Nuclear RNAs: RNA Sequences, Structure, and

Modifications. In *Struct. Funct. Major Minor Small Nucl. Ribonucleoprotein Part.*, pages 1–37. Springer Berlin Heidelberg, 1988.

D. Reinberg, G. Orphanides, R. Ebright, S. Akoulitchev, J. Carcamo, H. Cho, P. Cortes, R. Drapkin, O. Flores, I. Ha, J. A. Inostroza, S. Kim, T. K. Kim, P. Kumar, T. Lagrange, G. LeRoy, H. Lu, D. M. Ma, E. Maldonado, A. Merino, F. Mermelstein, I. Olave, M. Sheldon, R. Shiekhattar, N. Stone, X. Sun, L. Weis, K. Yeung, and L. Zawel. The RNA polymerase II general transcription factors: Past, present, and future. In *Cold Spring Harb. Symp. Quant. Biol.*, volume 63, pages 83–103, 1998.

Reinberg D and Roeder RG. Factors involved in specific transcription by mammalian RNA polymerase II. Transcription factor IIS stimulates elongation of RNA chains. *J. Biol. Chem.*, 262:3331–3337, 1987.

D. Reiness. Elongation Factor-dependent Transcript Shortening by Template-engaged RNA Polymerase 11". *J. Biol. Chem.*, 267(6):3795–3800, 1992.

H. S. Rhee and B. F. Pugh. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, 483(7389):295–301, 2012.

M. Rienzo and A. Casamassimi. Integrator complex and transcription regulation: Recent findings and pathophysiology. 1859(10):1269–1280, 2016.

J. R. Rodriguez, C. W. Pikielny, and M. Rosbash. In vivo characterization of yeast mRNA processing intermediates. *Cell*, 39(3 PART 2):603–610, 1984.

R. G. Roeder. The role of general initiation factors in transcription by RNA polymerase II. 21(9):327–335, 1996.

A. E. Rougvie and J. T. Lis. The RNA polymerase II molecule at the 5 end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell*, 54(6): 795–804, 1988.

B. Ruskin, A. R. Krainer, T. Maniatis, and M. R. Green. Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. *Cell*, 38(1):317–331, 1984.

S. Sainsbury, C. Bernecky, and P. Cramer. Structural basis of transcription initiation by RNA polymerase II. 16(3):129–143, 2015.

A. Sandelin, P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizaki, and D. A. Hume. Mammalian RNA polymerase II core promoters: Insights from genome-wide studies. 8(6):424–436, 2007.

J. M. Santos-Pereira and A. Aguilera. R loops: new modulators of genome dynamics and function. *Nat. Rev. Genet.*, 16(10):583–597, 2015.

H. Santos-Rosa, R. Schneider, A. J. Bannister, J. Sherriff, B. E. Bernstein, N. C. Emre, S. L. Schreiber, J. Mellor, and T. Kouzarides. Active genes are tri-methylated at K4 of histone H3. *Nature*, 419(6905):407–411, 2002.

M. Saponaro, T. Kantidakis, R. Mitter, G. P. Kelly, M. Heron, H. Williams, J. Söding, A. Stewart, and J. Q. Svejstrup. RECQL5 controls transcript elongation and suppresses genome instability associated with transcription stress. *Cell*, 157(5): 1037–1049, 2014.

A. Saunders, J. Werner, E. D. Andrulis, T. Nakayama, S. Hirose, D. Reinberg, and

J. T. Lis. Tracking FACT and the RNA polymerase II elongation complex through chromatin in vivo. *Science (80-.).*, 301(5636):1094–1096, 2003.

S. Saxonov, P. Berg, and D. L. Brutlag. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U. S. A.*, 103(5):1412–1417, 2006.

L. Schaeffer, R. Roy, S. Humbert, V. Moncollin, W. Vermeulen, J. H. Hoeijmakers, P. Chambon, and J. M. Egly. DNA repair helicase: A component of BTF2 (TFIIH) basic transcription factor. *Science (80-.).*, 260(5104):58–63, 1993.

M. Schlackow, T. Nojima, T. Gomes, A. Dhir, M. Carmo-Fonseca, and N. J. Proudfoot. Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs. *Mol. Cell*, 65:25–38, 2017.

M. Schmid and T. H. Jensen. Controlling nuclear RNA levels. 19(8):518–529, 2018.

D. R. Schmidt and S. L. Schreiber. Molecular association between ATR and two components of the nucleosome remodeling and deacetylating complex, HDAC2 and CHD4. *Biochemistry*, 38(44):14711–14717, 1999.

D. E. Schones, K. Cui, S. Cuddapah, T. Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell*, 132(5):887–898, 2008.

S. C. Schroeder, B. Schwer, S. Shuman, and D. Bentley. Dynamic association of capping enzymes with transcribing RNA polymerase II. *Genes Dev.*, 14(19):2435–2440, 2000.

Schwalb B, Michel M, Zacher B, Fruehauf K, Demel C, Tresch A, Gagneur J, and Cramer P. TT-seq maps the human transient transcriptome. *Science (80-.).*, 352 (6290):1225–1228, 2016.

E. A. Sekinger, Z. Moqtaderi, and K. Struhl. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol. Cell*, 18(6):735–748, 2005.

S. Shah, Y. Takei, W. Zhou, E. Lubeck, J. Yun, C. H. L. Eng, N. Koulena, C. Cronin, C. Karp, E. J. Liaw, M. Amin, and L. Cai. Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. *Cell*, 174(2):363–376.e16, 2018.

R. M. Sheridan, N. Fong, A. D'Alessandro, and D. L. Bentley. Widespread Backtracking by RNA Pol II Is a Major Effector of Gene Activation, 5 Pause Release, Termination, and Transcription Elongation Rate. *Mol. Cell*, 73(1):107–118.e4, 2019.

A. Shetty, S. P. Kallgren, C. Demel, K. C. Maier, D. Spatt, B. H. Alver, P. Cramer, P. J. Park, and F. Winston. Spt5 Plays Vital Roles in the Control of Sense and Antisense Transcription Elongation. *Mol. Cell*, 66(1):77–88.e5, 2017.

A. Shilatifard, W. S. Lane, K. W. Jackson, R. C. Conaway, and J. W. Conaway. An RNA polymerase II elongation factor encoded by the human ELL gene. *Science (80-.).*, 271(5257):1873–1876, 1996.

T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, and Y. Hayashizaki. Cap analysis gene expression for

high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.*, 100(26):15776–81, 2003.

M. K. Shivji, X. Renaudin, Ç. H. Williams, and A. R. Venkitaraman. BRCA2 Regulates Transcription Elongation by RNA Polymerase II to Prevent R-Loop Accumulation. *Cell Rep.*, 22(4):1031–1039, 2018.

K. Sidaway-Lee, M. J. Costa, D. A. Rand, B. Finkenstadt, and S. Penfield. Direct measurement of transcription rates reveals multiple mechanisms for configuration of the Arabidopsis ambient temperature response. *Genome Biol.*, 15(3):R45, 2014.

R. Simic, D. Lindstrom, H. Tran, K. Roinick, P. Costa, A. Johnson, G. Hartzog, and K. Arndt. Chromatin remodeling protein Chd1 interacts with transcription elongation factors and localizes to transcribed genes. *EMBO*, 22(8):1846–1856, 2003.

R. H. Singer and D. C. Ward. Actin gene expression visualized in chicken muscle tissue culture by using in situ hybridization with a biotinated nucleotide analog. *Proc. Natl. Acad. Sci. U. S. A.*, 79(23 I):7331–7335, 1982.

J. R. Skaar, A. L. Ferris, X. Wu, A. Saraf, K. K. Khanna, L. Florens, M. P. Washburn, S. H. Hughes, and M. Pagano. The Integrator complex controls the termination of transcription at diverse classes of gene targets. *Cell Res.*, 25(3):288–305, 2015.

K. Skourti-Stathaki, N. J. Proudfoot, and N. Gromak. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol. Cell*, 42(6):794–805, 2011.

K. Skourti-Stathaki, K. Kamieniarz-Gdula, and N. J. Proudfoot. R-loops induce re-

pressive chromatin marks over mammalian gene terminators. *Nature*, 516(7531): 436–439, 2014.

S. T. Smale and D. Baltimore. The "initiator" as a transcription control element. *Cell*, 57(1):103–113, 1989.

S. T. Smale and J. T. Kadonaga. The RNA Polymerase II Core Promoter. *Annu. Rev. Biochem.*, 72(1):449–479, 2003.

E. Smith and A. Shilatifard. Transcriptional elongation checkpoint control in development and disease. *Genes Dev.*, 27:1079–1088, 2013.

B. Sobhian, N. Laguette, A. Yatim, M. Nakamura, Y. Levy, R. Kiernan, and M. Benkirane. HIV-1 Tat Assembles a Multifunctional Transcription Elongation Complex and Stably Associates with the 7SK snRNP. *Mol. Cell*, 38(3):439–451, 2010.

G. Sokpor, R. Castro-Hernandez, J. Rosenbusch, J. F. Staiger, and T. Tuoc. ATP-dependent chromatin remodeling during cortical neurogenesis. 12(APR):226, 2018.

S. Solier, M. C. Ryan, S. E. Martin, S. Varma, K. W. Kohn, H. Liu, B. R. Zeeberg, and Y. Pommier. Transcription poisoning by topoisomerase i is controlled by gene length, splice sites, and miR-142-3p. *Cancer Res.*, 73(15):4830–4839, 2013.

M. Sopta, R. W. Carthew, and J. Greenblatt. Isolation of three proteins that bind to mammalian RNA polymerase II. *J. Biol. Chem.*, 260(18):10353–60, 1985.

J. Soutourina. Transcription regulation by the Mediator complex. 19(4):262–274, 2018.

J. Soutourina, S. Wydau, Y. Ambroise, C. Boschiero, and M. Werner. Direct interaction of RNA polymerase II and mediator required for transcription in vivo. *Science (80-.).*, 331(6023):1451–1454, 2011.

M. Spingola, L. Grate, D. Haussler, and A. Manuel. Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. *RNA*, 5(2):221–234, 1999.

B. Stadelmayer, G. Micas, A. Gamot, P. Martin, N. Malirat, S. Koval, R. Raffel, B. Sobhian, D. Severac, S. Rialle, H. Parrinello, O. Cuvier, and M. Benkirane. Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. *Nat. Commun.*, 5(1):1–11, 2014.

R. Stein, A. Razin, and H. Cedar. In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. *Proc. Natl. Acad. Sci. U. S. A.*, 79(11 I):3418–3422, 1982.

T. A. Steitz and J. A. Steitz. A general two-metal-ion mechanism for catalytic RNA. *Proc. Natl. Acad. Sci. U. S. A.*, 90(14):6498–6502, 1993.

M. Stoeckius, D. Grün, and N. Rajewsky. Paternal RNA contributions in the Caenorhabditis elegans zygote. *EMBO J.*, 33(16):1740–1750, 2014.

R. Straussman, D. Nejman, D. Roberts, I. Steinfeld, B. Blum, N. Benvenisty, I. Simon, Z. Yakhini, and H. Cedar. Developmental programming of CpG island methylation profiles in the human genome. *Nat. Struct. Mol. Biol.*, 16(5):564–571, 2009.

A. Suraweera, Y. Lim, R. Woods, G. W. Birrell, T. Nasim, O. J. Becherel, and M. F.

Lavin. Functional role for senataxin, defective in ataxia oculomotor apraxia type 2, in transcriptional regulation. *Hum. Mol. Genet.*, 18(3), 2009.

T. Szczepińska, K. Kalisiak, R. Tomecki, A. Labno, L. S. Borowski, T. M. Kulinski, D. Adamska, J. Kosinska, and A. Dziembowski. DIS3 shapes the RNA polymerase II transcriptome in humans by degrading a variety of unwanted transcripts. *Genome Res.*, 25(11):1622–1633, 2015.

M. N. Szentirmay and M. Sawadogo. Sarkosyl block of transcription reinitiation by RNA polymerase II as visualized by the colliding polymerases reinitiation assay. *Nucleic Acids Res.*, 22(24):5341–5346, 1994.

H. Takahashi, T. J. Parmely, S. Sato, C. Tomomori-Sato, C. A. Banks, S. E. Kong, H. Szutorisz, S. K. Swanson, S. Martin-Brown, M. P. Washburn, L. Florens, C. W. Seidel, C. Lin, E. R. Smith, A. Shilatifard, R. C. Conaway, and J. W. Conaway. Human mediator subunit MED26 functions as a docking site for transcription elongation factors. *Cell*, 146(1):92–104, 2011.

H. Takahashi, A. Ranjan, S. Chen, H. Suzuki, M. Shibata, T. Hirose, H. Hirose, K. Sasaki, R. Abe, K. Chen, Y. He, Y. Zhang, I. Takigawa, T. Tsukiyama, M. Watanabe, S. Fujii, M. Iida, J. Yamamoto, Y. Yamaguchi, Y. Suzuki, M. Matsumoto, K. I. Nakayama, M. P. Washburn, A. Saraf, L. Florens, S. Sato, C. Tomomori-Sato, R. C. Conaway, J. W. Conaway, and S. Hatakeyama. The role of Mediator and Little Elongation Complex in transcription termination. *Nat. Commun.*, 11(1):1–20, 2020.

J. C. Tanny, H. Erdjument-Bromage, P. Tempst, and C. D. Allis. Ubiquitylation of hi-

stone H2B controls RNA polymerase II transcription elongation independently of histone H3 methylation. *Genes Dev.*, 21(7):835–847, 2007.

K. Tantale, F. Mueller, A. Kozulic-Pirher, A. Lesne, J. M. Victor, M. C. Robert, S. Capozi, R. Chouaib, V. Bäcker, J. Mateos-Langerak, X. Darzacq, C. Zimmer, E. Basyuk, and E. Bertrand. A single-molecule view of transcription reveals convoys of RNA polymerases and multi-scale bursting. *Nat. Commun.*, 7(1):1–14, 2016.

S. S. Teves, C. M. Weber, and S. Henikoff. Transcribing through the nucleosome. 39 (12):577–586, 2014.

M. J. Thomas, A. A. Platas, and D. K. Hawley. Transcriptional fidelity and proofreading by RNA polymerase II. *Cell*, 93(4):627–637, 1998.

J. M. Tome, N. D. Tippens, and J. T. Lis. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nat. Genet.*, 50(11):1533–1541, 2018.

T. Tsukiyama, C. Daniel, J. Tamkun, and C. Wu. ISWI, a Member of the SliW2KNF2 ATPase Family, Encodes the 140 kDa Subunit of the Nucleosome Remodeling Factor. Technical report, 1995.

C. R. Vakoc, M. M. Sachdeva, H. Wang, and G. A. Blobel. Profile of Histone Lysine Methylation across Transcribed Mammalian Chromatin. *Mol. Cell. Biol.*, 26(24): 9185–9195, 2006.

S. B. Van Oss, C. E. Cucinotta, and K. M. Arndt. Emerging Insights into the Roles of the Paf1 Complex in Gene Regulation. 42(10):788–798, 2017.

A. Vannini and P. Cramer. Conservation between the RNA Polymerase I, II, and III Transcription Initiation Machineries. 45(4):439–446, 2012.

N. L. Vastenhouw, Y. Zhang, I. G. Woods, F. Imam, A. Regev, X. S. Liu, J. Rinn, and A. F. Schier. Chromatin signature of embryonic pluripotency is established during genome activation. *Nature*, 464(7290):922–926, 2010.

J. P. Venables, C.-S. Koh, U. Froehlich, E. Lapointe, S. Couture, L. Inkel, A. Bramard, É. R. Paquet, V. Watier, M. Durand, J.-F. Lucier, J. Gervais-Bird, K. Tremblay, P. Prinos, R. Klinck, S. A. Elela, and B. Chabot. Multiple and Specific mRNA Processing Targets for the Major Human hnRNP Proteins. *Mol. Cell. Biol.*, 28(19):6033–6043, 2008.

M. Vermeulen, K. W. Mulder, S. Denissov, W. W. Pijnappel, F. M. van Schaik, R. A. Varier, M. P. Baltissen, H. G. Stunnenberg, M. Mann, and H. T. M. Timmers. Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. *Cell*, 131(1):58–69, 2007.

G. Villamil, L. Wachutka, P. Cramer, J. Gagneur, and B. Schwalb. Transient transcriptome sequencing: computational pipeline to quantify genome-wide RNA kinetic parameters and transcriptional enhancer activity. *bioRxiv*, page 659912, 2019.

N. Viphakone, I. Sudbery, L. Griffith, C. G. Heath, D. Sims, and S. A. W. Correspondence. Co-transcriptional Loading of RNA Export Factors Shapes the Human Transcriptome. *Mol. Cell*, 75:310–323, 2019.

S. M. Vos, L. Farnung, M. Boehning, C. Wigge, A. Linden, H. Urlaub, and P. Cramer.

Structure of activated transcription complex Pol II–DSIF–PAF–SPT6. *Nature*, 560 (7720):607–612, 2018a.

S. M. Vos, L. Farnung, H. Urlaub, and P. Cramer. Structure of paused transcription complex Pol II–DSIF–NELF. *Nature*, 560(7720):601–606, 2018b.

T. Wada, T. Takagi, Y. Yamaguchi, D. Watanabe, H. Handa, T. Wada, T. Takagi, and Y. Yamaguchi. Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro. Technical Report 24, 1998.

A. Wagschal, E. Rousset, P. Basavarajaiah, X. Contreras, A. Harwig, S. Laurent-Chabalier, M. Nakamura, X. Chen, K. Zhang, O. Meziane, F. Boyer, H. Parrinello, B. Berkhout, C. Terzian, M. Benkirane, and R. Kiernan. Microprocessor, Setx, Xrn2, and Rrp6 co-operate to induce premature termination of transcription by RNAPII. *Cell*, 150(6):1147–1157, 2012.

E. Wahlet. Purification and Characterization of a Mammalian Polyadenylate Polymerase Involved in the 3' End Processing of Messenger RNA Precursors*. *J. Biol. Chem.*, 266(5):3131–3139, 1991.

L. Wan, H. Wen, Y. Li, J. Lyu, Y. Xi, T. Hoshii, J. K. Joseph, X. Wang, Y. H. E. Loh, M. A. Erb, A. L. Souza, J. E. Bradner, L. Shen, W. Li, H. Li, C. D. Allis, S. A. Armstrong, and X. Shi. ENL links histone acetylation to oncogenic gene expression in acute myeloid leukaemia. *Nature*, 543(7644):265–269, 2017.

R. Wan, C. Yan, R. Bai, G. Huang, and Y. Shi. Structure of a yeast catalytic step i spliceosome at 3.4 Å resolution. *Science (80-.)*., 353(6302):895–904, 2016.

E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

R. Wang, D. Zheng, L. Wei, Q. Ding, and B. Tian. Regulation of Intronic Polyadenylation by PCF11 Impacts mRNA Expression of Long Genes. *Cell Rep.*, 26(10): 2766–2778.e6, 2019.

L. Warfield, S. Ramachandran, T. Baptista, D. Devys, L. Tora, and S. Hahn. Transcription of Nearly All Yeast RNA Polymerase II-Transcribed Genes Is Dependent on Transcription Factor TFIID. *Mol. Cell*, 68(1):118–129.e5, oct 2017. ISSN 10974164. doi: 10.1016/j.molcel.2017.08.014. URL `http://dx.doi.org/10.1016/j.molcel.2017.08.014`.

C. M. Weber, S. Ramachandran, and S. Henikoff. Nucleosomes are context-specific, H2A.Z-Modulated barriers to RNA polymerase. *Mol. Cell*, 53(5):819–830, 2014.

J. Weltner, D. Balboa, S. Katayama, M. Bespalov, K. Krjutškov, E. M. Jouhilahti, R. Trokovic, J. Kere, and T. Otonkoski. Human pluripotent reprogramming with CRISPR activators. *Nat. Commun.*, 9(2643), 2018.

R. C. Wilkins and J. T. Lis. Dynamics of potentiation and activation: GAGA factor and its role in heat shock gene regulation. Technical Report 20, 1997.

M. E. Wilkinson, C. Charenton, and K. Nagai. RNA Splicing by the Spliceosome. *Annu. Rev. Biochem.*, 89(1), 2020.

E. M. Wissink, A. Vihervaara, N. D. Tippens, and J. T. Lis. Nascent RNA analyses: tracking transcription and its regulation. *Nat. Rev. Genet.*, 2019.

K. H. Wong, Y. Jin, and K. Struhl. TFIIH Phosphorylation of the Pol II CTD Stimulates Mediator Dissociation from the Preinitiation Complex and Promoter Escape. *Mol. Cell*, 54(4):601–612, 2014.

Y. M. Woo, Y. Kwak, S. Namkoong, K. Kristjánsdóttir, S. H. Lee, J. H. Lee, and H. Kwak. TED-Seq Identifies the Dynamics of Poly(A) Length during ER Stress. *Cell Rep.*, 24(13):3630–3641.e7, 2018.

C. H. Wu, Y. Yamaguchi, L. R. Benjamin, M. Horvat-Gordon, J. Washinsky, E. Enerly, J. Larsson, A. Lambertsson, H. Handa, and D. Gilmour. NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. *Genes Dev.*, 17 (11):1402–1414, 2003.

D. Yadav, K. Ghosh, S. Basu, R. G. Roeder, and D. Biswas. Multivalent Role of Human TFIID in Recruiting Elongation Components at the Promoter-Proximal Region for Transcriptional Control. *Cell Rep.*, 26(5):1303–1317.e7, 2019.

T. Yamada, Y. Yamaguchi, N. Inukai, S. Okamoto, T. Mura, and H. Handa. P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. *Mol. Cell*, 21(2):227–237, 2006.

Y. Yamaguchi, T. Takagi, T. Wada, K. Yano, A. Furuya, S. Sugimoto, J. Hasegawa, and H. Handa. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell*, 97(1):41–51, 1999.

Y. Yamaguchi, N. Inukai, T. Narita, T. Wada, and H. Handa. Evidence that Negative Elongation Factor Represses Transcription Elongation through Binding to a DRB Sensitivity-Inducing Factor/RNA Polymerase II Complex and RNA. *Mol. Cell. Biol.*, 22(9):2918–2927, 2002.

C. Yan, R. Wan, and Y. Shi. Molecular mechanisms of pre-mRNA splicing through structural biology of the spliceosome. *Cold Spring Harb. Perspect. Biol.*, 11(1): a032409, 2019.

F. Yan, D. R. Powell, D. J. Curtis, and N. C. Wong. From reads to insight: A hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.*, 21(1):1–16, 2020.

C. Yang, E. Bolotin, T. Jiang, F. M. Sladek, and E. Martinez. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, 389(1):52–65, 2007.

Q. Yang, X. Liu, T. Zhou, J. Cook, K. Nguyen, and X. Bai. RNA polymerase II pausing modulates hematopoietic stem cell emergence in zebrafish. *Blood*, 2016.

Z. Yang, J. H. Yik, R. Chen, N. He, K. J. Moon, K. Ozato, and Q. Zhou. Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. *Mol. Cell*, 19(4):535–545, 2005.

K. L. Yap, S. Li, A. M. Muñoz-Cabello, S. Raguz, L. Zeng, S. Mujtaba, J. Gil, M. J. Walsh, M.-M. Zhou, R. Kurokawa, and et Al. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Mol. Cell*, 38(5):662–74, 2010.

K. Yen, V. Vinayachandran, K. Batta, R. T. Koerber, and B. F. Pugh. Genome-wide nucleosome specificity and directionality of chromatin remodelers. *Cell*, 149(7): 1461–1473, 2012.

K. Yoshida and Y. Miki. Role of BRCA1 and BRCA2 as regulators of DNA repair, transcription, and cell cycle in response to DNA damage. *Cancer Sci.*, 95(11): 866–871, 2004.

M. Yu, W. Yang, T. Ni, Z. Tang, T. Nakadai, J. Zhu, and R. G. Roeder. RNA polymerase II-associated factor 1 regulates the release and phosphorylation of paused RNA polymerase II. *Science (80-.).*, 350(6266):1383–1386, 2015.

S. Yunger, L. Rosenfeld, Y. Garini, and Y. Shav-Tal. Single-allele analysis of transcription kinetics in living mammalian cells. *Nat. Methods*, 7(8):631–633, 2010.

J. Zeitlinger, A. Stark, M. Kellis, J.-W. Hong, S. Nechaev, K. Adelman, M. Levine, and R. A. Young. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat. Genet.*, 39(12):1512–6, 2007.

C. Zhang and Z. F. Burton. Transcription factors IIF and IIS and nucleoside triphosphate substrates as dynamic probes of the human RNA polymerase II mechanism. *J. Mol. Biol.*, 342(4):1085–1099, 2004.

D. W. Zhang, A. L. Mosley, S. R. Ramisetty, J. B. Rodríguez-Molina, M. P. Washburn, and A. Z. Ansari. Ssu72 phosphatase-dependent erasure of phospho-Ser7 marks on the RNA polymerase II C-terminal domain is essential for viability and transcription termination. *J. Biol. Chem.*, 287(11):8541–8551, 2012.

H. Zhang, F. Rigo, and H. G. Martinson. Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site. *Mol. Cell*, 59(3):437–448, 2015.

Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137, 2008.