The role of enhancers in regulating the transcription cycle

Martin S. C. Larke

St Cross College

MRC Molecular Haematology Unit

Weatherall Institute of Molecular Medicine

University of Oxford

A thesis submitted for the degree of Doctor of Philosophy of the University of Oxford

Michaelmas Term 2018

Declaration

I, Martin Larke, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Enhancers are key determinants in the spatiotemporal control of gene expression which underlies cellular differentiation and disease; however, the dominant mechanisms through which they exert their effects on genes remain poorly understood. In order to achieve fine control over nascent transcript levels, transcription is highly regulated at many stages. It has been suggested that enhancers may act on many specific stages in transcription to modulate transcript production. The idea that initiation of transcription is the principal barrier in achieving changes in gene expression has been challenged by the discovery of promoter proximal pausing of RNA polymerase II (Pol 2 pausing) at the majority of protein coding genes in addition to further downstream regulatory steps.

In this thesis I sought to determine the predominant mechanism by which enhancers bring about differential gene expression *in vivo* by examining knockout models of well-defined enhancers at the alpha and beta globin gene clusters. Using a wide variety of complementary methodologies, I demonstrate that these two enhancer clusters promote transcription initiation whilst having minimal effects on downstream transcriptional kinetics. In addition, I build on these findings to show that genome-wide differential transcription initiation best correlates with differential gene expression and is likely the predominant regulatory step in transcription. Pol2 pausing is detectable genome-wide but scales according to initiation frequency suggesting this process is an inherent kinetic property of mammalian promoters, but its role in gene regulation may be less significant than previously thought.

This study highlights the role that enhancers play in regulation of specific stages in transcription to achieve tissue-specific gene expression. These findings may be applicable

to the broader study of metazoan transcription, given the highly conserved nature of transcriptional mechanisms and will enable a better understanding of how eukaryotic gene expression is altered development and disease.

Acknowledgements

I am extremely grateful to have had the opportunity to work towards my thesis under the supervision of Professors Doug Higgs and Jim Hughes. They have provided me with patient and considerate supervision throughout my studies whilst giving me the freedom to develop as a scientist.

I wish to thank my co-workers within the WIMM and wider University who have helped and guided me through my DPhil. I am also thankful to the MRC and the WIMM prize studentship scheme for funding and facilitating my studies.

I particularly wish to thank those who have taught and helped me in generating material and data for this project: Dr Jackie Sharpe, Dr Jackie Sloane-Stanley, Mrs Sue Butler, Dr Kevin Clarke, Dr Bryony Crompton, Dr Mira Kassouf and Dr Damien Downes.

To my partner Laura, my parents Nick and Lynda, and my sister Sarah: thank you for supporting me throughout this process and tolerating my attempts to enthuse you with details of my research. You have taught me a great detail about how to read an audience!

I finally wish to acknowledge those who have provided helpful suggestions and guidance during the writing of this thesis: Professor Jim Hughes, Professor Doug Higgs, Dr Rob Beagrie, Dr Marieke Oudelaar and Dr Laura Larke.

Table of contents

Declaration **2**
Abstract **3**
Acknowledgements **5**
Table of contents **6**
List of figures **10**
Table of abbreviations **12**
Chapter 1 - Introduction **15**
1.1 Transcriptional rate alters RNA output 15
1.2 Methods to identify transcriptional kinetics 18
1.2.1 Indirect 18
1.2.2 Direct 19
1.3 The transcription cycle and the role of enhancers in regulating it 24
1.3.1 Initiation 25
1.3.2 Pol2 pausing 25
1.3.3 Splicing 30
1.3.4 Termination 31
1.4 General features of transcription in higher eukaryotes 32
1.4.1 Transcription initiation occurs in different modes 32
1.4.2 Promoter proximal pausing of Pol2 may be ubiquitous 33
1.4.3 Antisense transcription 33
1.5 Outline of the thesis 34
Chapter 2 – Materials and Methods **37**
2.1 Wet lab protocols 37
2.1.1 Growth of K562 cells, pharmacological treatment and harvesting 37
2.1.2 Growth of mouse foetal livers and harvesting of erythroid cells 37
2.1.3 RNA extraction and quality control 38
2.1.4 mNET-seq 39
2.1.5 polyA+ RNA seq library preparation 39
2.1.6 Pol2 ChIP-seq 40
2.1.7 scaRNA-seq 42
2.1.8 ATAC-seq 45
2.1.9 Capture-C 45
2.2 Analysis 46
2.2.1 Published data used in this thesis 46
2.2.2 Summary of alignment and pre-processing for all techniques 47
2.2.3 Generation of observed and annotated TSS list 48
2.2.4 Generation of non-coding RNA list 48
2.2.5 Generation of splice sites list 49
2.2.6 Read coverage relative to list of coordinates (TSS) 49
2.2.7 scaRNA-seq stages in analysis 50

2.2.7.1 Normalisation and visualisation of individual genes 50
2.2.7.2 observed TSS calling using Homer 50
2.2.7.3 Intersecting the observed TSS with annotated TSS 51
2.2.7.4 Observed and annotated TSS motif enrichment analysis 51
2.2.7.5 Read density ratio calculations 51
2.2.8 Processing of the annotated UCSC gene list - stages in analysis 52
2.2.8.1 Whole gene coordinates 52
2.2.8.2 Whole gene minus first 300bp 52
2.2.9 Plotting distances between observed versus annotated TSS 52
2.2.10 ChIP-seq 52
2.2.10.1 Travelling ratio 53
2.2.10.2 Visualisation on genome browser and normalisation 53
2.2.10.3 Pol2 read density calculation 53
2.2.11 polyA+ RNA-seq 53
2.2.11.1 Visualisation on genome browser and normalisation 53
2.2.11.2 Expression calculation (RPKM) 53
2.2.11.3 Intronic RNA read isolation 54
2.2.11.4 Coverage over gene bodies (ngsplot) 54
2.2.11.5 Intronic RNA read density ratio calculations 54
2.2.12 Splice site analysis 54
2.2.13 ATAC-seq 55
2.2.13.1 Generation of consensus peaks 55
2.2.13.2 Visualisation on the genome browser and normalisation 55
2.2.14 mNET-seq visualisation and normalisation 55
2.2.15 Differential expression analysis 55
2.2.15.1 scaRNA-seq 55
2.2.15.2 Intronic RNA-seq 56
2.2.15.3 polyA+ RNA-seq 56
2.2.15.4 ATAC-seq 56
2.2.16 Pearson correlation analysis of replicates (ChIP-seq, ATAC-seq) 57
2.2.17 Capture C 57
2.2.18 4sU-seq - Expression calculation (RPKM) 57
2.2.19 Scatter plots and hexbin plots 57
Chapter 3 – Quantitating promoter proximal transcription 59
3.1 Introduction 59
3.1.1 Existing methods to identify transcription initiation genome wide 59
3.1.2 Our changing view of promoter proximal transcriptional regulation 64
3.2 Results 65
3.2.1 scaRNA-seq resolves initiation and pausing simultaneously at the single genes 65
3.2.2 Bioinformatic calling and validation of tissue-specific TSS 69
3.2.3 Average views of transcription using observed TSS differ from annotated TSS 74
3.2.4 Comparison of scaRNA-seq to GRO-cap/GRO-seq 80

3.2.5 Comparison of scaRNA-seq RNA end distribution to mNET-seq and GRO-seq 85
3.2.6 The relationship between Pol2 pausing factors and the +1 nucleosome 90
3.3 Discussion 92
3.3.1 scaRNA-seq a combinatorial method to investigate promoter proximal transcription 92
3.3.2 Implications for our understanding of promoter proximal transcriptional regulation 93
Chapter 4 – The prevalence of Pol2 pausing *in vivo* 96
4.1 Introduction 96
4.1.1 Determining the extent of promoter proximal Pol2 pausing 96
4.2 Results 97
4.2.1 The travelling ratio 97
4.2.2 The relationship between TR and gene expression 98
4.2.3 Does the TR correlate better with nascent transcription? 99
4.2.4 Total Pol2 versus gene expression 103
4.2.5 Resolution limits of ChIP-seq for studying transcriptional kinetics 105
4.2.6 Using scaRNA to analyse initiation and pausing 106
4.2.7 A single gene analysis of the effects of DRB on scaRNA production 110
4.2.8 The effects of DRB on Intronic RNA at the metagene level 111
4.2.9 The effects of DRB on Intronic RNA at the single gene level 112
4.2.10 Delineating response to DRB at a single gene level 113
4.2.11 GRO-seq reveals promoter proximal and distal effects of DRB 116
4.2.12 GRO-seq reveals a DRB induced block in elongation at +500bp relative to the TSS 119
4.2.13 Positions of 1st splice sites correlate with DRB induced loss of nascent transcription 120
4.2.14 Decreases in GRO-seq signal upon DRB treatment correlate with nucleosomal barriers to elongation 122
4.2.15 Determining the extent of promoter proximal pausing of Pol2 *in vivo* 127
4.3 Discussion 129
4.3.1 Measuring Pol2 pausing *in vivo* 129
4.3.2 Implications for the field 130
Chapter 5 – The role of enhancers in regulating the transcription cycle 133
5.1 Introduction 133
5.1.1 The role of enhancers in pausing 133
5.1.2 Control of transcriptional elongation at alpha and beta globin 133
5.2 Results 135
5.2.1 RNA-seq reveals targeted nature of alpha globin enhancer deletion 135
5.2.2 Analysis of promoter proximal chromatin accessibility in enhancer knockouts 136
5.2.3 Analysis of Pol2 levels in the alpha and beta globin enhancer knockouts 142
5.2.4 Quantification of changes in Pol2 using TR 145
5.2.5 The major enhancers of alpha and beta globin direct transcription initiation 146
5.2.6 The beta globin chromatin compartment wildtype versus bHom 151
5.3 Discussion 152
Chapter 6 – Discussion 158
6.1 Quantitating promoter proximal transcription 158

6.2 The prevalence of Pol2 pausing *in vivo* 162
6.3 The role of enhancers in regulating the transcription cycle 166
6.4 Summary of conclusions 170
Chapter 7 – References 173

List of figures

Introduction

Figure 1.1) The alpha and beta globin loci in mouse....17
Figure 1.2) A simplified overview of the GRO-seq method. 21
Figure 1.3) A simplified overview of the mNET-seq method. 22
Figure 1.4) A schematic of the Start-seq method 24
Figure 1.5) An overview of the proposed mechanism of Pol2 pausing and pause release....27

Chapter 3 – Quantitating promoter proximal transcription

Figure 3.1) A schematic of the CAGE method. 60
Figure 3.2) A schematic of the GRO-cap method. 63
Figure 3.3) MNase-seq read coverage in K562 cells. 66
Figure 3.4) A schematic representing the key stages in scaRNA-seq....67
Figure 3.5) *HSPA1A* and *MYC* promoter proximal regions 68
Figure 3.6) Initiator motif enrichment 71
Figure 3.7) Observed versus annotated TSS position (all genes) 72
Figure 3.8) Observed versus annotated TSS position (unique genes). 73
Figure 3.9) Example of TSS skew....74
Figure 3.10) scaRNA-seq metaplot – observed and annotated TSS 78
Figure 3.11) Nucleosome occupancy (MNase) versus scaRNA 3' ends....79
Figure 3.12) GRO-cap and GRO-seq metaplot – observed and annotated TSS 82
Figure 3.13) Nucleosome occupancy (MNase) and GRO-seq metaplot 85
Figure 3.14) Metaplots of mNET-seq, GRO-seq and scaRNA-seq 3' RNA ends 86
Figure 3.15) A simplified overview of the ChIP-exo method 87
Figure 3.16) Pol2 ChIP-exo relative to observed and annotated TSS....89
Figure 3.17) Metaplots of ChIP-seq for factors thought to regulate Pol2 90

Chapter 4 – The prevalence of Pol2 pausing *in vivo*

Figure 4.1) Pol2 TR values for all UCSC genes in K562 cells 98
Figure 4.2) TR versus RNA expression for all unique annotated UCSC genes....99
Figure 4.3) RNA-seq read gene coverage metaplots. 100
Figure 4.4) Comparison of RNA-seq methods. 101
Figure 4.5) TR versus polyA+ RNA (all reads) 102
Figure 4.6) Examples of TR at different genes 103
Figure 4.7) Expression versus Pol2 density 104
Figure 4.8) Pol2 and scaRNA over observed TSS 106
Figure 4.9) scaRNA-seq metaplots – observed and annotated TSS....109
Figure 4.10) scaRNA-seq read density ratio (RDR). RDR values are for treatment versus control. 111
Figure 4.11) The effects of DRB on polyA+ (all reads) and polyA+ (intronic reads only). 112
Figure 4.12) Intronic read density ratio (RDR). RDR values are for treatment versus control 113
Figure 4.13) scatter plot of scaRNA versus intronic RNA (RDR's). Each gene is one yellow dot. 114
Figure 4.14) Effects of DRB on the *Myc* gene. 115
Figure 4.15) Effects of DRB on GRO-seq over annotated and observed TSS 119
Figure 4.16) GRO-seq full RNA fragments plotted relative to annotated TSS 120
Figure 4.17) UCSC genes 1st splice site versus TSS position 121
Figure 4.18) open chromatin in HEK293 cells.DNase-seq over all annotated TSS....122
Figure 4.19) GRO-seq (+/- DRB) and open chromatin....123
Figure 4.20) As in figure 4.19 except Y axis for GRO-seq data scaled for visualisation....124
Figure 4.21) As in 4.19 and 4.20 but with annotated TSS coordinates....126
Figure 4.22) Scatter plot of logFC intronic RNA versus logFC for observed "active" genes. 128

Chapter 5 – The roles of enhancer in regulating the transcription cycle

Figure 5.1) Differential gene expression analysis of wildtype versus aHom 136
Figure 5.2) Pearson correlation (R^2) analysis of ATAC-seq data....137

Figure 5.3) Open chromatin profile of the alpha and beta globin loci 138
Figure 5.4) Open chromatin peaks at the alpha and beta globin genes 139
Figure 5.5) Differential accessibility analysis of open chromatin peaks 141
Figure 5.6) Pearson correlation (R^2) of Pol2 ChIP-seq data 141
Figure 5.7) Pol2 ChIP-seq at the i) alpha and ii) beta globin loci 144
Figure 5.8) TR values in wildtype and enhancer knockouts 145
Figure 5.9) Capture C in bHom versus wildtype 152

Table of abbreviations

3’	Three prime
5’	Five prime
aHom	Alpha globin homozygous R1, R2 enhancer deletion
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
ATP	Adenosine triphosphate
Bam	Binary alignment map
bHom	Beta globin homozygous R2 (HS-2), R3 (HS-3) enhancer deletion
bp(s)	Base pair(s)
BRD4	Bromodomain-containing-protein 4
BRep	Biological replicate
CAGE	Capped analysis of gene expression
CCNT2	Cyclin T2
ChIP-exo	ChIP exonuclease
ChIP-nexus	chromatin immunoprecipitation experiments with nucleotide resolution through exonuclease, unique barcode and single ligation
ChIP(-seq)	Chromatin immunoprecipitation (sequencing)
CIP	Calf intestinal phosphatase
DEPC	Diethyl pyrocarbonate
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DNase(-seq)	Deoxyribonuclease (sequencing)
DRB	5,6-Dichloro-1-beta-D-ribofuranosylbenzimidazole
DSIF	DRB sensitivity inducing factor
EDTA	Ethylenediaminetetraacetic acid
EGTA	ethylene glycol-bis(β-aminoethyl ether)-*N*,*N*,*N*′,*N*′-tetraacetic acid
Epo	Erythropoietin
eRNA	Enhancer RNA
EST	Expressed sequence tags
FACS	Flow assisted cell sorting
FBS	Foetal bovine serum
FISH	Fluorescent in situ hybridisation
FP	Flavopiridol
GRO-seq	Global run on sequencing
GWAS	Genome wide association studies
HEK293	Human embryonic kidney
HeLa	Henrietta Lacks
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
JMJD6	Jumonji C-domain-containing protein 6
Kb	Kilobase
m7G	7-Methylguanosine
mES	Mouse embryonic stem cells

MNase(-seq)	Micrococcal nuclease (sequencing)
mNET-seq	Mammalian native elongating transcript sequencing
NaAc	Sodium acetate
NELF	Negative elongation factor
NET-seq	Native elongating transcript sequencing
nt(s)	Nucleotide(s)
NTP	Nucleoside tri phosphate
P-TEFb	Positive transcription elongation factor beta
PAF1	Polymerase associated factor
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PI	Pausing index
Pol2 pausing	RNA polymerase 2 pausing
polyA(+/-)	Poly adenylated (plus/minus)
PPEP	Promotor proximal enrichment of polymerase
PRO-seq	Precision run on sequencing
qPCR	Quantitative PCR
R	Regulatory element
RCF	Relative centrifugal force
RIN	RNA integrity number
RIPA	Radioimmunoprecipitation assay buffer
RNA	Ribonucleic acid
RPKM	Reads per kilobase per million reads
rpm	Revolutions per minute
RPMI	Roswell Park Memorial Institute
RT	Room temperature
RT-qPCR	Reverse transcription quantitative PCR
S2	Schneider 2
Sam	Sequence alignment map
scaRNA(-seq)	Short capped RNA (sequencing)
SCF	Stell cell factor
SDS	Sodium dodecyl sulphate
SNP	Single nucleotide polymorphism
SUPT5H	Suppressor of tyrosine 5 homolog human
T.E	Tris EDTA
TBE	Tris borate edta
TES	Transcription end site
TR	Travelling ratio
TRep	Technical replicate
UCSC	University of California Santa Cruz
WT	Wildtype

Chapter 1

Introduction

Chapter 1 - Introduction

Spatiotemporal control of gene expression enables the development of complex multicellular organisms, containing hundreds of distinct cell types with a wide variety of functions. There are tens of thousands of enhancers in the genome which direct development and differentiation by upregulating gene expression, yet how individual enhancers alter the output of their target genes remains unknown. Therefore, a major current goal of biology is to understand, mechanistically, how regulatory elements modulate transcriptional kinetics and give rise to the large dynamic ranges of RNA expression underlying specific cellular functions.

1.1 Transcriptional rate alters RNA output

The process of transcription is finely controlled to achieve correct spatiotemporal expression patterns of thousands of genes during the development and lifespan of complex organisms. The rate of transcription is a key factor in determining the final levels of transcripts available for translation into proteins[1]. It has been shown through several decades of crystallographic and genetic studies that the transcription start sites of genes and enhancers are decorated with a plethora of protein factors which coordinate to activate gene expression. Core promoter sequences, present near the 5' end of genes, are thought to be hubs for the binding of protein transcription factors (TFs), which are thought to help drive the formation of the pre-initiation complexes that recruits RNA polymerase II (Pol2), allowing the initiation of transcription[2,3]. Therefore, the simplest way to alter transcriptional rate would appear to be through more efficient recruitment of transcription factors which would increase the rates of transcriptional initiation. This is known as "transcriptional activation by recruitment"[4]. However, it has been known for nearly 30 years that eukaryotic transcription can be affected by *cis*-acting regulatory elements located

at long distances from a gene on the linear chromosome, adding an additional potential layer of regulation to gene expression in multicellular organisms.

The best understood of these *cis*-acting elements is the enhancer[5,6]. Enhancers bind transcription factors and components of the pre-initiation complex and are often physically juxtaposed with target promoters through the three dimensional conformation of the genome[3,6]. This has been shown in high resolution for many genomic loci across metazoans using chromatin confirmation capture (3C) techniques[7–10]. The physical looping of enhancers to promoters has led to the suggestion that enhancers may function to increase the local concentration of activating proteins and transcriptional machinery at a given promoter, thereby increasing the transcriptional rate[11]. However, at present, which proteins are involved and what aspects of transcription they might control is not well understood. It has been proposed that the multiprotein complex "Mediator" may act as a "bridge" between enhancers and promoters, recognising enhancer-bound TFs and core promoters at target genes. This is done by physically juxtaposing these elements and leading to more efficient formation of preinitiation complexes, or by activating Pol2 for transcription elongation through phosphorylation of its carboxy terminal domain[2,12,13]

Work using high throughput *in vitro* reporter assays and targeted genetic manipulation of enhancers *in vivo* has proven that transcriptional output of a given promoter is affected by enhancer elements. Classic examples such as the alpha and beta globin genes are driven by the combined effects of multiple enhancer elements to achieve tissue-specific gene expression[7,14] (Figure 1.1). These two gene loci have served as paradigms for gene regulation for several decades, providing numerous general insights into this process including the existence of distal regulatory elements. Both the alpha globin and beta globin enhancers have been show to act additively with regards to gene expression; however,

transcription of beta but not alpha globin has been reported to be regulated predominantly through Pol2 pause release[7,14]. Therefore a parallel investigation into the mechanism of action of these two sets of enhancers may enable a deeper understanding of how transcription outputs are "fine-tuned" through distinct stages in transcription and how this process can vary between genes[7].

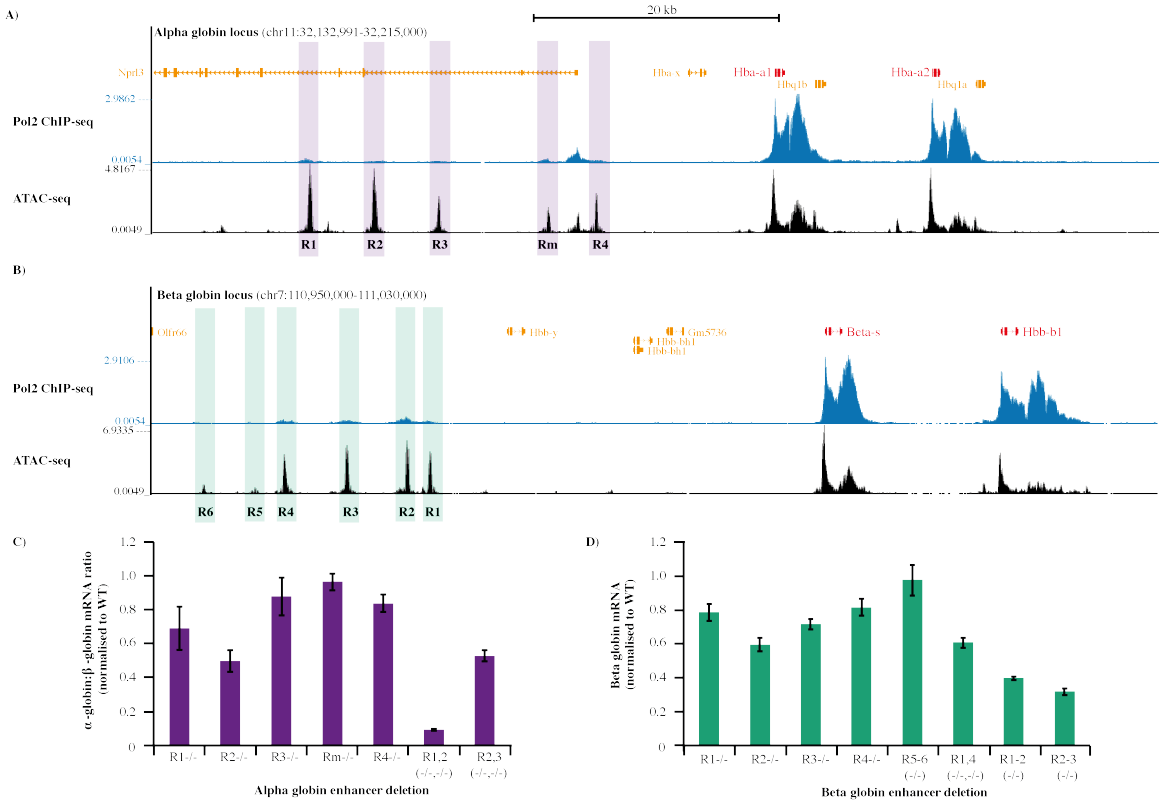


Figure 1.1) The alpha and beta globin loci in mouse. All data is from primary mouse foetal liver tissues in "intermediate" erythropoiesis (see methods for definitions). A) Alpha globin loci, with the two copies of alpha globin highlighted red and regulatory enhancer elements (R1-R4) highlighted as purple boxes. Traces show Pol2 ChIP-seq on the top row and ATAC-seq on the bottom row. B) Beta globin loci, with the two copies of beta globin gene highlighted red and regulatory enhancer elements (R6-R1) highlighted as green boxes. Traces show Pol2 ChIP-seq on the top row and ATAC-seq on the bottom row. C) Modified from Hay et al 2016[7]; changes in nascent RNA level upon individual and combinatorial enhancer deletions of alpha globin. Values are ratios of alpha globin versus beta globin expression normalised to wildtype expression of alpha globin. D) Changes in mRNA level upon individual and combinatorial enhancer deletions of beta globin adapted from Bender et al 2012[14]. Values are beta globin expression versus wildtype beta globin expression. Error bars represent standard error from a minimum of three replicates.

Mutations in enhancer elements which alter transcription factor binding are believed to be a primary mechanism underlying complex disease risk, as single nucleotide polymorphisms (SNP's) identified in genome wide association studies (GWAS) are often clustered in putative enhancers[15]. There are numerous examples of mutations in functionally validated enhancers which cause disease[15–18]. What remains elusive however is a firm mechanistic understanding of how enhancers normally drive transcription at their target promoters. In order to understand this, we must first consider the stages in transcription which can be regulated before evaluating the effect that enhancers may have on them.

1.2 Methods to identify transcriptional kinetics

1.2.1 Indirect

In order to understand the kinetics of transcript production in cells, the field has turned to two broad categories of techniques. The first of these does not focus on analysing the RNA product of the cells but rather the protein complexes which transcribe them, including Pol2. The assumption in this style of work is that the levels of particular Pol2 core subunits across genes provide a proxy for transcriptional activity and therefore gene expression.

The general method used to detect levels of Pol2 in the context of chromatin is known as chromatin immunoprecipitation (ChIP) and works by fixing proteins to DNA, sonicating the fixed chromatin, retrieving bound fragments by immunoprecipitation with an antibody specific for the protein of interest, followed by elution and qPCR or sequencing of the bound DNA fragments[19].

One of the earliest applications of ChIP to Pol2 for the purposes of studying transcriptional changes was at the beta globin locus in chickens, where it was revealed that Pol2

distribution across the body of the beta globin gene was specific to different stages of differentiation. Pol2 was found to be spread across the gene in nuclei isolated from mature erythrocytes but localised to the 5' end of genes in nuclei from immature cells[20]. In Drosophila S2 cells Pol2 ChIP was also used to demonstrate that Pol2 is predominantly localised to the 5' end of the un-induced heat shock protein 70 (*Hsp70*) gene but becomes uniformly distributed at high levels across the gene body after heat shock[21]. These studies demonstrated that transcriptional activity may not necessary correlate with total Pol2 levels as both beta globin and *Hsp70* could be expressed at low levels while having high levels of Pol2 engaged with DNA near their promoters. In the last decade, Pol2 ChIP-seq has suggested that Pol2 is not evenly distributed across genes, with a higher Pol2 density in promoter proximal regions when compared to downstream gene bodies[22]. This suggests that the rate of Pol2 elongation may be regulated at points beyond transcription initiation[23,24].

ChIP has evolved over the years to analyse an increasing number of DNA fragments in single experiments, ultimately resulting in ChIP-seq. This method can identify binding sites for a protein of interest genome wide[25]. However, despite providing interesting insights into the localisation of Pol2 across genes, it is not clear what a particular Pol2 profile represents in terms of transcriptional activity of the protein at a given gene. In addition, traditional ChIP-seq protocols require high numbers of cells (0.5-5 x10^7) per experiment, often preventing the use of primary tissues in analysis. However, numerous low-input versions of the ChIP-seq protocol have recently been developed[26].

1.2.2 Direct

Remarkably, one direct method for analysing transcription has dominated this field for nearly 40 years: the nuclear run on assay[27]. In essence, this method is a way of *in vitro*

labelling engaged Pol2 and involves isolation of nuclei under conditions causing transcriptional arrest due to loss of free nucleotides. This is then followed by resumption of transcription by Pol2 molecules which are engaged with the DNA template and have nascent RNA still bound in their active site. This is achieved by the addition of simple reaction buffer containing a mixture of labelled and unlabelled NTPs. Nascent RNA produced in this reaction can be identified via its label and hybridisation to gene specific probes[27].

The refinement and adaptation of the nuclear run on assay led to the development of the genomic run on assay, capable of surveying transcription elongation rates genome-wide in yeast using microarrays[28]. This ultimately led to the era of global run on sequencing (GRO-seq), developed by the Lis laboratory, which replaced radiolabelled nucleotides with a mixture of non-labelled and the biotinylated BrU nucleotide, allowing massive enrichment of the nascent RNA fraction followed by sequencing (Figure 1.2)[29]. A further development of this method replaced the mixture of NTPs with four biotinylated nucleotides, limiting run on reactions to a few base pairs and allowing precision run on sequencing (PRO-seq)[30], yielding the first method which could profile positions of Pol2 accumulation genome-wide at base pair resolution.

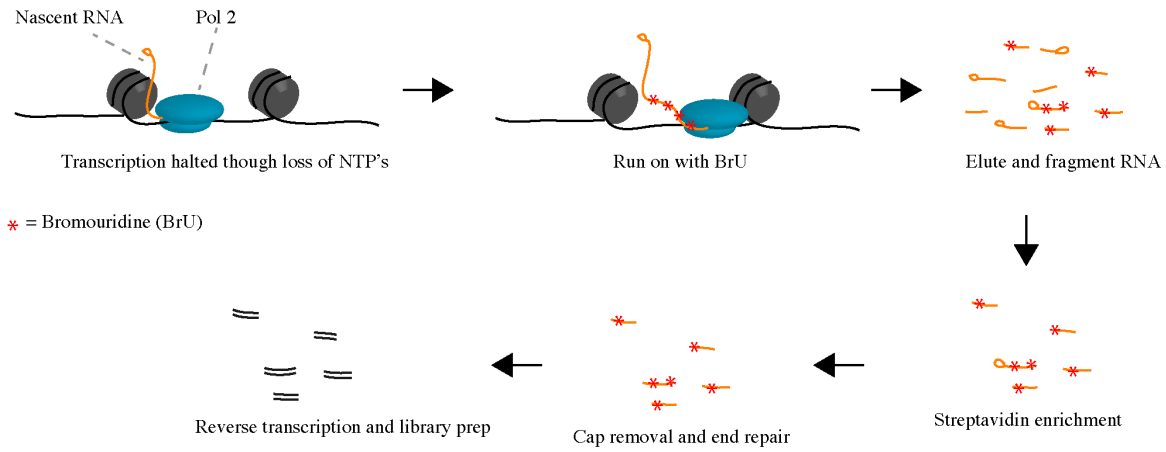


Figure 1.2) A simplified overview of the GRO-seq method. Pol2 is halted in the process of transcription due to loss of nucleotides; the addition of a simple buffer containing 5-bromouridine 5′-triphosphate (BrU) allows resumption of transcription for a short run <100nt of bases before BrU incorporation inhibits elongation. RNA is extracted and subjected to biotinylation, followed by streptavidin pulldown. Labelled RNA is fragmented, decapped and made into a sequencing library.

Although widely used, these run on methods are non-physiological in nature and rely on major assumptions: i) Pol2 must be transcriptionally engaged with the template and nascent RNA 3' end, allowing for the resumption of elongation; ii) the isolation of nuclei or permeabilisation of cells required to perform the assay does not perturb the cell in a way that may affect resumption of transcription in a manner consistent with *in vivo*, for example that the concentrations of nucleotides added to allow transcription to resume are physiological; iii) the addition of a biotinylated nucleotide to the RNA molecule does not impact its abundance in the cell for example through artificial stabilisation of unstable RNA. Although efforts are made to control for some of these factors, the GRO-seq and PRO-seq assays remain technically challenging and it is debatable how closely nascent RNA generated by this method represents *in vivo* transcription.

Due to the caveats of GRO-seq/PRO-seq, there has been a recent development of alternative methods to study Pol2 elongation in a more physiological way. One such method is native elongating transcript sequencing (NET-seq), first developed and applied to yeast by the Churchman and Weissman laboratory[31] and later refined to allow specific pulldown of chromatin bound by Pol2 through immunoprecipitation, using specific antibodies by the Proudfoot laboratory as Mammalian NETseq (mNET-seq)[33]. This method is unique in its ability to specifically isolate Pol2 associated RNA and uses micrococcal nuclease (MNase) digestion to degrade RNA and DNA not protected by the Pol2 molecule itself. Elution, followed by sequencing of the protected RNA from its 3' end, allows the identification of the position occupied by the Pol2 active site at base pair resolution and

does not require extensive manipulation of cells or nuclei (Figure 1.3). This method was initially limited by the requirement of high cell number inputs (>10^8) but data presented in this thesis suggests the assay produces interpretable single gene profiles of Pol2 positions from modest cell inputs ($5x10^6$), meaning it is applicable to a wide variety of cell types, including primary cells.

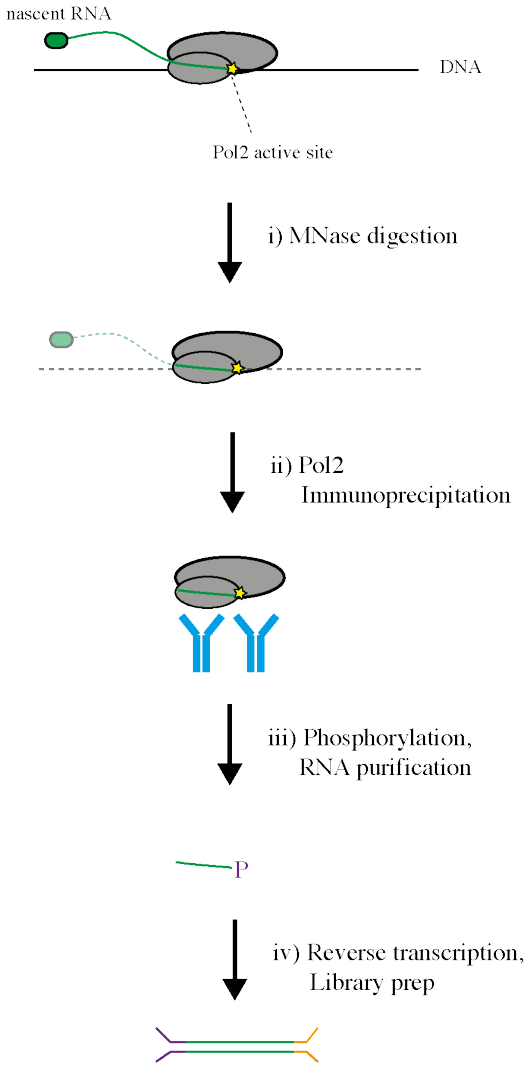


Figure 1.3) A simplified overview of the mNET-seq method. Pol2 engaged with DNA and with RNA. i) MNase digestion degrades RNA and DNA not protected by protein. ii) Pol2 associated RNA, DNA complexes are isolated by immunoprecipitation. iii) RNA is extracted and phosphorylated to allow directional adapter ligation. cDNA is produced and made into sequencing libraries.

An alternative strategy which shows great promise in elucidating the promoter proximal regulation of transcription has been Start-seq[34,35]. Short capped RNA's (scaRNA) which are produced by Pol2 in early elongation can be sequenced from their 3' ends, allowing the promoter proximal positions of Pol2 accumulation to be resolved[35] (Figure 1.4). scaRNAs have also been shown to stand proxy for Pol2 which accumulates near the 5' end of genes

and is assumed to be transiently paused. In addition, they are believed to be rapidly degraded by the exosome upon release from chromatin, meaning they provide an excellent marker for nascent and active promoter proximal transcription[36].

However, there are limitations to this method: previously, only very short <=100nt RNAs have been isolated. Given there is evidence that the +1 nucleosome may provide a barrier to elongation[31,37–39] and the fact that this feature lies at ~190bps downstream of the transcription start site (TSS) [39], this would preclude analysis of RNA distributions around the +1 nucleosome. Another limitation is the use of single end sequencing, resulting in 5' or 3' ends of RNA being sequenced separately and thus losing any single molecule power to interrogate the ends of individual RNA molecules simultaneously. One assay which could quantitatively measure 5' and 3' ends of RNA in one experiment would enable the analysis of transcription initiation and pausing at the same time genome-wide, providing a powerful tool for relating changes in gene expression to changes in the Pol2 pausing.

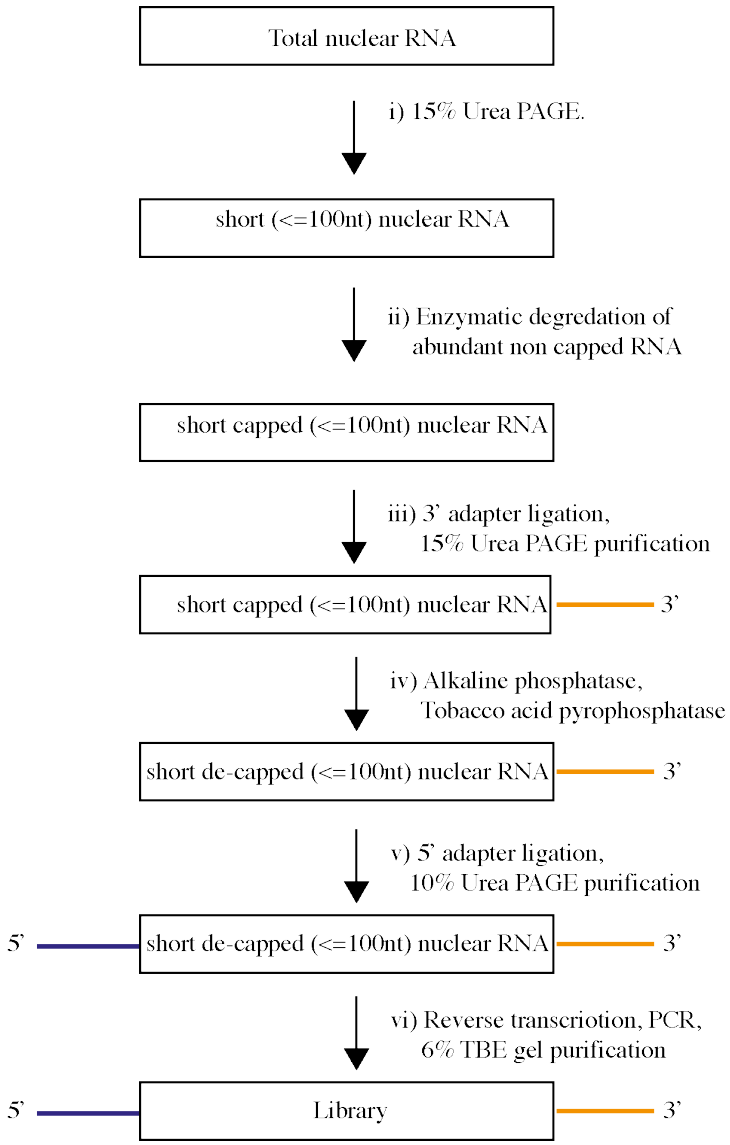


Figure 1.4) A schematic of the Start-seq method modified from Nechaev et al 2010[35]

1.3 The transcription cycle and the role of enhancers in regulating it

Based on our current understanding, transcription at a given gene can be regulated at three broad stages: initiation (opening of promoter regions and recruitment of Pol2), elongation (Pol2 pausing and productive elongation beyond pausing) and termination. Whether any individual stage has a more general impact on the final levels of a transcript is unknown but it has been postulated that initiation and Pol2 pausing are highly regulated[4,23,24]. These interconnected stages have been referred to as the transcription cycle due to the assumption that changes in any one stage will impact the others[40].

1.3.1 Initiation

This stage in transcription is arguably the best understood. High resolution crystal structures of many components which are thought to comprise the core machinery responsible for initiation of transcription have been solved, allowing detailed mechanistic insights to emerge from this process[2]. Elaborate structures of Pol2 in the process of initiation and early elongation have recently been produced by the Cramer laboratory[41,42], yielding unprecedented details of this process. In brief, transcription factors bind to a promoter region resulting in the recruitment and formation of the pre-initiation complex (PIC). Pol2 is recruited to the PIC and nascent RNA chain synthesis begins from the transcription initiation site. Initiation factors dissociate from Pol2 allowing formation of the elongation complex. This process has been expertly reviewed in Sainsbury et al[2]. The stoichiometry between promoter, chromatin and associated transcription factors is believed to be a key determining factor in driving gene changes in gene expression *in vivo*. The expression or mis-expression of a few transcription factors can entirely alter cell fate[43].

Initiation of transcription has been suggested to be a major point of regulation with recruitment of transcription factors and Pol2 to promoters a potential bottle neck in transcript synthesis. This has been supported by experiments such as artificial recruitment of activator proteins to promoters which have been shown to be sufficient to drive activation of gene expression[44,45].

Enhancer elements have been widely implicated in driving initiation as it has been shown that often enhancer elements are physically juxtaposed in three dimensional space to the promoters they regulate, providing a mechanism to increase the local concentration of activating transcription factors at promoters[7,8].

1.3.2 Pol2 pausing

The idea that the rate of Pol2 elongation could be regulated is not new, originally emerging from investigations into the mechanism of action of an ATP analogue (5,6-Dichloro-1-beta-D-ribofuranosylbenzimidazole (DRB)) on transcription by Darnell and colleagues. They showed that synthesis of heterogenous nuclear RNA was reduced by 95% when adenovirus-infected HeLa cells were treated with DRB[46]. They further showed that this reduction in total RNA levels was not due to a defect in initiation but rather involved steps following initiation since shorter, truncated adenoviral transcripts were produced upon DRB treatment[27].

After pioneering ChIP-PCR studies Drosophila, showing clustering of Pol2 near the transcription start sites of genes despite a lack of transcript production, it was hypothesised that if a cellular mechanism to enforce or relive this block existed it could be utilised to fine tune the transcriptional rate[20].

A series of intricate biochemical experiments using *in vitro* transcription systems were key in identifying a DRB-sensitive factor (P-TEFb) which allowed production of longer RNA transcripts[47]. P-TEFb was proposed to be one target of DRB as it possesses a kinase component, CDK9, which requires ATP; DRB as an ATP analogue was suggested to compete for ATP binding. In the few years following this discovery, the DRB sensitivity inducing factor (DSIF) and negative elongation factor (NELF)[48,49] were purified and shown to be required to establish efficient pausing and pause release *in vitro*. These factors were shown to require phosphorylation by CDK9 to allow release of Pol2 pausing and were therefore indirect targets of DRB[3].

These data resulted in a proposed mechanism for the promoter proximal pausing of Pol 2, briefly summarised into three stages (Figure 1.5):

i) Pol2 binds to open promoter regions and begins RNA synthesis.

ii) Pol2 pauses just downstream of the TSS (+20-60bp) due to the coordinated action of negative elongation factors (DSIF and NELF).

iii) The release of Pol2 pausing is achieved through phosphorylation of DSIF, NELF and the Pol2 CTD by the positive transcription elongation factor (P-TEFb via CDK9). Pol2 thus becomes elongation competent and continues transcription across genes[47–49].

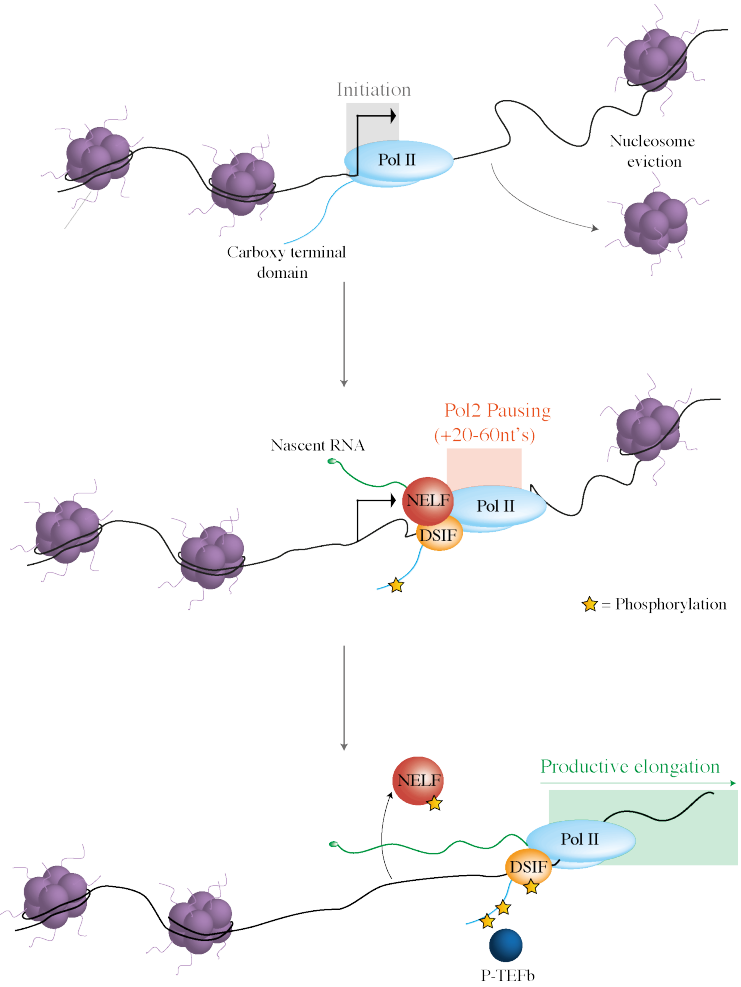


Figure 1.5) An overview of the proposed mechanism of Pol2 pausing and pause release based on *in vitro* studies[47–49].

More recent work utilising ChIP-seq to map protein locations *in vivo* has shown that DSIF, NELF and Pol2 commonly co-localise on promoters and also that promoter proximal enrichment of Pol2 occurs on the majority of genes. These findings hint that the *in vitro*

described mechanism of Pol2 pausing may exist at most Pol2 bound promoters *in vivo*[22]. It is important to note, however, that ChIP-based methodologies are inherently limited in their ability to study transcriptional kinetics, firstly because fixation reactions occur over several minutes, reducing temporal resolution as Pol2 molecules may move on the DNA template during this time window thus blurring the signal. Secondly, the resolution of ChIP-seq is limited by the size of DNA fragments that are immunoprecipitated[17,50]. Given that Pol2 pausing is thought to occur within 100bps from the TSS [24], resolving populations of Pol2 at closely connected stages such as initiation and Pol2 pausing is extremely challenging. Nevertheless, plotting Pol2 ChIP-seq data around annotated TSS has suggested that Pol2 accumulates just downstream of the TSS at mammalian promoters across much of the genome[22].

The use of an alternative inhibitor, Flavopiridol (FP) (which has the same mode of action as DRB[51]) in combination with GRO-seq, has suggested that a block to elongation can be induced near to the promoter of nearly all Pol2 transcribed genes in mouse embryonic stem cells (mES). This observation has strengthened the suggestion that Pol2 pausing may represent an obligate rate-limiting step in transcription at all protein coding genes[12].

However, at present there is limited high resolution data probing how levels of pausing change for a given change in gene expression. In addition, point mutations in DSIF and NELF give a variety of tissue-specific non-lethal phenotypes[52], whilst reduction in levels of these proteins has been shown to cause relatively modest changes in Pol2 distribution across genes[22] and gene expression[53], arguing against a key role for these proteins in regulating the overall levels of transcription[52].

There is some evidence implicating distal regulatory elements in the establishment and release of paused Pol2 at promoter regions. In Drosophila, 3C which profiles the 3-D

organisation of chromatin in the nucleus, has been performed in combination with Pol2 ChIP-seq at "early" and "mid" embryogenesis stages. This demonstrated that enhancer promoter loops are preformed early in development with an enrichment of Pol2 at the enhancer of differentially expressed but stably looped genes. It was suggested that an initial transcription factor-binding event could induce loop formation, recruitment of Pol2 at the enhancer and pausing. Subsequent transcription factor binding or looping to additional transcription factor bound enhancers could enable the release of paused Pol2[10].

In an opposing mechanism, enhancers co-bound by bromodomain-containing-protein 4 (BRD4) and Jumonji C-domain-containing protein 6 (JMJD6) have been shown to function as "anti-pause" enhancers. ChIP-seq experiments in human (HEK293T) cells show strong co-localisation of JMJD6 and BRD4 at intergenic regions bearing classical enhancer marks such as H4Ac and H3Ac. Both proteins were shown to be capable of activating P-TEFb by releasing it from the 7SK inhibitory complex. 3C revealed looping between these co-bound enhancers and coding genes, providing insights into a mechanism of enhancer directed pause release[54].

Finally, and most recently, a mechanism whereby recruitment of PAF1 to enhancer regions causes increases in pause release and upregulation of gene expression has been proposed by the Shilatifard laboratory based on the idea that PAF1 may play regulatory roles in Pol2 pause release[55,56].

There are several considerations to be taken into account when reviewing the published literature on the putative role of enhancers in regulating pausing. Firstly, one must be sure that the enhancer being tested is functional in the cell type being analysed and has a measurable contribution to nascent transcription at its target gene. Secondly, the experiments used to assay promoter proximal transcription must have a sufficiently high

level of resolution to analyse individual stages in early transcription, i.e. initiation and Pol2 pausing; this presents problems for many published analyses. Another consideration is the effect of the enhancer deletion on global transcription. Removal of a particular enhancer(s) must have minimal effects on the global levels of RNA, to ensure that any change in transcription at the target gene is not indirect (due to perturbed expression of other transcriptional regulators or due to a developmental defect). Thus a study of a true functional enhancer in its correct active tissue context has yet to be performed at high resolution.

1.3.3 Splicing

In the same year that Darnell and colleagues demonstrated DRB-induced blocks to elongation, splicing of RNA transcripts was identified[57,58]. As splicing occurs co-transcriptionally, this suggests the possibility that the rate of splicing affects the rate of transcription and so acts as the regulatory step. Recent work has suggested that Pol2 may pause over splice sites, providing the potential for regulation, although active manipulation of this process *in vivo* to alter transcript levels has not yet been demonstrated[32,33].

Live cell imaging of splicing of introns in the beta globin gene in HEK293 cells revealed that splicing takes on average 20-30 seconds, but it is not known how generally true these measurements are to other genes or whether the rate of splicing of a given gene differs between tissues, suggesting some form of regulation[59]. Interestingly, in an experiment following activation of the beta globin gene in the inducible G1E erythroid cell line, some of the enhancers of beta globin were found in close proximity with the 5' end of the gene throughout a time course whereas other enhancers were localised in different positions along the gene body during transcription suggesting they "track" with Pol2. However, the exact mechanisms through which this may be achieved is unclear and due to the population

nature of the study it is unclear whether this phenomena occurs on individual genes in individual cells[60].

Splicing may impact RNA levels through RNA stability a notable example of this is the differing stabilities of protein coding and non-coding transcripts such as enhancer RNA's (eRNA). eRNA's are thought to be unstable and short lived, whereas protein coding RNA's are more stable; it has been shown that protein coding genes are enriched for 5' splice sites when compared to non-coding transcripts. As the 5' splice site is recognised and co-transcriptionally bound by the spliceosome, it has been suggested this association could prevent transcript degradation by the exosome, thus increasing transcript levels[61,62].

Although evidence is limited, it is feasible that enhancer elements could direct splicing of genes into alterative isoforms, or increase the rate of splicing by providing higher levels of spliceosome machinery to a given gene. In fact, the first enhancer element identified (the SV40 transcriptional enhancer) has been shown to promote alternative splicing when inserted into mini gene constructs expressed *in vivo*[63].

1.3.4 Termination

It has recently been proposed that the rate of transcription termination may be regulated in a manner similar to promoter proximal Pol2, as the use of DRB in Hela cells caused premature transcription termination[64]. However, it is not known how to consider these findings since, if DRB causes inhibition of promoter proximal pause release, then the kinetics of downstream stages in the transcription cycle could be uninterpretable or at least perturbed to some degree.

It has also been shown in yeast that promoters and terminator regions can be physically juxtaposed, providing a potential mechanism for promoter proximal factors to exert effects

on termination[65]. Due to the ability of the chromatin fibre to form loops, it is entirely possible that enhancers as well as promoters may be in physical contact with termination regions, providing the potential for regulation.

In summary, the rate of transcription could potentially be altered at any stage in the transcription cycle and as yet there is no clear evidence as to which if any is the predominant point of regulation.

1.4 General features of transcription in higher eukaryotes

Over the past three decades, our view of the transcription of genes has dramatically changed. Once perceived to be a relatively simplistic process whereby activator proteins bind to a promoter, recruit components of the pre-initiation complex and Pol2 leading to nascent RNA synthesis[4], it is now thought that modulation of a given gene output is a highly regulated process. Several features of "general" transcription have been gleaned from meta-analysis of data around annotated TSS position.

1.4.1 Transcription initiation occurs in different modes

It is believed that transcription does not simply initiate at a precise location (as would be expected to ensure high fidelity copying of the template into RNA); rather, it occurs broadly in two different modes depending on the promoter in question. The first mode is the focused or "punctate" initiation, associated with motifs such as the TATA box and Initiator sequences, or sequences having poorly phased nucleosome density around the site of transcription initiation. The second mode is the dispersed or "broad" initiation. This type of promoter is often found at housekeeping genes with well positioned nucleosomes around their initiation sites and are associated with CpG islands[66,67].

1.4.2 Promoter proximal pausing of Pol2 may be ubiquitous

The connection between levels of Pol2 pausing and gene expression is a complex one with highly and lowly expressed genes exhibiting a characteristic accumulation of Pol2 at their promoters. There is no clear correlation between gene expression levels and with high levels of promoter proximal Pol2[22,23].

However, a few studies have suggested that Pol2 pausing not only occurs *in vivo* but that it represents an obligate regulatory step in the transcription of all protein coding genes in higher eukaryotes[12,23,24]. The primary evidence for this is twofold: 1) the *in vivo* co-localisation of DSIF, NELF and Pol2 to promoters which harbour high levels of promoter proximal Pol2 density, mirroring what was seen *in vitro*[22,47–49]; 2) the use of DRB and other inhibitors of P-TEFb results in relative promoter proximal accumulation of RNA transcripts when compared to the downstream gene body genome wide[12].

1.4.3 Antisense transcription

In 2008, four separate groups described widespread antisense transcription occurring upstream of annotated promoters in mammalian cells: 1) the Lis laboratory, with GRO-seq in IMR90 cells revealed antisense transcription proceeded effectively beyond the promoter region[29]; 2) the Sharp laboratory, which first identified antisense transcription genome-wide in mES cells[34]; 3) the Jensen laboratory detected widespread antisense transcription in HeLa cells in experiments where the exosome was depleted in HeLa cells; and, 4) the Kinzler laboratory developed a novel bisulphate sequencing based strategy to reveal abundant antisense RNA in primary peripheral blood mononuclear cells[68,69]. There is much debate about the potential role of this type of transcription and it remains to be determined whether this phenomenon is important for transcriptional regulation genome-wide or is merely a by-product of the process of transcription[70,71].

1.5 Outline of the thesis

In this thesis I seek to explore the relationship between transcriptional initiation, Pol2 pausing and the levels of RNA *in vivo*.

My first question asks how common is Pol2 pausing *in vivo*? As yet, the extent of Pol2 pausing has not been determined without the use of chemical inhibition through DRB or drugs such as Flavopiridol (FP), both of which have a very similar mechanism of action[51]. The assumption inherent to these previous studies is that DRB or FP specifically blocks pause release (as demonstrated in early *in vitro* work) without affecting other cellular processes. It has been demonstrated that DRB can inhibit transcription initiation through inhibiting another kinase, CDK7[51]; DRB (an ATP analogue) may well inhibit other cellular process, confusing more general effects with effects mediated through Pol2 pausing. In this work I aim to determine how frequently Pol2 pausing acts as a regulatory step in transcription *in vivo*. Experiments have been performed with and without the use of DRB to understand whether changes in the regulation of promoter proximal transcript production brings about differential gene expression.

Secondly, I aim to determine at which stage in the transcription cycle enhancers act. I have dissected both the alpha and beta globin loci and their associated dominant, tissue-specific regulatory elements to determine if removal of such enhancers causes reduction in the levels of transcription through increased Pol2 pausing. It is of interest that the beta globin gene has been suggested to accumulate paused Pol2 and the enhancer elements of this gene have been proposed to act through transcriptional elongation (presumably through release of Pol2 pausing)[14,20].

My work will add to our understanding of transcription in eukaryotes in general and also help to elucidate the molecular mechanism(s) through which enhancers regulate expression of their target genes.

Chapter 2

Materials and methods

Chapter 2 – Materials and Methods

2.1 Wet lab protocols

2.1.1 Growth of K562 cells, pharmacological treatment and harvesting

K562 cells were maintained in RPMI media supplemented by 15% FBS. Cell density was maintained below 10^6 cells/ml. Culture density and viability were determined using the nucleocounter platform (chemometec).

For experiments utilising DRB aliquots of $5x10^6$ cells were taken from culture, spun @ 500g for 5 mins and washed 1x in room temperature PBS. Pellets were resuspended in 5ml of warm media with either DRB (Sigma) at a concentration of 100μM or DMSO for 3 hours. Cells were spun @ 500g for 5 mins and washed 1x ice cold PBS. For scaRNA-seq and RNA-seq libraries cell pellets were lysed in TRI reagent (Sigma) according to the manufacturers protocol.

For mNET-seq cell pellets were snap frozen in a dry ice ethanol bath. All samples were stored at -80°C until use.

All K526 experiments were performed in isogenic duplicate.

2.1.2 Growth of mouse foetal livers and harvesting of erythroid cells

Mouse foetal livers from day 12.5 embryos were harvested and disaggregated in expansion media (StemPro (ThermoFisher)) supplemented with Epo (1U/ml), SCF (50ng/ml), Dexamethosone (1μM)). Cells from individual foetal livers were grown for 6 days, maintaining a constant erythroblast cell count of <= 1 $x10^6$. Cells were counted using the nucleocounter before selection for CD117 positive cells using magnetic assisted cell separation LS columns (130-042-401 - Milltenyi) following the manufacturers protocol. Cells were left to recover for 6h in expansion media @ $1x10^6$ cells/ml. Aliquots of $5x10^6$

cells were harvested as the '0h' timepoint corresponding to early erythropoiesis whilst the remaining cells were re-suspended in differentiation media (which is expansion media with the following changes: increased levels of Epo (5U/ml) and transferrin (0.5mg/ml), SCF and Dexamethasone removed). To induce erythroid differentiation cells were grown for 24h before harvest.

Cell morphology of the cells at 0h or 24h timepoints was examined by cytospins and immunophenotype determined by flow assisted cell sorting (FACS) (Attune NxT – Thermo Fisher Scientific) with markers for CD117 and Ter119. Early erythropoiesis was defined as (CD117+,Ter119-) and intermediate as (CD117-, Ter119+) and cells isolated were confirmed as having these appropriate markers[72,73].

Material was prepared for RNA-seq, scaRNA-seq and mNET-seq in an identical manner to K562 cells. For ChIP-seq, material was prepared by harvesting $5x10^6$ cells, pelleting (500g, 5 mins RT) before resuspension in 9ml of RPMI media (+10% FBS) and addition of 1ml fixation solution (50mM HEPES pH 8.0, 1mM EDTA pH 8.0, 0.5mM EGTA pH 8.0, 100mM NaCl, 10% formaldehyde). Samples were incubated on the roller at room temperature for 10 mins and fixation reactions quenched by the addition of 1.25ml of fresh glycine solution (1M). Samples were incubated on the roller for 5 mins before spinning down at (300g 5 mins 4°C), washing 1 x in 10mls of ice-cold PBS and snap freezing the pellet in a dry ice ethanol bath. All samples were stored at -80°C until use.

All experiments on mouse foetal liver were performed in biological triplicate.

2.1.3 RNA extraction and quality control

For polyA+/- RNA-seq, samples were extracted from TRIZOL and DNase treated using the Direct-Zol RNA Mini Prep Plus (Zymo) according to manufacturer's protocol except with an increased DNAse treatment time of 30 mins.

For scaRNA-seq samples were extracted from TRI Reagent using Phase Lock Gel heavy tubes (Quanta bio) according to the manufacturers protocol followed by ethanol precipitation and resuspension in 12.5µl DEPC treated water.

For all samples RNA quality was assessed using the Tapestation (Agilent). All samples tested had RIN scores of 9.6 or above.

2.1.4 mNET-seq

mNET-seq was performed by Dr T Nojima as described in[33] but with a starting input cell number of 5 x 10^6 cells. Library size range and concentration was determined using the Tapestation (Agilent) and Qubit (Thermo Fisher Scientific) respectively before normalisation to 6nM for sequencing.

2.1.5 polyA+ RNA seq library preparation

2.0mg of Total RNA was depleted of ribosomal and globin RNA species with the Globin-Zero Gold rRNA Removal Kit (Illumina). Samples were resuspended in 50µl DEPC treated water. polyA+ selection was performed using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB) according to the manufacturers protocol and isolated polyA+ RNA was resuspended in 5µl DEPC treated water.

Strand specific polyA+ libraries were made using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB) according to the manufacturers protocol and utilising a 15 minute fragmentation step and 12 cycles of PCR. Library size range and concentration

was determined using Tapestation and Qubit respectively before normalisation to 6nM for sequencing.

2.1.6 Pol2 ChIP-seq

Cell pellets were removed from -80°C and lysed by the addition of 200µl cell lysis buffer (5mM PIPES pH 8.0, 85mM KCl, 0.5% IGEPAL-CA 630) and incubation on ice for 20 mins. Nuclei were pelleted (4°C, 900RCF, 10 mins), supernatant discarded and nuclei lysed in 100µl nuclear lysis buffer (50mM Tris-HCl pH 7.5, 10mM EDTA pH 8.0, 1% SDS) for 5 mins on ice. Chromatin was sonicated for 8 mins using a Covaris S220 sonicator (Duty cycle = 2%, Int = 3.0, Cycles/burst = 200, Power mode = frequency sweeping, Duration = 120s, Temp = 6°C). Insoluble material was pelleted by centrifugation (4°C, 20,000RCF, 20 mins), supernatant removed and 10x diluted with 1ml RIPA buffer without SDS (10mM Tris-HCl pH 7.5, 1mM EDTA pH 8.0, 0.5mM EGTA pH 8.0, 1% Triton X 100, 0.1% Sodium deoxycholate). 100µl of the diluted chromatin was frozen to use as an "input" control. 110µl of a 1:1 Protein A and Protein G Dynabead slurry (Thermo) was prepared by combining equal volumes of the Protein A and G beads and washing twice in low salt RIPA buffer (0.1% SDS) using a magnetic Eppendorf rack to separate beads from supernatant. The remaining sonicated chromatin was pre-cleared 2x by incubation on a rotating platform (4°C, 10RPM, 1 hour) with 5µl of the A+G bead slurry. 10µl of Pol2 antibody (N20 – sc899c, Santa Cruz) was conjugated to 100µl of the bead slurry by incubation on a rotating platform (4°C, 10RPM, 1 hour). The precleared chromatin and antibody conjugated beads were combined and incubated overnight on a rotating platform (4°C, 10RPM). In the afternoon of the following day (24 hour incubation) washing of the immunoprecipitate was performed using 0.5ml of the following buffers at 4°C on a magnetic Eppendorf rack:

Buffer	No. of washes	Composition

Low salt RIPA	2x	Tris-HCl pH 7.5, EDTA pH 8.0 (1mM), EGTA pH 8.0 (0.5mM), Triton X-100 (1%), SDS (0.1%), Sodium Deoxycholate (0.1%), NaCl (140mM)
High salt RIPA	2x	As above except NaCl final concentration of (0.5M)
LiCl RIPA	1x	As above except omit NaCl and replace with LiCl (250mM)
T.E	2x	Tris-HCl pH 8.0 (10mM), EDTA pH 8.0 (1mM)

For each wash step, tubes were spun briefly to avoid beads sticking to lid, placed in magnetic rack and allowed to clear for ~ 1min. Supernatant was removed and discarded, before removing tubes from the magnetic rack and gently resuspending in 0.5ml of relevant buffer. Samples were incubated on the rotating platform @ 4°C for 4 mins. During the final wash a fresh solution of elution buffer (Tris-HCl (pH 7.5) (20mM), EDTA (pH 8.0) (5mM), NaCl (50mM), 1% SDS) was prepared for each immunoprecipitate and its corresponding input sample (0.6ml total per ChIP experiment). Tubes were removed from the magnet, resuspended in 0.15ml of elution buffer and transferred to a fresh Eppendorf tube. 0.2ml of elution buffer was added to each frozen input sample the remaining elution buffer stored at room temperature until use the next day. 2µl of RNAse A (10mg/ml) was added to the beads and 4µl to the corresponding input sample. Beads and inputs were incubated @ 37°C on a thermomixer with shaking function (1,000rpm, 30s on, 30s off (interval mix)) for 30 mins. 1µl of Proteinase K (20mg/ml) was added to the beads and 2µl to the inputs. Temperature was increased to 65°C and samples left overnight to incubate on the thermomixer. The following morning samples were quick spin down and samples with beads transferred to the magnetic rack. Beads were cleard for ~1min and supernatant transferred to a new Eppendorf. The beads were resuspended in 0.15ml of elution buffer, incubated for 5 mins @ 65°C and magnetic separation and supernatant transfer repeated. Combined supernatant or the input sample were extracted using phenol chloroform (using phase lock tubes according to the manufacturers protocol) followed by ethanol precipitation using 2µl of the GlycoBlue co-precipitant (20mg/µl), 30µl NaAc pH 5.2

(~1/10 volume) and 0.9ml (~ 3 volumes) of 100% EtOH. Samples were incubated for 1hour @ -80°C. Spun at 20,000g, 30 mins @ 4°C, washed 1x in 1ml 75% Ethanol and pellets resuspended in 70µl 0.1 x T.E. Library preparation was performed using the NEB Ultra DNA library preparation kit for Illumina and using 15 cycles of PCR followed by an Ampure XP bead (Beckman) clean-up. Library size range and concentration was determined using Tapestation and Qubit respectively before normalisation to 6nM for sequencing.

2.1.7 scaRNA-seq

A 10% TBE-Urea gel (Thermo Fisher Scientific) was prepared according to the manufacturer's instructions (saving 100ml 1xUltrapure TBE for gel staining) and was pre-run for 15-20 mins at 180V. 4µl of the NEB ssRNA low molecular weight ladder was denatured in 36µl of NEB RNA loading dye for 4 mins @ 90°C in a thermal cycler before snap cooling on ice. 12.5µl of extracted RNA (see 2.1.3) in DEPC water (see RNA extraction and quality control) was mixed with 12.5µl of Novex TBE-Urea Sample Buffer (2X) (Thermo Fisher Scientific) and denatured for 3 mins @ 70°C before snap cooling on ice. The ladder was loaded in duplicate in lanes at either end of the gel and the sample(s) loaded between the ladders, skipping at least one lane between each loaded lane. The gel was run for 50 mins @ 180V (just before the lower size marker dye migrated to the bottom of the gel). The gel was then stained for 30 mins using SYBR gold solution (Thermo Fisher Scientific) in 100ml 1x Ultrapure TBE. The regions of each sample lane corresponding to RNA 0-300nt in length was excised from the gel using scalpels and gel pieces placed into a 2ml Eppendorf tube with holes pierced in the bottom with a .22-gauge needle. The 2ml Eppendorf tube is then placed inside a 15ml falcon tube and the lid secured with parafilm. Samples are then spun for (5mins, 4°C @ 4000g) in a bench top centrifuge. The remaining pieces of gel left in the Eppendorf were crushed through pierced holes using a plunger from

a 2ml syringe. The tube was inverted and flicked to dislodge and transfer any remaining gel pieces into the falcon. The falcon tube containing crushed gel was weighed and two volumes of elution buffer (Tris–HCl pH 7.5 20mM, NaAc 0.25M, EDTA pH 8.0 10mM, SDS w/v 0.25%) added. The sample was vortexed and spun (1 min, @ 4000g) in a bench top centrifuge. The sample was then frozen for 15 mins @ -80°C. The samples were removed from the freezer and placed in a rack at room temperature overnight to elute the RNA. 2mls of vacuum centrifuge grease (Corning) were added to a fresh 15ml falcon. The samples and fresh falcon were spun (5min, room temperature, @ 4000g). The supernatant was transferred from the sample falcon into to the fresh falcon using a P1000 pipette with a wide bore pipette tip (with care taken to minimize transfer of gel pieces to the fresh tube). 0.5 volumes of elution buffer were added to the gel pieces in the original falcon tube. Gel and elution buffer were mixed by vortexing before spinning again (5 mins, room temperature, @ 4000g). The supernatant was then transferred to the fresh falcon tube. A phenol chloroform extraction was performed on the supernatant using the phase lock light tubes (according to the manufacturers protocol, several tubes were required per sample extraction). An ethanol precipitation was performed on the extracted aqueous phase using 3 volumes of ethanol, 1/10 volume 3M NaAc and 1µl glycoblue per 2ml Eppendorf tube. Samples were incubated @ -80°C for 1 hour before spinning (20,000g 30 mins @ 4°C). Supernatant was discarded and the pellet and walls of the tube washed with 1ml ice cold 75% ethanol (in DEPC treated nuclease free water). Samples were spun (5 mins, 20,000g, 4°C), supernatant removed before a final quick spin to collect residual ethanol. Residual ethanol was removed with P10 pipette. The pellets for each sample are resuspended in sequence (6µl DEPC was added to the first pellet before pipetting to resuspend, this sample was then transferred to the tube with the next pellet from this sample, continuing until all pellets per sample are re-suspended in the 6µl volume. 3’ adapter ligation was performed

for 1 hour and according the NEB small RNA for Illumina protocol. The reaction was topped up with 180µl DEPC treated nuclease free water and a phenol chloroform extraction performed using phase lock light tubes followed by and ethanol precipitation (Samples were incubated @ -80°C). The pellet was resuspended in 25.5µl DEPC treated nuclease free water. A phosphatase treatment was performed using calf intestinal phosphatase (CIP, NEB) to remove fragmented RNA molecules from downstream library preparation for 30 mins @ 37°C. The sample was topped up with 180µl DEPC treated nuclease free water and a phenol chloroform extraction performed (identical to post adapter ligation). The pellet was resuspended in 26µl DEPC treated nuclease free water. A decapping reaction was performed in a reaction volume of 30µl for 30 mins @ 37°C using CAP-clip (Tebu bio). The sample was again phenol chloroform extracted with the pellet resuspended in 17.5µl DEPC treated nuclease free water. The SR RT primer (NEB small RNA kit for Illumina) was hybridised to the 3' adapter and the 5' adapter ligated by adding 2µl T4 RNA ligase buffer, 5µl PEG 8000, 1µl SR RT Primer and following the manufacturers protocol (NEB small RNA for Illumina library prep): "Hybridise the reverse transcription primer" followed by "ligate 5' SR adaptor…". A reverse transcription reaction was performed in a reaction volume of 50µl by adding Superscript III First Strand Synthesis System was used to generate cDNA by adding reagents directly to the 5' adapter ligated samples (total volume of 50µl) and incubating for 1 hour at 50°C (following the manufacturers protocol). The reverse transcribed samples were subjected to PCR for 15 cycles by adding 50µl of the PCR mastermix from the NEB small RNA for Illumina library prep kit (NEB), 5µl of the Universal adapter and 5µl of a unique indexed adapter (see NEB library pooling guidelines for suitable adapter combinations for three sample sequencing runs). Library clean-up was performed using Ampure XP beads (Beckman) (using 110µl per PCR reaction and following manufacturer's instructions). Libraries were eluted from beads into 23µl 1x T.E.

Library size range and concentration was determined using Tapestation and Qubit respectively before normalisation to 6nM for sequencing.

scaRNA-seq libraries were produced in isogenic duplicate for K562 cells (all conditions) and biological triplicate for mouse foetal liver (0h and 24h wildtype and enhancer knockouts).

2.1.8 ATAC-seq

ATAC-seq was performed exactly as in Buenrostro et al [74] on 5 x10^5 cell aliquots. All ATAC-seq experiments were performed in biological triplicate. Library size range and concentration was determined using Tapestation and Qubit respectively before normalisation to 6nM for sequencing.

2.1.9 Capture-C

Experiments were performed by Dr D Downes. 5 x10^6 mouse foetal liver cells were fixed using a final concentration of 2% formaldehyde in PBS, spun down 500g 5 mins and placed into lysis buffer before freezing at -80°C. 3C libraries were prepared using *Dpn*II and standard methods as previously described[16] with the following modifications: no douncing was performed, all spins were performed at 300RCF, and after ligation intact nuclei were pelleted (15 min, 300RCF), supernatant was discarded, and nuclei were resuspended in 300 µl 1x T.E (Sigma) for phenol chloroform extraction.

Digestion efficiency was determined by RT-qPCR with TaqMan and custom oligonucleotide, and ligation efficiency qualitatively determined by gel electrophoresis. Only 3C libraries with >70 % digestion efficiency was used. 3C libraries were sonicated to 200bp in a Covaris S220 and indexed with NEB Next Illumina library Prep reagents (NEB).

Enrichment for specific viewpoints was performed with 70mer biotinylated oligonucleotides designed using CapSequm[8] (http://apps.molbiol.ox.ac.uk/CaptureC/cgi-bin/CapSequm.cgi). Double capture was performed in multiplexed reactions with pools of oligonucleotides targeting either promoter proximal or promoter distal *DpnII* fragments following the described method[8] with each oligonucleotide at a working concentration of 2.9nM. Captured 3C libraries were sequenced on the NextSeq platform (Illumina) with 150 bp paired-end reads.

2.2 Analysis

2.2.1 Published data used in this thesis

N.B: BRep = biological replicate, TRep = technical replicate

K562 cell data

Assay	GEO or SRP ascension number	SAMN ascension number (with SRX number if required to distinguish replicates)
MNase-seq	GSE78984	BRep1: SAMN04538241 (PE)
GRO-seq	GSE60456	BRep1: SAMN02991551 (SE)
ChIP-seq (SUPT5) (Antibody = A300-868A)	GSE92144	BRep1: SAMN06121986 (SE) BRep2: SAMN06122354 (SE)
ChIP-seq (CCNT2)	GSE31477	TRep1 + 2: SAMN01001086 (SRR502498, SRR502499) (SE)
ChIP-seq (NELF-E) (Antibody = sc-32912)	GSE31477	TRep1: SAMN01000932 (SE) TRep2: SAMN01000932 (SE)
ChIP-exo (Pol2) (Antibody = sc899)	GSE108323	BRep1: SAMN07546015 (SE) BRep2: SAMN07546016 (SE)
ChIP-exo (Pol2) (Antibody = sc899)	SRP019209	BRep1: SAMN01942539 (SE) BRep2: SAMN01942540 (SE)
GRO-cap	GSE60456	BRep1: SAMN02991547 (SE)
ChIP-seq (Pol2) (Antibody = sc899)	GSE108323	BRep1: SAMN07546015 (SRR6386683) (SE) BRep2: SAMN07546016 (SRR6386687) (SE)
4sU-seq	GSE75792	BRep3, TRep1: SAMN04325065 (SRR5243204) PE

HEK293 cell data

Assay	GEO or SRP ascension number	SAMN ascension number (with SRX number if required to distinguish replicates)
TSS-seq	SRP000403	BRep1: SAMN00001377 (SE)
GRO-seq Control	GSE97072	BRep1: SAMN06645256 (SE) BRep2: SAMN06645255 (SE)
GRO-seq DRB	GSE97072	BRep1: SAMN06645254 (SE) BRep2: SAMN06645253(SE)
DNase-seq	GSE32970	TRep2: SAMN01730094 (SRR576373) TRep1: SAMN01730094 (SRR5763734)

Mouse foetal liver aHom cell data

Assay	GEO or SRP ascension number	SAMN ascension number (with SRX number if required to distinguish replicates)
ATAC-seq	GSE78800	BRep1: SAMN04526274 (PE) BRep2: SAMN04526275 (PE) BRep3: SAMN04526277 (PE)

2.2.2 Summary of alignment and pre-processing for all techniques

N.B. PCR duplicate filtering performed with samtools/0.1.19 rmdup[75], removal of reads overlapping splice sites or non-coding RNA performed using bedtools/2.25.0 intersect[76] and alignments performed using bowtie/1.1.2[77] or rna-star (STAR)[78].

Method	Aligner (and flags used)	PCR duplicate filtering (Y/N)	Removal of non-coding RNA (Y/N)	Removal of splice sites
MNase-seq	Bowtie/1.1.2 (-m 2 --maxins 2000 --best --strata --chunkmb 256)	Y	N	N
scaRNA-seq	Bowtie/1.1.2 (-m 2 -p 4 --maxins 1000 --best --strata --chunkmb 256)	N	Y	Y
GRO-cap	Bowtie/1.1.2 (-m 2 -p 1 --maxins 1000 --best --strata --chunkmb 256)	N	Y	Y

GRO-seq	Bowtie/1.1.2 (-m 2 -p 1 --maxins 1000 --best --strata --chunkmb 256)	N	N	N
mNET-seq	Bowtie/1.1.2 (-m 2 -p 1 --maxins 1000 --best --strata --chunkmb 256)	Y	Y	Y
polyA+ RNA-seq	rna-star/2.4.2a (--outFilterMultimapNmax 2)	N	N	N
intronic RNA-seq	Bowtie/1.1.2 (-m 2 -p 1 --maxins 1000 --best --strata --chunkmb 256)	N	N	N
ChIP-exo	Bowtie/1.1.2 (-m 2 -p 1 --maxins 1000 --best --strata --chunkmb 256)	Y	N	N
ChIP-seq	Bowtie/1.1.2 (-m 2 -p 1 --maxins 1000 --best --strata --chunkmb 256)	Y	N	N
ATAC-seq	Bowtie/1.1.2 (-m 2 -p 1 --maxins 1000 --best --strata --chunkmb 256)	Y	N	N
4sU-seq	Bowtie/1.1.2 (-m 2 -p 1 --maxins 1000 --best --strata --chunkmb 256)	N	N	N

2.2.3 Generation of observed and annotated TSS list

UCSC annotated genes for genome build hg19 were downloaded from the table browser (http://genome.ucsc.edu/cgi-bin/hgTables), in bed format describing the whole annotated gene. A custom Perl script was written to merge genes which had exactly the same chromosome, start, stop and strand retaining individual ID's. The TSS were extracted as the 5' most coordinate of each gene taking account for strand. These coordinates were compared to observed TSS (derived from scaRNA-seq data in the same tissue – see 2.2.7.3 observed TSS calling using Homer[79]) to find coordinates which fell within 500bp. The UCSC TSS which had a scaRNA-seq TSS were retained and used as the "Annotated" TSS list. The homer derived TSS coordinate for each UCSC gene was used as the "Observed TSS).

2.2.4 Generation of non-coding RNA list

Lists of Ensembl non coding gene coordinates (for mm9 and hg19 genes respectively) were downloaded from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables), using the following filtering terms; “miRNA”, “tRNA”, “snoRNA”, “Mt_tRNA”, “Mt_rRNA”, “rRNA”, “snoRNA”, “snRNA”. Lists of UCSC non-coding gene coordinates were also downloaded using the following filtering terms; “snRNA”, “miRNA”, “tRNA”, “snoRNA”, “Mt_tRNA”, “Mt_rRNA”, “rRNA”, “snoRNA”, “snRNA”. All coordinates were combined into a consensus list for either the mm9 or hg19 genome and windowed by 100nt up or downstream using bedtools slop. The lists were used in read filtering with bedtools intersect.

2.2.5 Generation of splice sites list

UCSC annotated genes were downloaded from the table browser (http://genome.ucsc.edu/cgi-bin/hgTables), in gtf format and splice sites extracted for each gene using Awk from the exon coordinates in the file.

2.2.6 Read coverage relative to list of coordinates (TSS)

Homer tag directories were made from aligned (post processed) bam or bed files using makeTagDirectory[79] and specific flags depending on the sample in question:

Assay	Homer makeTagDirectory flags used
MNase, ChIP-seq	-format bed -keepAll -genome hg19
scaRNA (5’ RNA), TSS-seq, GRO-cap, ChIP-exo	-format bed -sspe -keepAll -genome hg19
scaRNA (3’ RNA), GRO-seq, mNET-seq	-format bed -sspe -keepAll -genome hg19 -flip

Coverage values relative to annotated or observed TSS coordinates, in 1bp bins, were generated using Homer annotatePeaks[79] and specific flags depending on sample in question:

Assay	Homer annotatePeaks flags used
MNase-seq, ChIP-exo, ChIP-seq	-hist 1 -size 1000 -pc 3 -raw
GRO-cap, GRO-seq, scaRNA-seq, mNET-seq, TSS-seq, ChIP-exo	-hist 1 -size 1000 -pc 3 -raw -len 1

Histograms were plotted in from coverage values using Microsoft excel.

2.2.7 scaRNA-seq stages in analysis

2.2.7.1 Normalisation and visualisation of individual genes

From the aligned reads, properly paired reads (bitwise flag =3) were extracted using samtools/0.1.19 view and converted to BEDPE format using bedtools/2.25.0 bamtobed. The 5' and 3; most coordinates of the aligned read pairs were used to estimate the start and end coordinates of the RNA insert and produce a bed file of these values. Co-ordinates overlapping non-coding RNA, annotated splice sites were removed using bedtools/2.25.0 intersect. The 5' and 3' coordinates of each RNA insert were windowed by 2nt in each direction to aid in visualisation. Strand specific normalised bedGraph tracks were produced representing the windowed starts and ends of each RNA insert using bedtools/2.25.0 genomecov based on coverage of the 7SK RNA transcribed by Pol III[80]. bedGraph files were converted to bigWigs using ucsctools/1.0 bedGraphToBigWig and data were visualised on the UCSC genome browser.

2.2.7.2 observed TSS calling using Homer

Homer findPeaks -style tss was used to call TSS coordinates from scaRNA-seq or TSS-seq Tag directories representing only the 5' ends of RNA inserts.

A consensus list of scaRNA-seq TSS peaks in control K562 cells was produced by calling TSS coordinates from isogenic duplicates and combining and condensing overlappers using Awk and bedtools/2.25.0 merge.

A consensus list of observed mouse foetal liver TSS coordinates was produced by calling TSS in each of the biological triplicates for the early and late erythroid stages (n=6 peaks calls) and combining and condensing overlappers using Awk and bedtools/2.25.0 merge.

The TSS lists were called as observed if they fell within 500bp's of an annotated UCSC TSS coordinate (see 2.2.3 Generation of annotated UCSC TSS list).

2.2.7.3 Intersecting the observed TSS with annotated TSS

The coordinates of each observed TSS were windowed and distances to the nearest annotated UCSC TSS determined using bedtools/2.25.0 closest taking account of strand when plotting distances. Observed TSS with a UCSC within 500bp's were retained and used as "observed" TSS in all downstream analysis

2.2.7.4 Observed and annotated TSS motif enrichment analysis

Motif enrichment around observed or annotated TSS was performed using a custom Perl script. Briefly the script uses the coordinates of the TSS to retrieve sequence information for a user specified window around the coordinate. The sequence is scanned in a user specified bin size to identify specific user specified motifs (also using the reverse of the user specified motif to account for strand). The sum of the motif occurrence is plotted for each bin, using the 5' end of the motif as the point plotted (i.e. a 2nt motif ends at the TSS would be plotted at -2 relative to the TSS.

2.2.7.5 Read density ratio calculations

The observed TSS were windowed upstream by 100nt and downstream by 300nt using bedtools/2.25.0 slop (taking account for strand). Read coverage of the TSS windows was calculated from the merged reads the isogenic duplicate. Read counts were converted to

density by dividing by the size of the window (400). Densities multiplied by a normalisation factor (1/ (read coverage over 7SK non-coding RNA)) resulting in DRB treatment values being multiplied by 0.56. Finally, a ratio of DRB versus control density (read density ratio) was calculated and plotted for each TSS.

2.2.8 Processing of the annotated UCSC gene list - stages in analysis

2.2.8.1 Whole gene coordinates

UCSC annotated genes for genome build hg19 were downloaded from the table browser (http://genome.ucsc.edu/cgi-bin/hgTables), in bed format describing the whole annotated gene. A custom Perl script was written to merge genes which had exactly the same chromosome, start, stop and strand retaining individual ID's.

2.2.8.2 Whole gene minus first 300bp

The whole gene coordinates were modified to remove the first 300bp using bedtools/2.25.0 slop.

2.2.9 Plotting distances between observed versus annotated TSS

The distances between observed and annotated TSS (see 2.2.3) were plotted as a histogram using a custom Perl script, taking account of strand and summing the occurrences of each distance in 1bp bins.

2.2.10 ChIP-seq

Aligned bam files were merged in each condition and read per million normalised bigwig files were made produced using bedtools (genomecov) and ucsctools bedGraphToBigWig (1.0). bigWigs were visualised on the UCSC genome browser.

2.2.10.1 Travelling ratio

TR was calculated as in[22] using a custom Perl script. A slightly larger upstream TSS window of 100bp as opposed to 30bp was used and only unique genes (See 2.2.8.1 Whole gene coordinates) longer than 600bp analysed.

2.2.10.2 Visualisation on genome browser and normalisation

Aligned and post processed bam files from biological or isogenic replicates were merged using samtools merge. bedGraph tracks were generated from merged files using bedtools/2.25.0 genomecov. Tracks were normalised using reads per million normalisations. bedGraph files were converted to bigWigs using ucsctools/1.0 bedGraphToBigWig and data were visualised on the UCSC genome browser.

2.2.10.3 Pol2 read density calculation

Read coverage over UCSC genes was calculated using bedtools/2.25.0 coverage using aligned bam files. Density was calculated by dividing the read coverage by the size of the gene.

2.2.11 polyA+ RNA-seq

2.2.11.1 Visualisation on genome browser and normalisation

Aligned bam files were merged in each condition and strand specific read per million normalised bigwig files were made produced using bedtools (genomecov) and ucsctools bedGraphToBigWig (1.0). bigWigs were visualised on the UCSC genome browser.

2.2.11.2 Expression calculation (RPKM)

Aligned bam files were merged in each condition and strand specific coverage of the genes was calculated using bedtools/2.25.0 coverage and divided by the size of the gene in kilobases and then by the number of million mapped reads in the sample.

2.2.11.3 Intronic RNA read isolation

After alignment and pre-processing isogenic replicates were merged per condition and reads that intersected UCSC annotated introns (obtained from the UCSC table browser) were isolated using bedtools intersect (2.25.0) the resultant reads were further processed by removing reads overlapping UCSC annotated exons (obtained from the UCSC table browser).

2.2.11.4 Coverage over gene bodies (ngsplot)

Gene body coverage plots were generated using ngsplots/2.47.1[81] and the following flags: -G hg19 -D refseq -R genebody -RB 0.05 -MQ 0 -SE 0 -L 5000 -VLN 0 -MW 2.5

2.2.11.5 Intronic RNA read density ratio calculations

300bp from the start of annotated genes associated with an observed TSS (see 2.2.8.2) was trimmed, taking account for strand, using bedtools/2.25.0 slop. Read coverage was computed using bedtools/2.25.0 coverage over the trimmed genes and density calculated by dividing the count by the size of the trimmed gene in bps. The intronic read density ratio was computed as a read density in treatment versus control conditions.

2.2.12 Splice site analysis

UCSC genes for genome build hg19 were downloaded in gtf format from the UCSC table browser (http://genome.ucsc.edu/cgi-bin/hgTables). Distances of the 5' most splice site for each gene were extracted and distances to the annotated TSS determined using a custom

Perl script. Distances of the 5' splice sites were plotted relative to the annotated TSS as a histogram in 1bp bins.

2.2.13 ATAC-seq

Aligned bam files were merged in each condition and read per million normalised bigwig files were made produced using bedtools (genomecov) and ucsctools bedGraphToBigWig (1.0). bigWigs were visualised on the UCSC genome browser.

2.2.13.1 Generation of consensus peaks

Peak calling was performed using macs2/2.0.10 callpeaks[82] with the flag (-atac). Peak calls for individual replicates for specific genotypes (Wt, aHom, bHom) were merged (n=9 peak calls total) and a consensus peak call produced using Awk and bedtools/2.25.0 merge.

2.2.13.2 Visualisation on the genome browser and normalisation

Aligned bam files were merged in each condition and strand specific read per million normalised bigwig files were made produced using bedtools (genomecov) and ucsctools bedGraphToBigWig (1.0). bigWigs were visualised on the UCSC genome browser.

2.2.14 mNET-seq visualisation and normalisation

Aligned bam files were merged in each condition and normalised bigwig files representing 3' ends of RNA on each strand were made produced using bedtools (genomecov -3 -strand) and ucsctools bedGraphToBigWig (1.0). bigWigs were visualised on the UCSC genome browser. Normalisation was performed relative to the 7SK RNA gene.

2.2.15 Differential expression analysis

2.2.15.1 scaRNA-seq

Observed windowed TSS in mouse (see 2.2.7.2) were windowed upstream by 100nt and downstream by 300nt using bedtools/2.25.0 slop (taking account for strand). Read coverage

of the TSS windows was calculated for each replicate of mouse foetal liver at the two time points (0h and 24h). Differential gene expression was calculated using EdgeR, filtering genes with < 1 count per million. Significantly differential expressed genes were identified with P values <= 0.05 and a false discovery rate of <= 0.05.

2.2.15.2 Intronic RNA-seq

300bp from the start of annotated genes associated with an observed TSS was trimmed (see 2.2.8.2), taking account for stand, using bedtools/2.25.0 slop. Strand specific read density was computed using bedtools/2.25.0 coverage for each replicate of mouse foetal liver, at the two time points (0h and 24h). Differential gene expression was calculated using EdgeR[83], filtering genes with < 1 read count per million. Significantly differential expressed genes were identified with P values <= 0.05 and a false discovery rate of <= 0.05.

2.2.15.3 polyA+ RNA-seq

Strand specific read density over observed UCSC genes -300bp from annotated start (see 2.2.8.2) was computed using bedtools/2.25.0 coverage for each replicate of mouse foetal liver, at the two time points (0h and 24h). Differential gene expression was calculated using EdgeR, filtering genes with < 1 count per million. Significantly differential expressed genes were identified with P values <= 0.05 and a false discovery rate of <= 0.05.

2.2.15.4 ATAC-seq

Read density over consensus ATAC peaks (see 2.2.13.1) was computed using bedtools/2.25.0 coverage for each replicate of mouse foetal liver in each genotype (WT, aHom, bHom). Differential gene expression was calculated using EdgeR, filtering peaks with < 1 count per million. Significantly differential expressed peaks were identified with P values <= 0.05 and a false discovery rate of <= 0.05.

2.2.16 Pearson correlation analysis of replicates (ChIP-seq, ATAC-seq)

Pearson correlation between ChIP-seq or ATAC bam files was determined using deeptools/2.2.2[84] multiBamSummary bins using flags (--centerReads) and plotted using deeptools/2.2.2 plotCorrelation and flags (--skipZeros --colorMap BrBG -c pearson -p heatmap --plotFileFormat pdf –removeOutliers)

2.2.17 Capture C

Analysis performed by Dr D Downes. Reads were mapped and analysed using CCseqBasic4 (github.com/Hughes-Genome-Group/CCseqBasic4) as previously described[32] with custom settings (-m 2 -v 3). After primary analysis replicates were compared using CaptureCompare (github.com/Hughes-Genome-Group/CaptureCompare). CaptureCompare normalises *cis* reporter counts for each DpnII fragment per 100,000 *cis* reporters and generates mean counts for replicates.

2.2.18 4sU-seq - Expression calculation (RPKM)

Strand specific coverage of the genes was calculated from aligned reads in bam format using bedtools/2.25.0 coverage and divided by the size of the gene in kilobases and then by the number of million mapped reads in the sample.

2.2.19 Scatter plots and hexbin plots

Scatter plots were made manually in Microsoft excel. Hexbin plots were produced from tab delimited files of paired values (e.g. TR versus Expression) using a custom python script.

Chapter 3

Quantitating promoter proximal transcription

Chapter 3 – Quantitating promoter proximal transcription

3.1 Introduction

3.1.1 Existing methods to identify transcription initiation genome wide

Since the completion of the Human Genome Project, annotation of the reference sequence to identify regions harbouring genes and regulatory elements has been the focus of large scale consortia such as GENCODE[85]. An important aspect of this work has been the identification of transcription start sites (TSS), as the promoter regions proximal to these positions have important regulatory functions in gene expression[86].

The first method to offer TSS mapping in high throughput was cap analysis of gene expression (CAGE)[87]. This method makes use of the fact that eukaryotic RNAs transcribed by RNA polymerase II contain methyl 7 guanosine (m^7G) caps at their 5' ends[88]. The RNA is first reverse transcribed using random primers. The RNA cap and 3' RNA ends are biotinylated. Single stranded RNA (e.g. Non RNA/DNA hybrids are removed by digestion with RNAse I). Remaining RNAs which have been reverse transcribed completely to the cap are isolated by streptavidin pulldown. Second strand synthesis is performed and the double stranded DNA products "tags" produced. These tags are concatemerised before Sanger sequencing and mapping to the genome. Thus, in one experiment, thousands of TSS can be profiled simultaneously (Figure 3.1).

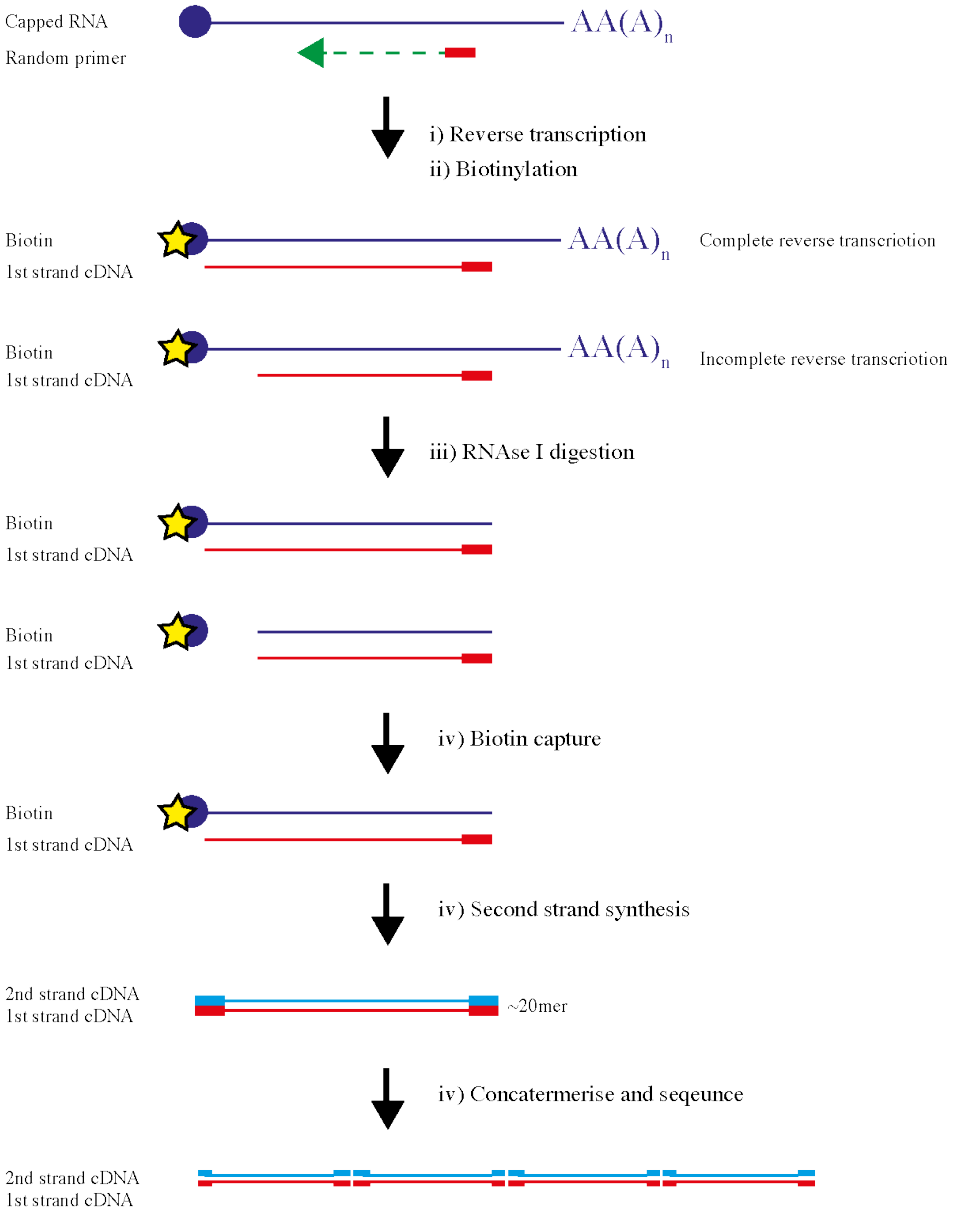


Figure 3.1) A schematic of the CAGE method. Capped RNAs have a random primer hybridised to them to allow reverse transcription. RNA cap is biotinylated. RNAse I treatment degraded single stranded RNA leaving only fully reverse transcribed RNA hybridised to cDNA. Streptavidin pulldown of biotin and second strand synthesis generates cDNA. cDNA is concatenated before Sanger sequencing,

One of the features of CAGE tags distributions noted by its authors was the variability in transcription start site positions relative to previously sequenced full length mouse cDNAs from the FANTOM consortium[87]. The authors suggested that between 11-27% of CAGE

tags added a 100nt extension to the previously defined TSS position but concluded that their methodology likely represented a more accurate map of TSS. A comparative analysis between CAGE defined TSS and predicted GENCODE TSS position (based on cDNA mapping) showed a relatively low percentage of overlap (65%). This was attributed to limited coverage of the CAGE data sets in terms of the tissues surveyed which, at least for the GENCODE consortia, were derived from 24 whole cell line polyA+ CAGE experiments[85], but could also be due to the poor efficiency of the protocol in identifying TSS accurately.

Further development of CAGE resulted in nanoCAGE, with key distinctions being lower input amounts of RNA used to generate libraries (50ng as opposed to 25mg in CAGE), the use of next generation sequencing to increase sensitivity of detection and a novel approach to mark the position of the m7G cap by adding a run of G's to this position with a template switching reverse transcriptase[89]. However, the authors showed that the template switching activity of the reverse transcriptase is not specific to capped RNA and a significant percentage (8%) of all reads mapped to rRNA which do not possess a cap, suggesting this method may introduce false positives into gene annotation by marking truncated or non-capped RNA species.

More recently, methods have been developed to mark the 5' ends of capped RNA by enzymatic degradation and conversion into a free phosphate group which can act as a substrate for adapter ligation. Previously described in the introduction as a method to profile promoter proximal pausing, the "Start-seq" method can also be used to identify TSS *in vivo* with high resolution (~100nt), due to size selection of <=100nt RNA on polyacrylamide gels prior to 3' and 5' adapter ligations[35,36]. This method uses single end Illumina sequencing from the 5' or the 3' ends of the insert, first mapping the insert to the

genome and then plotting only the 5' or 3' most base. High density of these 5' or 3' ends in a small window ~150bp is indicative of an active TSS or pausing sites (Figure 1.4).

An alternative approach for TSS identification, which relies on isolating nascently transcribed RNA by virtue of an incorporated biotinylated nucleotide, is GRO-cap. Also known as 5'GRO-seq[61,90]. This method, like Start-seq, enzymatically degrades the m7G cap and uses adapter ligation onto the exposed free phosphate group to mark the 5' end (initiation site of the RNA). This allows base pair identification of TSS genome wide (Figure 3.2). However, as the generation of nascent RNA involves nuclear isolation or cell permeabilization followed by *in vitro* run on, it is unknown whether this method gives a truly *in vivo* view of transcription initiation as many of the factors and chromatin structure expected to influence and regulate the process of initiation or elongation may be perturbed.

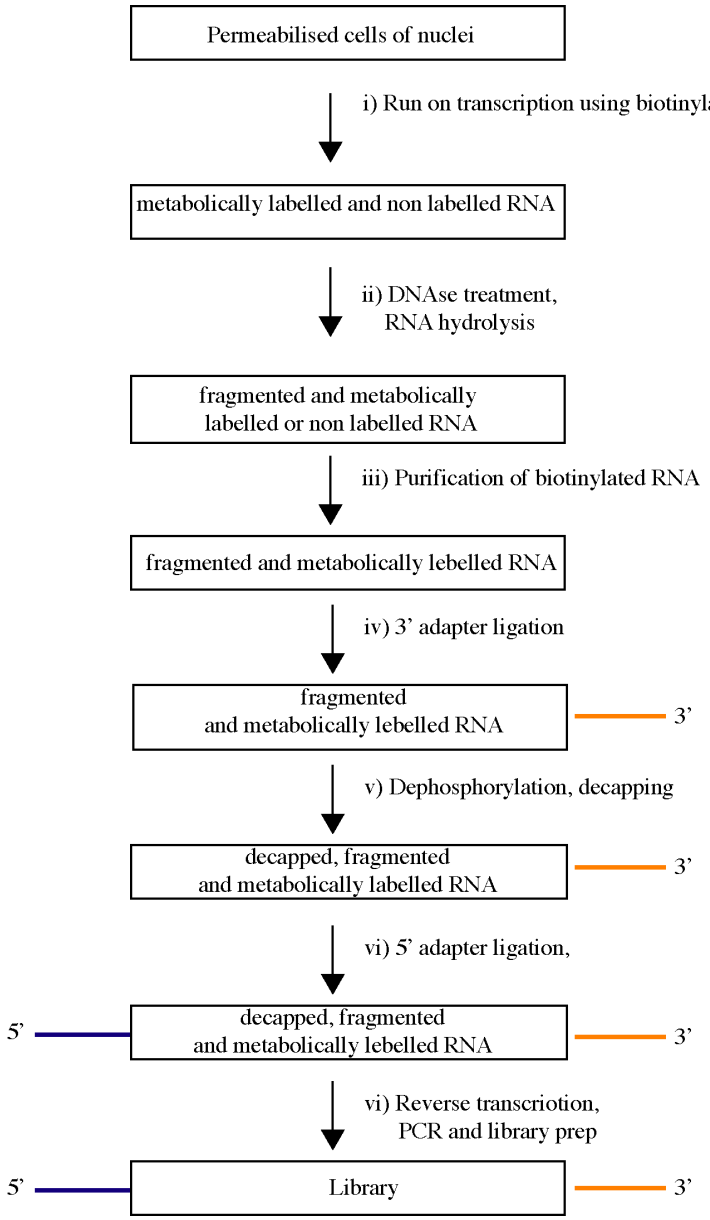


Figure 3.2) A schematic of the GRO-cap method.

These methods have gleaned many insights into transcription at metazoan promoters such as the propensity for divergent transcription initiation in mammals, high levels of Pol2 pausing and the presence of thousands of eRNA's with as yet undefined functionality or roles[36,61].

However non tissue specific TSS are always used in these analyses of promoter proximal transcription[22,29,30,34,53,71,91]. Due to the uncertainly in accuracy of TSS annotation and quantitative levels of transcription occurring from each active TSS[87], we cannot be sure

that our current view of general transcription at promoters and enhancers is not skewed. At the very least, the use of annotated TSS dilutes signal in genome wide analysis as not all TSS are active but yet all TSS coordinates are used in meta analyses. Therefore, there is a need to develop and apply accurate tissue-specific TSS mapping methods and resolve transcription around the dominant active TSS at individual genes.

3.1.2 Our changing view of promoter proximal transcriptional regulation

As discussed in the introduction, our current understanding of the regulation of promoter proximal transcription has undergone a sea change in the past few decades, with an increasing appreciation of the complexity of transcription regulation through recognition of distinct stages in the transcription cycle and the role that enhancers may play in regulating them.

Pol2 pausing has emerged as a potential obligate regulatory step in transcription through studies which use pause release inhibitors such as DRB to reveal accumulations of promoter proximal RNA and Pol2 upon treatment, indicative of a post initiation block to elongation[12,64].

However, these studies have limited power to resolve distinct promoter proximal stages in the transcription cycle simultaneously and the inhibitors commonly used to test the potential prevalence of Pol2 pausing regulation have unknown specificity *in vivo*. In order for us to understand whether Pol2 pausing truly represents an *in vivo* regulatory step we must be able to detect it at high-resolution, in the context of other stages in the transcription cycle and ultimately examine how this phenomenon changes in combination with changes in gene expression.

3.2 Results

3.2.1 scaRNA-seq resolves initiation and pausing simultaneously at the single genes

In order to investigate the relationship between transcription initiation, Pol2 pausing and elongation, I modified the Start-seq assay in three ways: i) I increased the size of scaRNA molecules selected on denaturing PAGE gels from a range of 0-100nt to 0-300nt; ii) I adopted paired-end sequencing of scaRNA, allowing us to simultaneously map sites of transcription initiation and Pol2 pausing quantitatively and at a single molecule level; iii) we removed chromatin isolation during the preparation.

i) Our first modification allows us to examine the distributions of 3' RNA ends (for example positions of Pol2 accumulation such as Pol2 pausing) farther away from the TSS than in previous work. This is important as the +1 nucleosome may play a role in regulation of promoter proximal elongation and establishment of Pol2 pausing[37,39,92] and is located ~150bps downstream of annotated TSS.

   This was confirmed by re-analysis of published micrococcal nuclease followed by sequencing (MNase-seq) in K562 cells which reveal regions of DNA bound by proteins (Figure 3.3).

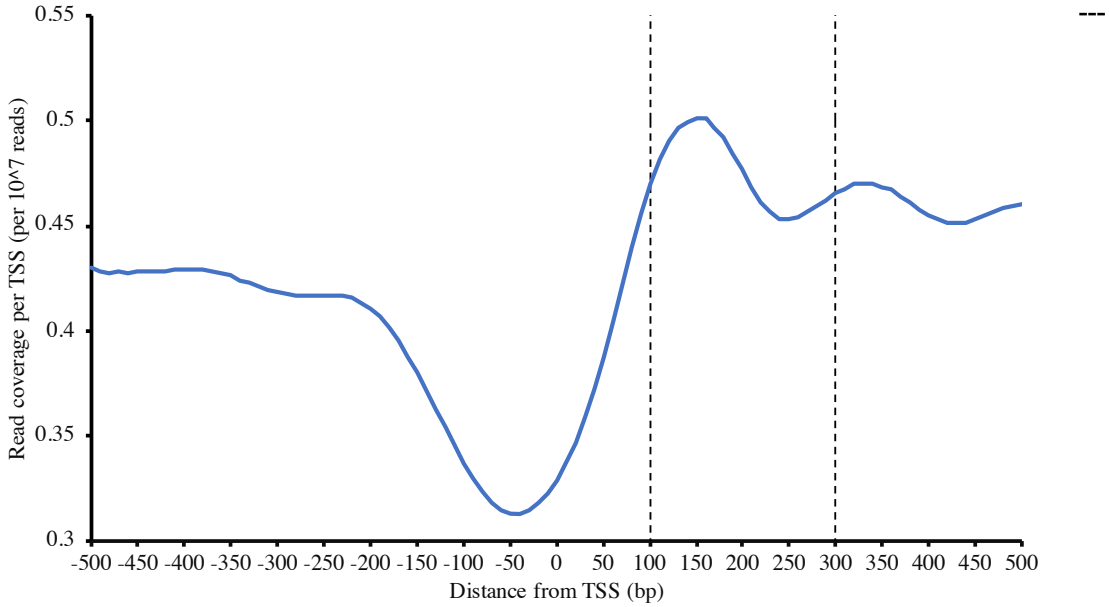


Figure 3.3) MNase-seq read coverage in K562 cells. Data plotted relative to UCSC annotated TSSs. Genes on the plus or minus strand are oriented in the same direction. Dotted lines indicate the length of RNA selected by Start-seq (100nts) and scaRNA-seq (300nts).

ii) The second modification to Start-seq was the adoption of paired-end sequencing methodology, enabling the simultaneous identification of the 5' and 3' ends of each RNA molecule. Regions which show enrichment for 5' or 3' ends in a narrow window are likely to be TSSs or Pol2 pause sites, respectively.

iii) We performed scaRNA-seq on total cellular RNA as opposed to chromatin bound RNA as our input cell numbers ($5x10^6$) were significantly lower than the 10^8 cells used in traditional Start-seq experiments[35], meaning material losses were of a greater concern and also given the observation that the majority of short capped RNA species remain stably associated with chromatin in mES cells, meaning greater recovery of scaRNA may be achieved without loss of the ability to resolve nascent " chromatin associated" transcription events[38]

I refer to this modified procedure as short capped RNA-seq (scaRNA-seq), a schematic representing key stages in the procedure is shown in Figure 3.4.

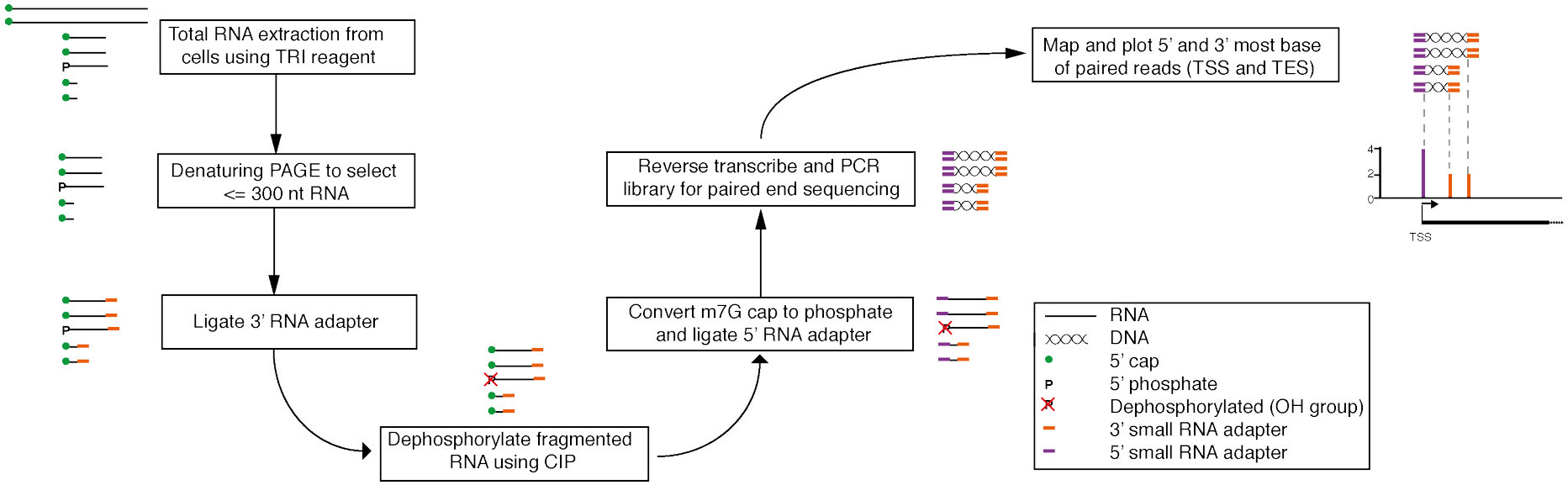


Figure 3.4) A schematic representing the key stages in scaRNA-seq. Total cellular RNA is extracted and RNA <= 300nt's is size selected on denaturing PAGE gels, removing the majority of mature cytoplasmic RNA and enriching for nascent RNA species. A 3' sequencing adapter is ligated. 5' phosphate on uncapped or fragmented RNA are removed to prevent downstream ligation to adapters. The 5' ends of capped RNA are de-capped to expose a 5' phosphate only on the first nucleotide of each RNA. A 5' sequencing adapter can then be ligated to the exposed phosphate specifically on de-capped RNA and libraries are made by reverse transcription followed by PCR amplification. Libraries are sequenced and mapped reads used to reconstruct the sequence of the original RNA molecules. The plotting of the 5' and 3' most bases of these molecules determines the distributions of transcription start (TSS) and transcription end sites (TES) respectively.

I chose to use the chronic myelogenous leukaemia (K562) cell line as a model to validate scaRNA-seq as a wealth of published genomics data is available from this cell line, allowing comprehensive comparison to other methods for profiling promoter proximal transcription.

I produced scaRNA-seq libraries from $5x10^6$ K562 cells and mapped the data to the reference genome. I also generated polyA+ RNA sequencing data in K562 cells to determine the extent of transcription genome-wide. To determine whether scaRNA-seq could resolve TSS and locations of promoter proximal pausing I visualised the well-defined *Myc and HSPA1A* (Heat Shock Protein Family A (Hsp70) Member 1A) genes (Figure 3.5), which have both served as paradigms for promoter proximal pausing[93–95].

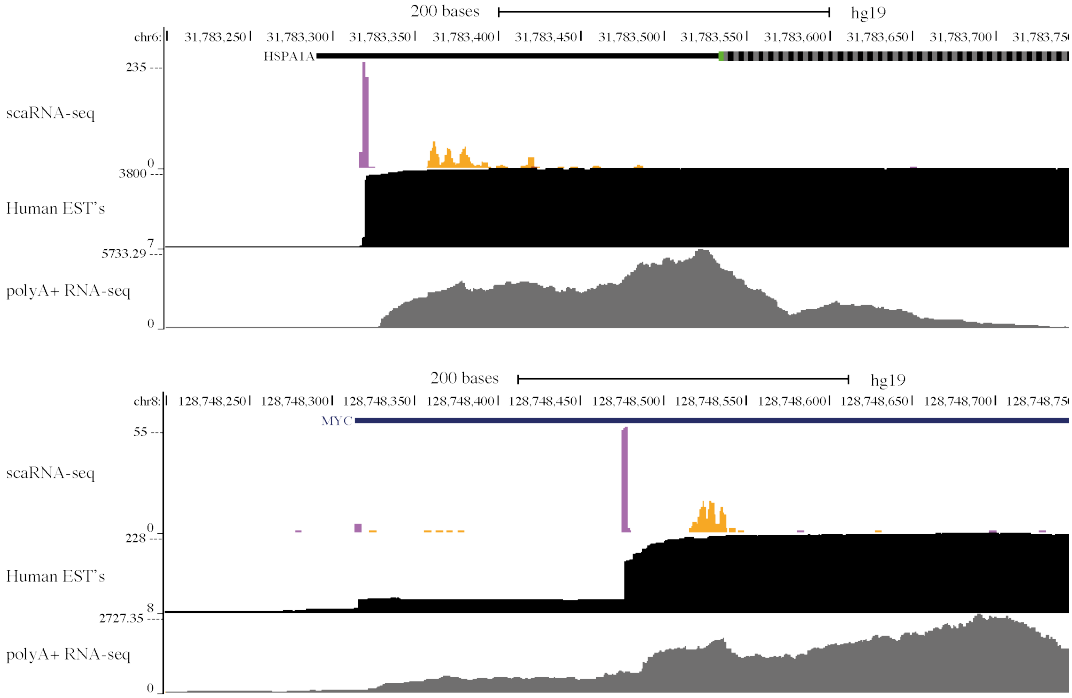


Figure 3.5) *HSPA1A* and *MYC* promoter proximal regions. Data from K562 cells. Top row = scaRNA-seq; 5' RNA ends shown in purple, 3' RNA ends shown in orange. Middle row = EST (Expressed Sequence Tags). Bottom row = read counts for poly+ RNA-seq. Y axis indicates raw read counts for scaRNA-seq, RPKM normalised read counts for poly+ RNA-seq and total numbers of isoforms for ESTs.

Comparison with polyA+ RNA-seq and EST evidence (but importantly not gene annotation) shows the 5' RNA ends (purple) marking the site of transcription initiation with base pair resolution and a downstream accumulation of 3' RNA ends (orange) consistent with promoter proximal pausing at these genes. Interestingly, the *MYC* gene shows evidence of a major and a minor initiation site; this is also supported by the relative quantities of EST's and polyA+ RNA-seq in this cell type and observations from *in vitro* transcription assays[94].

The *HSPA1A* gene shows some evidence of mis-annotation of the TSS or an alternative initiation site a few tens of bps downstream from the annotated TSS and this is supported by EST and RNA-seq data. This analysis shows that scaRNA-seq achieves the same base pair strand specificity as Start-seq libraries, where the 5' or the 3' ends of RNAs are assayed independently. However using scaRNA-seq this information is achievable in one assay and importantly the position of the transcript's initiation can be linked to its pausing.

3.2.2 Bioinformatic calling and validation of tissue-specific TSS

Due to the observed discrepancies between the annotated and observed TSS at the *Myc* and *HSPA1A* genes, I investigated whether mis-annotation of TSSs is a common feature of annotated genes. I began by calling TSS from our scaRNA data using the Homer package (findPeaks) which identifies TSS from short RNA sequencing methods[79]. This analysis revealed 18,331 "observed TSSs" in K562 cells.

To test whether the observed TSS were more accurate than the UCSC annotated TSSs, I searched for the Initiator motif, which is known to be associated with the precise position of transcription initiation, is conserved across species and has been previously shown to be enriched in Start-seq data[35,96]. This analysis shows a strong enrichment of this motif in our

scaRNA-seq data with 22.1% of TSS possessing the motif as opposed to 2.9% of UCSC "annotated TSS" (Figure 3.6i). To further validate the ability of Start-seq to detect genuine TSSs in vivo, I searched for the TATA box binding protein motif (TATAWR) in both annotated and observed TSS datasets (Figure 3.6ii), revealing an enrichment of the TATA box binding protein motif at the expected -28-32bp position relative to the TSS. The enrichment is low in both data sets but clearer when using scaRNA-seq derived TSS. It has been noted by others that promoter and TSS associated motifs such as the TATA box and Initiator motif which were defined in early *in vitro* studies are often poorly enriched *in vivo*[96]. The conserved TATA box motif (TATAAR), often described as a core eukaryotic promoter element, was found at only 3.5% of focussed human promoters[96]. This suggests the "grammar" of transcription initiation is poorly understood. Nevertheless, the detection of the conserved initiator motif at 22.1% of active TSSs in our data indicates that this motif is a common feature of active genes and that our assay detects these regions with higher accuracy than in existing annotations.

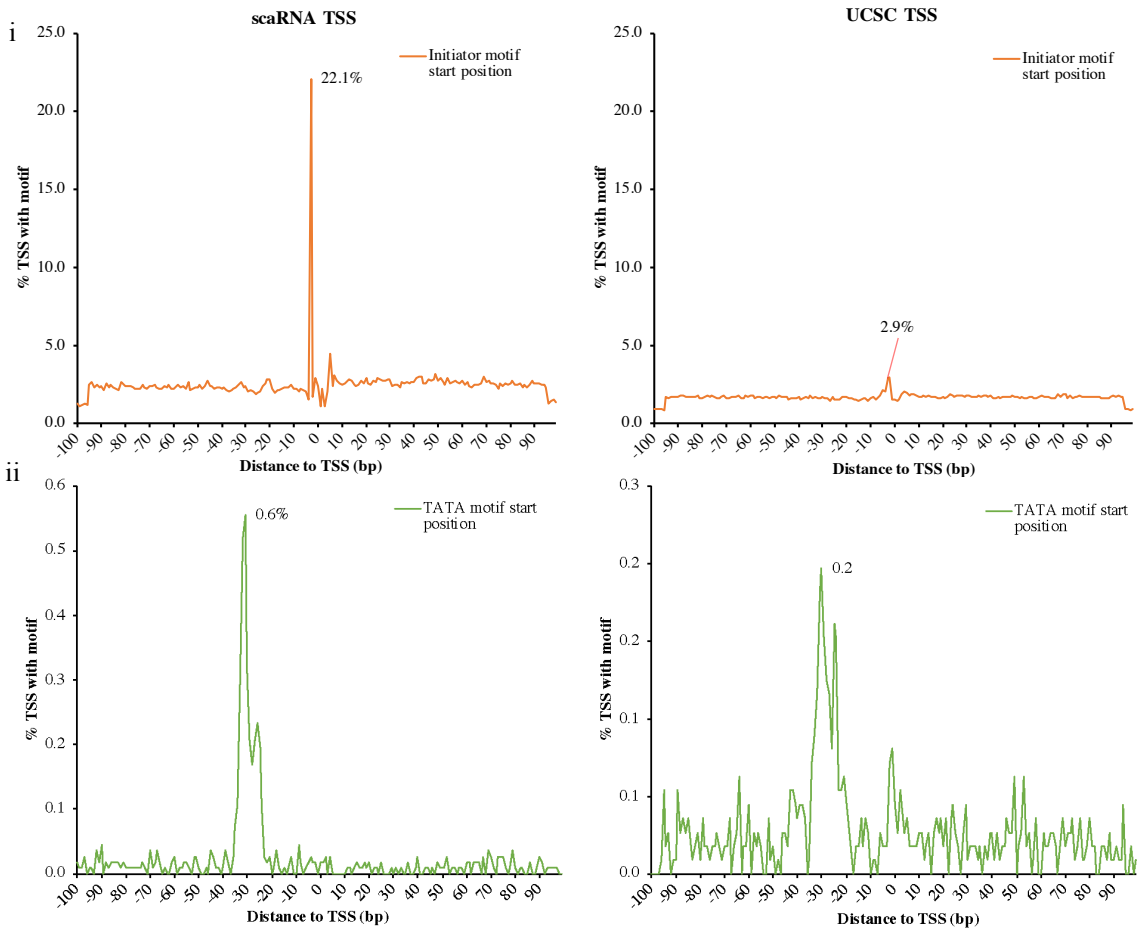


Figure 3.6) Initiator (i) and TATA box binding protein (ii) motif enrichment. Data plotted relative to scaRNA-seq derived (observed) and UCSC (annotated) TSS. Peaks correspond to the leftmost base in the expected motif (BBCABW = Initiator) and (TATAWR = TATA box binding protein) where B = C, G or T, W = A or T and R = A or G.

To investigate how the scaRNA-seq derived TSS coordinates relate to the commonly annotated starts of genes, I identified the annotated TSS within 500bp +/- relative to each observed TSS. If multiple annotated TSS were found then all hits were reported. This gave 18,930 annotated genes for 18,329 observed TSS. I plotted the distance of the annotated TSS relative to the observed TSS (Figure 3.7). This reveals that observed TSSs are often found downstream of their annotated positions in this cell type.

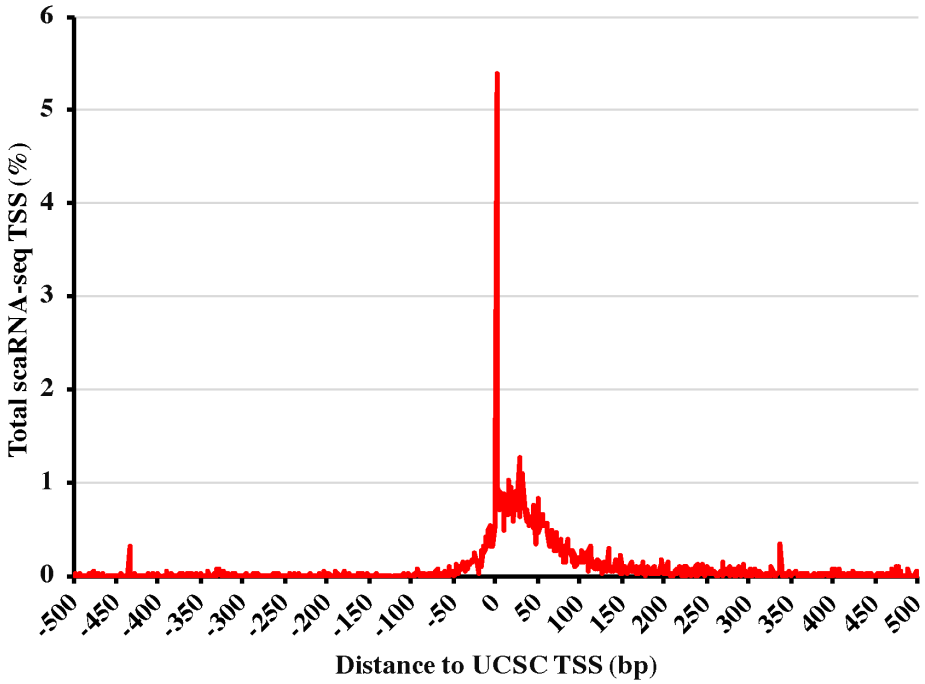


Figure 3.7) Observed versus annotated TSS position (all genes). Observed TSS derived from scaRNA-seq in K562 cells. Distance to UCSC TSS plotted for observed TSS which fell within 500bp (-/+) of an annotated UCSC TSS in a strand specific manner. Percentage (%) of total observed TSS is shown on the Y axis.

To ensure that our analysis of observed TSSs versus annotated TSSs was not skewed by annotated lists of genes containing "duplicate" isoforms in which different gene ID's have exactly the same chromosomal coordinates, I collapsed all duplicate gene isoforms, retaining individual IDs. This reduced the total number of annotated genes from 82,960 to 60,741. I then reanalysed the annotated versus observed positions of TSS. This revealed 11165 annotated TSS which fell within 500nts of an observed TSS and an almost identical skewed distribution of observed versus annotated TSS (Figure 3.8).

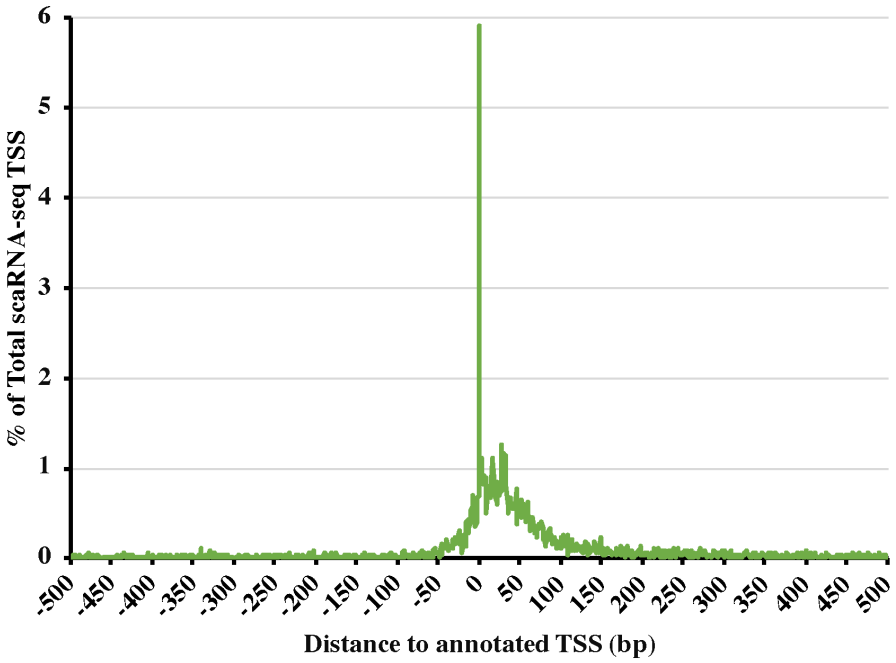


Figure 3.8) Observed versus annotated TSS position (unique genes). Observed TSS derived from scaRNA-seq in K562 cells. Distance to UCSC TSS plotted for observed TSS which fell within 500bp (-/+) of an annotated UCSC TSS in a strand specific manner. Percentage (%) of total observed TSS is shown on the Y axis.

I hypothesise that because cDNA libraries used in genome annotation projects are generated by sequencing and mapping of random RNA fragments generated from mRNA (i ived from different tissues[97], transcript isoforms with differing transcription start sites may be present (for example the *MYC* gene – Figure 3.5). Typically, the isoform with the furthest upstream start site is assumed to be the primary start site but no account of quantity of each transcript is taken into consideration. Therefore, the most 5' mapped position of all isoforms originating from a gene is systematically chosen as "***the*** annotated TSS"[85,98]. If a downstream TSS is the dominant one *in vivo*, signal arising from this TSS will conflate with that of the minor upstream TSS giving the impression of diffuse initiation rather the ii) punctate signals, for example at the *MYC* gene (Figure 3.9).

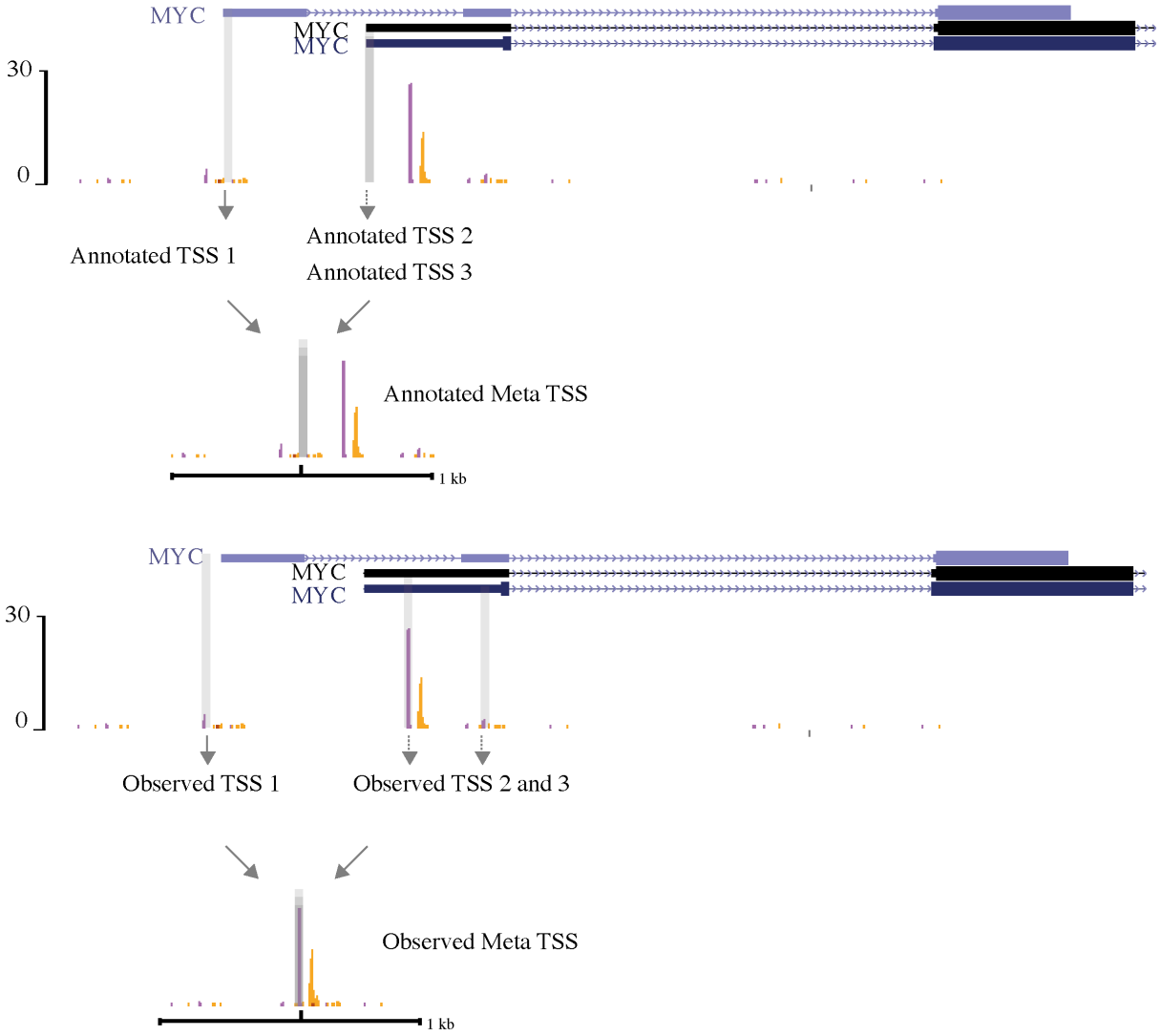


Figure 3.9) Example of TSS skew. Data is from K562. Counts on the Y axis are reads. Gene isoforms are UCSC annotated genes. i) Shows a meta analysis using annotated TSS coordinates. ii) shows meta analysis of the same data using observed TSS (scaRNA derived).

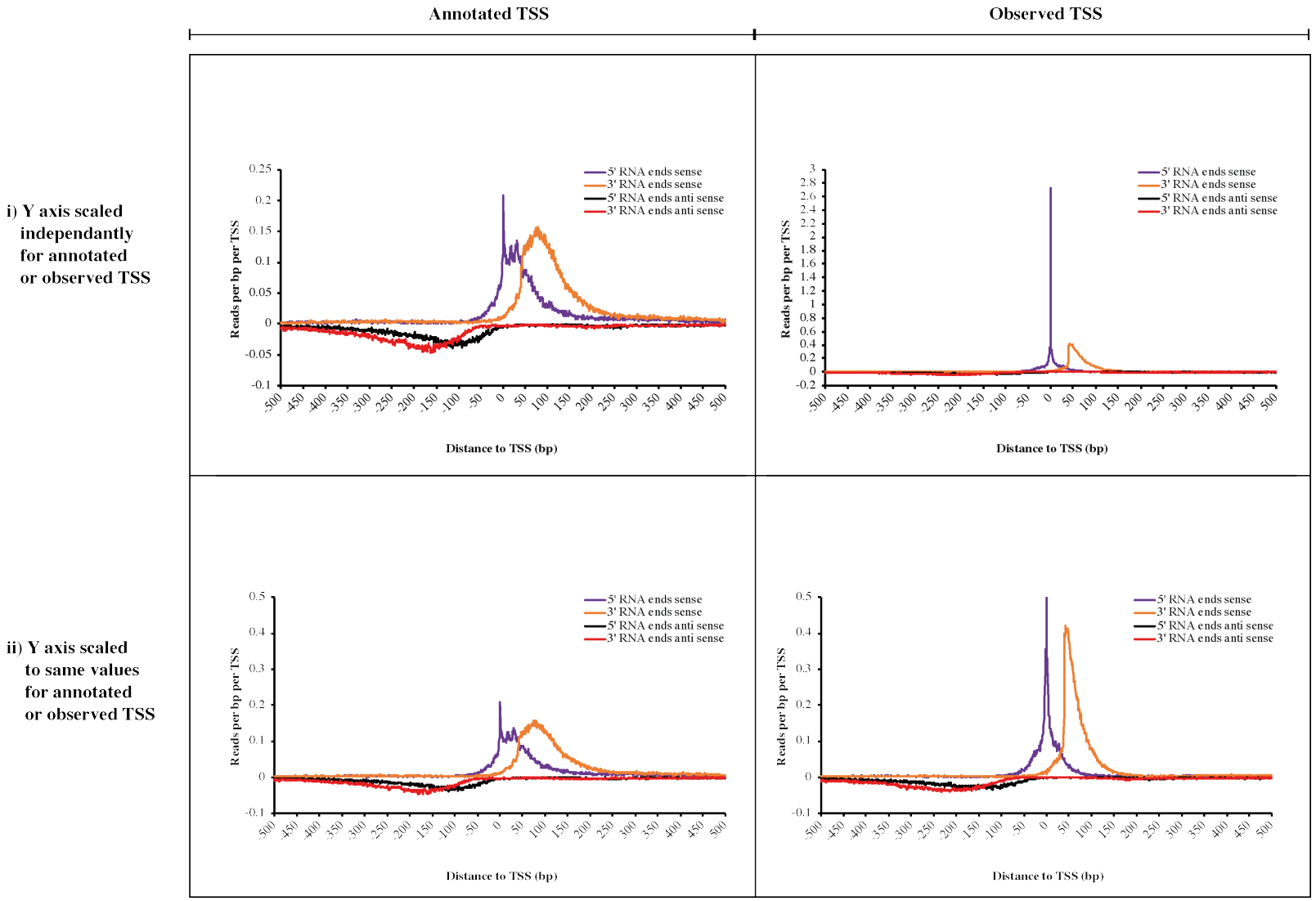
3.2.3 Average views of transcription using observed TSS differ from annotated TSS

Our understanding of promoter proximal transcription has been shaped by average views of RNA sequencing data plotted around annotated TSS lists (metaplots). The development of the GRO-seq assay by the Lis laboratory and application to human IMR90 cells

postulated that both pausing and divergent initiation are widespread features of promoters across the genome as a result of these metaplot style analyses[29].

Work by the Young and Adelman laboratories sequencing transcription start site associated RNAs in mES[34] and Drosophila S2 cells[35] respectively also support the idea of divergent transcription occurring at in metazoans and suggests that in the sense direction transcription initiates diffusely between 0->+50bp relative to the TSS[34].

Given the apparent skew in positions of observed versus annotated TSS, I wondered whether using observed TSS would alter the profiles seen in metaplots of RNA-seq data. Therefore, I plotted scaRNA-seq data against all annotated or observed TSS (Figure 3.10).


Annotated TSS
Observed TSS
i) Y axis scaled independantly for annotated or observed TSS
ii) Y axis scaled to same values for annotated or observed TSS
5' RNA ends sense
3' RNA ends sense
5' RNA ends anti sense
3' RNA ends anti sense
Reads per bp per TSS
Distance to TSS (bp)

For figure legend P.T.O.

Figure 3.10) scaRNA-seq metaplot – observed and annotated TSS. 5' ends of RNA molecules on the sense strand (purple), 3' ends of RNA on the sense strand (orange). 5' ends of RNA molecules on the anti-sense strand (black). 3' ends of RNA molecules on the anti-sense strand (red). Left hand side and right-hand side of the figure shows data aligned against annotated UCSC TSS coordinates or observed TSS coordinates (called from scaRNA-seq data and intersected with a UCSC gene) respectively. Row i) shows the data for each set of TSS (observed or annotated) with the Y axis scaled independently for each TSS annotation. Row ii) shows the same data as row i) but with the Y axis scaled to the same values for each set of TSSs. In both rows the Y axis represents read counts per bp per TSS normalised to 10^7 reads.

The annotated TSS metaplot (Figure 3.10, left hand side) reveals aspects of transcription at human promoters previously detected by others using combinations of nascent RNA sequencing methods[29,61]. These aspects include evidence of promoter proximal pausing (3' RNA ends accumulation between +40->+150bp downstream of the TSS with a maxima at around +70bp): diffuse initiation and, on average, high levels of antisense transcription relative to sense transcription.

The metaplot over the observed TSS (Figure 3.10 right hand side) looks very different. Using these TSSs, the majority of 5' ends of the RNA molecules align precisely to the 0bp position. The amount of diffuse 5' end signal around the TSS is lower, suggesting that what has been interpreted as common diffuse initiation around TSS may be an artefact encountered by using annotated TSS, which causes artefactual spreading of 5' signal around the TSSs. Nevertheless, transcription initiation may occur diffusely on some genes as indicated by the broad low-level signal of 5' RNA ends between -50 and +50bp relative to the TSS.

As the position of antisense transcription initiation is thought to occur over a range of distances relative to the sense TSS[61], when antisense initiation and pausing signals (5' and

3’ RNA ends) are plotted relative sense they appear as a diffuse signal extending from 0bp to -300bp relative to the TSS. When sense strand initiation signal is plotted adjusting for skew in annotated vs observed position, this results in the appearance of much lower level antisense transcription. We would need to further investigate the levels of antisense versus sense transcription on a single gene level to determine how antisense and sense transcription are related to levels of gene expression.

The current analysis reveals a punctate accumulation of 3’ ends of RNA molecules at the +42bp position on the sense strand with distributions tailing off towards +150bp. This is consistent with Pol2 pausing occurring commonly at this position. Importantly, the tailing of 3’ RNA ends fits well with the profile of nucleosome occupancy (Figure 3.11).

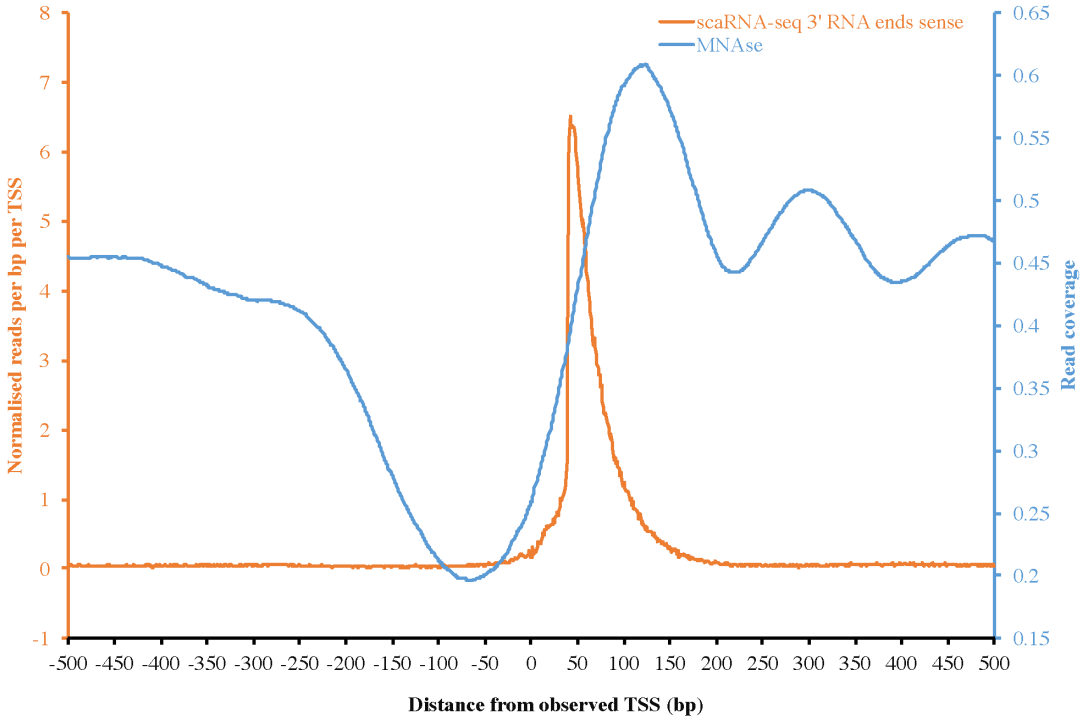


Figure 3.11) Nucleosome occupancy (MNase) versus scaRNA 3’ ends. Data plotted is for the sense strand only and is plotted relative to observed TSS which had an annotated UCSC TSS within 500bp (-/+).

This strongly suggests that the +1 nucleosome may represent a barrier to elongation: the level of 3’ RNA end accumulation (considered to be paused Pol2) upstream of this position

may represent the balance between initiating Pol2, pausing and Pol2 navigating the +1 nucleosome.

These results, together with single gene views at *MYC* and *HSPA1A* (Figure 3.5), suggest that scaRNA-seq is a powerful all-in-one method of examining the early stages of the transcription cycle and highlights significant issues when utilising non-tissue specific annotation coordinates of TSS. These results significantly alter our view of general transcription.

I reveal that the ubiquitous use of TSSs derived from multiple tissues in genome annotation projects has resulted in a loss of resolution when examining general transcription around promoters due to systematic inaccuracy in estimating the in vivo 5' end of a transcript. This has generated a subtle but significant scaling issue, which forces us to reconsider our general view of promoter proximal transcription in several ways:

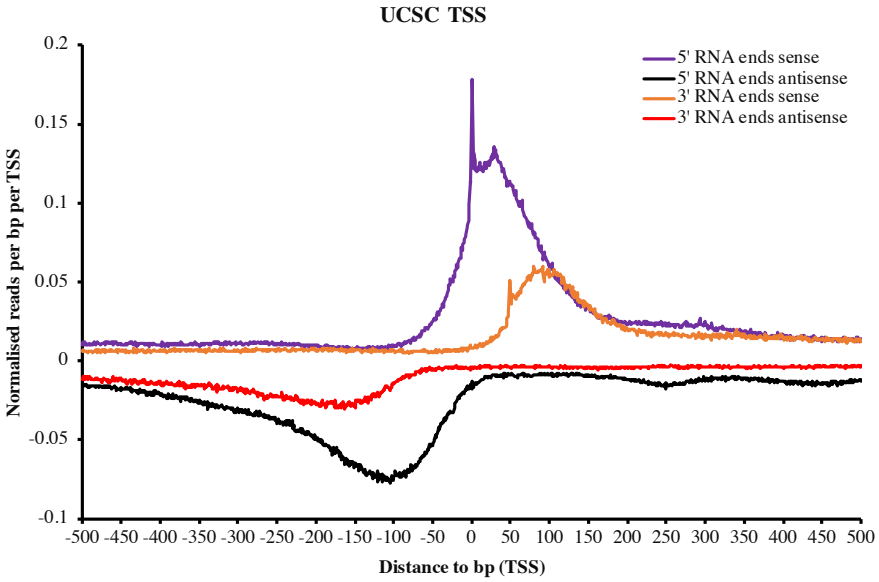
i) Transcription initiation is generally punctate rather than diffuse at the majority of genes.

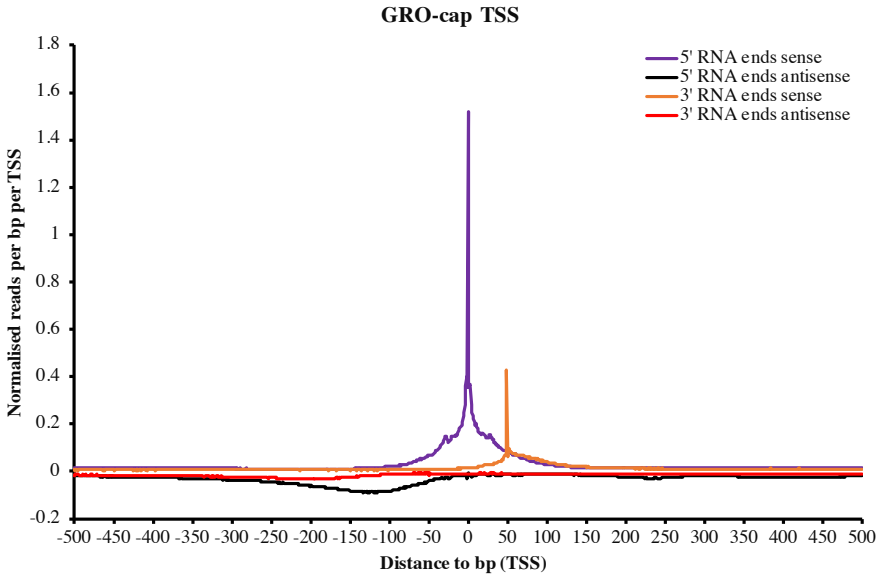
ii) Pol2 pausing begins very precisely at around +40 (in line with early *in vitro* work) and elongation after this position may be further limited by the +1 nucleosome.

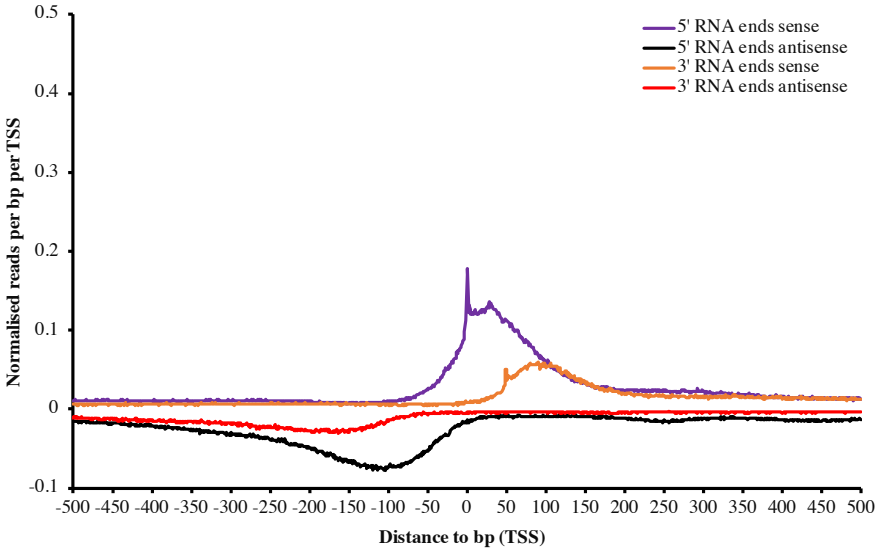
iii) The average level of antisense transcription initiation as compared to sense transcription may have been wrongly estimated using inaccurate annotated TSS coordinates.

3.2.4 Comparison of scaRNA-seq to GRO-cap/GRO-seq

Given the important changes to the view of general transcription when using scaRNA-seq data, I wanted to investigate whether I could recapitulate our findings using existing published datasets, which assays either the 5' ends of RNA (initiation) or 3' ends of nascent RNA (pausing) in independent experiments, in order to test whether scaRNA-seq can yield this information in one experiment. I chose to reanalyse published GRO-cap and GRO-seq data in K562 cells[61] as GRO-cap allows us to map active TSS and GRO-seq allows us to determine Pol2 pausing across genes. I called TSS from GRO-cap data in an identical manner to the scaRNA-seq data and then plotted GRO-cap and GRO-seq data relative to either the GRO-cap derived (GRO-cap TSS) or existing annotated UCSC TSS (Figure 3.12)


UCSC TSS
5' RNA ends sense
5' RNA ends antisense
3' RNA ends sense
3' RNA ends antisense
Normalised reads per bp per TSS
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-500 -450 -400 -350 -300 -250 -200 -150 -100 -50 0 50 100 150 200 250 300 350 400 450 500
Distance to bp (TSS)


GRO-cap TSS
5' RNA ends sense
5' RNA ends antisense
3' RNA ends sense
3' RNA ends antisense
Normalised reads per bp per TSS
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-500 -450 -400 -350 -300 -250 -200 -150 -100 -50 0 50 100 150 200 250 300 350 400 450 500
Distance to bp (TSS)


UCSC TSS (scaled Y axis)
5' RNA ends sense
5' RNA ends antisense
3' RNA ends sense
3' RNA ends antisense
Normalised reads per bp per TSS
0.5
0.4
0.3
0.2
0.1
0
-0.1
-500 -450 -400 -350 -300 -250 -200 -150 -100 -50 0 50 100 150 200 250 300 350 400 450 500
Distance to bp (TSS)

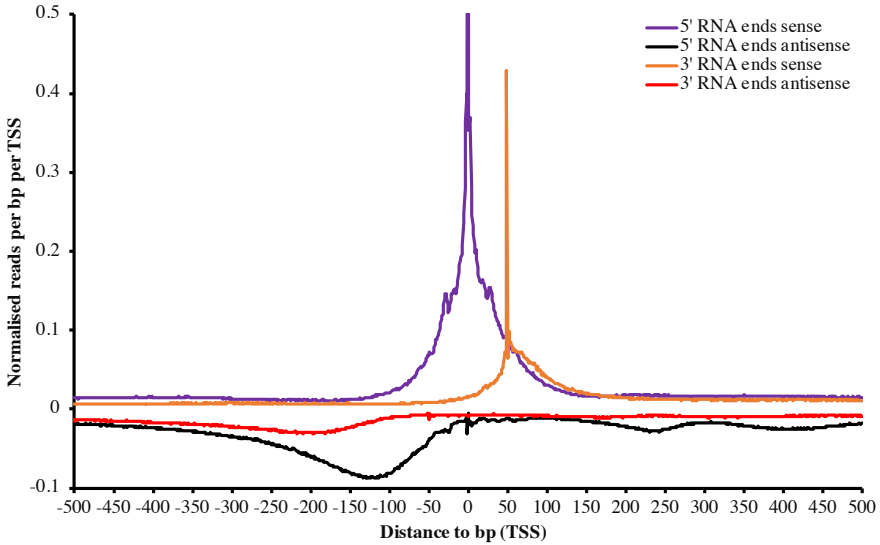

GRO-cap TSS (scaled Y axis)
5' RNA ends sense
5' RNA ends antisense
3' RNA ends sense
3' RNA ends antisense
Normalised reads per bp per TSS
0.5
0.4
0.3
0.2
0.1
0
-0.1
-500 -450 -400 -350 -300 -250 -200 -150 -100 -50 0 50 100 150 200 250 300 350 400 450 500
Distance to bp (TSS)

Figure 3.12) GRO-cap and GRO-seq metaplot – observed and annotated TSS. Annotated (UCSC) TSS (left) and Observed (GRO-cap derived) TSS (right). 5' ends of RNA molecules from GRO-cap data sets on the sense strand (purple), 3' ends of GRO-seq RNA on the sense strand (orange). 5' ends of RNA molecules on the anti-sense strand (black). 3' ends of RNA molecules anti-sense strand (red). Top row shows Y axis scaled independently for each TSS annotation set, bottom row shows the same scaling of the Y axis. In both rows the Y axis indicates reads per bp per TSS per 10^7 reads.

This analysis reveals a very similar skew of TSS positions and effects on meta-analysis when using observed GRO-cap TSS as when I use scaRNA-seq derived TSS:

i) The signal associated with transcription initiation (5' RNA ends sense) appears far higher at 0bp when using GRO-cap TSS due to correct plotting of 5' ends around TSS and removal of genes which dilute the signal, as they are present in the annotation list but not active in this tissue. The initiation signal becomes more centralised with less evidence of diffuse transcription initiation when using GRO-cap TSS. In GRO-cap data there appears to be more diffuse transcription initiation around the TSS relative to scaRNA-seq; this may result from performing the assays under non-physiological conditions, i.e. transcription initiation may be occurring across the broader region of open chromatin around the TSS as a result of loss of fidelity in Pol2 initiation.

ii) The 3' RNA ends on the sense strand have a maximum value at 49bp (7bp beyond the maximum value in scaRNA-seq data and highlighting that there is an accumulation of RNA ends at this position consistent with promoter proximal pausing. The position of this accumulation is slightly farther downstream than when using scaRNA-seq. I suspect this is due to the fact that GRO-seq run on reactions do not always terminate at the first incorporation of a modified nucleotide and often a run on of several base pairs may occur[23,29]. Alternatively, this could result from enriching for biotinylated RNA, where RNA

which contains a run of several modified nucleotides is recovered with greater efficiency than species with fewer incorporated modified nucleotides.

iii) The view of antisense transcription does not change dramatically when using observed TSS; however, the relative maximum heights of sense and antisense 5' RNA end signals appear changed suggest that using observed TSSs changes scaling in metaplots.

This analysis mirrors our scaRNA-seq analysis and suggests that tissue specific TSS annotation generated by scaRNA-seq and GRO-cap consistently differ from the annotated TSS positions.

A notable difference between the GRO-seq data and scaRNA data when aligned against observed TSS is the apparent initiation of transcription on the antisense strand over the region downstream of the TSS in GRO-seq data (see Figure 3.12, bottom right). I hypothesise that under run on conditions, spurious transcription initiation may be occurring on open chromatin. To test this I plotted MNase seq data relative to the observed TSS as defined by GRO-cap (Figure 3.13).

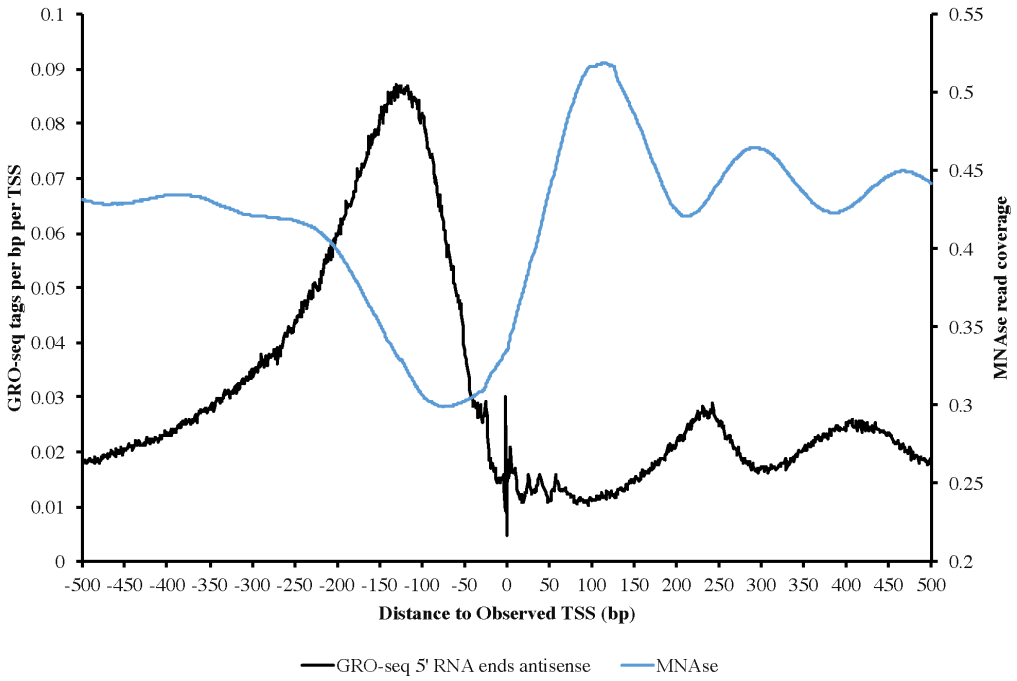


Figure 3.13) Nucleosome occupancy (Mnase) and GRO-seq metaplot. Data plotted relative to observed TSS (GRO-cap derived) in K562.

This analysis reveals enrichment of signals associated with transcription initiation on the antisense strand exactly at the positions where Mnase-seq signal drops (i.e. open chromatin). This highlights the potential non-physiological effects of performing run on assays on isolated nuclei or chromatin and suggests under GRO-seq run on conditions transcription initiation is occurring non-specifically, to a small degree, in open chromatin.

3.2.5 Comparison of scaRNA-seq RNA end distribution to mNET-seq and GRO-seq

As I was size selecting scaRNA and enzymatically modifying the resultant material, I wanted to ensure that the distributions of 3' RNA ends seen in our meta-analysis (Figure 3.10) were not a result of RNA degradation or size selection during library prep (for example preferential amplification of shorter versus longer RNA species), and rather represent positions of transcriptionally engaged Pol2 with nascent RNA loaded into its active site.

Mammalian native elongating transcript sequencing (mNET-seq)[32,33] was performed by Dr Taka Nojima in 5x10^6 K562 cells, allowing us to isolate RNA 3' ends which are associated with Pol2. I aligned mNET-seq data against our scaRNA-seq observed TSS and for comparison also plotted GRO-seq data relative to these positions as this assay detects nascent RNA[29]. Mapping of 3' RNA ends of mNET-seq and GRO-seq data relative to the observed TSS reveals an excellent fit with the 3' end of the scaRNA and suggests that, on average, the scaRNA represents nascent elongating RNAs engaged with Pol2 (Figure 3.14).

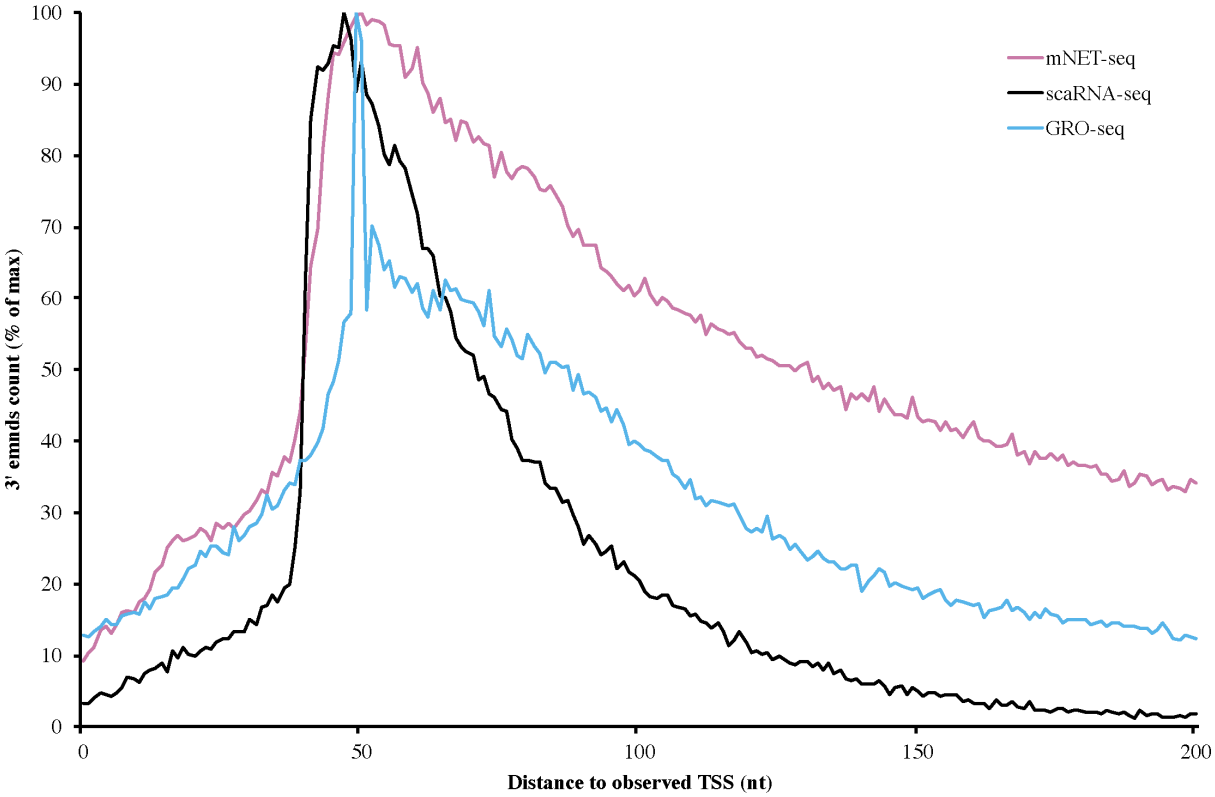


Figure 3.14) Metaplots of mNET-seq, GRO-seq and scaRNA-seq 3' RNA ends. Sense strand data only plotted relative to observed TSS.

Further support for the biological significance of the 3' end distributions identified in scaRNA-seq can be found by analysis of Pol2 ChIP-exo data in K562 cells[99,100]. ChIP-exo is a modified version of the ChIP-seq protocol with the key distinction being that after the immunoprecipitation of chromatin bound by a protein of interest, Lambda exonuclease is

used to digest one strand of the DNA bound to the protein in a directional manner. The Lambda exonuclease digests DNA until it is impeded by the bound protein. By mapping the sense and antisense strands of each digested DNA protein bound fragment, one can determine the distribution of the exonuclease digest stop points. The region occluded between the sense and antisense strand distributions represents the precise position of protein binding (Figure 3.15).

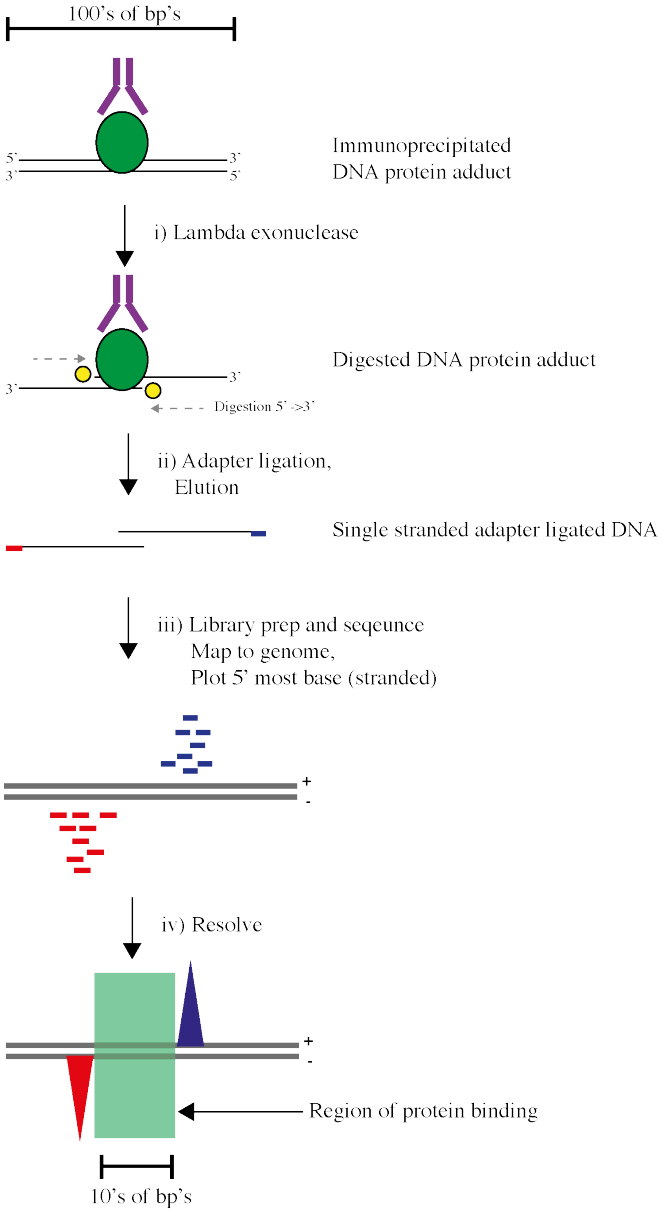


Figure 3.15) A simplified overview of the ChIP-exo method.

I re-analysed published ChIP-exo data in K562 relative to the annotated TSS and observed TSS (Figure 3.16). This analysis reveals a clearer fit of ChIP-exo data with the observed TSS as opposed to the annotated TSSs as shown by a tightening of the distribution, increasing the resolution with which we can resolve the expected position of Pol2 (between the maximum peak on the sense strand and minimum peak on the antisense). We can clearly see that the maximal region of Pol2 occupancy overlays the distribution of 3' RNA ends 40bp's downstream form the TSS, suggesting that *in vivo* this position is most likely where Pol2 pauses near the promoter, in line with earlier estimates. Intriguingly, we can also see there is a signal associated with the TSS (around 0bp) when using scaRNA-seq data. This probably represents Pol2 in the process of initiating transcription, which was previously undetectable due to mis-annotation of TSS and subsequent loss of resolution.

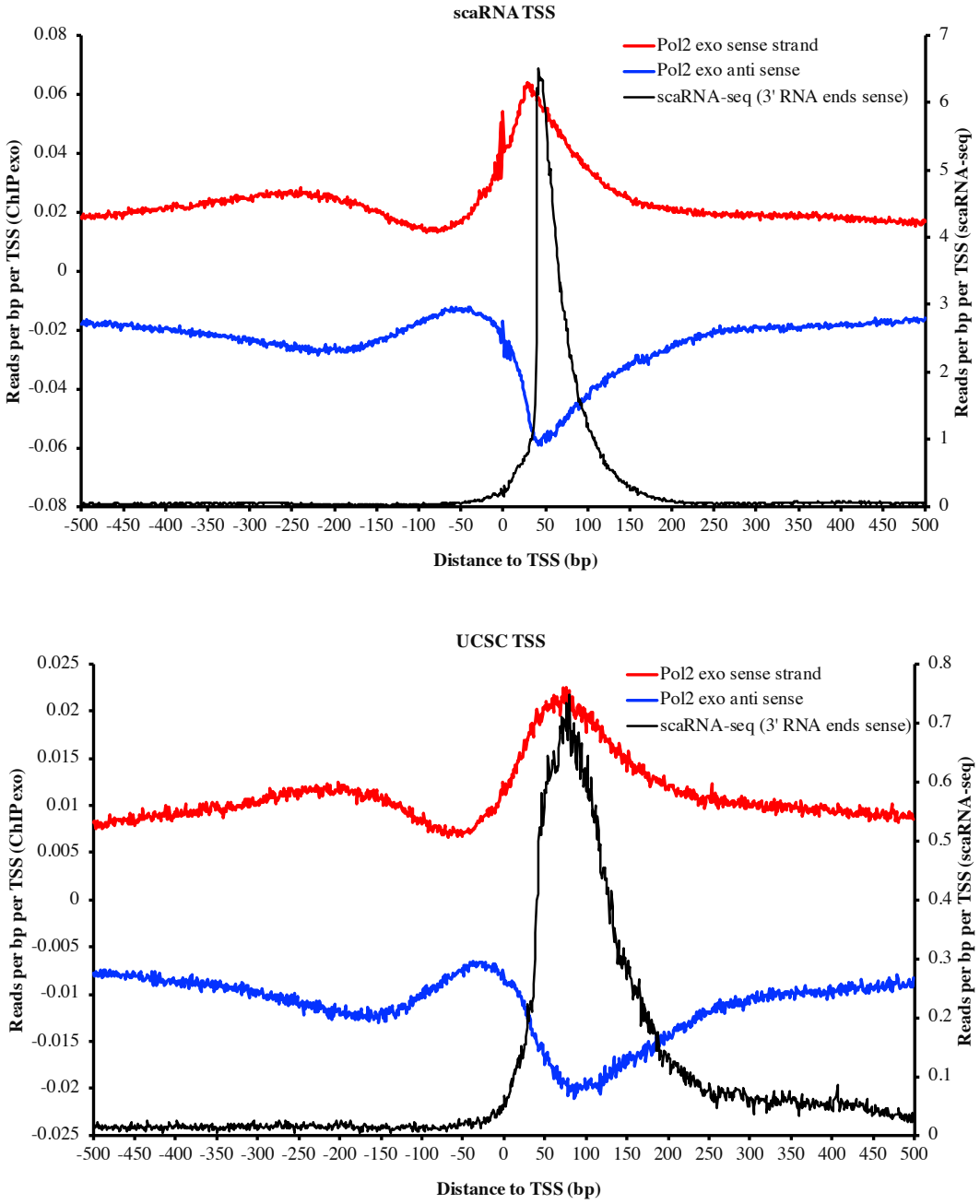


Figure 3.16) Pol2 ChIP-exo relative to observed and annotated TSS. Exonuclease digest stop points are plotted for sense (red) and anti-sense strands (blue). scaRNA-seq 3' RNA ends on the sense strand are plotted in black.

3.2.6 The relationship between Pol2 pausing factors and the +1 nucleosome

Pol2 pausing is thought to occur via the coordinated action of negative elongation factors (DSIF and NELF) on the early elongating Pol2 complex through interaction with nascent RNA and the Pol2 clamp domain[48,49,101,102]. Pausing is thought to be released when P-TEF-b is recruited to promoters and phosphorylates Pol2, NELF and DSIF. It is also thought that the +1 nucleosome may play a role in regulating pausing[23,103].

Understanding the distributions of these proteins *in vivo,* and nucleosome occupancy across promoters, may shed light on the interplay between these factors and, given the changes seen when analysing scaRNA and ChIP-exo data comparing observed versus annotated TSS, I wanted to investigate this. I used publicly available ChIP-seq data for SUPT5 (a component of DSIF), NELF-E (the RNA binding component of the NELF complex) and CCNT2 (a component of the P-TEFb complex), as well as MNase data for K562. I plotted datasets relative to annotated or observed TSS and scaled each data set to the maximum signal in the region -500 to +500 around the TSS (Figure 3.17).

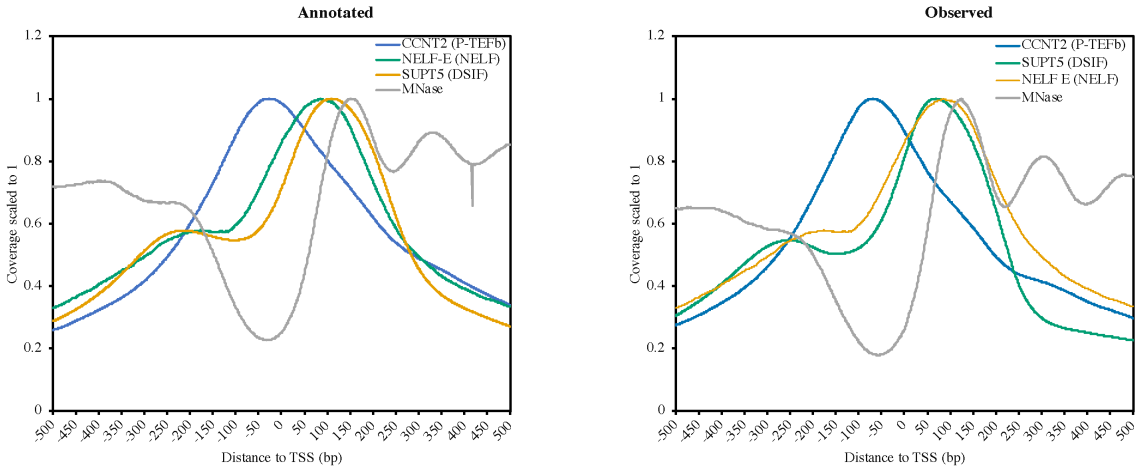


Figure 3.17) Metaplots of ChIP-seq for factors thought to regulate Pol2. SUPT5 (Orange), NELF-E (Green) and CCNT2 (Blue) and closed chromatin – nucleosome positioning (MNase-seq - Grey) over annotated (UCSC) or observed (scaRNA-seq TSS intersecting UCSC). Protein complexes which individual proteins are components of are shown in brackets for each subunit. Coverage values are normalised by

millions of reads and scaled to the value of the bp with the highest signal in the window +/-500bp's relative to the TSS (scaled to max signal).

Using either the annotated or observed TSS produces a plot with the maximum signal associated with the negative elongation factor complex components (SUPT5 and NELF-E) at around +50bp. Interestingly, P-TEFb appears maximally occupied just upstream of the TSS and in the case of both annotations. P-TEFb has been suggested to act on DSIF and NELF directly; however, it has been noted that the use of the CDK9 inhibitor DRB or analogue sensitive versions of this protein cause defects in the initiation of transcription[51,104,105] . These were assumed to be due to an indirect effect on Pol2 pause release but perhaps CDK9 has direct roles in initiation. Localisation directly over or upstream of the TSS would be suggestive of this but further experiments would be required to confirm a role.

The use of the observed versus annotated TSS changes the position of the maximal signal of each factor assayed. Using the annotated TSS (SUPT5, NELF-E and CCNT2) have maximal signals at +109, +85 and -27 respectively. With the observed TSSs, these change to +71, +41 and -69. This would mean maximum the NELF signal at observed TSSs precisely co localises with the 3'RNA ends as revealed by scaRNA-seq, this is in line with the NELF-E protein's role in RNA binding[106]. DSIF is located at the leading edge of the 3' RNA ends and both factors co-localise with the +1 nucleosome which has a maximum signal of +151 or +124bp in the annotated and observed annotations respectively. These positions are consistent with the role of DSIF and NELF in recognition and binding to an early Pol2 elongation complex downstream of the TSS as suggested by *in vitro* studies[47–49] and recent crystal structures[41,42]. How nucleosome and pausing factor positioning relate to transcriptional regulation remains to be understood. Of note, the +1 nucleosome position

fits well with the decreasing density of 3' RNA ends beyond +40bp's (Figure 3.13), suggesting nucleosome bound DNA may play a potential role in delimiting elongation just after initiation.

This analysis along with our previous analysis of Pol2 ChIP-exo and nascent RNA position suggests that Pol2, the 3' ends of nascent RNA and Pol2 pausing factors all co-localise within or around +40bp relative to the TSS and this is therefore likely the most common position of pausing *in vivo*. This estimate also fits with single gene examples such as *MYC* and *HSPA1A* (Figure 3.5).

3.3 Discussion

3.3.1 scaRNA-seq a combinatorial method to investigate promoter proximal transcription

I have described a method (scaRNA-seq) to identify sites of transcriptional initiation and promoter proximal regions where RNA ends accumulate. I show that mapping TSSs using scaRNA-seq enriches for the Initiator motif at a far higher level than in previously annotated lists of TSS. Using scaRNA-seq I can derive thousands of TSS *de novo* using a simple peak calling approach utilising the Homer package[79]. I also reveal that annotated TSS lists do not accurately represent sites of transcriptional activity; this is due to a systematic skew in annotation of TSS upstream of their start positions *in vivo*. In genome annotation projects the 5' most coordinate of a putative gene isoform is assigned as the TSS, regardless of the frequency of its use[85,98,107]. This has inadvertently introduced a systematic bias in previously reported genome annotation projects by calling the most 5' coordinates of TSSs rather than the most frequently used TSS coordinates *in vivo*.

3.3.2 Implications for our understanding of promoter proximal transcriptional regulation

The general view of transcription at mammalian promoters is that Pol2 pausing and transcription initiation are often diffuse and occur over wide regions relative to the TSS (-50 to +50nt and +20 to +60nt respectively). It is also believed that upstream antisense transcription is relatively common compared to sense transcription[24,29,34,61,108]. Our analyses reveal that due to the systematic bias in the assignment of annotated TSS coordinates, the current view of average transcription has been skewed. By adjusting for this bias using TSSs derived *in vivo* from scaRNA data and realignment of published nascent RNA, ChIP-seq, ChIP-exo, MNase-seq against these observed TSSs, I modify our view of promoter proximal transcription in several important ways.

First, we see little evidence of broad initiation; instead, most 5' RNA ends align to a few bps around the site of maximal transcription initiation. This is in contrast to previous work classifying promoters as either focussed (punctate) or diffuse (broad), which suggested that these two types of promoter are distinct and common across the genome[108]. More recent work has, however, suggested that the two modes of initiation may occur on the same promoters, suggesting the distinction between diffuse and focused transcription initiation may vary on individual promoters and may not be used to imply specific functions of TSSs genome wide[109].

Second, we find evidence of Pol2 pausing at a very precise location (+40bp); this is in line with early work *in vitro* and later genome-wide studies[23,24]. Our analysis adds a high-resolution view of this process that has not been achievable *in vivo* before due to blurring of signal brought about by the analysis of TSSs from a wide range of tissues. I show that

3’ RNA ends co-localise with Pol2 pausing factors (SUPT5 and NELF-E) and are located in the predicted precise location of Pol2 binding (as shown through ChIP-exo). I also reveal that RNA elongation is potentially limited by the +1 nucleosome, highlighting a decay in the signal of 3’ RNA ends beyond +40bp towards +150bp; this is an excellent fit with the idea that the +1 nucleosome acts as a barrier to elongation[38,39,108].

Third, we reveal that using observed TSS causes changes to the apparent level of punctate sense strand transcription initiation relative to antisense transcription initiation[29,34]. However further analysis on a single gene level may help to elucidate the relationship between sense and antisense RNA production genome-wide since it remains unclear whether this type of transcription is important as a general regulatory mechanism or a by-product of sense transcription.

One key point to note is that scaRNA-seq and all other methods for mapping TSS (CAGE, GRO-cap, Start-seq) rely on selection of m7G capped RNA and therefore uncapped RNA species are not detected. Until recently it was assumed all eukaryotic RNA was either uncapped in the case of rRNA or capped by the m7G moiety. However, recent work has revealed the presence of a large fraction of non-canonically capped RNA in human cells[110,111]. The function of these transcripts is yet to be determined. It would be interesting to examine the promoter proximal transcription properties of these genes using a modified scaRNA-seq protocol to see if they differ from canonically described m7G capped RNA promoters.

Chapter 4

The prevalence of Pol2 pausing *in vivo*

Chapter 4 – The prevalence of Pol2 pausing *in vivo*

4.1 Introduction

4.1.1 Determining the extent of promoter proximal Pol2 pausing.

Our modified meta-analyses using tissue-specific TSS reveal that transcriptionally engaged Pol2 often accumulates at the +40bp position and co-localises with DSIF and NELF. This is in line with the classic model of P-TEFb mediated pausing, occurring commonly genome wide[22]. However, there is much debate about the prevalence of this phenomenon[23,24], as depletion or mutation of DSIF and NELF *in vivo* has been shown to result in a variety of tissue-specific, non-lethal phenotypes[52]. Knockdown of these proteins causes relatively modest changes in Pol2 distribution across genes[22] and gene expression[53], casting some doubt on the role of these proteins in transcription[52] *in vivo*.

The nucleoside analogue 5,6-dichlorobenzimidazole 1-β-D-ribofuranoside (DRB) is thought to inhibit Pol2 pause release by competitively inhibiting the kinase activity of CDK9. This prevents the phosphorylation of both Pol2 and the negative elongation factors NELF and DSIF. This phosphorylation allows Pol2 to become "elongation competent," promoting the transition from paused to productive elongation[51,52]. Through the use of DRB or Flavopiridol (FP), another inhibitor of CDK9, others have suggested Pol2 pausing is a ubiquitous regulatory step in Drosophila[38], mouse[12] and human cells[64].

An additional complicating factor in determining the prevalence of Pol2 pausing is that in order to identify and quantify changes in pausing, previous studies have relied on a metric known as the travelling ratio (TR) or pausing index (PI), which is a relative measurement of Pol2 density at the promoter compared to the gene body, and this defines paused genes as having a TR of above 1[22,23]. This method has been applied to data obtained by ChIP-seq and GRO-seq and such studies conclude that up to 90-95% of genes are

paused[12,22,23,64,90]. However, it has also been shown that the TR does not correlate with gene expression[22] and the thresholds which determine whether a gene is considered to be "paused" or not are determined somewhat arbitrarily. These methods are confounded by the poor resolution or large promoter proximal windows over which read density is calculated. These regions contain the TSS, proposed site of Pol2 pausing (+20-60bp) and other promoter proximal features such as the +1 nucleosome. Comparing read density changes over these regions to regions lying downstream is often interpreted as changes in pausing but it could equally be due to changes in initiation or navigation of the +1 nucleosome by the transcriptional machinery.

Given the importance of defining the key stages in the transcription cycle for understanding differential gene expression, we need to distinguish between Pol2 pausing and other stages of early transcriptional regulation to truly identify the prevalence of Pol2 pausing *in vivo*.

4.2 Results

4.2.1 The travelling ratio

The most commonly used metric to assess the extent of promoter proximal enrichment of Pol2 (PPEP) is the ratio of promoter proximal to promoter distal Pol2 density. Changes in TR are routinely used to infer changes in Pol2 pausing and have been used to investigate the potential roles of several proteins such as MYC, DSIF, NELF and PAF1 in this process[22,56]. More recently, changes in the pausing ratio have been used to suggest that enhancers regulate the release of Pol2 pausing[56].

To investigate whether PPEP is common in K562 cells, I calculated the TR for unique UCSC annotated genes from published ChIP-seq data, using a Pol2 antibody (N20) which detects all forms of Pol2, yielding 49,279 TRs **(**Figure 4.1). I identify PPEP (TR > 1) at

78% (38178/49279) of genes. 46% (22810/23820) of genes have PPEP greater than or equal to 2 (TR >= 2); this indicates that PPEP is common across the genome in line with previous estimates using this metric[23].

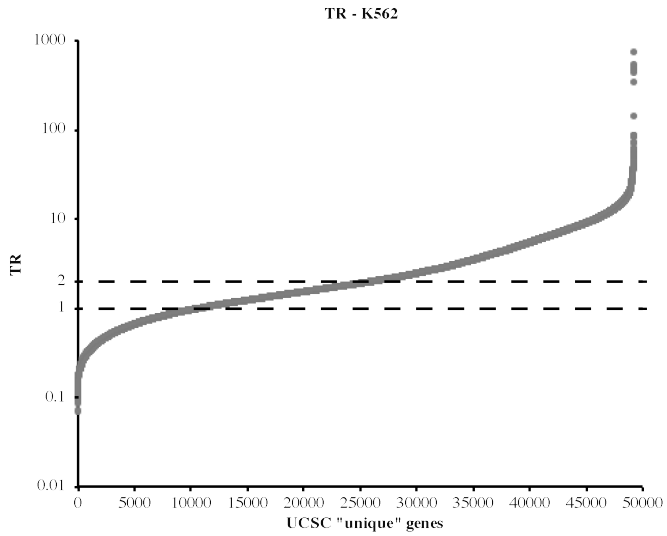


Figure 4.1) Pol2 TR values for all UCSC genes in K562 cells. TR values of 1 (equal promoter versus gene body read density) or 2 (double promoter versus gene body read density) are indicated as dotted lines.

4.2.2 The relationship between TR and gene expression

I investigated whether the TR is related to gene expression. If Pol2 pausing is common and leads to lower gene expression one might expect to see that genes with high TR would in general also have low expression values. I determined RPKM values as a measure of expression from our polyA+ RNA-seq data from K562 cells and plotted TR versus expression for all unique annotated UCSC genes as a density scatter (hexbin) plot (Figure 4.2).

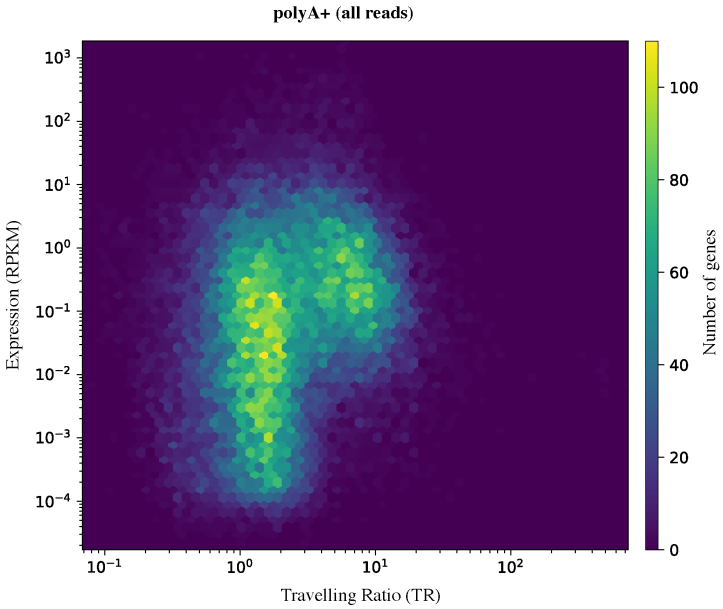


Figure 4.2) TR versus RNA expression for all unique annotated UCSC genes (n= 49,279). Pearson correlation coefficient (R^2) = 0.24.

There appears to be two groups of genes: one with high TR and mid-range to high expression (0.1-10 RPKM) and another group with lower TR and a wider range of expression values. High TR values do not correlate with low expression as one would expect if PPEP was a rate limiting step in transcription. It is more common for genes with high expression levels to have high TR values, indicating that PPEP is associated with higher levels of transcription. This would suggest recruitment or initiation, not Pol2 pausing, are the main limiting steps in the transcription cycle.

4.2.3 Does the TR correlate better with nascent transcription?

Nascent transcription more faithfully reflects transcription than steady state levels of RNA and therefore I compared PPEP with nascent RNA. I realigned RNA-seq data using Bowtie (a non-splicing aware mapper), essentially treating our data as though it were genomic DNA, a strategy employed by other nascent RNA seq methods such as GRO-seq[29] and mNET-seq[32,33]. I also limited the maximum distance between mapped reads in a pair to 1000bps; this enriches for pre-mRNA as mature mRNA has undergone splicing, bringing pieces of transcribed RNA together which may be located several kb apart in the linear

genome. I isolated reads mapping to introns but not exons by sequential intersection with annotated lists of introns and exons from the UCSC table browser, leaving only intronic RNA which may better represent nascent, unspliced or pre-mRNA transcripts. A similar strategy has recently been used to study the effect of transcription and splicing inhibitors in Hela cells[112]. I plotted the original aligned data and the intronic enriched data against all refseq genes using the tool ngsplots[81] (Figure 4.3). I had to use the refseq gene list as ngsplots does not contain the UCSC gene list. However, refseq gene predictions are aligned to the UCSC genome build before being retained as "UCSC genes", meaning the refseq gene list is a slightly less comprehensive list of gene predictions than UCSC. Nevertheless, it still contains around 90% of the total number of UCSC genes[98].

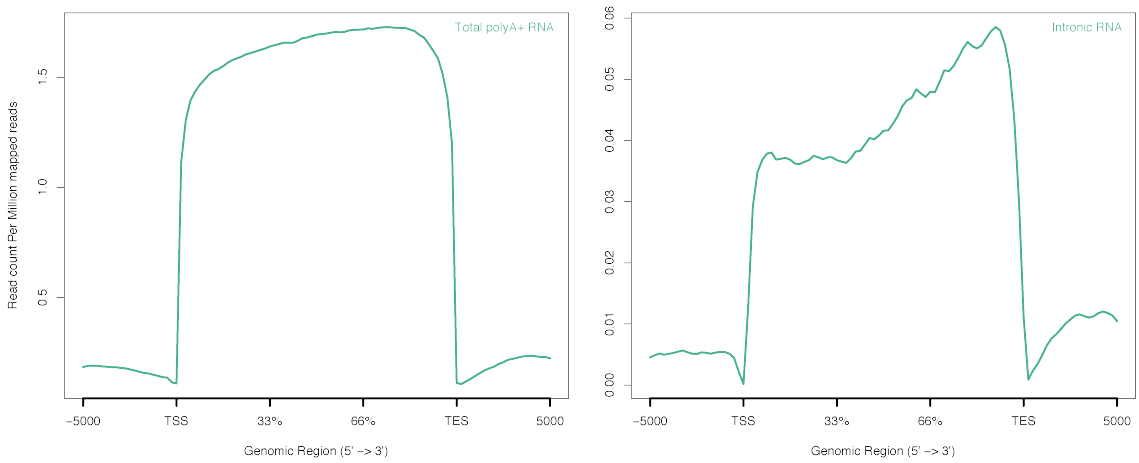


Figure 4.3) RNA-seq read gene coverage metaplots. (Left) polyA+ RNA-seq (all reads). (Right) polyA+ RNA-seq (intronic reads only) over all refseq annotated genes.

Although read coverage is lower in the intronic RNA-seq data, we can still clearly see coverage over all annotated gene regions, indicating that these data provide a measure of "nascent" transcription using simple bioinformatic extraction from existing RNA-seq datasets. To further investigate this, I analysed published 4sU RNA-seq in K562 cells, which identifies nascent transcription by sequencing transcripts metabolically labelled with

4-thiouridine (4sU) *in vivo*[113]. I compared gene expression (RPKM) between 4sU and poly+ RNA (all reads) and polyA+ RNA (intronic reads only) over all UCSC unique UCSC genes (Figure 4.4)

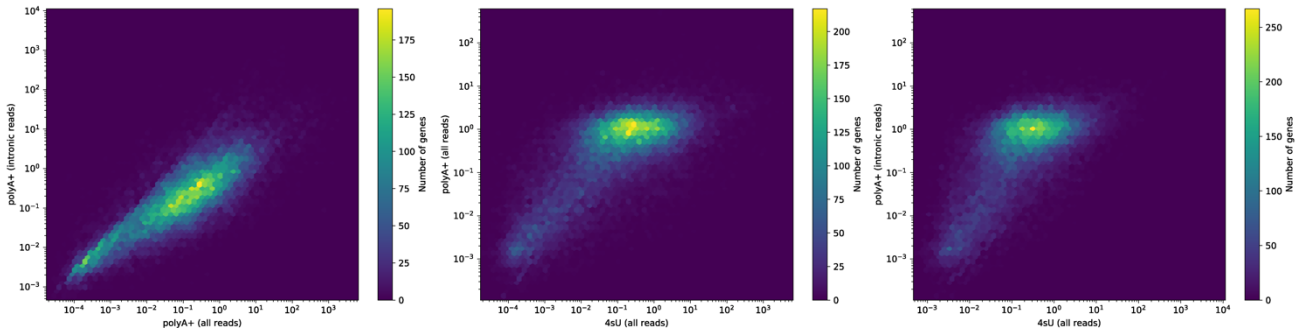


Figure 4.4) Comparison of RNA-seq methods. Gene expression values plotted against each other for each unique UCSC gene. From left to right; polyA+ (intronic reads) versus polyA+ (all reads), polyA+ (all reads) versus 4sU, polyA+ (intronic reads) versus 4sU. R^2 values left to right; 0.85, 0.74 and 0.68.

In this analysis, all data comparisons yielded a positive correlation between expression values with high R^2 values (0.85, 0.74 and 0.68), indicating good correlation and low variance in the data. The clearest correlation with the least variance is polyA+ (intronic reads) versus polyA+ (all reads), confirming that intronic read extraction does not reduce the ability to detect changes in gene expression. Comparisons between 4sU and polyA+ (all reads or intronic reads) are very similar, indicating that intronic RNA can be used as a measure of nascent transcription in a similar manner to 4sU.

I performed this analysis on polyA+ selected RNA samples to evaluate the production of full length (pre-mRNA); however, it would be interesting to compare the correlation of intronic reads extracted from total RNA-seq containing polyA+/- RNA, as this material may have a greater enrichment for shorter non poly adenylated transcripts which represent RNA in the process of transcription. Thus, a more "nascent" view of transcription may be achieved.

To determine if TR was a better indicator of nascent RNA production, I plotted the TR's derived from Pol2 ChIP-seq data against the intronic RNA expression (RPKM) for each gene (Figure 4.5).

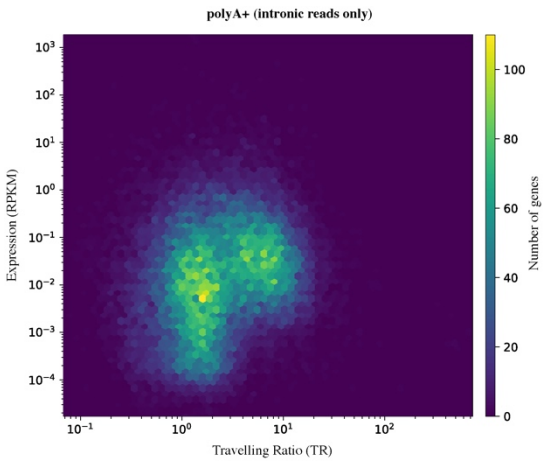


Figure 4.5) TR versus polyA+ RNA (all reads). All unique UCSC genes displayed. Pearson correlation coefficient (R^2) = 0.21

As for polyA+ RNA (Figure 4.2) there is no correlation between expression and TR, but there appear to be two groups of genes separated by high/low travelling ratios and either medium to high, or a wider range of, expression values. These results suggest that whether we assay total polyA+ RNA or polyA+ intronic, the TR does not predict the level of gene expression.

This result is perhaps not surprising as the relationship between the ratio of PPEP and gene body Pol2 versus levels of expression is a complex one, with examples of highly expressed genes showing a range of levels of PPEP (Figure 4.6).

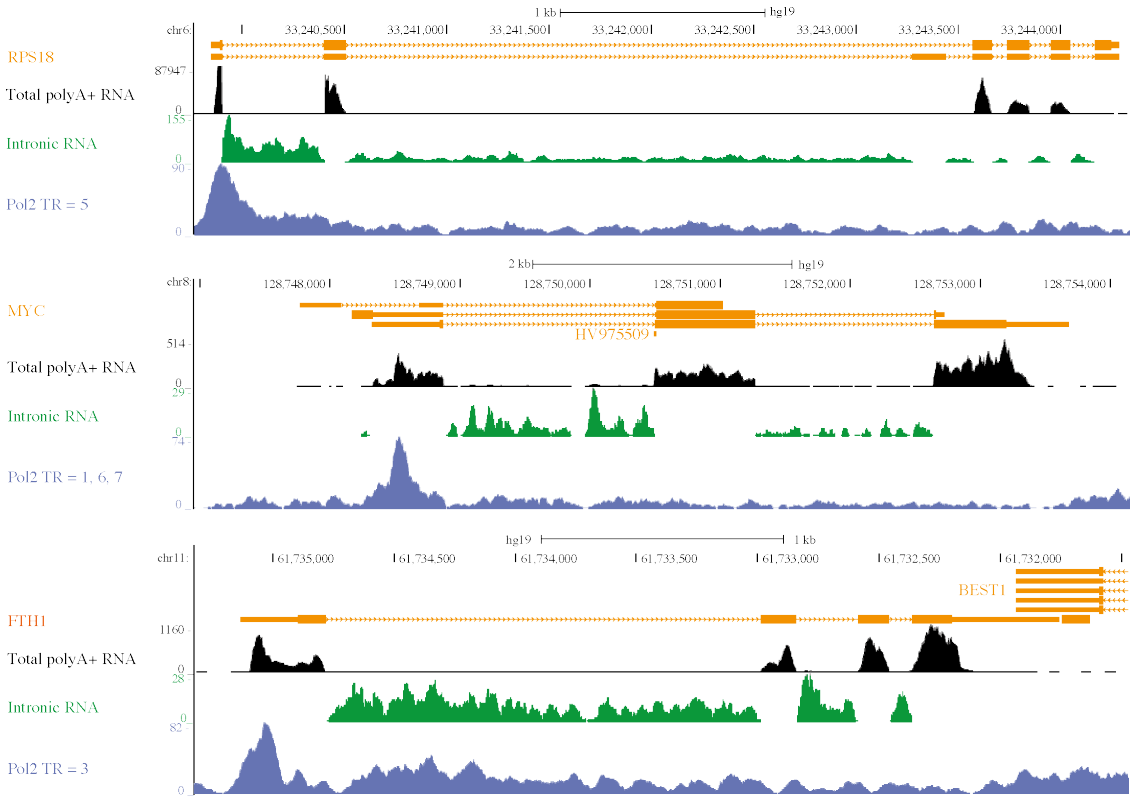


Figure 4.6) Examples of TR at different genes. All traces are displayed with raw (un normalised) read counts. Traces show polyA+ (all reads) - (top), polyA+ (intronic reads only) – (middle) and Pol2 ChIP-seq (bottom). TR values calculated from Pol2 ChIP-seq are indicated by the trace name.

Dividing the density of Pol2 at the promoter versus the gene body obscures the relationship between Pol2 density and transcription, and therefore TR is not related to the level of gene expression.

4.2.4 Total Pol2 versus gene expression

I wanted to explore the relationship between Pol2 density and RNA expression levels more deeply to determine how quantitatively Pol2 ChIP-seq could represent gene expression. I calculated total Pol2 density (reads/bp) across all unique UCSC annotated genes as well as expression values (RPKM) from polyA+ RNA-seq (all reads) and polyA+ (intronic reads

only). To reduce noise from the data I removed genes with extremely low (<100) or extremely high (> 10000) read counts per gene before plotting as a scatter density (hexbin) plot (Figure 4.7).

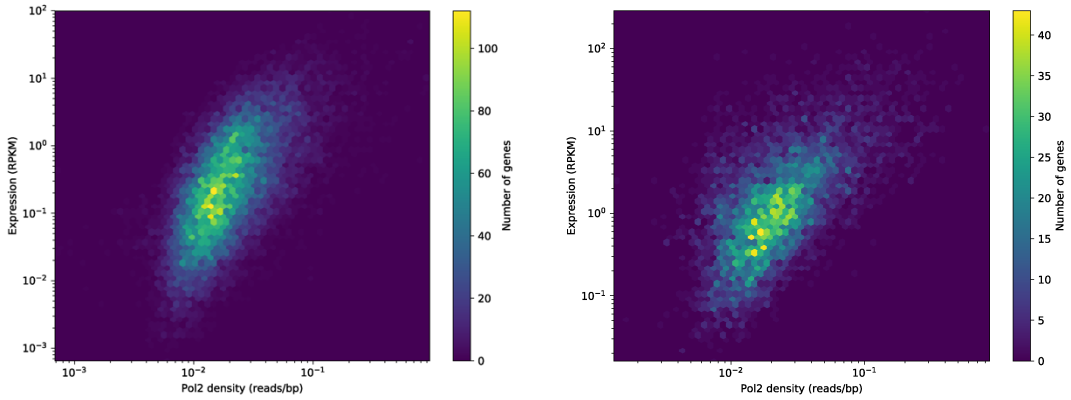


Figure 4.7) Expression versus Pol2 density. (Left) Pol2 density versus expression (RPKM) from polyA+ (all reads), $R^2 = 0.63$. (Right) Pol2 density (reads/bp) versus expression (RPKM) from polyA+ (intronic reads only), $R^2 = 0.60$.

We now find a positive correlation between total level of Pol2 and expression of polyA+ RNA (all reads) as well as intronic reads extracted from polyA+ data. The R^2 values (both >0.6) suggest some variance showing that similar levels of RNA expression can have varying enrichments of ChIP-seq and *vice-versa*. A relatively small range of Pol2 density (0.01 → 1 reads/bp) fits with a large range of RNA expression (0.0001→ 1000 RPKM) suggesting the sensitivity and dynamic range of ChIP-seq may be limited compared to RNA-seq.

Another important parameter influencing the ability to detect gene expression is RNA stability. As different RNAs decay at different rates, the levels of RNA are not all related to levels of Pol2[1]. This means comparing these types of data directly may never be truly quantitative as ChIP-seq could never provide a read out of transcript stability.

4.2.5 Resolution limits of ChIP-seq for studying transcriptional kinetics

We see a good correlation between RNA expression and levels of Pol2 over genes, and show the PPEP affects a majority of genes. However, there are significant caveats to using ChIP-seq when analysing the fine detail of transcription. The most significant of these is resolution. Theoretically, ChIP-seq can resolve proteins that are located tens of bp apart as it is limited only by the size of the DNA fragment immunoprecipitated[17,50]. However, in practice, a maximum resolution of 175bp is believed to be the best achievable by standard ChIP methods based on advice from antibody manufacturers (https://www.abcam.com/protocols/chip-tips-and-tricks-written-by-dr-andy-bannister). Given that initiation, Pol2 pausing and navigation of the +1 nucleosome all occur within ~200bp of each other on the linear DNA template[23,61], it is clear that we cannot delineate precise positions of these events using ChIP. To highlight this further, I produced a meta-analysis of 3' ends of sense strand scaRNA and Pol2 density over observed TSS. The resulting ChIP-seq plots is clearly inadequate to resolve the precise positions of initiation, Pol2 pausing, upstream antisense transcription and the +1 nucleosome (Figure 4.8).

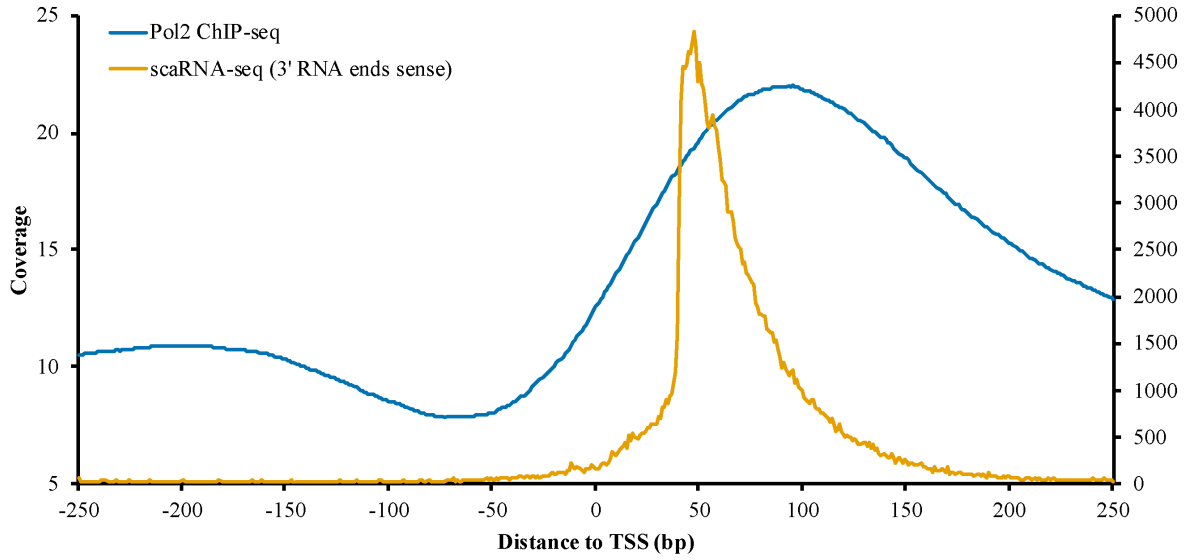


Figure 4.8) Pol2 and scaRNA over observed TSS. Only sense strand data plotted relative to observed (scaRNA-seq derived) TSS in K562.

Peaks of Pol2 enrichment are visible, upstream of the TSS consistent with antisense transcription at ~-250→-200bp and just downstream of the TSS ~+100nt. Coincident with the TSS, there is also a significant level of Pol (higher that the maximum peak on the upstream antisense region); however, we cannot clearly distinguish a peak at the TSS itself. This signal is merged with that of the downstream Pol2 accumulation at the position of promoter proximal pausing[23]. Our earlier analysis (Figure 3.16) suggested that, using data derived from high resolution Pol2 ChIP-exo, we see evidence of a distinct Pol2 peak at the TSS, indicative of Pol2 loading. Recent work using the ChIP-exo variant ChIP-nexus reveals distinct peaks of Pol2 can be resolved tens of base pairs apart near the TSS and that these distributions are sensitive to transcription inhibition by triptolide (an inhibitor of transcription initiation[51]). This suggests dynamic variation in Pol2 density at these positions is linked to transcriptional regulation, but this cannot be resolved with standard ChIP-seq[104].

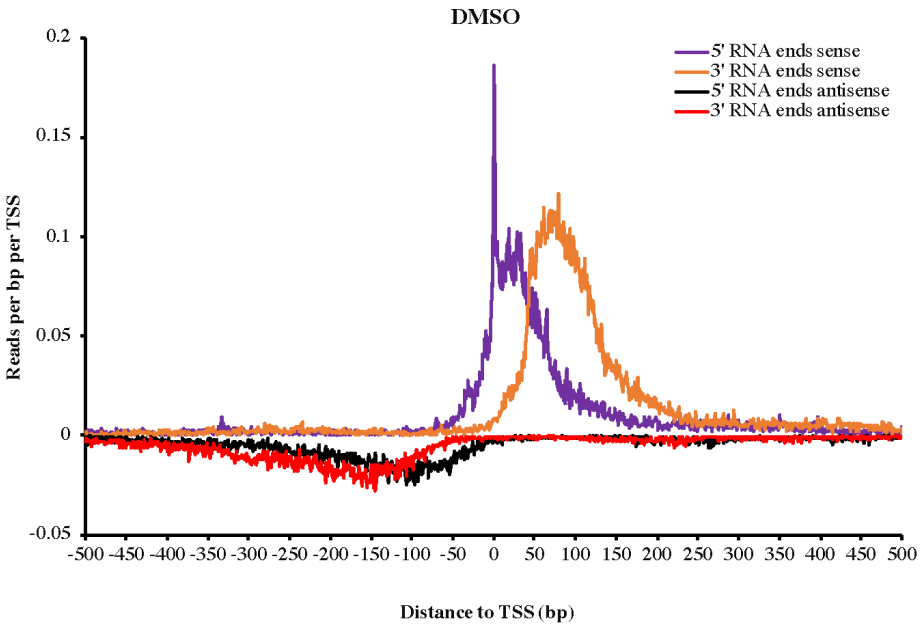
In conclusion, Pol2 ChIP-seq provides some predictive power in determining levels of gene expression but the TR is limited and can only be used to determine if PPEP versus downstream enrichment is common, which appears to often be the case. However, the resolution limits of ChIP-seq means that it is not possible to distinguish between high levels of Pol2 at the TSS or immediately downstream or upstream of this position. To analyse the early regulatory stages of the transcription cycle in detail, it is necessary to resolve them with far greater accuracy and at improved signal-to-noise ratios than ChIP-seq provides.

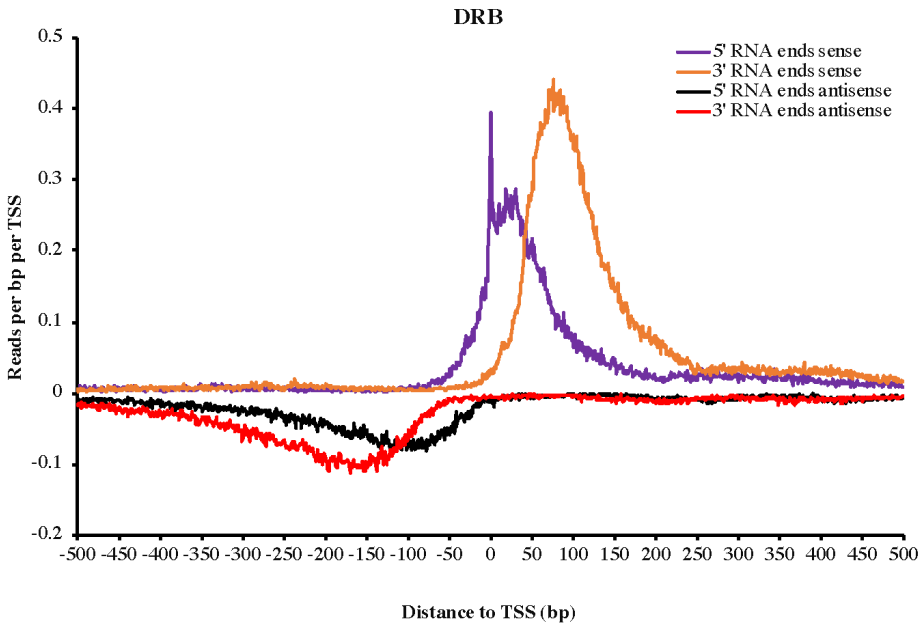
4.2.6 Using scaRNA to analyse initiation and pausing

To determine if scaRNA could be used to compare the roles of initiation and Pol2 pausing in the regulation of gene expression, I revisited classical transcription inhibition experiments using DRB. This reagent blocks the transition of pause release into productive elongation. When transcription is paused, the levels of scaRNA should increase at a given gene.

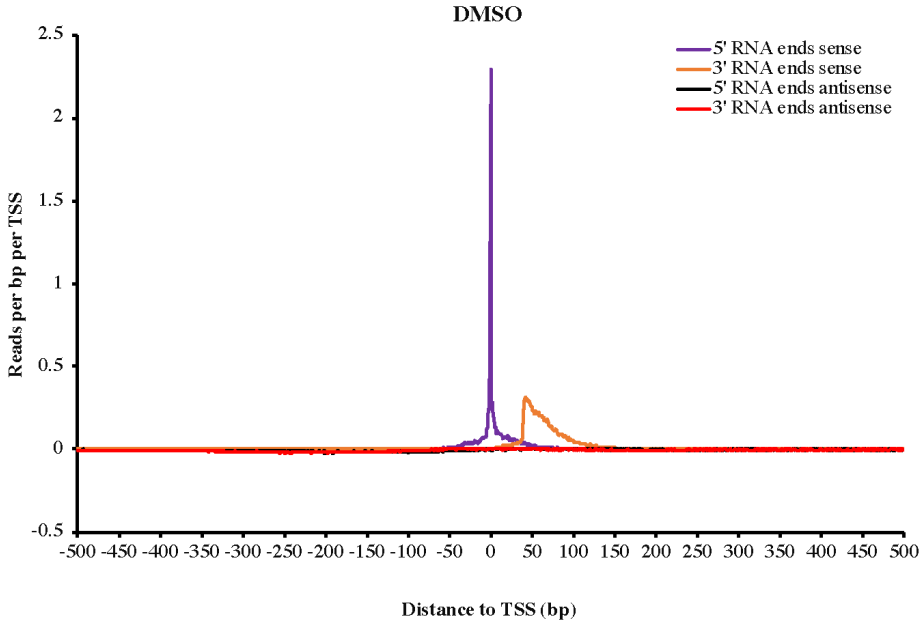
I treated K562 cells either with DRB (100μM) or control (DMSO) conditions for 3 hours. I treated the cells with this concentration of DRB as this has been previously shown to rapidly (30 mins) inhibit transcriptional elongation of GAPDH in Hela cells[64]. A range of treatment times up to 6 hours have been used in previous studies and I chose 3 hours as I wanted to complete transcriptional response to DRB whilst trying to minimise negative secondary effects brought about by global perturbation of transcription[64,112]. Given the changes to metaplots seen in our earlier analysis (Figure 3.10), I examined scaRNA coverage relative both to observed TSSs (called from this data set in the control condition) and UCSC annotated TSSs (Figure 4.9).

UCSC annotated TSS





scaRNA-seq derived TSS



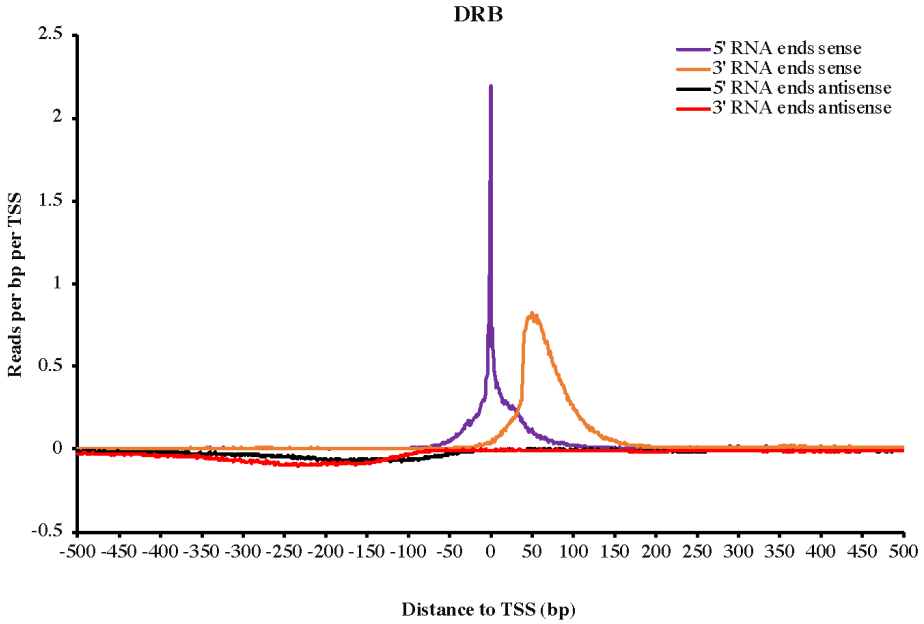


Figure 4.9) scaRNA-seq metaplots – observed and annotated TSS. 5' ends of RNA molecules on the sense strand (purple), 3' ends of RNA on the sense strand (orange). 5' ends of RNA molecules on the anti-sense strand (black). 3' ends of RNA molecules on the anti-sense strand (red). Data are plotted relative to annotated (top row) and observed (bottom row) TSS. Treatment and control condition are shown on right hand side (DRB) and left-hand side (DMSO) respectively.

This reveals a decrease in the maximum height of 5' RNA ends at the TSSs and an increase in the density of 3' RNA ends on the sense strand under treatment conditions. This indicates a block in promoter proximal transcription elongation consistent with inhibited pause release. Notably, this effect appears far more pronounced using the annotated TSSs which may be due to inaccurate positioning of 5' read ends due to the systematic skew in the annotation of many TSSs described in the previous chapter. Calling TSSs *in vivo* corrects for this skew and results in a more accurate profile of transcription initiation, pausing and the effect of DRB on these early stages of the transcription cycle.

One of the most unexpected features of this analysis is the lack of a significant reduction in read density beyond the predicted position of Pol2 pausing (~40bp relative to the observed TSS); rather, there is a gradual tailing off of 3' ends similar to that seen in the DMSO control. If pausing is occurring at this +40bp and CDK9 inhibition affects pause release, then why do we not see an immediate and punctate reduction in read density? There are several hypotheses which may explain these findings:

1) DRB has a general effect on elongation over the whole promoter proximal region (0-150bp), resulting in an equal change in 5' and 3' signals across the region assayed but does not cause a specific increase in Pol2 pausing which would manifest as an increase in <=40nt long transcripts.

2) Pausing occurs over a range of distances downstream of the TSS. Therefore, DRB is not having a specific effect on the +40bp position but rather over this range of positions. Again, this would result in an equal change in 5' and 3' signals across the region.

4.2.7 A single gene analysis of the effects of DRB on scaRNA production

By examining scaRNA read coverage over individual genes in treated or untreated cells, I can directly measure the levels of scaRNA and determine if DRB causes an increase or decrease in scaRNA. This allowed us to determine if the main effect of DRB *in vivo* are upon initiation (less scaRNA after treatment) or promotor proximal elongation (more scaRNA in treatment). More scaRNA could be due to a block in Pol2 pause release at a specific position or a more general increase in scaRNA over the whole promoter proximal regions (non-specific promoter proximal elongation block).

I called observed TSSs in the control condition (ensuring each TSS had an annotated UCSC TSS within +/-500bp). I then drew windows (-100 to +300bp) relative to the centre of the TSS call and determined scaRNA-seq read density. For each TSS I compared the read density under treatment and control conditions, producing what I refer to as a "read density ratio" (RDR) for each TSS.

As DRB affects the total level of RNA in the cell[114], I cannot use normalisation based on total read numbers. I therefore needed to ensure that I could normalise our data relative to a gene which is not predicted to be affected by DRB. I normalised RDR values relative to read coverage over the 7SK RNA as this RNA is transcribed by Pol III[80] and is therefore not thought to be regulated by CDK9. I plotted RDR values for each TSS corresponding to an observed gene (Figure 4.10).

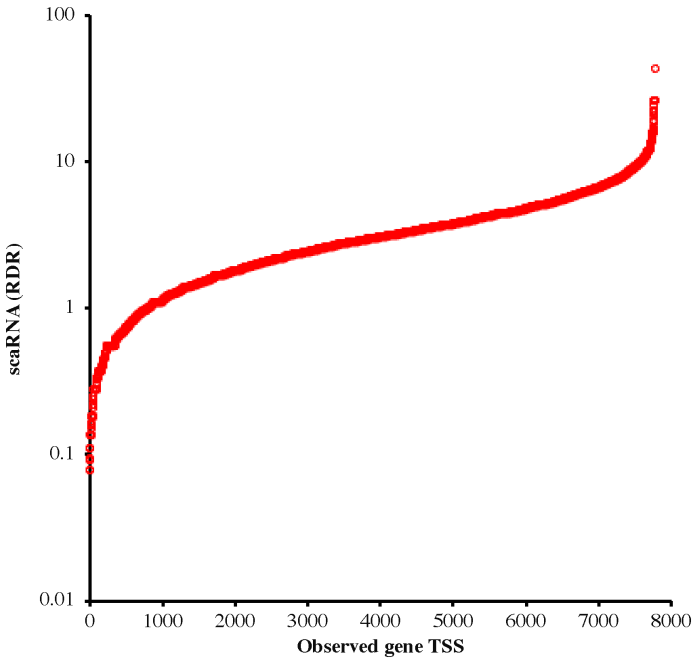


Figure 4.10) scaRNA-seq read density ratio (RDR). RDR values are for treatment versus control. Each red circle is one gene.

The RDR calculations show that 88% (6949/7876) of observed gene TSS have an RDR above 1, indicating that DRB does indeed induce increases in scaRNA density. If a more stringent threshold of 2 is taken (i.e. double the read density in DRB treatment), 69% (5407/7876) of genes are affected. This indicates DRB acts to cause genome wide increases in promoter proximal read density but does not generally inhibit transcription initiation. However, DRB induced increases in read density beyond +40bp (Figure 4.9) could indicate either that DRB is blocking transcriptional elongation beyond the predicted position of pausing or that pausing occurs over the whole +40 to 150bp region.

4.2.8 The effects of DRB on Intronic RNA at the metagene level

To investigate the effects of DRB on read coverage downstream of the promoter region, I produced metaplots of polyA+ RNA (all reads) and polyA+ (intronic reads only) over observed gene TSSs using ngsplots (Figure 4.11).

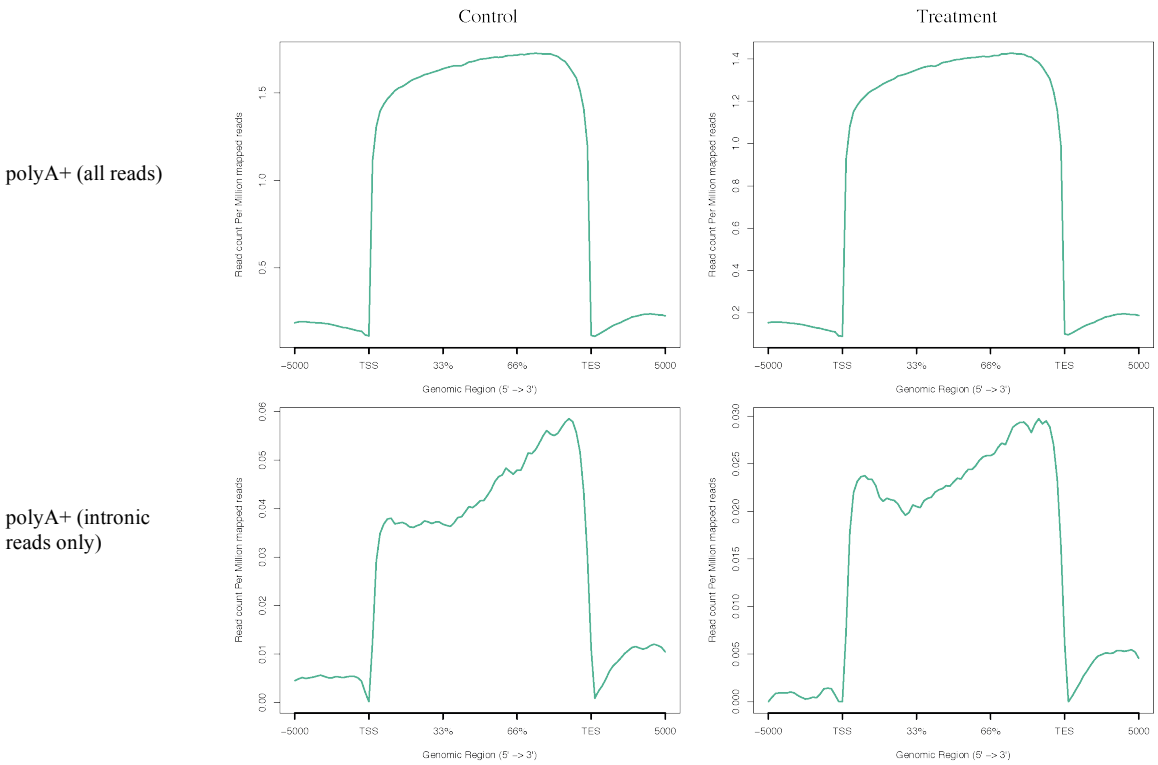


Figure 4.11) The effects of DRB on polyA+ (all reads) and polyA+ (intronic reads only). Data from K562 cells. Read coverage represents combined coverage of all refseq annotated genes scaled to the same length and normalised by total millions of mapped reads.

We show a DRB induced reduction in millions of mapped reads which is clearer in the intronic RNA plots. We can also see a relative increase in read density near to the TSS in the intronic RNA data, which is not evident in the total polyA+ plots. This suggests intronic RNA reads better reflect the expected effects of DRB on transcriptional elongation at the metagene level.

4.2.9 The effects of DRB on Intronic RNA at the single gene level

Our meta-analysis of intronic RNA appears to show an effect on read density near to the TSS. However, as this analysis is an average of all genes, we cannot determine the effect of DRB on individual genes. It may be that at some genes DRB is causing increases in RNA levels but these are outnumbered by genes where this is not the case.

To investigate this, I calculated read coverage over each UCSC gene body (defined as +300 relative to the TSS, to the end of the annotated gene) that had an observed TSS assigned to it. I then calculated the RDR of treatment versus control (normalising relative to polyA+ [all reads] which mapped to the 7SK RNA gene) and plotted the values for each gene (Figure 4.12). This analysis reveals 93% (7341/7876) of genes have an intronic RDR < 1, indicating a reduction in elongation (as judged by read density) is occurring at nearly all genes upon DRB treatment. This is consistent with DRB's purported role as an inhibitor of transcriptional elongation.

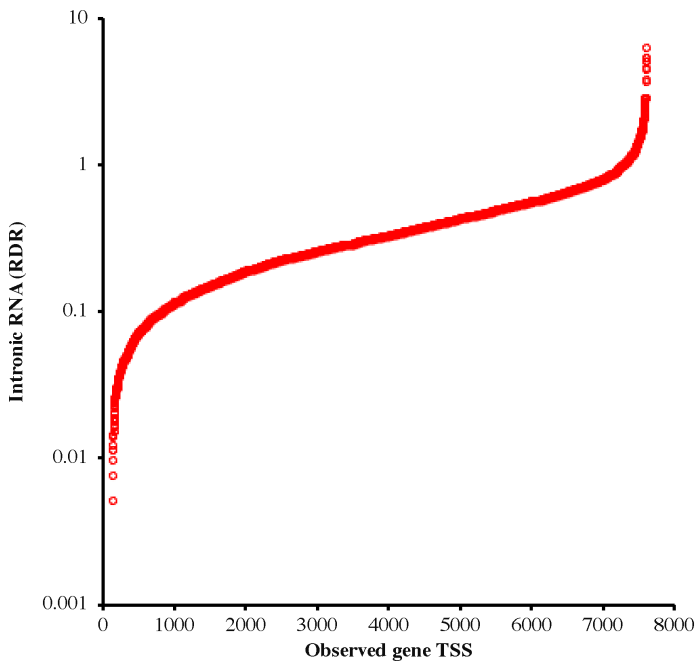


Figure 4.12) Intronic read density ratio (RDR). RDR values are for treatment versus control. Each red circle is one gene.

4.2.10 Delineating response to DRB at a single gene level

Our findings so far suggest that DRB treatment causes an increase in the levels of scaRNA at promoters across the genome and also that intronic RNA production is reduced at most genes. To understand the relationship between the levels of scaRNA and intronic RNA production at a single gene level, I paired the RDR (change in read density treatment versus control) for scaRNA with the associated intronic RNA for each gene and plotted these values against each other. This revealed 7,368 paired values (Figure 4.13).

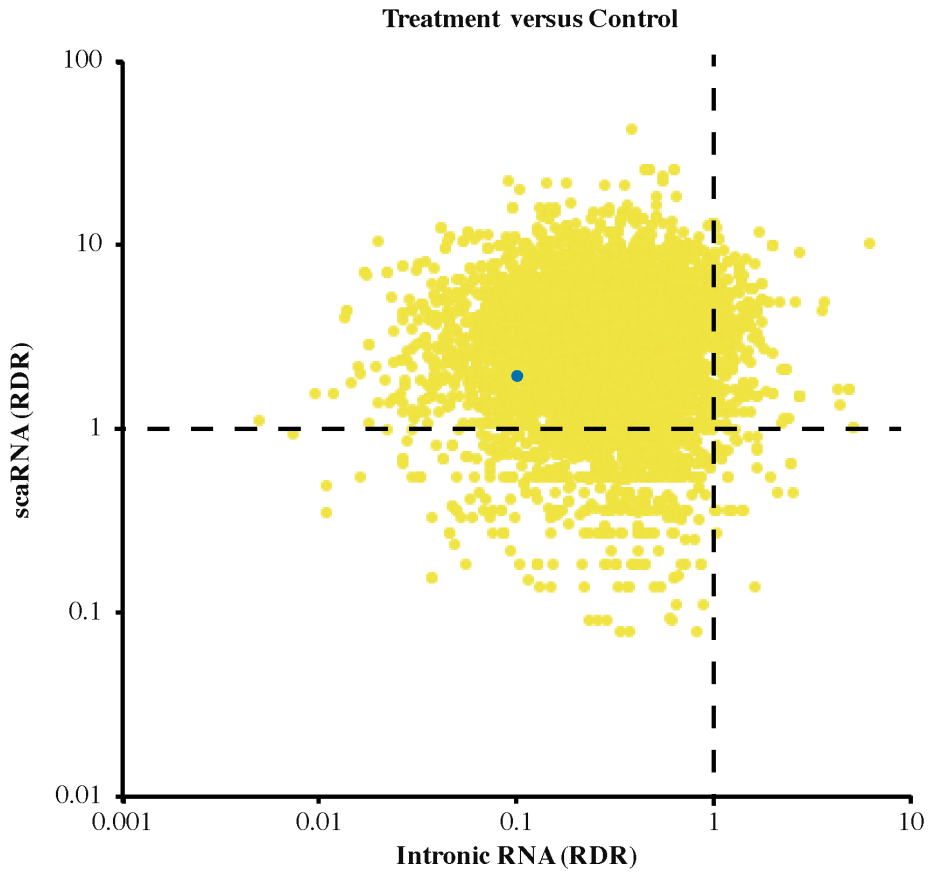


Figure 4.13) scatter plot of scaRNA versus intronic RNA (RDR's). Each gene is one yellow dot. The ***MYC*** gene is indicated as a blue dot.

86% (6339/7368) of genes had a scaRNA RDR <1 and intronic RNA RDR > 1, indicating an increase in promoter proximal reads with a concomitant reduction in intronic RNA reads. These genes appear to be affected by DRB in a classical fashion, consistent with a block in elongation or an increase in pausing, leading to a reduction in full length RNA and an increase in promoter proximal short RNA transcripts. An example of one such gene is *MYC*, which has been described as being regulated by Pol2 pausing. *MYC* has a scaRNA RDR of 1.9 and an intronic RDR of 0.1, placing it in the top 6% of genes affected by DRB and validating our approach for detecting DRB induced changes in transcriptional elongation (Figure 4.14).

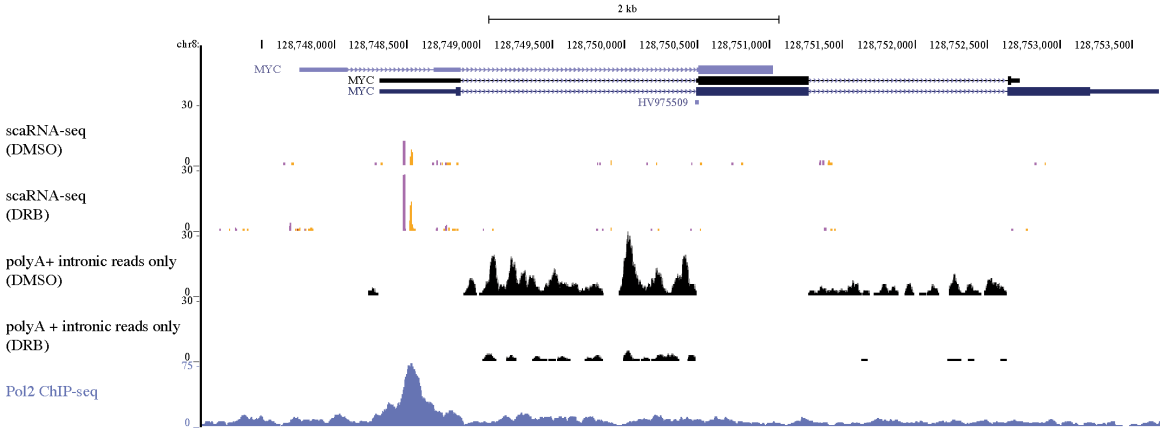


Figure 4.14) Effects of DRB on the *MYC* gene. Showing scaRNA and intronic RNA under control and treatment conditions and Pol2 ChIP-seq in K562 cells. polyA+ (intronic reads only) data and scaRNA data are normalised to the 7SK RNA. Pol2 ChIP-seq is normalised by total millions of mapped reads.

Many genes are affected in a different manner to *MYC*. For example, 10% (740/7368) of genes respond to DRB with a reduction in both short and full-length RNA (scaRNA RDR < 1, intronic RNA RDR < 1) indicating a reduction in initiation and production of full length transcripts. This is consistent with an effect of DRB on transcription initiation either directly or indirectly.

52% (3844/7368) of genes are "strongly" affected by DRB having both a twofold increase in scaRNA and a twofold decrease in intronic RNA. Although this threshold is arbitrary, it indicates that large changes in scaRNA and intronic RNA density are common upon DRB treatment.

Although the majority of genes exhibit a response which is in line with DRB inhibiting release of Pol2 pausing, we do not see a clear correlation between changes in scaRNA levels and intronic RNA levels suggesting this relationship is complex. This may be due to the effect that CDK9 inhibition has on other stages in transcription such as termination[64], making the specific effects of this type of treatment and therefore the prevalence of

specifically promoter proximal CDK9 regulation difficult to disentangle from global effects.

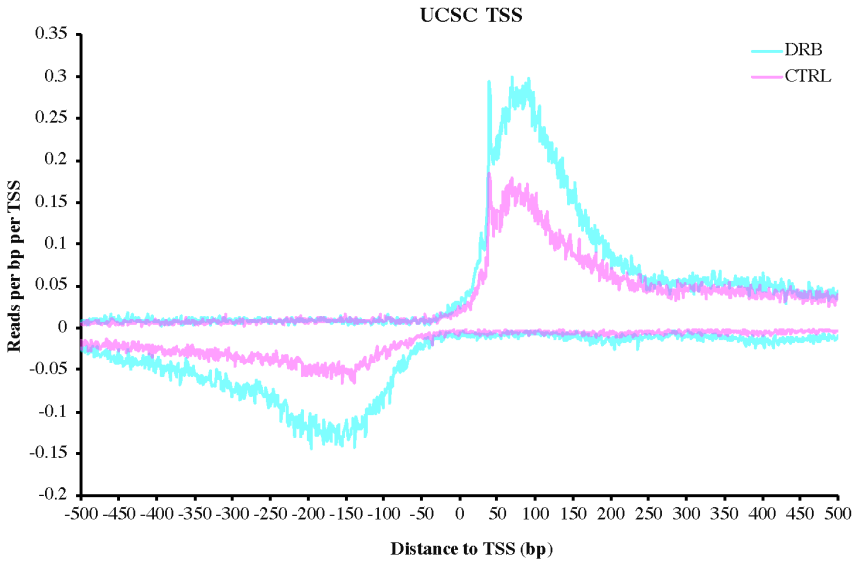
4.2.11 GRO-seq reveals promoter proximal and distal effects of DRB

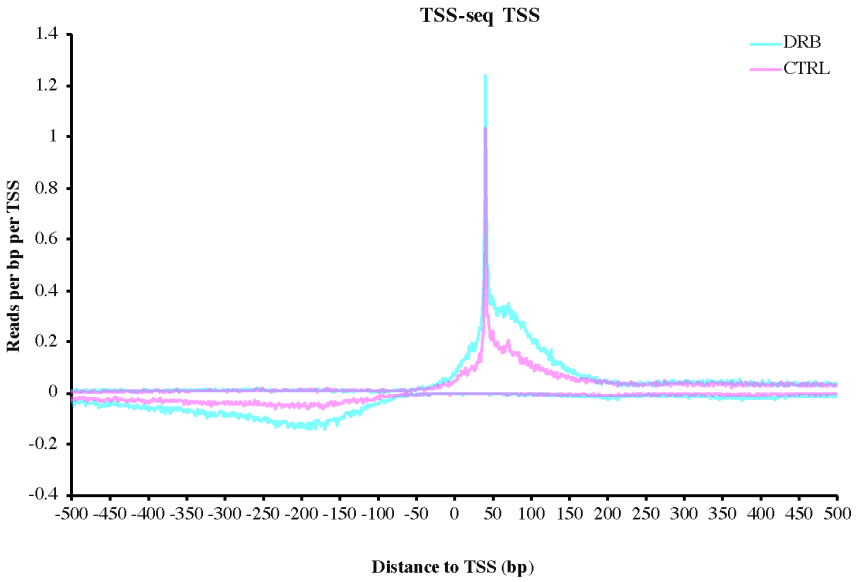
Others have investigated the effects of CDK9 inhibition on transcriptional elongation using GRO-seq. They have concluded that Pol2 pausing is a widely observed step in transcriptional elongation as promoter proximal increases in RNA read density are evident when cells are treated with DRB[12]. This conclusion was drawn from TR calculations performed on GRO-seq data; but due to the limitations of ChIP-seq TR in predicting gene expression, and in light of the high-resolution view afforded to us by scaRNA-seq, I revisited these conclusions.

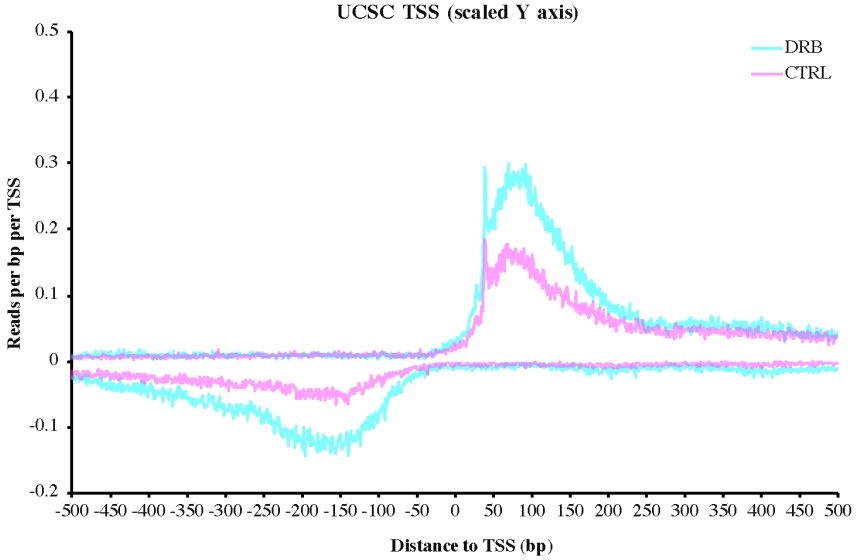
Our analysis of scaRNA under DRB treatment yielded the unexpected finding that scaRNA read density increases not only up to the point of Pol2 pausing as shown by ChIP-exo, but also beyond this point and up to +150bp relative to the TSS (Figure 3.10). This indicates that either pausing occurs over a large region +40 to 150bp, or that DRB influences other aspects of transcription occurring near to the promoter and not only on Pol2 pausing. Either or both of these effects would cause increased levels of scaRNA at the expense of full length RNA, causing TR calculations to indicate increased PPEP (commonly interpreted as increased Pol2 pausing).

To investigate these possibilities, I re-analysed a published GRO-seq dataset from DRB treated (100mM for 2 hours) and control HEK293T cells. I also obtained a TSS-seq data set in the same cell type in control conditions[115], allowing us to derive *in vivo* TSSs and avoid the skew caused by using annotated TSS datasets. TSS-seq is a short capped RNA sequencing method which allows us to derive TSS using the same peak calling approach as with scaRNA data. I called TSS positions from the TSS-seq data set using Homer and

assigned peaks to the nearest annotated TSS. I selected only TSS positions that fell within 500bp -/+ of an annotated UCSC gene TSS in a strand specific manner. I then plotted the 3' most base (corresponding to the final incorporated nucleotide) of each GRO-seq read relative to the observed TSS positions and the annotated TSSs for the assigned genes (Figure 4.15).


UCSC TSS
DRB
CTRL
Reads per bp per TSS
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-500 -450 -400 -350 -300 -250 -200 -150 -100 -50 0 50 100 150 200 250 300 350 400 450 500
Distance to TSS (bp)


TSS-seq TSS
DRB
CTRL
Reads per bp per TSS
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-500 -450 -400 -350 -300 -250 -200 -150 -100 -50 0 50 100 150 200 250 300 350 400 450 500
Distance to TSS (bp)


UCSC TSS (scaled Y axis)
DRB
CTRL
Reads per bp per TSS
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-500 -450 -400 -350 -300 -250 -200 -150 -100 -50 0 50 100 150 200 250 300 350 400 450 500
Distance to TSS (bp)

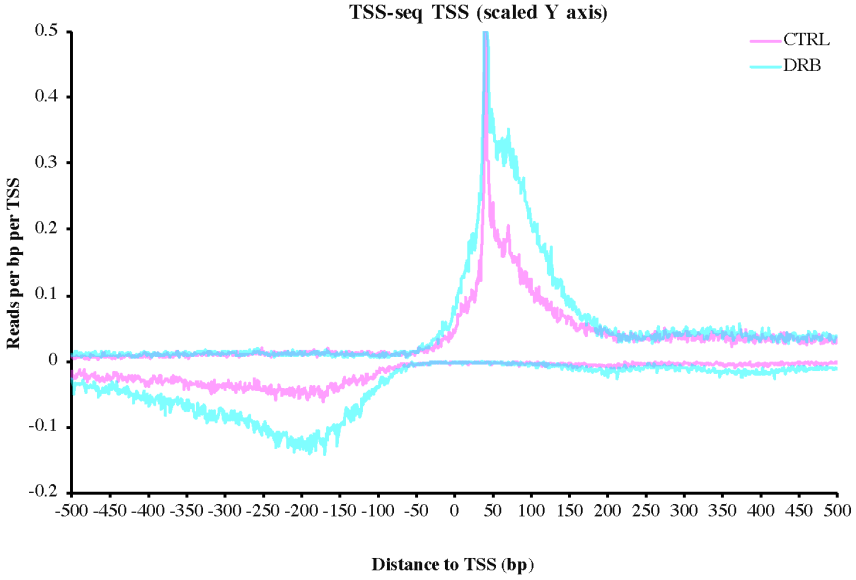

TSS-seq TSS (scaled Y axis)
CTRL
DRB
Reads per bp per TSS
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-500 -450 -400 -350 -300 -250 -200 -150 -100 -50 0 50 100 150 200 250 300 350 400 450 500
Distance to TSS (bp)

Figure 4.15) Effects of DRB on GRO-seq over annotated and observed TSS. GRO-seq 3' read ends plotted relative to annotated (UCSC) (left hand side) or observed (TSS-seq) (right hand side) TSS in treatment (DRB) (cyan) or control (DMSO) (magenta) conditions. The top and bottom rows show alternative scalings of the Y axis. The top row is a scaling to show all the data and the bottom row shows the Y axis scaled to the same proportions for comparison between treatment and control.

Although the TSSs and 3' ends of the RNAs are derived from a different methodology, these data show very similar results to those derived from our scaRNA data (Figure 4.9), with a sharp accumulation of 3' read ends at the +40-45bp position. Crucially, we also show that DRB in these published experiments does not block elongation specifically at this position as expected, but rather DRB treatment increases read density across the region of +40→150bp downstream of the TSS, encompassing the pause site and the region adjacent to the +1 nucleosome. The changes appear more subtle with the TSS-seq derived TSSs compared to annotated TSSs. This is most likely due to the fact that reads are correctly aligned to the TSS position (generally 3' of the annotated TSS), reducing blurring of signal as seen with scaRNA-seq data.

4.2.12 GRO-seq reveals a DRB induced block in elongation at +500bp relative to the TSS

To investigate the effect of DRB further downstream of the TSS, I extended the analysis to a region of -500 → + 2000bp around the TSS, focusing on transcription of the sense strand relative to observed TSSs which map near to the annotated UCSC genes. I also plotted the full GRO-seq RNA fragments rather than the 3' bases only so that this analysis is most similar to what has been previously performed and reported in the literature[12] (Figure 4.16).

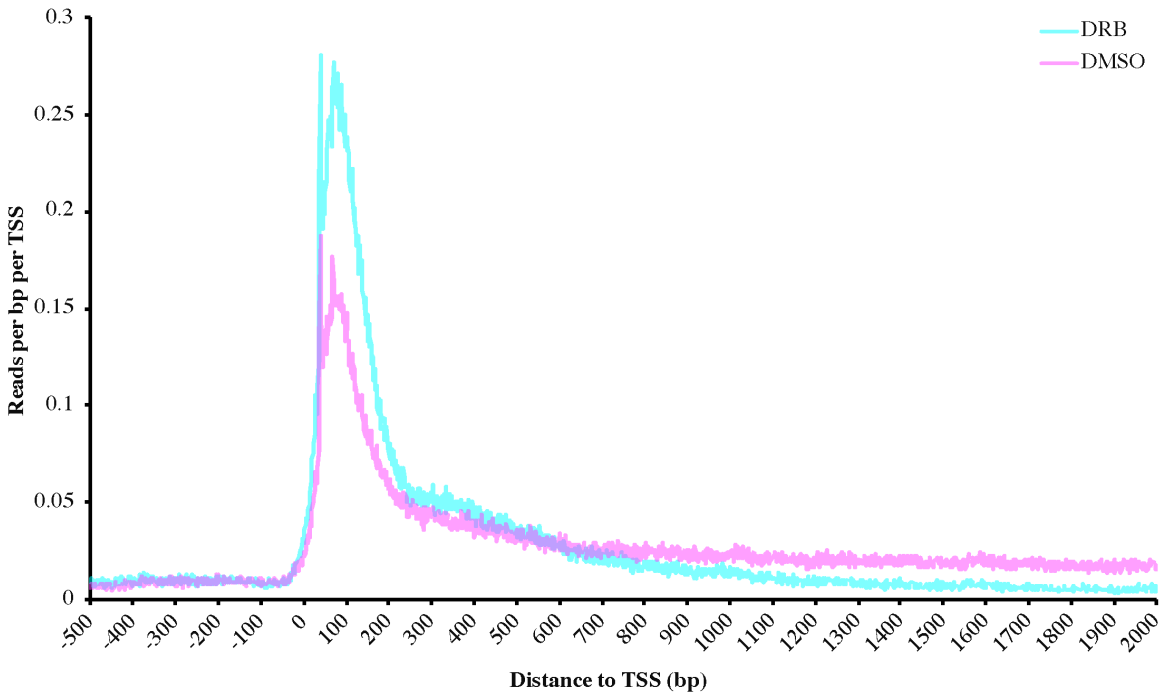


Figure 4.16) GRO-seq full RNA fragments plotted relative to annotated TSS. Treatment (DRB) shown in cyan and control (DMSO) shown in magenta.

This analysis shows a progressive reduction in read density in the bodies of genes which is particularly marked beyond +500bp relative to the TSS. This is striking and difficult to reconcile with the idea that DRB is specifically affecting Pol2 elongation at a pause site near the promoter since read density does not drop off significantly until several hundred base pairs away from the TSS.

4.2.13 Positions of 1st splice sites correlate with DRB induced loss of nascent transcription

I wondered which potential processes could be affected by DRB that would cause a reduction in gene body RNA read density. Primary candidates would be splicing or the displacement of nucleosomes by transiting Pol2. To investigate whether the positioning

of splice sites could underlie the profile seen in Figure 4.14, I plotted the positions of all annotated 1st splice sites for UCSC genes relative to their annotated TSS (Figure 4.17).

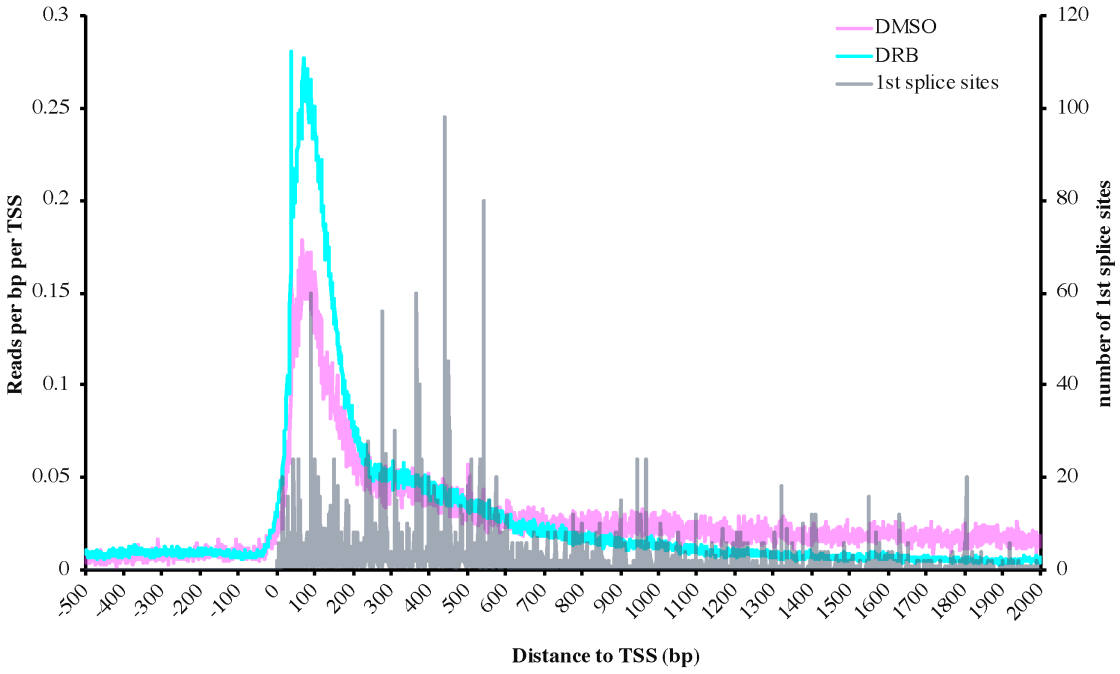


Figure 4.17) UCSC genes 1st splice site versus TSS position. GRO-seq data for DRB treated of control cells is also shown - treatment (DRB) shown in cyan and control (DMSO) shown in magenta. Position of annotated 1st (most 5') splice site shown as a grey histogram in 1bp bins. Plot adjusted to account for strand. Primary Y axis (left hand side) shows average read coverage, secondary Y axis (right hand side) shows number of annotated splice sites at each position.

This analysis reveals, unsurprisingly, that there is a clustering of 1st (most 5') splice sites within the first 600bps downstream of the TSS, coincident with the region which shows the strongest downstream decrease in elongation. This suggests that DRB's effects on splicing across the population of genes may partly explain why there is a progressive reduction in GRO-seq reads beyond +150bp relative to the TSS and towards the ends of genes.

4.2.14 Decreases in GRO-seq signal upon DRB treatment correlate with nucleosomal barriers to elongation

I next considered if the progressive reduction in GRO-seq signal across genes was, at least in part, due to a failure of Pol2 to navigate nucleosomes during transcription. This seemed possible because the majority of GRO-seq signal is contained between +40 and +150bp, suggesting a barrier to nascent transcription at this point which would be consistent with *in vivo* nucleosome positioning.

MNase-seq data were not available for HEK293 cells, however, DNase-seq data which is generally used to reveal positions of open chromatin, was available through the Encode project. I analysed the DNAse-seq data, plotting it relative to observed (TSS-seq derived) TSSs in HEK293 cells, showing that the positions of the phased +1 to +3 nucleosomes, as well as the promoter associated open chromatin site, could be clearly discerned in a similar manner to MNase-seq (Figure 4.18).

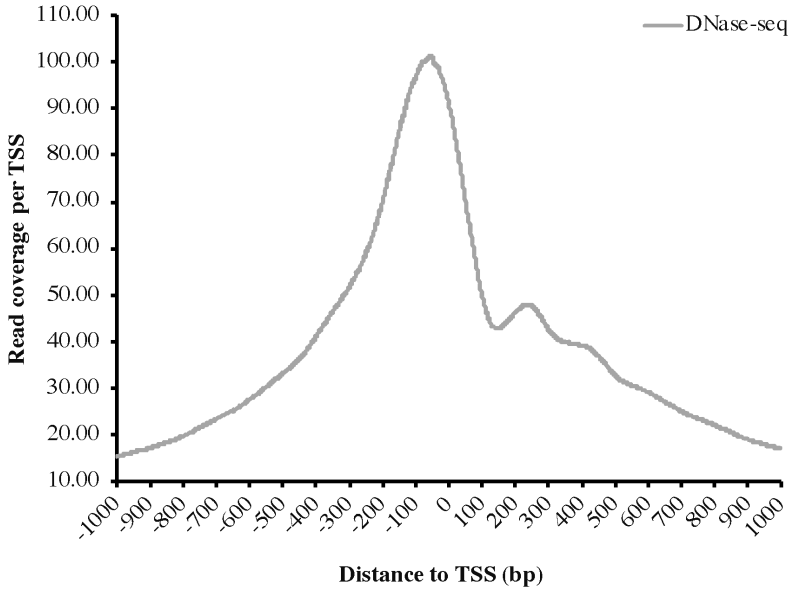


Figure 4.18) open chromatin in HEK293 cells. DNase-seq over all annotated TSS.

Since DNase-seq data is effective at detecting the position of phased nucleosomes, I compared this data to GRO-seq data pre- and post-DRB treatment (Figure 4.19).

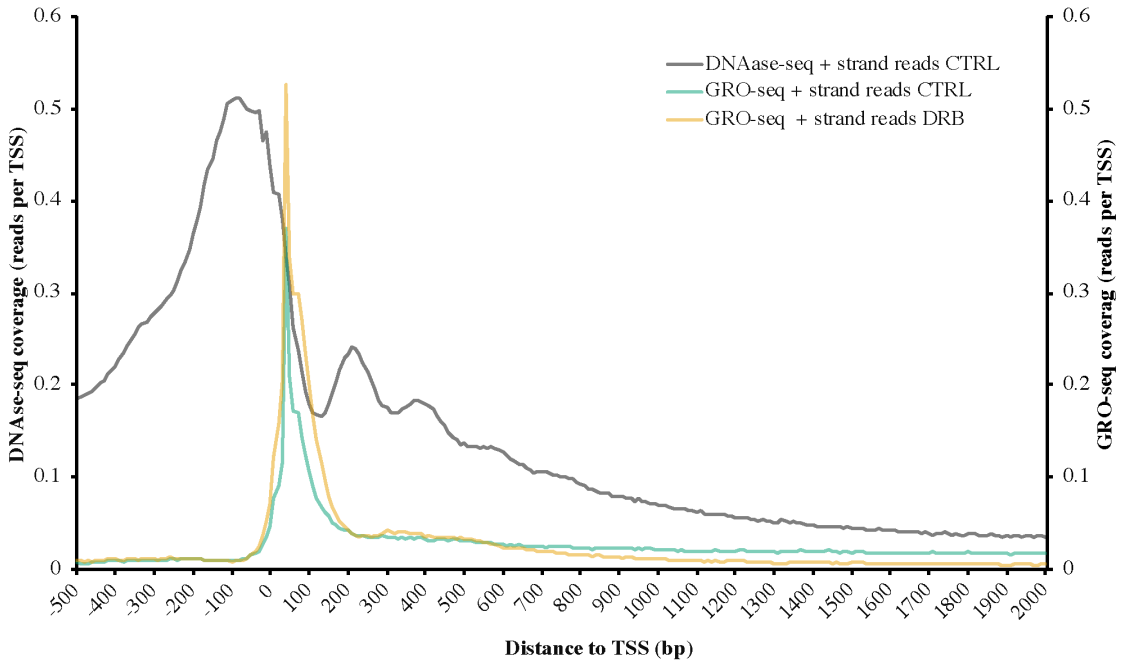


Figure 4.19) GRO-seq (+/- DRB) and open chromatin. GRO-seq and DNase-seq read coverage over observed (TSS-seq derived) TSS. Data plotted is from the plus strand only.

In agreement with scaRNA-seq and MNase-seq analysis (Figure 3.11), the extent of promoter proximal GRO-seq 3' RNA ends drops coincidently with the drop in open chromatin downstream of the TSS at ~+150bp, further suggesting that in control conditions the +1 nucleosome may be a major barrier to transcriptional elongation.

The high levels of GRO-seq signal in the promoter proximal region (0-150bp) means it is difficult to visually assess changes in GRO-seq signal, particularly over the region of the +2 and +3 nucleosomes (~+150→+500bp). Adjusting the scaling of the Y axis of the GRO-seq data reveals that under DRB treatment, and compared to the untreated control, further accumulations of 3' ends in the GRO-seq reads can build up between the positions of the +1 and + 2 nucleosomes and between the +2 and +3 nucleosomes, associated with an overall tailing off of GRO-seq signal through the body of the genes (Figure 4.20).

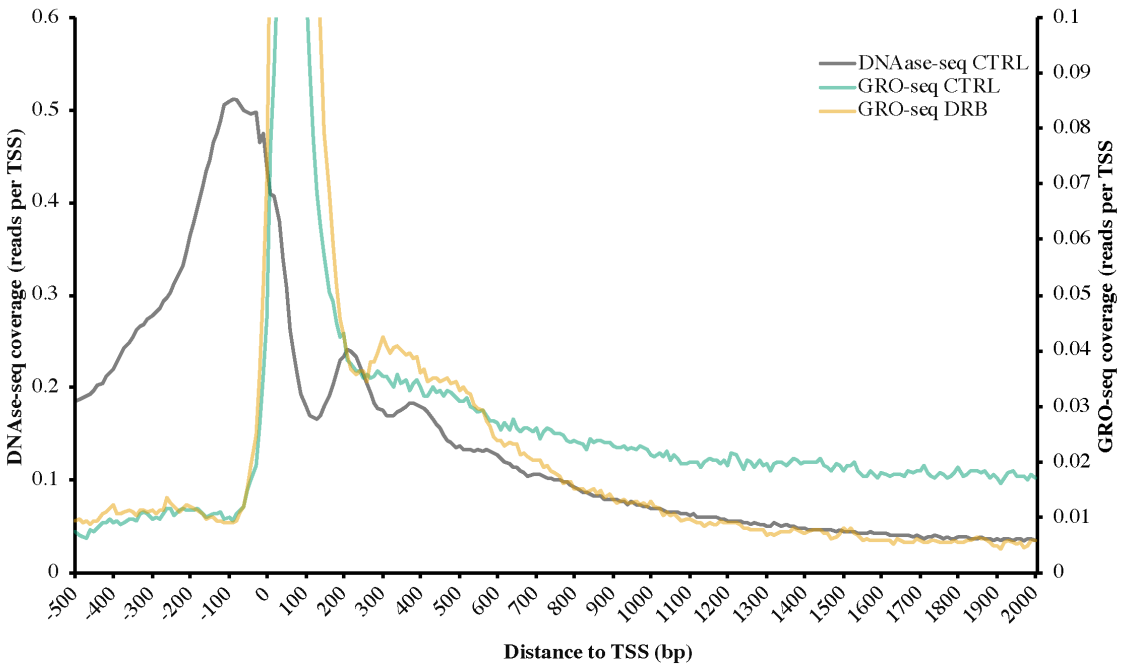
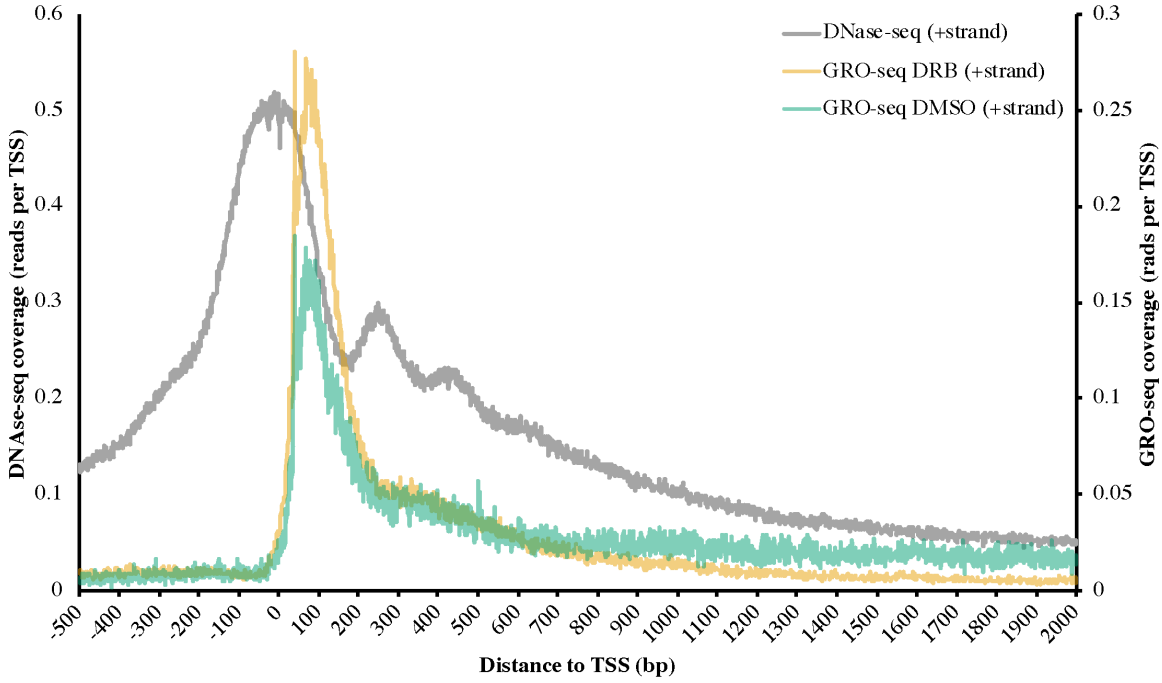


Figure 4.20) As in figure 4.19 except Y axis for GRO-seq data scaled for visualisation. Maximum Y axis value reduced from 0.6 to 0.1 to visualise GRO-seq read build up over nucleosomes.

Taken together, this suggests that under normal conditions the +1 nucleosome is a major barrier to elongation and this is increased further under DRB treatment, with progressive nucleosomes appearing to act as "brakes" on transcriptional elongation. I hypothesise that this may be due to DRB inhibiting active processes such as nucleosome removal and this may be a significant contributor to the reduced levels of full length transcripts seen with DRB treatment.

Performing this analysis using the annotated UCSC gene TSS coordinates rather than the TSS derived from TSS-seq, we generate a similar profile, with the majority of GRO-seq read density bounded by the +1 nucleosome. However, the accumulation of reads at the +2 and +3 nucleosomes is not clearly visible, again highlighting that *in vivo* derived TSS coordinates offer a clearer picture of general transcription when performing meta-analyses

and that the subtle effects of DRB on transcription have been obscured by non-tissue specific TSS annotations (Figure 4.21).



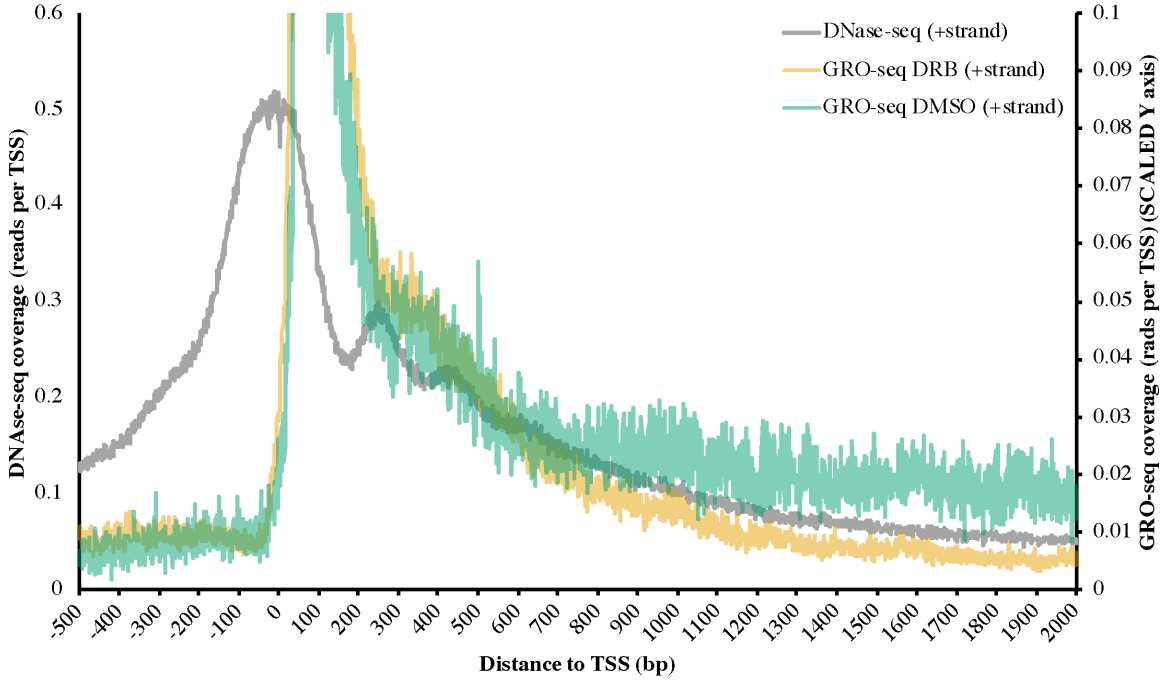


Figure 4.21) As in 4.19 and 4.20 but with annotated TSS coordinates. Top figure shows Y secondary axis automatic scaling to maximum signal, bottom figure shows secondary Y axis manual scaled to 0.1 (to aid in visualisation).

Examining the GRO-seq data reveals two effects of DRB consistent with our scaRNA and intronic RNA data: an increase in promoter proximal read density and a reduction of read density downstream in the gene body. Previously, the promoter proximal effect was attributed to blocking of Pol2 pause release, but examination of the position of Pol2 pausing +40bp shows that DRB does not block elongation at this point but rather increases transcription over the whole promoter proximal region and up to the +1 nucleosome. The second effect was previously thought to be a direct result of a block in Pol2 pause release and therefore assumed to occur immediately after this position. However, we see that further reduction in read density occurs well into the body of genes (beyond +500nt), suggesting different mechanisms of action independent of the promoter proximal effect contribute to the reduction in elongation.

Thus I have shown by two independent methods that the effects of DRB on transcription are complex and cannot be explained entirely by inhibition of CDK9 leading to a block in pause release. Contrary to the conclusions from *in vitro* experiments, it appears that either DRB can directly inhibit targets other than CDK9 which have important roles in transcription, or the inhibition of CDK9 by DRB is specific but CDK9 has far more extensive roles in transcription than simply pause release. Due to its ability to act as a competitive ATP analogue, I hypothesise that other active processes which require chemical energy in this form could be perturbed by the use of DRB. For example, the reason that we see an increase in promoter proximal reads under treatment conditions may not be due only to blocking pause release but rather failure to traverse the +1 nucleosome and therefore premature transcription termination. Inhibitors of "pause release" such as

DRB may not truly reflect the extent of Pol2 pausing *in vivo* due to their uncertain mode of action.

These findings have significant implications for the field as CDK9 inhibitors such as DRB or FP have been used to estimate the prevalence[12] and stability of Pol2 pausing[38], as well as to suggest that CDK9 dependant checkpoints exist at the 3' end of genes[64]. In summary, there is limited data on the prevalence of Pol2 pausing *in vivo* as a regulatory mechanism without the confounding effects of these drugs.

4.2.15 Determining the extent of promoter proximal pausing of Pol2 *in vivo*

To determine how commonly promoter proximal Pol2 pausing acts as a mechanism to regulate differential gene expression *in vivo*, I turned to murine erythropoiesis as a model system. This primary cell model provides access to two distinct stages (0h and 24h) of differentiation corresponding to early and intermediate stages of erythropoiesis[7]. Gene expression changes throughout this transition, providing thousands of genes for which we can examine changes in scaRNA and intronic RNA levels without the confounding effect of using elongation inhibitors.

I therefore produced scaRNA-seq and polyA+ (intronic reads only) RNA-seq data sets at the two timepoints in biological triplicate. I then called TSS from the scaRNA-seq data at the 0h and 24h stages of differentiation and calculated the logFC in scaRNA over the TSS (-100→+300) using an approach which normalises for sample variation, in a manner similar to standard RNA-seq experiments. I then identified statistically significant (p= <=0.05) changes in gene expression. In parallel, I measured the polyA+ (intronic reads only) logFC over the gene bodies (+300 to gene end) using a similar approach[83]. I then

examined the relationship between significantly changing scaRNA and intronic RNA levels in differentiating primary cells (Figure 4.22).

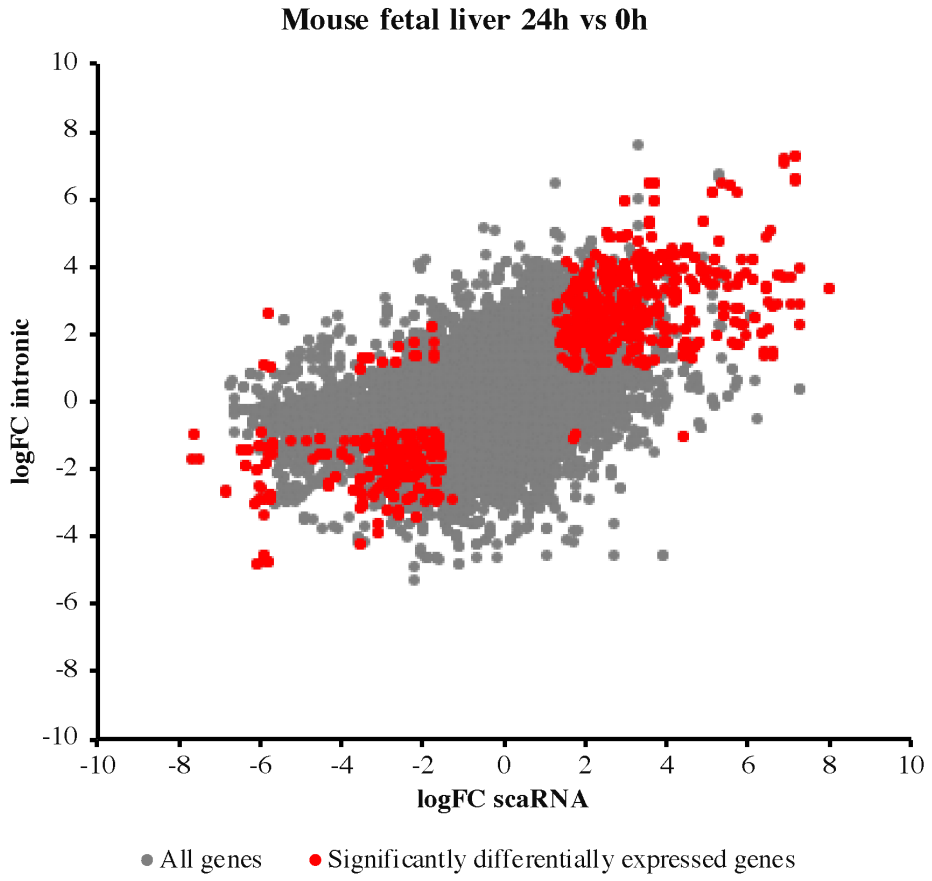


Figure 4.22) Scatter plot of logFC intronic RNA versus logFC for observed "active" genes.

Significantly differentially expressed genes are shown in red.

From this analysis there is a positive correlation between increasing scaRNA level and intronic RNA level, which is the opposite of what would be predicted if promoter proximal pausing was a major mechanism of gene regulation. In this case, scaRNA levels should increase as gene transcription decrease and vice versa. This suggests that most differential gene expression changes in this *in vivo* system occur due to alteration in the level of transcription initiation. This causes an associated increase in the levels of full-length transcripts. There is considerable spread around the positive correlation, suggesting that

the relationship between scaRNA and intronic RNA production is influenced by other modulating factors.

Previous findings have often relied on TR style calculations or the use of inhibitors of transcription initiation or elongation to estimate the prevalence of Pol2 pausing *in vivo*. Using an unbiased quantitative approach which allows us to examine the levels of short capped RNA present at genome-wide TSSs and the levels of intronic (nascent) RNA that these genes produce, I have shown that TR has limited power to predict gene expression. Furthermore, the inhibitors used in these experiments are not specific to promoter proximal transcription in the range previous described as Pol2 paused (+20-60bp).

I conclude that Pol2 pausing is not commonly a rate limiting step in differential gene expression *in vivo* and that transcription initiation appears to be the predominant point of regulation in the transcription cycle of mammalian genes.

4.3 Discussion

4.3.1 Measuring Pol2 pausing *in vivo*

In this chapter I examined the widely used method (TR) to evaluate Pol2 elongation and pausing. I showed that this metric is a poor indicator of gene expression and, in fact, Pol2 ChIP-seq is only modestly quantitative. The resolution issues with Pol2 ChIP-seq also mean that it is not possible to distinguish between initiating, paused or elongating Pol2 over the small regions (~300bp) used to calculate the TR and so any changes in this ratio in previous work may be wrongly interpreted as being related to pausing when, in fact, they may be due to other changes in promoter proximal transcription.

Here I moved away from TR calculations and used a direct comparison of scaRNA and intronic RNA at a single gene level. I showed, in line with previous findings, both *in vitro* and *in vivo,* that DRB induces a block in elongation just downstream of TSSs and often a reduction in the amount of intronic RNA being produced. However, when interpreting these changes at a single gene level I see no correlation between increases in scaRNA and reductions in intronic RNA, suggesting the effects of DRB on transcription are complex.

In line with this finding I highlight the fact that DRB causes two independent effects on nascent transcription: i) a promoter proximal effect, with increased read coverage across the region of +40 to 150bp; and, ii) a progressive reduction in read density which is particularly evident at +500bp downstream of the TSS, which correlates with a high density of splice sites and strongly positioned nucleosomes. Overall, these data indicate that DRB causes a variety of effects on transcription which cannot be simply explained by the blocking of a punctate Pol2 pausing step near the TSS of most genes.

Finally and in contrast to the current view of the field, I find that when I examine natural differential transcription of thousands of genes during development, I see little evidence of promoter proximal regulation of transcription via Pol2 pausing. This suggests that the accumulation of scaRNAs at promoters may represent the balance between initiating Pol2 and Pol2 which is yet to navigate the +1 nucleosome.

4.3.2 Implications for the field

The view that Pol2 pausing is a ubiquitous point of regulation in the transcription cycle was borne out of experiments utilising DRB and FP (another ATP analogue), which competitively inhibits CDK9 in similar manner to DRB[51].

It has previously been reported that the response to DRB can differ based on the gene in question with some genes exhibiting late DRB inducible blocks in elongation and also with many genes exhibiting a DRB sensitive checkpoint in elongation at their 3' ends[64].

I have shown that *in vivo* DRB treatment causes a variety of effects that could either be explained by its non-specificity for CDK9 and inhibition of other active transcriptional processes, or that CDK9 may have broad roles in transcription. The prevalence estimates for Pol2 pausing may therefore be inaccurate but most importantly the extent to which promoter proximal Pol2 pausing is regulated to modulate gene expression may have been significantly over estimated.

Earlier work to estimate the prevalence of Pol2 pausing used TR to show that promoter proximal enrichment of Pol2 (PPEP) was common across the genome and commonly co-localised with DSIF and NELF suggesting the *in vitro* described mechanism for pausing and release may operate on promoters genome-wide *in vivo*[22,23]. I show that although PPEP is common, much of this Pol2 may not be subject to regulation during erythropoiesis. It is possible that a lack of ability to distinguish between initiating versus paused Pol2 when performing TR calculations is the primary reason for this view. If the transition of Pol2 between initiation and Pol2 pausing is inefficient as suggested by recent imaging studies[116], one could explain changes in TR as changes in the amount of initiating Pol2 rather than paused Pol2, leading to a different hypothesis for transcriptional regulation genome-wide.

To test this directly and further explore our finding that transcription initiation seems to be the rate limiting step in transcription, in the next chapter I report an *in vivo* analysis of the first mammalian gene (beta globin) to be described as paused and contrast this to alpha globin, a gene not thought to be regulated by pausing but expressed in the same tissues to a similar level.

Chapter 5

The role of enhancers in regulating the transcription cycle

Chapter 5 – The role of enhancers in regulating the transcription cycle

5.1 Introduction

5.1.1 The role of enhancers in pausing

It has been proposed that enhancers can regulate transcription by modulating Pol2 pausing at promoters through several mechanisms. These include recruitment of components of the super-elongation complex to promoters, decoy of pausing factors from promoters to enhancers via eRNA binding, and roles in delivering pause release factors to cognate promoters[54,56,117,118].

However, enhancers defined in these studies have often not been tested functionally. In addition, changes in Pol2 pausing have been inferred from altered TR's generated from ChIP-PCR or ChIP-seq data[22]. Due to the resolution limits discussed previously, changes in the relative amounts of promoter proximal Pol2 (-30 +300) versus gene body Pol2 (+301 to end of gene) could be interpreted as increases of paused or initiating Pol2 with no way to distinguish between the two. Enhancers may have roles in pause release or alternative mechanisms that may give the effect of increased pause release by ChIP-seq. As an example, if enhancers have a role in chromatin remodelling at cognate promoters then deletion of enhancers could cause an increase of promoter proximal Pol2 versus gene body Pol2; but, this would not be due to Pol2 pausing but to the inherent kinetic bottleneck of the +1 nucleosome being enforced. Here I sought to determine the contribution of enhancers to regulation of transcription through transcription initiation or Pol2 pausing in a more direct way by studying well defined enhancer elements at the alpha globin and beta globin loci.

5.1.2 Control of transcriptional elongation at alpha and beta globin

The haemoglobin beta gene (Hbb) was the first mammalian gene to be described as paused[93]. More recent work using RNA FISH and ChIP-qPCR suggests that the six enhancers of this gene (HS1-6, referred to as a locus control region or more recently a superenhancer[7,119]) act to promote elongation. When HS2 and HS3 are homozygously deleted in combination (referred to in this study as bHom) there is a reduction of approximately 75% in expression, indicating these two enhancers in combination have the strongest effect on beta globin gene expression *in vivo* (Figure 1.1). However, this study used ChIP-PCR to detect changes in the levels of Pol2 at the promoter and downstream in the gene, and the authors concluded that transcriptional elongation was defective in the enhancer knockout. Given the resolution limits of ChIP-seq described in this thesis, I revisited these finding using high resolution scaRNAs as a measure of initiation and promoter proximal pausing.

Our lab has previously developed and analysed the effect of targeted homozygous deletion of two (R1 and R2) of the four alpha globin enhancers (also considered a superenhancer[7,119]) in combination (referred to here as aHom) .These two elements together are responsible for driving over 90% of nascent alpha globin production and are therefore the major tissue specific enhancers of this gene *in vivo*[7] (Figure 1.1).

aHom primary murine foetal liver cells and embryos have undergone extensive functional characterisation in previous work by our laboratory. They have been validated as *bona fide* enhancer elements by transgenic reporter assays by assessing the impact of their deletion on nascent alpha globin expression *in vivo* and by phenotypic assessment of murine haematology and development[7]. Likewise, the enhancers of beta globin have been similarly characterised over the years[14,120].

Here I investigated in detail the effect of the aHom and bHom enhancer deletions using a combinatorial approach focusing on gene expression, levels of pausing and initiation, Pol2 distribution and quantity, chromatin accessibility and 3-D structure, with the aim of revealing the mechanism by which these two sets of enhancers regulate their target gene expression.

5.2 Results

5.2.1 RNA-seq reveals targeted nature of alpha globin enhancer deletion

Previous work studying the effects of enhancers on the transcription cycle used putative enhancers which often regulate genes of unknown importance in the cell type being studied[54,56,117]. In order to understand the impact that enhancers have on the transcription cycle of a gene, it is important that perturbation of the target gene does not alter the homeostasis of the host cell. For example, if the gene being targeted is a transcription factor, observations might be confounded by secondary effects on lineage fidelity, cellular differentiation or basic homeostasis such as cell cycle dynamics or apoptosis.

Previous work from our lab has concluded that the combined output of both copies of the alpha globin genes in *cis* are significantly downregulated in the enhancer knockout[7]. To investigate the specific effects of the alpha globin enhancer on the total transcriptional output in mouse primary foetal liver cells, I performed differential gene expression analysis, including all annotated UCSC genes using polyA+ RNA-seq libraries depleted of globin RNA (Figure 5.1).

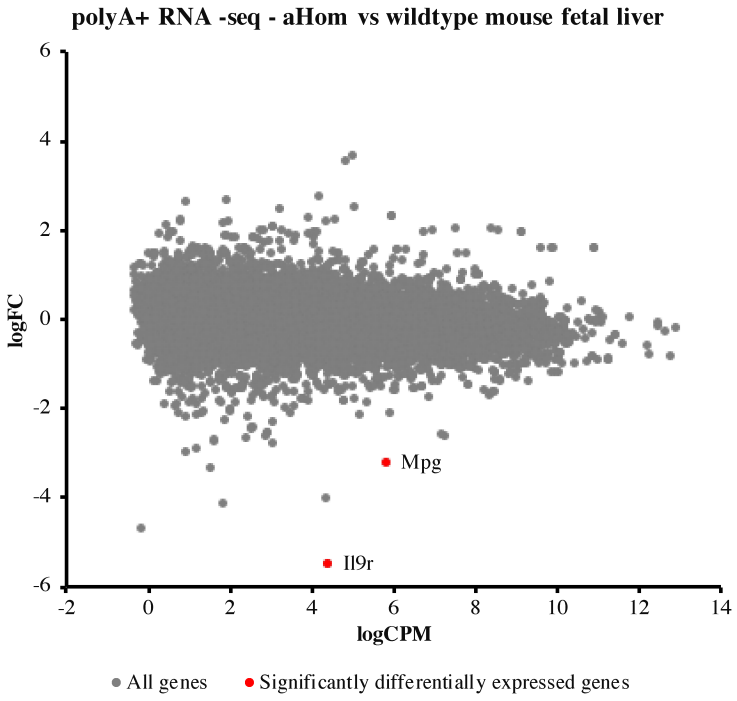


Figure 5.1) Differential gene expression analysis of wildtype versus aHom. RNA-seq (polyA+ all reads) from differentiated mouse foetal liver cells. Significantly differentially expressed genes are shown as red dots; non-significantly expressed genes are shown as grey dots.

I show that other than alpha globin itself, only 2/19400 other genes are significantly differentially expressed. This suggests that the alpha globin enhancer deletion has no significant effect on the global transcriptome and is therefore an excellent model to study the targeted effects of enhancers on cognate gene transcription, as compared to previous studies which may have been confounded by the secondary effects of deleting enhancers with unknown effects in global transcription[54,56].

Both of the genes detected as significantly differentially expressed in the total RNA data set are located adjacent to the alpha globin cluster and interact with the alpha globin enhancers[7,8]. Their reductions in expression are likely explained by the fact that they are normally directly upregulated by the action of alpha globin enhancers as a bystander effect.

5.2.2 Analysis of promoter proximal chromatin accessibility in enhancer knockouts

To understand how the enhancers affect the alpha globin genes, I first studied the effect of the enhancer deletions on the chromatin accessibility of the alpha and beta globin gene

promoters: lower accessibility could be a result of lower levels of transcription factor recruitment and therefore initiation of transcription, reducing nascent RNA production. I performed ATAC-seq to reveal open chromatin in primary mouse foetal liver tissues using wildtype, aHom and bHom genotypes. These analyses were reproducible (Figure 5.2) and so I combined biological triplicates for each of the conditions.

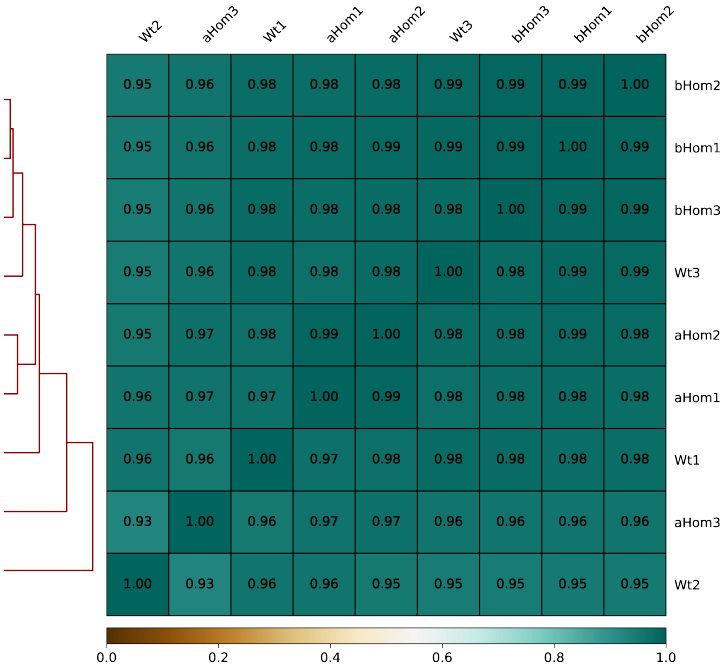


Figure 5.2) Pearson correlation (R^2) analysis of ATAC-seq data.

Visualisation of the alpha and beta globin loci (Figure 5.3) reveals the targeted enhancer deletions at both genes (shown in purple).

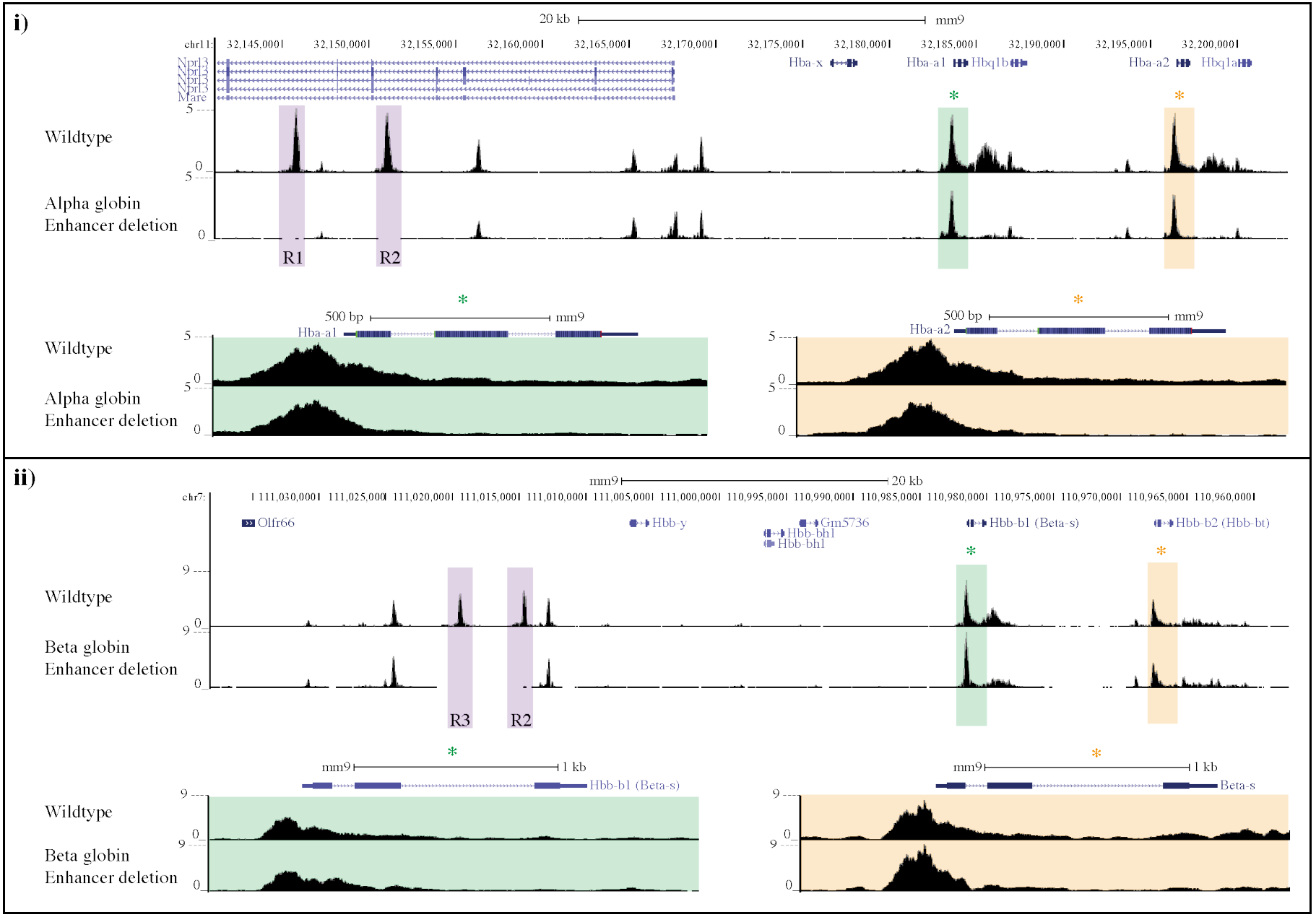

i)
20 kb
mm9
chr11:
32,145,000
32,150,000
32,155,000
32,160,000
32,165,000
32,170,000
32,175,000
32,180,000
32,185,000
32,190,000
32,195,000
32,200,000
Nprl3
Mare
Hba-x
Hba-a1
Hbq1b
Hba-a2
Hbq1a
Wildtype
Alpha globin
Enhancer deletion
R1
R2
500 bp
mm9
Hba-a1
Hba-a2
ii)
mm9
20 kb
chr7:
111,030,000
111,025,000
111,020,000
111,015,000
111,010,000
111,005,000
111,000,000
110,995,000
110,990,000
110,985,000
110,980,000
110,975,000
110,970,000
110,965,000
110,960,000
Olfr66
Hbb-y
Gm5736
Hbb-bh1
Hbb-bh1
Hbb-b1 (Beta-s)
Hbb-b2 (Hbb-bt)
Wildtype
Beta globin
Enhancer deletion
R3
R2
1 kb
Hbb-b1 (Beta-s)
Beta-s

Figure 5.3) Open chromatin profile of the alpha and beta globin loci. ATAC-seq of alpha (i) and beta globin (ii) loci under wildtype and enhancer knockout (aHom and bHom) conditions. Tracks are normalised by millions of mapped reads and are produced from a merged biological triplicate for each genotype. Regulatory elements (R) under investigation indicated by purple boxes for each loci.

In both wildtype and enhancer knockout conditions, both alpha and beta globin appear to have high levels of hypersensitivity across the genes, showing that the enhancer deletion does not significantly affect chromatin accessibility at these genes; both genes remain hypersensitive and therefore accessible to the transcriptional machinery.

To investigate in more detail whether the enhancer deletions had a significant effect on hypersensitivity particularly near the 5' end of the genes associated with the TSSs and the promoter regions, I peak called open chromatin in both wildtype and enhancer knockout conditions and produced a consensus list of peaks. I removed peaks on chromosome X, Y and those associated with mitochondrial DNA as I did not want to calculate false differential peak changes due to these common confounding parameters in ATAC-seq libraries[121].

I called peaks using MACS2 and visually determined which specific peak best overlapped the respective alpha or beta globin 5' gene ends (Figure 5.4).

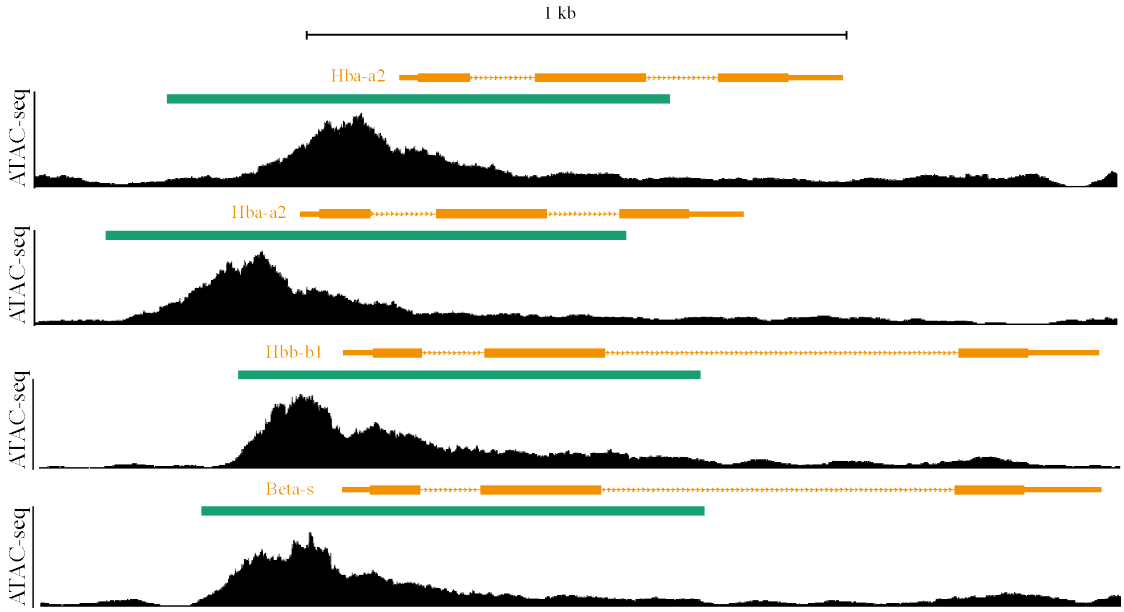


Figure 5.4) Open chromatin peaks at the alpha and beta globin genes. ATAC peaks associated with the 5' ends of alpha and beta globin genes. Regions called as a peak shown in green; genes shown in orange. ATAC-seq and peak calls were performed under wildtype conditions.

Although these regions are large when compared to the size of the genes, they cover the TSSs and upstream promoter regions, allowing us to analyse differential hypersensitivity over the start of these genes as opposed to across the entire gene bodies. I then counted reads which fell over consensus peaks from each replicate and performed a differential analysis using EdgeR, allowing us to produce a statistically robust analysis, as visualising individual genes using merged replicates may confound our analysis despite their high correlation across the genome (Figure 5.2).

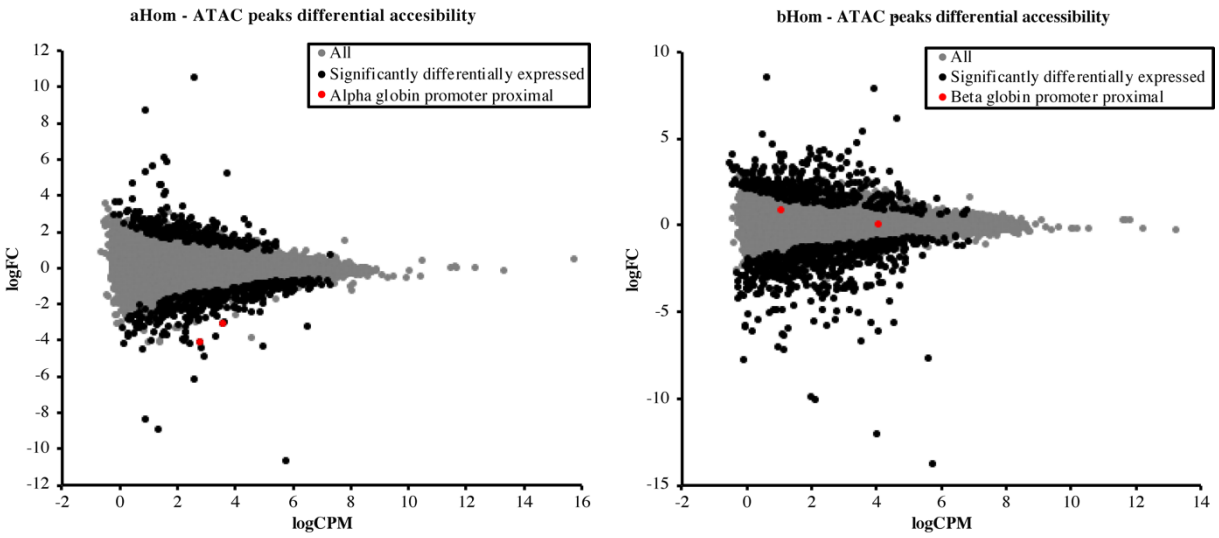


Figure 5.5) Differential accessibility analysis of open chromatin peaks. aHom versus wildtype. ATAC-seq performed in differentiated mouse foetal liver. All genes = grey dots. Significantly differentially expressed genes = black dots. specific peaks over alpha and beta promoter proximal region =red dots. N.B alpha globin promoter peaks are significantly differentially expressed whilst beta globin's are not.

This analysis reveals that we see minimal changes in hypersensitivity between wildtype to aHom; only 1% of peaks are significantly differentially expressed in the aHom or bHom versus wildtype. This indicates that there are very few loci affected by the enhancer deletions, again confirming their highly targeted nature and minimal effect on global chromatin accessibility. Interestingly, this analysis suggests that the peaks associated with alpha globin are differently hypersensitive, whereas the peaks at beta globin are not. From visual inspection of the 5' end of the genes (Figure 5.3) we can see a subtle loss of ATAC-seq specifically over the TSS of alpha globin genes, whereas at beta globin no similar visual change is evident.

It is tempting to interpret this as differential function of enhancers at alpha and beta globin, but given that both loci overall retain the majority of their wildtype accessibility, it is clear that the enhancers do not play a major role in maintaining overall accessibility.

The significant changes at alpha globin may relate specifically to a loss of hypersensitivity over the TSS, but this is difficult to interpret as the resolution of ATAC is too limited to study this feature in isolation without conflating the signal with that of the upstream promoter.

5.2.3 Analysis of Pol2 levels in the alpha and beta globin enhancer knockouts

I next investigated the effect of the enhancer deletions on the levels of Pol2 at the alpha and beta globin promoters as it has been suggested that enhancers recruit components of the pre-initiation complex and Pol2 and deliver them to promoters, providing a mechanism by which the loss of an enhancer may reduce transcription[3]. It has also been suggested, in the case of beta globin, that the enhancer elements work by promoting the transition from initiation to elongation[120]. I wanted to test this using newer ChIP-seq technology as a failure to elongate would manifest as a loss of signal in the body of the gene with a maintenance or increase in the levels of promoter proximal Pol2 depending on whether Pol2 undergoes rounds of abortive or stable pausing.

I performed ChIP-seq for Pol2 using an antibody targeting the unmodified N terminus of Rbp1 in wildtype and alpha and beta globin enhancer knockout conditions. Our biological replicates were reproducible (Figure 5.6) and so I merged them for each condition.

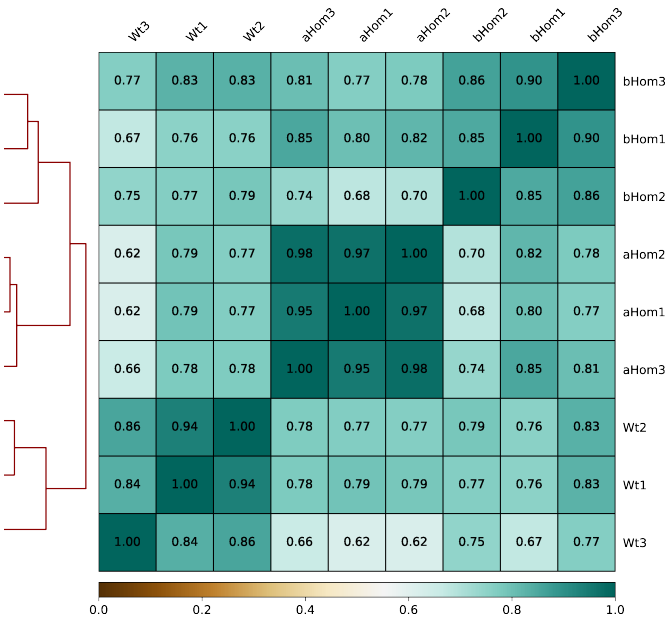


Figure 5.6) Pearson correlation (R^2) of Pol2 ChIP-seq data.

Visualisation of the alpha and beta globin loci both containing two copies of each gene (Figure 5.7) reveals remarkably little change in the levels of Pol2 recruited to the TSSs, with the suggestion that, at alpha globin, the levels of promoter proximal Pol2 may be slightly higher when its enhancer is removed. The signal for Pol2 in the gene body is reduced at both loci and in both copies of the genes.

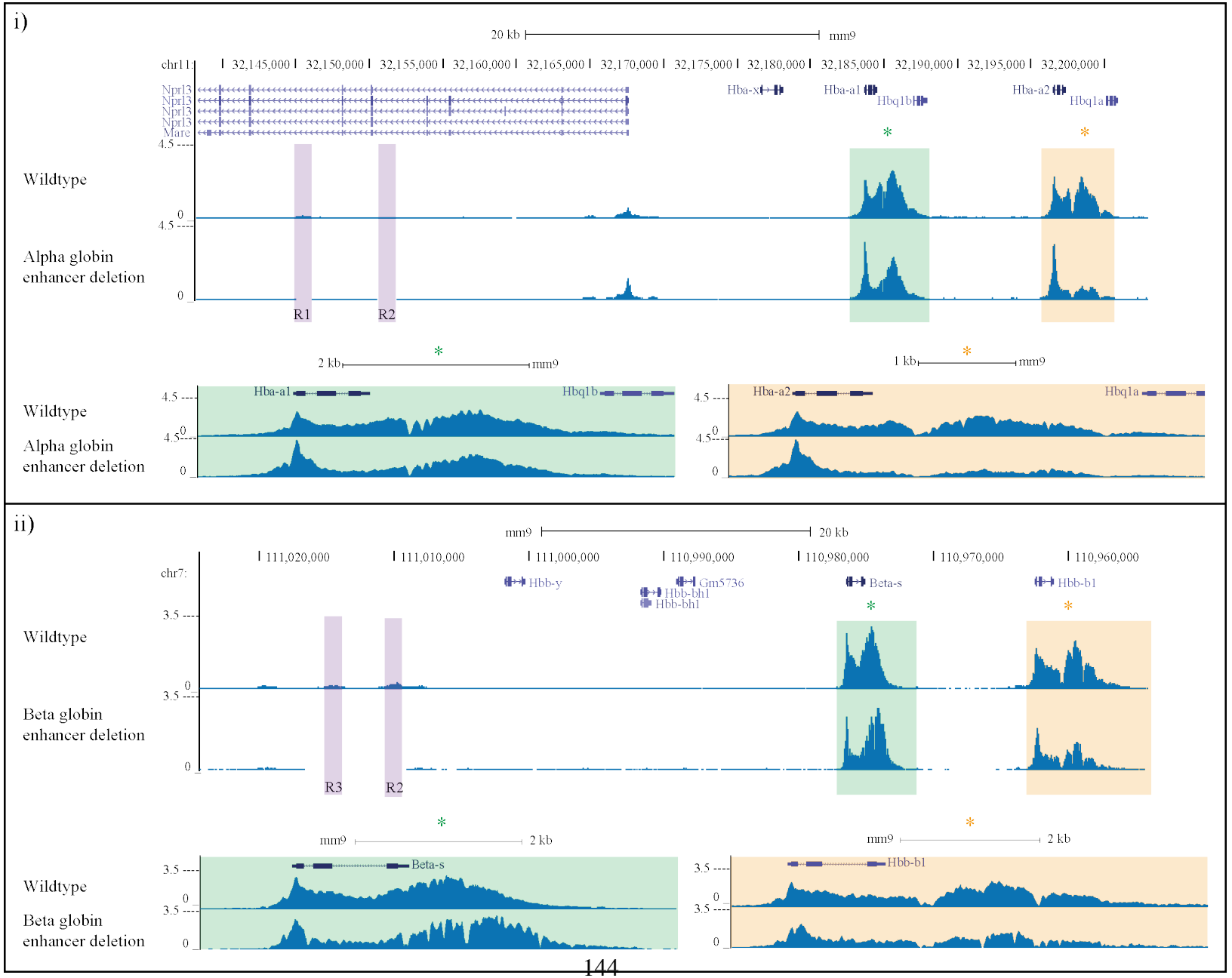

i)
20 kb
mm9
chr11:
32,145,000
32,150,000
32,155,000
32,160,000
32,165,000
32,170,000
32,175,000
32,180,000
32,185,000
32,190,000
32,195,000
32,200,000
Nprl3
Nprl3
Nprl3
Nprl3
Mare
Hba-x
Hba-a1
Hbq1b
Hba-a2
Hbq1a
Wildtype
Alpha globin enhancer deletion
4.5
0
4.5
0
R1
R2
2 kb
mm9
1 kb
mm9
Hba-a1
Hbq1b
Hba-a2
Hbq1a
Wildtype
Alpha globin enhancer deletion
ii)
mm9
20 kb
111,020,000
111,010,000
111,000,000
110,990,000
110,980,000
110,970,000
110,960,000
chr7:
Hbb-y
Gm5736
Hbb-bh1
Hbb-bh1
Beta-s
Hbb-b1
Wildtype
Beta globin enhancer deletion
3.5
0
3.5
0
R3
R2
mm9
2 kb
mm9
2 kb
Beta-s
Hbb-b1
Wildtype
Beta globin enhancer deletion

Figure 5.7) Pol2 ChIP-seq at the i) alpha and ii) beta globin loci. Individual copies of alpha globin or beta globin highlighted in green and blue; regulatory elements (R) under investigation are indicated by purple boxes for each loci. Top ChIP-seq track in i) or ii) is wildtype, bottom track is enhancer knockout for each loci. Tracks are normalised by total millions of mapped reads.

5.2.4 Quantification of changes in Pol2 using TR

To quantify the changes in PPEP, I calculated the Pol2 TR. Mean TR values for wild type and enhancer KO are shown in Figure 5.8.

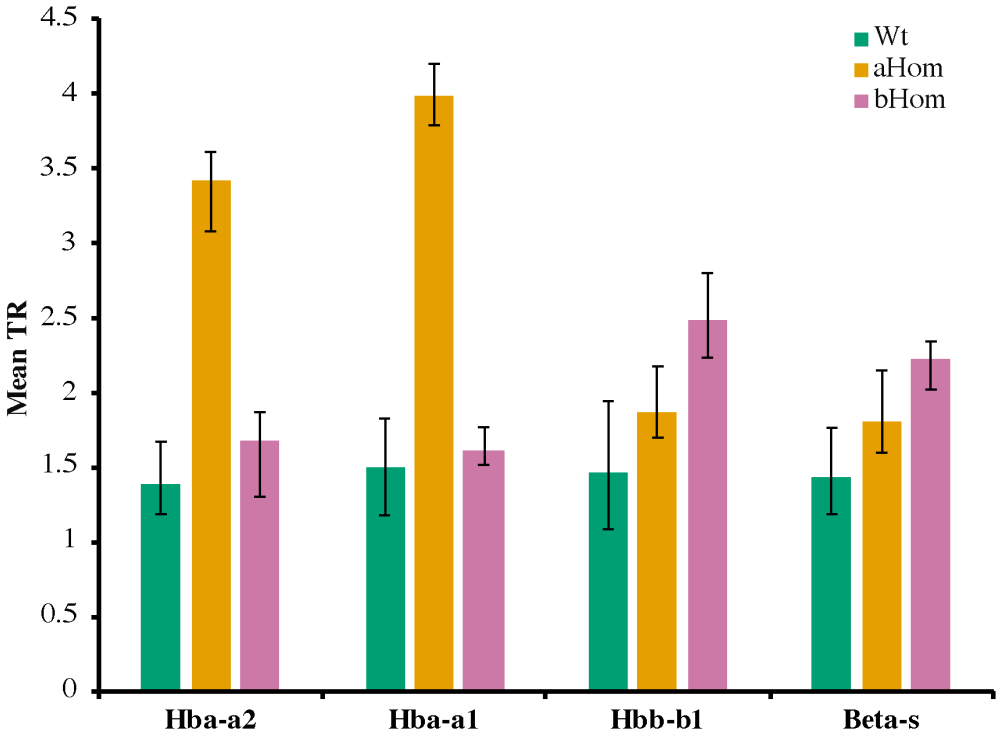


Figure 5.8) TR values in wildtype and enhancer knockouts. TR values in wildtype (Wt = green), alpha globin enhancer deletion (aHom = orange) and beta globin enhancer deletion (bHom = pink). Mean TR from n= 3 experiments are plotted. Error bars represent the range of values used to calculate mean. Hba-a2 and Hba-a1 are the two copies of alpha globin and Hbb-b1 and Beta-s are the copies of beta globin.

In the aHom deletion, the TR has approximately doubled whereas in the bHom TR appears increased but not as dramatically. This suggests that both the enhancer knockouts have affected the balance of promoter proximal to gene body Pol2, but exactly what this change

means in terms of identifying the specific stage of transcription which is being affected is unclear. Of note, although I have shown the TR to have no quantitative power in predicting gene expression, both alpha and beta globin would be classed as paused under wildtype conditions based on the arbitrary threshold of TR > 1 (indicating a higher density of Pol2 near the promoters than in the gene body)[22,23].

The change in TR could be explained by a loss of Pol2 pause release[56]. Inspection of the genes (Figure 5.7) shows little change in the height and distribution of the promoter proximal Pol2 signal, indicating change in TR is due to loss of signal in the gene body (loss of elongating Pol2), consistent with the ATAC data (Figure 5.3). The resolution of ChIP-seq is too limited to distinguish between a failure to release Pol2 pausing or a general failure to elongate past the promoter proximal region (abortive transcription). In addition, given our earlier analysis showing TR has no correlation with gene expression, it is not possible to relate changes in TR with changes in gene expression. We can however say with confidence that Pol2 is recruited to hypersensitive (open) promoter proximal regions of alpha and beta globin in the context of enhancer deletions, and also that changes in TR are a result of loss of gene body signal rather than increase in promoter proximal density.

Therefore, any loss in gene expression at alpha and beta globin is not due to a failure to recruit Pol2. If Pol2 is being recruited to the promoters then why do we see a loss of transcript? This could be due to a failure to initiate transcription or a failure to progress to elongation downstream of the promoter proximal region after initiation.

5.2.5 The major enhancers of alpha and beta globin direct transcription initiation

Our data suggest that promoters remain open and able to recruit Pol2 in the context of the enhancer knockouts. I next wanted to determine the effects of the enhancer deletions on

the transcription cycle, at high resolution, to understand whether enhancer deletions cause a failure to initiate transcription or to navigate a downstream step in the transcription cycle. scaRNA-seq was used to establish the effects of the deletions on promoter proximal transcription (<=300 away from the TSS). If the enhancers promote transcriptional elongation through Pol2 pause release or other promoter proximal regulatory steps we would expect to see increases in the amount of scaRNA at their promoters. However, if the enhancers direct initiation of transcription we would expect the level of scaRNA to decrease in the knockouts.

We also investigated the effect of the enhancer deletions on downstream kinetics of elongation of alpha and beta globin RNA. For this, scaRNA-seq and mNET-seq were performed by myself and Dr Taka Nojima respectively on wildtype and alpha and beta globin enhancer deletion mouse foetal liver cells (Figure 5.9).

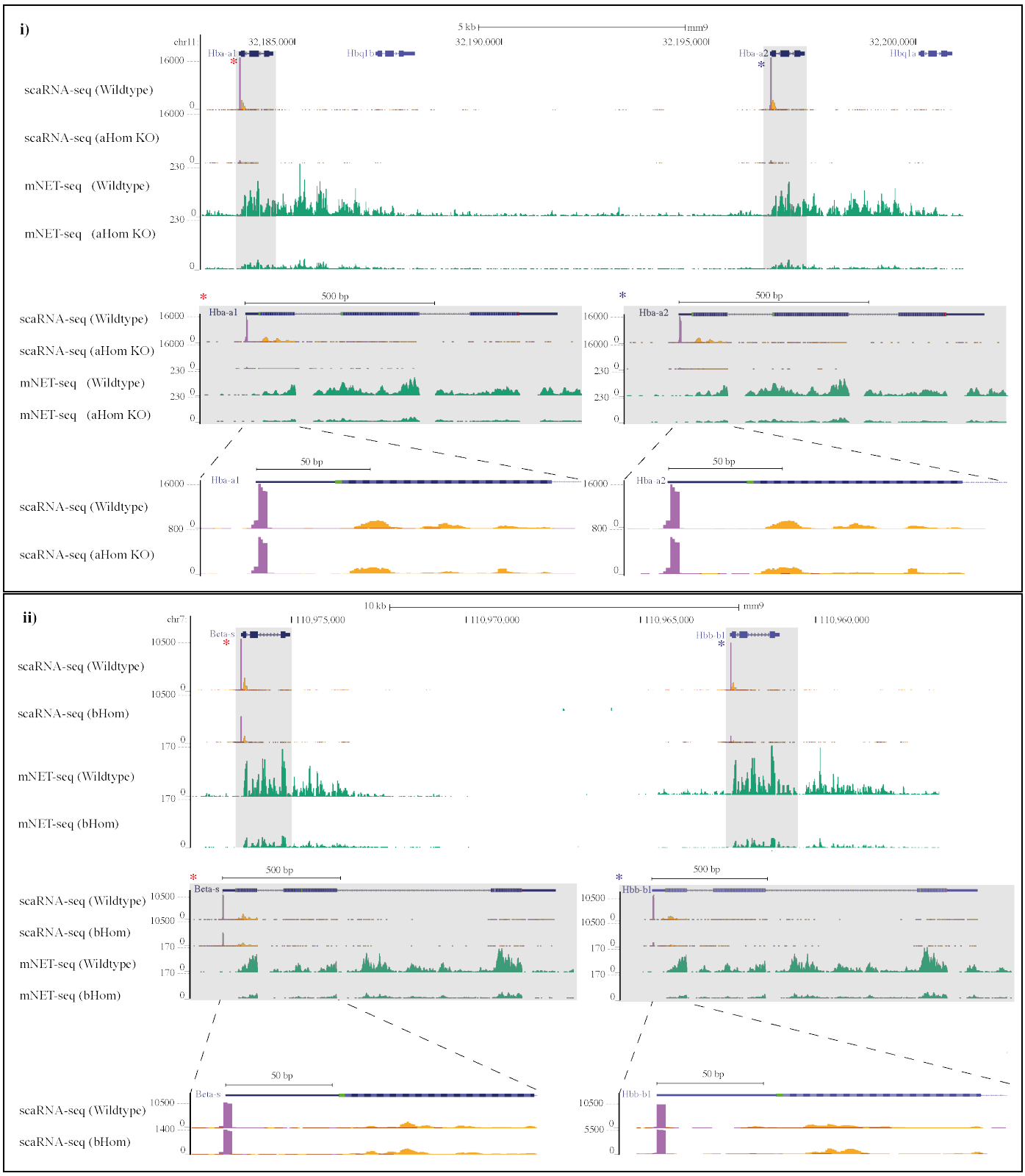


Figure 5.9) scaRNA-seq and mNET-seq in wildtype, enhancer knockouts. In scaRNA tracks, purple = 5' RNA ends, orange = 3' RNA ends. mNET-seq tracks = green. Each copy of alpha (i) or beta globin genes (ii) at each locus are highlighted in grey boxes with red or blue asterisks denoting each unique copy of the gene. A close up of scaRNA-seq data near each genes TSS is shown in the bottom row of each panel.

Firstly, in the wildtype cells, both alpha and beta globin initiate transcription at their annotated TSS; we also see an accumulation of 3' RNA ends just downstream of the TSS (+50bp), consistent with the position of canonically described Pol2 pausing[23,24]. Of interest, at alpha globin we can also see several downstream accumulations of 3' RNA ends prior to the first exon intron boundary. mNET-seq reveals the same distribution, suggesting it is not a function of scaRNA but represents the 3' ends of nascent RNAs which are engaged with Pol2. This suggests there may be several unexpected regulatory or kinetic bottle necks at this promoter, which are undetectable using ChIP-seq due to resolution limits.

Comparison of the wildtype and alpha globin enhancer deletion scaRNA-seq traces (Figure 5.9i) reveals a dramatic reduction in the height of the initiation peak (~95%). The reduction mirrors the reductions in nascent alpha globin production previously detected by our lab in this model[7]. Rxamination of the mNET-seq traces reveals remarkably little change in the shape of the profile; rather, the signal across the whole gene is reduced, indicating a reduction in the amount of Pol2 transcribing the gene in line with reduced initiation. Importantly, in the alpha globin enhancer deletion scaRNA still shows the same characteristic accumulations of 3' ends near the TSS, suggesting that at least at this gene, Pol2 pausing may be an obligate property of the template rather than a dynamic regulatory step which modulates final transcript levels. Taken together these data strongly suggest that the R1/R2 enhancers of alpha globin act to direct transcription initiation, not Pol2 pause release.

Examination of the beta globin loci appears similar to alpha globin (Figure 5.9ii). Large reductions in the level of scaRNA are coupled with a dramatic loss of mNET-seq signal across the gene. This strongly suggests that beta globin is also regulated at the level of transcription initiation.

These results are surprising as beta globin has previously been shown to have paused Pol2 near its promoter (being one of the earliest mammalian examples of this phenomena) and it has also been suggested that the enhancers of beta globin act to promote elongation, potentially through Pol2 pause release[14]. I can explain these findings in the context of the new data presented here. The published work used ChIP-PCR to examine levels of Pol2 at the start of the gene and in exon 2. Examination of our ChIP-seq data reveals that one would expect to see a loss of signal in exon2 of beta globin in the enhancer knockouts and therefore the authors were right to suggest a defect in elongation. The authors' estimate of the loss of nascent RNA in the enhancer knockouts (~75%) mirrors what I see with scaRNA and NET-seq; however, our RNA data provide significantly higher resolution than the published ChIP-qPCR data. This allows us to disentangle a failure to initiate from a failure to elongate and shows that a loss of Pol2 signal in the gene body is due to a lack of initiation of transcription, not a failure to navigate promoter proximal pausing.

The resolution limits of ChIP-based technologies have hampered the interpretation of these data. I show that beta globin accumulates Pol2 near its promoter (ChIP-seq and mNET-seq) and in enhancer deletions there is a loss of gene body Pol2 which increases the TR at the beta genes. scaRNA levels are reduced in the enhancer knockout, suggesting a defect in initiation, not pause release. This leads us to two hypotheses: either Pol2 is undergoing a round of transcription initiation before aborting and reinitiating, or Pol2 is stably paused but the rate of initiation determines the rate of pause release and therefore lower rates of initiating Pol2 mean less Pol2 clearing the promoters. Both of these hypotheses are supported by computational modelling of Pol2[122]

As scaRNA is believed to be rapidly degraded upon release from chromatin[38] any increase

in abortive transcription would manifest as a total reduction in scaRNA reads as is the case in both knockout models whilst Pol2 levels at the TSS would be anticipated to remain approximately the same. To distinguish between these scenarios, direct measurement of turnover of nascent scaRNA would be required and is technically not yet feasible.

5.2.6 The beta globin chromatin compartment wildtype versus bHom

I have previously shown that aHom deletion has minimal impact on chromatin interactions between alpha globin promoters and the remaining enhancer elements at this locus[7]. Thus, an intact cluster of enhancers (referred to as a 'super enhancer'[119]) is not required for the formation or maintenance of the alpha globin chromatin compartment.

I next wanted to investigate the effect of the R2/R3 enhancer deletions on chromatin structure at beta globin to determine if enhancers at this locus play a different role in establishing or maintaining an active chromatin compartment. I also wanted to ensure (as in the case of alpha globin) that the effect of removing these enhancer elements is not in part due to a compromised three-dimensional chromatin confirmation and rather due to the direct *cis*-acting effects of the enhancers.

Capture-C was performed by Dr Damien Downes in bHom mouse foetal liver cells and compared to our previously generated Capture-C data in wild type foetal liver (Figure 5.10).

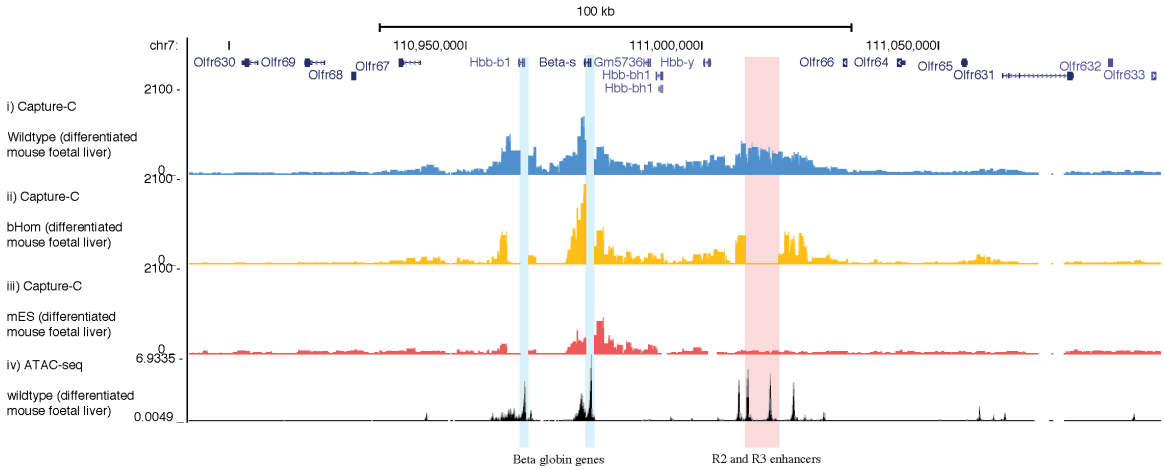


Figure 5.9) Capture C in bHom versus wildtype. i) wildtype (differentiated mouse foetal liver), ii) bHom (differentiated mouse foetal liver), iii) wildtype mES, and iv) wildtype ATAC-seq (differentiated mouse foetal liver).

We conclude that the beta globin chromatin compartment is approximately 100kb in size (roughly comparable to alpha globin at 70Kb[7]) and extends from around the 5' ends of Olfr67 to the 5' end of Olfr66. In the enhancer deletion the compartment is still visible with interactions between non-deleted enhancers and the two copies of beta globin (Hbb-b1 and Beta-s) persisting. This suggests that deletion of the HS2/HS3 enhancers at beta globin has not grossly altered the structure of the chromatin compartment with individual enhancer elements remaining hypersensitive and interacting with target promoters. Therefore, changes in gene expression associated with this enhancer deletion are not likely due to impaired overall chromatin structure and relate specifically to the functions of the deleted elements on transcription of the beta genes themselves.

5.3 Discussion

The results presented in this chapter aimed to answer a fundamental question in gene regulation: how do enhancers affect expression of their target genes? Through stepwise analysis of Pol2 occupancy, chromatin accessibility, scaRNA production, nascent elongating Pol2 (mNET-seq) and chromatin confirmation I have addressed this question at two well-defined genomic loci.

Our analysis differs from previous work by using a wide range of recently developed techniques to study gene regulation. Our analysis of Pol2 ChIP-seq through to differential analysis and TR, reveals that although deletion of the key enhancers at the alpha and beta globin clusters alters the density of Pol2 in the gene body versus the promoter, this cannot be explained by Pol2 pausing alone. The promoters of alpha and beta globin genes still recruit significant levels of Pol2 in the absence of the enhancers, suggesting that enhancers do not necessarily stimulate the recruitment of Pol2 at these genes in the mouse clusters. By analysing open chromatin around the TSSs of alpha and beta globin I was also able to show that the enhancer elements do not significantly contribute to opening the promoter regions of these genes. However, the resolution of ATAC-seq may not be sufficient to conclusively interrogate the TSSs and upstream promoter regions independently, as common peak calling approaches such as MACS2 call this region as one peak. As accessibility of the promoter regions and TSSs may be regulated differently to enable the stepwise formation of the PIC before Pol2 recruitment, and since enhancers could be regulating either of these processes, I cannot definitively conclude that enhancers of the alpha and beta globin genes do not regulate hypersensitivity of their cognate TSSs specifically. However, I can conclude that the enhancer deletions do not cause a complete loss of accessibility over the regions associated with regulation of gene expression and initiation of transcription.

We used high resolution analysis of RNA to interrogate enhancer function in unprecedented detail. The most recent studies on the role of enhancers in regulating RNA production have used changes in TR based on ChIP-seq or GRO-seq data to stand proxy for changes in pausing[10,54,56]; in our earlier analysis I revealed that the TR does not correlate with expression and also that changes in the TR may result from changes in initiation not pausing and therefore currently published work becomes hard to interpret. I took a direct approach by studying the production of scaRNA near the TSS, revealing that enhancer deletion at alpha and beta globin causes a loss of these molecules and therefore cause a defect in initiation. mNET-seq reveals that the distribution of Pol2 across the alpha and beta globin genes does not change in the enhancer knockouts but rather decreases uniformly. This suggests that, after initiation of transcription, the enhancers of alpha and beta globin do not play a significant role in maintaining Pol2 processivity. Previous studies showed some enhancers of beta globin may track with the elongating Pol2 complex[60], suggesting a role for enhancers in maintaining Pol2 processivity and therefore transcript levels; however, we see no evidence for this.

Both alpha and beta globin display characteristic accumulation of scaRNA 3' ends consistent with Pol2 pausing. In the enhancer deletions, the levels of RNA drop dramatically at both genes, yet remarkably the characteristic profile of 3' RNA end accumulation remains. This suggests that 3' end accumulation is directly linked to transcript production and that the enhancers do not act to release paused Pol2 since such a mechanism would increase scaRNA levels in the absence of full length transcript production. We therefore show at these two genes that pausing occurs but is not the rate limiting step in gene expression. This is surprising given that beta globin provided one of the first examples of a paused gene. From the data presented here, it appears that the enhancers are instead required to initiate transcription after recruitment.

The idea that Pol2 may inefficiently elongate at promoter proximal regions, giving rise to higher Pol2 densities in these regions when compared to gene bodies, is supported by computational modelling of Pol2 densities on genes, assuming regulation at the points of initiation, elongation and processivity on genes[122]. Recent work using live cell imaging of Pol2 in MRC-5 sv40 cells (human lung) concluded that most Pol2 inefficiently elongates often aborting transcription after initiation but before elongation through the gene body[116]. Single molecule footprinting in Drosophila has suggested Pol2 is rapidly turned over at promters[123] and recently it has been shown that a large fraction of promoter proximal Pol2 can be removed using a simple hypotonic salt wash, indicating it is unlikely to be stably engaged with the DNA template[124]. These studies suggest that promoter proximal pausing may be unstable and reflect Pol2 in the process of being recruited to promoters, initiating transcription and then rapidly aborting before passing the +1 nucleosome and disengaging with the template.[23,24,38].

Work defining pausing as a stable state has often involved the use of transcription initiation or elongation inhibitors with uncertain specificity *in vivo*[38,51]. The processes of initiation and Pol2 pausing have been suggested to affect each other and so the use of these drugs to target one stage in transcription may inadvertently perturb the other, giving a false readout[104,105]. Notable work presented using Start-seq in Drosophila, which concluded that pausing was stable, used triptolide to inhibit initiation and determine the release rates of paused complexes[38]. However, triptolide causes Pol2 degradation by ubiquitination as well as inhibiting Pol I and Pol II transcription. Disentangling specific rates of pausing is challenging in the context of a globally perturbed transcriptome[51]. Also of note, the levels of pausing at genes were defined by ranked TRs and therefore the effects on scaRNA at

these genes may have resulted from increased initiation and premature termination (in line with the DRB treatment of K562 cells discussed in this thesis).

Models of stable pausing in human (K562 cells) have been proposed by others who have used mNET-seq coupled with nascent RNA production using TT-seq (a variant of 4sU seq). From this approach the authors were able to determine the amount of RNA synthesised per bp per engaged Pol2, allowing rates of elongation to be directly measured and coupled to Pol2 occupancy. Using these models the authors concluded that pausing is not likely to be rate limiting at the majority of genes and would only be predicted to affect nascent transcript levels after the rate of initiation becomes rate limiting[105]. This work however maintains Pol2 pausing is a stable process as a central assumption.

An investigation into whether Pol2 pausing is stable or abortive is of paramount importance in understanding the fine kinetics of transcription initiation, how it relates to Pol2 pausing and which of these two stages is most utilised as a regulatory step *in vivo*.

The data presented in this thesis support the idea that Pol2 pausing occurs very frequently across the genome but that in an *in vivo* setting is not commonly rate limiting and initiation is the predominant regulatory step in transcription. In line with this, we have determined that enhancers regulate the initiation of transcription at both alpha and beta globin, most probably through a failure of Pol2 to become elongation competent after recruitment. Our data represent the highest resolution and most comprehensive study of a functional enhancer element to date and it is likely that insights gained from the globin loci are more generally applicable to transcription at other genes.

Chapter 6

Discussion

Chapter 6 – Discussion

6.1 Quantitating promoter proximal transcription

In this thesis I developed a modified version of the Start-seq protocol, "scaRNA-seq," to map promoter proximal transcription genome-wide. By identifying the genome-wide distributions of 5' RNA ends, it was possible to call "observed" TSS in a quantitative and tissue-specific manner. Comparing observed TSS positions using scaRNA-seq to previously annotated TSS, it became clear that there is a subtle but systematic skew in TSS coordinates called by UCSC. This systematic skew is explained by the process used to derive the annotated gene list. The UCSC gene list used in this thesis represents one of the earliest and most comprehensive gene databases[125]. The UCSC gene annotation cross references cDNA sequence information from multiple databases with experimentally determined protein sequences (UniProt) to predict "known genes"[98]. An individual protein sequence may have several putative RNA isoforms and so a score is assigned to the most likely candidate isoform based on the following equation:

Score = (mRNA length + date in months since January 1970 * 2 − number of mismatches * 50.)

Consequently, the longest and most recent isoforms are often assigned the higher scores and are therefore more likely to be the chosen representation of the gene. cDNA sequences used by UCSC are derived from a wide variety of tissues and developmental stages with no adjustment for abundance; therefore, low abundance long isoforms are favoured over high abundance short transcripts[107,126]. This effect is further confounded as the process of cloning cDNA for sequencing in large scale annotation projects such as FANTOM give preference to clone libraries with larger inserts, biasing annotation towards longer genes. Correcting for this skew in future experiments is a simple matter of generating tissue specific TSS maps using scaRNA-seq and using these to filter candidate RNA isoforms

based on proximity to the predominant active TSS in that cell type. With the lowered cost and feasibility of performing RNA-seq to identify transcripts, tissue-specific genome annotation is well within the capabilities of the modern researcher and should be considered an essential tool in understanding transcription.

The use of a tissue-specific TSS annotation in genomics has profound implications for our overview of transcription. Recent work using global run-on sequencing and ChIP-seq utilised UCSC annotated TSS to derive an "averaged" view of transcription at eukaryotic promoters, finding prevalent of i) broad transcription initiation, ii) Pol2 pausing from 20-60bp downstream of the TSS and iii) high levels of upstream antisense transcription relative to sense transcription[29,34,61,71,91,127]. By accurately defining and quantifying the predominant TSSs in an individual cell line (K562) using scaRNA-seq and mNET-seq, and then re-evaluating previously published datasets (ChIP-seq, MNase, ChIP-exo, GRO-seq) using this same cell line, we have produced a revised model of the transcription cycle in these cells. This interpretation may more truly reflect the process than using non-quantified data derived from many different cell types which may have "blurred" the mechanistic detail.

i) Broad initiation

Firstly, using the observed TSS as opposed to annotated TSS, we see far less evidence of broad transcription initiation. This change appears to be a function of scaling relative to the predominant TSS associated signal (precisely at 0bp). We adjust for a systematic skew in annotated TSS coordinates (which often results in TSS being annotated upstream of their *in vivo* locations) by calling the precise locations of observed TSS *in vivo*. By plotting 5' ends of scaRNA molecules relative to the correct TSS position, read distribution downstream of the TSS is shifted upstream, reducing the breadth of the signal associated

with transcription initiation in meta-analysis of all genome wide TSS. An excellent example of this is provided by the *MYC* gene (Figures 3.5 and 3.9), where a low level of transcription initiation can be seen at the annotated upstream TSS but the dominant signal lies ~100bp downstream at a previously unannotated TSS. This effect is commonly seen genome-wide as shown by the distance plots of observed versus annotated TSS (Figures 3.7 and 3.8). This common type of mis-annotation gives the impression of widespread, broad transcription initiation. We suggest that although broad initiation may occur over regions of 50-100nt around annotated TSS *in vivo* (Figure 3.10), punctate initiation where initiation occurs at one or a couple of base pairs[66] is much more commonly observed when the predominant TSSs are used to analyse the data.

Further work should include analysis of promoters known to lie within CpG islands as these have been suggested to initiate transcription broadly[66]. Comparison of scaRNA signal over annotated or observed TSS for CpG island promoters could tell us if this is the case or whether it is another result of using skewed TSS annotations.

ii) Promoter proximal pausing from 20 to 60bp downstream of the TSS

Previous estimates have suggested Pol2 pausing occurs over a range of +20 to +60bp based on the length of RNA synthesised in *in vitro* run on assays and measuring from maximal Pol2 ChIP-seq and GRO-seq signals around the TSS[23,29,52]. We refine these estimates by using *in vivo* derived TSS annotations and show that the maximum signal of 3' RNA ends occurs precisely +40bp downstream of the TSS (Figure 3.10). This coincides with the location of factors (NELF and DSIF) thought to regulate pausing as well as the high-resolution positioning of Pol2 revealed by ChIP-exo (Figure 3.17 and 3.18). We show that the position of the +1 nucleosome fits well with the leading edge of this pausing, consistent

with the idea that the +1 nucleosome may play a role in limiting the extent of promoter proximal transcription (Figure 3.11)[92,103].

Interestingly, when Pol2 pausing was originally described, *in vitro* assays were performed on non-chromatinised templates and so the role of the +1 nucleosome in relationship to Pol2 pausing could not be assessed[47–49]. It therefore remains to be determined to what extent the nucleosomal barrier contributes to a pausing effect. One way to explore this would be to limit nucleosome removal, for example through chemical inhibition of remodelling proteins whilst measuring scaRNA and Pol2 levels. In this way, one might be able to show that a Pol2 pausing-like profile could be induced.

iii) High levels of upstream antisense transcription relative to sense

Antisense transcription is known to occur at high levels relative to sense transcription genome-wide[29,34,66]. However, all published analyses to date have used annotated TSS coordinates to plot upstream antisense RNA distributions, thus skewing our view of this process. When using directly observed TSSs, the relative height of antisense transcription initiation relative to sense transcription initiation appears far less. However, due to the low level of antisense transcription on average indicated by our analysis, our scaRNA-seq method may need further optimisation to interrogate these rare populations of scaRNA and investigate further how they relate to levels of expression at linked sense genes.

Overall our results shed new light on transcription in eukaryotes and highlight that a tissue-specific annotation of TSS is an essential tool for studying transcription at individual genes and for deriving the general principles underlying the mechanisms of transcription. scaRNA-seq requires small amounts of input material ($5x10^6$ cells) which, unlike commonly used methods such as GRO-seq, are not manipulated before examining

transcription, providing a more faithful view of nascent promoter proximal transcription *in vivo*. We aim to further develop scaRNA-seq to reduce input cell requirements, simplify workflow and add unique molecular identifiers (UMIs) to each RNA molecule by ligation before library preparation. This will allow us to rapidly generate highly accurate quantitative maps of TSS in any cell type or organism where m7G RNA caps are common.

6.2 The prevalence of Pol2 pausing *in vivo*

In this thesis I addressed the question of how frequently Pol2 pausing occurs *in vivo*. Pol2 pausing is thought to be a key regulatory step in transcription and enhancers are thought to regulate this process, so answering this question is key to understanding the role of Pol2 pausing in *in vivo* gene expression. I began by investigating the effects of the commonly used transcription inhibitor DRB, initially aiming to use this compound to validate our method, expecting that DRB would cause an increase in very short 40nt transcripts due to a block in pause release. This was based on the fact that 3' RNA ends reached a maximum at the +40nt position which coincides with Pol2 and components of DSIF and NELF, which have been reported to regulate Pol2 pausing *in vitro*[47–49,95]. Unexpectedly, DRB did not cause an increase in 40nt transcripts but rather an increase in the scaRNA levels at promoters genome wide concomitant with reductions in intronic RNA levels. This is consistent with a block in promoter proximal elongation but suggested two hypotheses:

1) Pol2 pausing occurs at variable distances from the TSS (40-150nt) and DRB inhibits is release.

2) DRB causes an increase in promoter proximal read density but not simply though blocking pause release.

I saw no correlation between changes in scaRNA and changes in intronic (nascent) RNA under DRB treatment when analysing single genes (Figure 4.13), suggesting that hypothesis 1) is less likely, as one would expect if DRB was specific to pause release, the increase in scaRNA would negatively correlate with changes in intronic RNA at the single gene level.

I analysed published GRO-seq data in HEK293 cells and showed that DRB causes an increase in promoter proximal read density relative to gene body signal (in line with our scaRNA data); however, I noticed that the reduction in gene body read density occurred around +500bps from the TSS, far downstream of the reported position of pausing[23]. This is an important observation because if DRB specifically affects promoter proximal release of pausing as is generally thought, why do we not see an immediate drop in read density after the promoter proximal region? Further analysis revealed a noticeable phased pattern of GRO-seq signal fitting well with predicted nucleosome positioning. This suggests that DRB compromises Pol2 processivity: a phenomenon distinct from defective promoter proximal pause release.

Further support of the idea that DRB may directly affect transcription through nucleosomes can be found in a published analysis of the Hsp70 gene, an archetypal "paused" gene, which shows that when DRB treatment is used in combination with heat shock, the DRB-treated cells have a far higher level of "closed" chromatin before heat shock than their control counterparts (as revealed by accessibility to MNase), suggesting the basal level of chromatin accessibility may be lower under DRB treatment[128].

I suggest that DRB causes multiple independent effects on transcription which are not limited to pause release. Further work will be required to determine whether it is the action of DRB itself as an inhibitor or whether CDK9 has roles in transcription beyond the

regulation of Pol2 pause release. A suggestion of the latter can be found through analysis the genome-wide distribution of CCNT2 (a partner of CDK9 in the P-TEFb complex). CCNT2 appears to be maximally located upstream of the TSS in the region of PIC formation, suggesting CDK9 may directly bind the preinitiation complex (Figure 3.17).

Due to the uncertain effects and specificity of DRB on transcription and as yet uncharacterised roles of CDK9 *in vivo,* I suggest that previous estimates of Pol2 pausing as a regulatory mechanism may be incorrect. In order to determine the extent of regulation of this process in gene expression, I moved away from the possible confounding effects of CDK9 inhibitors. Instead, I used mouse erythropoiesis as a model of differential gene expression and showed that scaRNA levels generally increase in proportion to intronic RNA levels. This suggests that regulation of Pol2 pause release to affect gene expression may be uncommon and that initiation of transcription is the predominant regulatory step in transcription. As a further extension of this observation, it is believed that enhancers regulate differential gene expression and therefore, if differential gene expression is driven at initiation, it is possible that this is the predominant mode of enhancer action. Much work still remains, however, to determine how, mechanistically, initiation is modulated by enhancers.

I am aware that this is a contentious conclusion given that several studies have suggested promoter proximal pausing of Pol2 is a ubiquitous regulatory step in the transcription cycle[12,22,53,91]. However, I believe there are several technical reasons why this conclusion may have been drawn and an alternative model whereby regulation of transcription initiation is the dominant regulatory barrier in transcription fit our data better.

Low resolution data

Many studies suggesting Pol2 pausing is prevalent use ChIP-seq to measure levels of Pol2 and attribute changes in TR to changes in Pol2 pausing[22,23,91]. I have shown that the TR does not correlate with gene expression and a high TR could be interpreted as high levels of initiation but failure to elongate (as appears to be the case in the enhancer deletion models presented here). The resolution of ChIP is too limited to distinguish initiation from Pol2 pausing or Pol2 accumulation at nucleosomal barriers.

The use of transcription inhibitors

The desire to reconcile results of *in vitro* run on assays and the hypothesised mechanism of Pol2 pausing with the detection of DSIF, NELF and Pol2 colocalised at promoters genome wide *in vivo* led to the idea that Pol2 is paused at promoters genome-wide[22,52,91,129]. This hypothesis has been tested by the use of inhibitors of pause release such as DRB[12,38]. The idea that these compounds are specific inhibitors of Pol2 pause release stems from *in vitro* studies using run on assays on non-chromatinised templates[47–49]. I show, using previously published data, that DRB-induced changes in signal density coincide with chromatin structure around TSSs, in particular nucleosome positioning, suggesting an alternative mode of action of DRB. This effect would not have been detectable in early *in vitro* studies due to the lack of chromatin and, therefore, the idea that DRB is specific to Pol2 pause release has never been verified *in vivo*.

I suggest that the use of DRB inhibits elongation non-specifically. Indeed, DRB has been reported to inhibit CDK7 in addition to CDK9[51] and original work defining the effects of DRB on total RNA production noted that DRB inhibits the uptake of nucleotides and affects total RNA levels: associated data had to be normalised for this effect. This means that DRB may not only affect other cellular processes, but that its effect on total RNA production has never been quantitatively measured by recently developed methods[46,114].

The key fact remains that if inhibitors such as DRB are not specifically inhibiting pause release (which appears to be the case), then the prevalence of Pol2 pausing as a regulatory stage could have been considerably overestimated.

By examining changes in scaRNA as a measure of pausing and intronic RNA as a measure of elongation throughout the development of red blood cells, I show that there is limited evidence for scaRNA levels anti-correlating with intronic RNA levels (i.e. regulation of Pol2 pausing), but that the differential expression of most genes could be explained by altering rates of initiation.

6.3 The role of enhancers in regulating the transcription cycle

The role of enhancers in transcriptional regulation is poorly understood although several roles for such elements in the regulation of transcription have been proposed, including roles in initiation, release of Pol2 pausing, and transcriptional elongation[3,6,54,56,60]. We addressed the question of how enhancers modulate transcription at their target genes by comparing transcription of promoters in wildtype and enhancer deletion mutants at the alpha and beta globin loci. Beta globin is thought to be regulated by Pol2 pause release[93] and two of its enhancer elements (HS2 and HS3 - bHom) have been proposed to assist in the transition from paused to elongating Pol2[14].

We performed scaRNA-seq and mNET-seq in wildtype and beta globin enhancer knockout primary mouse foetal liver cells, revealing that homozygous deletion of the HS2 and HS3 enhancers results in loss of scaRNA and total mNET-seq signal across the beta globin gene. However, our Pol2 ChIP-seq and ATAC-seq data show that chromatin at the beta globin promoter is open and recruits Pol2 in the absence of the enhancers. This suggests that removing the enhancers specifically affects the transition of Pol2 from the promoter to an

actively transcribing form. This is intriguing as it suggests that the HS2 and HS3 enhancers are not responsible for recruiting Pol2 to the beta globin promoter[14]. It may be that Pol2 is recruited effectively to promoters but fails to initiate transcription due to a loss of activating partners, which are recruited to the enhancers, or that Pol2 can often be promoter proximal but non-functional.

Analysis of the alpha globin gene R1 R2 enhancer deletion revealed an almost identical effect on transcription. This addressed a perceived difference between these loci: the suggestion that beta globin but not alpha globin is regulated by Pol2 pausing[14,120,130]. At the alpha globin genes we see a characteristic accumulation of Pol2 ChIP-seq signal and 3' scaRNA ends just downstream of TSS. This suggests that like beta globin, alpha bears the hallmarks of Pol2 pausing. However, deletion of the enhancer elements causes a concomitant reduction in total scaRNA and mNET-seq signal, arguing against regulation of Pol2 pausing at this gene and supporting the idea of initiation as the defining step in transcriptional regulation.

The similar effect of enhancer deletions of alpha and beta globin on transcription of their target genes suggests that their regulation may be more similar than previously thought. Importantly, at both loci the changes in scaRNA and mNET-seq are positively correlated, meaning that genes can bear the signs of paused Pol2 but that reductions in their expression does not change the relative level of pausing versus initiation.

Pol2 pausing could in fact be an obligate kinetic barrier at which Pol2 accumulates near TSSs. Removal of this barrier is linked to the amount of initiated Pol2; in essence, the more Pol2 which has become elongation competent and synthesised a short stretch of RNA, then the more the barrier is removed. This competition style mechanism between features downstream of the TSS such as the +1 nucleosome and initiated Pol2 has been suggested

to be one reason why Pol2 pausing exists. As a by-product, this mechanism would block re-deposition of nucleosomes at the TSS, further facilitating initiation by maintaining open chromatin[23,92].

I can hypothesise as to why Pol2 may be recruited but not elongate efficiently at alpha and beta globin. One striking feature of eukaryotic gene regulation is the association of phosphorylation states of the Pol2 carboxy terminal domain (CTD) with different stages in transcription[131]. In order for Pol2 to begin transcription, it must become phosphorylated specifically at serine 5 residues in the CTD; thus conversion from a hypo-phosphorylated to a hyper-phosphorylated state is concomitant with initiation of transcription and has also been associated with splicing[33,132]. It therefore seems reasonable that if Pol2 were recruited but not phosphorylated it may explain the lack of transcription initiation as seen in our knockout models. A deletion of the whole locus controls region of beta globin in Murine Erythroleukemia (MEL) cells affected the levels of Pol2 phosphorylated at serine 5 residues in its CTD (associated with paused Pol2[131]) and was associated with a 95% reduction in beta globin transcripts[120]. This suggests that the regulatory elements of this locus are in some way controlling the transition from hypo phosphorylated (free Pol2) to initiating or paused (ser5 phosphorylated Pol2).

Why would a loss of serine 5 phosphorylation cause a failure to initiate transcription? A recent and very appealing mechanism to explain these findings may be phase separation of Pol2 and transcriptional machinery. Recent work has demonstrated that the Pol2 CTD can become isolated in live cells within liquid-liquid phase separated droplets[34,133,134]. It may be that a loss of serine 5 phosphorylation causes an inability for Pol2 to move out of these droplets, resulting in a loss of initiation. Enhancers bind transcription factors, which are believed to recruit the Mediator complex, which is required to phosphorylate the CTD of

Pol2. Failure to do so in the absence of the enhancers would result in failure of the Pol2 CTD to become phosphorylated and initiate transcription[134–137]. I intend to revisit this using ChIP-seq targeting Pol2 phosphoforms, Tata box binding protein (TBP) and Mediator in wild type and enhancer knockout models. Based on the current model derived from the data in this thesis, I predict that there will be reduced Mediator and phosphorylated Pol2 levels, but normal levels of TBP and unphosphorylated Pol2 consistent with recruitment of the PIC complex to the gene promoter but failure to initiate or elongate transcription via phosphorylation.

Previous studies have suggested that enhancers regulate Pol2 pause release. A recent study in human HCT116 cells suggested that activation of enhancers leads to the recruitment of PAF1 and the release of Pol2 pausing at target genes[56]. Work in Drosophila has suggested that enhancers looped to promoters are loaded with paused Pol2 which is released later in development[10], and work in human HEK293T cells suggested that enhancers can act as "anti-pause" enhancers by recruitment of the 7SK snRNA, which enables the delivery of P-TEFb (a pause release factor) to target promoters through chromatin looping[54]. One notable theme throughout this work is the use of ChIP-seq and the TR to measure changes in Pol2 pausing. As demonstrated in the alpha and beta globin loci, Pol2 can still be recruited to promoters in the absence of enhancers. This would create a high promoter proximal signal but low gene body signal in a Pol2 ChIP, which with the TR calculations may give the incorrect conclusion that Pol2 is paused. Therefore, the use of ChIP-seq for Pol2 in drawing conclusions about enhancer function may be inherently flawed.

Another feature of the studies in human cells is the use of poorly characterised enhancer elements, often defined by chromatin features such as a high ratio of H3K4me1 to H3K4me3 in addition to activating marks such as H3K27ac and distance from annotated

genes[3,6]. Without using a functionally tested enhancer element we cannot draw such general conclusions about enhancer function. It is clear that some elements (e.g. R3 and Rm in the alpha globin locus) which have the signatures of active enhancers have no effect on their target genes when deleted[7]. An enhancer must be shown to increase the expression of a target gene in *cis* beyond the basal level achieved by its core promoter in the context of a chromatin-based assay. Ideally, a specific enhancer deletion should also have minimal effects on global transcript levels, as in the case of the alpha globin R1/R2 deletions to reduce possible secondary effects of the enhancer deletion.

6.4 Summary of conclusions

The work presented in this thesis adds to our fundamental understanding of eukaryotic gene regulation in several ways. A variety of hypotheses regarding the mechanism of enhancer functions in regulating transcription have been proposed in recent years and include suggestions that enhancers may act on multiple stages in the transcription cycle[4,24,39,54,56,60,117]. The work in this thesis set out to determine on which stage in the transcription cycle enhancers can act to alter the transcriptional output of their target genes. I examined the effect of deleting functionally validated and clinically relevant enhancers *in vivo*. In the process of developing methodology to address this specific question, I adapted our approach to study transcriptional kinetics genome wide through erythropoiesis, thereby elucidating the predominant effect of tissue-specific enhancers on transcription as being, most likely, through regulation of initiation.

<u>The major findings of this study are as follows:</u>

- TSS annotation is systematically skewed. Adjusting for this skew alters our view of average transcription in three ways:

    - Transcription initiation is punctate not broad.
    - Pol2 pausing is most common at +40bp.
    - The scale of anti-sense transcription relative to sense may be far lower than previously thought.
- The extent of Pol2 pausing regulation may have been incorrectly estimated due to uncertain effects of "pause release" inhibitors such as DRB.
- Gene expression changes during development are largely driven by regulation of the rate of transcription initiation.
- The enhancers of alpha and beta globin regulate initiation of transcription.
- Promoters can recruit Pol2 and remain open in the absence of enhancers, suggesting a defect in initiation related to activation of Pol2 for transcription.
- Pol2 pausing may be an obligate kinetic barrier of promoters as shown at alpha and beta globin (the level of scaRNA is linked to the rate of initiation, not pausing).

I believe that our study represents the first *in vivo* primary cell study of how enhancers modulate transcription at target genes. I have used two well defined enhancer knockout models, revealing a unified mechanism of regulation at the level of transcription initiation of their target genes. I hypothesise that this effect is due to a failure of Pol2 to become hyperphosphorylated. As an extension to this work, I show that the majority of differential gene expression throughout differentiation (erythropoiesis) is regulated at the level of transcription initiation; therefore, if enhancers regulate differential gene expression, then their predominant mode of action should be through modulation of this stage in the transcription cycle. I hope that this work will add to our understanding of the mechanisms which govern differential gene expression and shed light on the mechanism through which distal regulatory elements affect this fundamental process.

Chapter 7

References

Chapter 7 – References

1. Das, S., Sarkar, D. & Das, B. The interplay between transcription and mRNA degradation in Saccharomyces cerevisiae. *Microb. Cell* **4**, 212–228 (2017).
2. Sainsbury, S., Bernecky, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 129–43 (2015).
3. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–86 (2014).
4. Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386**, 569–577 (1997).
5. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
6. Pennacchio, L. a, Bickmore, W., Dean, A., Nobrega, M. a & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14**, 288–95 (2013).
7. Hay, D. *et al.* Genetic dissection of the α-globin super-enhancer in vivo. *Nat. Genet.* **48**, 895–903 (2016).
8. Davies, J. O. J. *et al.* Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat. Methods* **13**, 74–80 (2015).
9. Andrey, G. *et al.* A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science* **340**, 1234167 (2013).
10. Ghavi-Helm, Y. *et al.* Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **512**, 96–100 (2014).
11. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–54 (2015).
12. Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* **3**, e02407 (2014).
13. Allen, B. L. & Taatjes, D. J. The Mediator complex: a central integrator of transcription. *Nat. Rev. Mol. Cell Biol.* **16**, 155–66 (2015).
14. Bender, M. A. *et al.* The hypersensitive sites of the murine β-globin locus control region act independently to affect nuclear localization and transcriptional elongation. *Blood* **119**, 3820–7 (2012).
15. Corradin, O. & Scacheri, P. C. Enhancer variants: Evaluating functions in common disease.

Genome Med. **6**, 1–14 (2014).

16. Hatton, C. S. *et al.* Alpha-thalassemia caused by a large (62 kb) deletion upstream of the human alpha globin gene cluster. *Blood* **76**, 221–7 (1990).
17. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497–502 (2007).
18. Miguel-Escalada, I., Pasquali, L. & Ferrer, J. Transcriptional enhancers: Functional insights and role in human disease. *Curr. Opin. Genet. Dev.* **33**, 71–76 (2015).
19. Furey, T. S. ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.* **13**, 840–852 (2012).
20. Gariglio, P., Bellard, M. & Chambon, P. Clustering of RNA polymerase B molecules in the 5' moiety of the adult beta-globin gene of hen erythrocytes. *Nucleic Acids Res.* **9**, 2589–98 (1981).
21. Gilmour, D. S. & Lis, J. T. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. *Mol. Cell. Biol.* **6**, 3984–9 (1986).
22. Rahl, P. B. *et al.* c-Myc regulates transcriptional pause release. *Cell* **141**, 432–45 (2010).
23. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–31 (2012).
24. Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 167–77 (2015).
25. Albert, I. *et al.* Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature* **446**, 572–6 (2007).
26. Dahl, J. A. & Gilfillan, G. D. How low can you go? Pushing the limits of low-input ChIP-seq. *Brief. Funct. Genomics* **17**, 89–95 (2018).
27. Goldberg, S., Weber, J. & Darnell, J. E. The definition of a large viral transcription unit late in Ad2 infection of HeLa cells: Mapping by effects of ultraviolet irradiation. *Cell* **10**, 617–621 (1977).
28. García-Martínez, J., Aranda, A. & Pérez-Ortín, J. E. Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. *Mol. Cell* **15**, 303–13 (2004).
29. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–8 (2008).
30. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**, 950–3 (2013).

31. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368–73 (2011).

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

33. Nojima, T. *et al.* Mammalian NET-seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* **161**, 526–540 (2015).

34. Seila, A. C. *et al.* Divergent transcription from active promoters. *Science* **322**, 1849–51 (2008).

35. Nechaev, S. *et al.* Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* **327**, 335–8 (2010).

36. Scruggs, B. S. *et al.* Bidirectional Transcription Arises from Two Distinct Hubs of Transcription Factor Binding and Active Chromatin. *Mol. Cell* **58**, 1101–12 (2015).

37. Jimeno-González, S., Ceballos-Chávez, M. & Reyes, J. C. A positioned +1 nucleosome enhances promoter-proximal pausing. *Nucleic Acids Res.* **43**, 3068–3078 (2015).

38. Henriques, T. *et al.* Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. *Mol. Cell* **52**, 517–28 (2013).

39. Studitsky, V. M., Nizovtseva, E. V, Shaytan, A. K. & Luse, D. S. Nucleosomal Barrier to Transcription: Structural Determinants and Changes in Chromatin Structure. *Biochem. Mol. Biol. J.* **2**, 1–4 (2016).

40. Shandilya, J. & Roberts, S. G. E. The transcription cycle in eukaryotes: from productive initiation to RNA polymerase II recycling. *Biochim. Biophys. Acta* **1819**, 391–400 (2012).

41. Vos, S. M., Farnung, L., Urlaub, H. & Cramer, P. Structure of paused transcription complex Pol II–DSIF–NELF. *Nature* **560**, 601–606 (2018).

42. Vos, S. M. *et al.* Structure of activated transcription complex Pol II–DSIF–PAF–SPT6. *Nature* **560**, 607–612 (2018).

43. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006).

44. Farrell, S., Simkovich, N., Wu, Y., Barberis, A. & Ptashne, M. Gene activation by recuitment of the RNA polymerase II holoenzyme. *Genes Dev.* **10**, 2359–2367 (1996).

45. Barberis, A. *et al.* Contact with a component of the polymerase II holoenzyme suffices for gene activation. *Cell* **81**, 359–68 (1995).

46. Sehgal, P. B., Darnell, J. E. & Tamm, I. The inhibition of DRB (5,6-dichloro-1-β-d-ribofuranosylbenzimidazole) of hnRNA and mRNA production in HeLa cells. *Cell* **9**, 473–480 (1976).

47. Marshall, N. F. & Price, D. H. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J. Biol. Chem.* **270**, 12335–8 (1995).

48. Yamaguchi, Y. *et al.* NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97**, 41–51 (1999).

49. Wada, T. *et al.* DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev.* **12**, 343–56 (1998).

50. Bardet, A. F. *et al.* Identification of transcription factor binding sites from ChIP-seq data at high resolution. *Bioinformatics* **29**, 2705–13 (2013).

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

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

53. Muse, G. W. *et al.* RNA polymerase is poised for activation across the genome. *Nat. Genet.* **39**, 1507–11 (2007).

54. Liu, W. *et al.* Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. *Cell* **155**, 1581–95 (2013).

55. Chen, F. X. *et al.* PAF1, a Molecular Regulator of Promoter-Proximal Pausing by RNA Polymerase II. *Cell* **162**, 1003–15 (2015).

56. Chen, F. X. *et al.* PAF1 regulation of promoter-proximal pause release via enhancer activation. *Science* **357**, 1294–1298 (2017).

57. Chow, L. T., Roberts, J. M., Lewis, J. B. & Broker, T. R. A map of cytoplasmic RNA transcripts from lytic adenovirus type 2, determined by electron microscopy of RNA:DNA hybrids. *Cell* **11**, 819–836 (1977).

58. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci.* **74**, 3171–3175 (1977).

59. Martin, R. M., Rino, J., Carvalho, C., Kirchhausen, T. & Carmo-Fonseca, M. Live-Cell Visualization of Pre-mRNA Splicing with Single-Molecule Sensitivity. *Cell Rep.* **4**, 1144–1155

(2013).

60. Lee, K., Hsiung, C. C., Huang, P., Raj, A. & Blobel, G. A. Dynamic enhancer – Gene body contacts during transcription elongation. *Genes Dev.* 1992–1997 (2015). doi:10.1101/gad.255265.114.

61. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* **46**, 1311–20 (2014).

62. Kaida, D. *et al.* U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**, 664–668 (2010).

63. Kadener, S., Fededa, J. P., Rosbash, M. & Kornblihtt, A. R. Regulation of alternative splicing by a transcriptional enhancer through RNA pol II elongation. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 8185–90 (2002).

64. Laitem, C. *et al.* CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II-transcribed genes. *Nat. Struct. Mol. Biol.* **22**, 396–403 (2015).

65. O'Sullivan, J. M. *et al.* Gene loops juxtapose promoters and terminators in yeast. *Nat. Genet.* **36**, 1014–1018 (2004).

66. Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* 1–17 (2018). doi:10.1038/s41580-018-0028-8

67. Gaertner, B. & Zeitlinger, J. RNA polymerase II pausing during development. *Development* **141**, 1179–1183 (2014).

68. Preker, P. *et al.* RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322**, 1851–4 (2008).

69. He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptomes of human cells. - supplementary online material. *Science* **322**, 1855–7 (2008).

70. Wu, X. & Sharp, P. A. XDivergent transcription: A driving force for new gene origination? *Cell* **155**, 990–996 (2013).

71. Andersson, R. *et al.* Human Gene Promoters Are Intrinsically Bidirectional. *Mol. Cell* **60**, 346–7 (2015).

72. McGrath, K. E., Catherman, S. C. & Palis, J. Delineating stages of erythropoiesis using imaging flow cytometry. *Methods* **112**, 68–74 (2017).

73. Edling, C. E. & Hallberg, B. c-Kit-A hematopoietic cell essential receptor tyrosine kinase. *Int. J.*

Biochem. Cell Biol. **39**, 1995–1998 (2007).

74. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–8 (2013).

75. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–9 (2009).

76. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–2 (2010).

77. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

78. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

79. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–89 (2010).

80. Diribarne, G. & Bensaude, O. 7SK RNA, a non-coding RNA regulating P-TEFb, a general transcription factor. *RNA Biol.* **6**, 122–128 (2009).

81. Shen, L., Shao, N., Liu, X. & Nestler, E. Ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics* **15**, 1–14 (2014).

82. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, (2008).

83. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).

84. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. DeepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* **42**, 187–191 (2014).

85. Harrow, J. *et al.* GENCODE: The Reference Human Genome Annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).

86. Hashimoto, S. *et al.* 5'-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.* **22**, 1146–9 (2004).

87. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci.* **100**, 15776–15781 (2003).

88. Furuichi, Y. Discovery of m(7)G-cap in eukaryotic mRNAs. *Proc. Jpn. Acad. Ser. B. Phys. Biol.*

Sci. **91**, 394–409 (2015).

89. Salimullah, M., Mizuho, S., Plessy, C. & Carninci, P. NanoCAGE: A high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harb. Protoc.* **6**, 96–111 (2011).

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

91. Zeitlinger, J. *et al.* RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat. Genet.* **39**, 1512–6 (2007).

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

93. Bentley, D. L. & Groudine, M. A block to elongation is largely responsible for decreased transcription of c-myc in differentiated HL60 cells. *Nature* **321**, 702–6 (1986).

94. Meulia, T., Krumm, A. & Groudine, M. Distinct properties of c-myc transcriptional elongation are revealed in Xenopus oocytes and mammalian cells and by template titration, 5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole (DRB), and promoter mutagenesis. *Mol. Cell. Biol.* **13**, 5647–58 (1993).

95. Wu, C.-H. *et al.* NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. *Genes Dev.* **17**, 1402–14 (2003).

96. Vo Ngoc, L., Wang, Y., Kassavetis, G. A. & Kadonaga, J. T. The punctilious RNA polymerase II core promoter. *Genes Dev.* **31**, 1289–1301 (2017).

97. Parkinson, J. & Blaxter, M. Expressed sequence tags: an overview. *Methods Mol. Biol.* **533**, 1–12 (2009).

98. Hsu, F. *et al.* The UCSC known genes. *Bioinformatics* **22**, 1036–1046 (2006).

99. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–19 (2011).

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

101. Bernecky, C., Herzog, F., Baumeister, W., Plitzko, J. M. & Cramer, P. Structure of transcribing mammalian RNA polymerase II. *Nature* **529**, 551–4 (2016).

102. Ping, Y. H. & Rana, T. M. DSIF and NELF interact with RNA polymerase II elongation complex and HIV-1 Tat stimulates P-TEFb-mediated phosphorylation of RNA polymerase II and DSIF

during transcription elongation. *J. Biol. Chem.* **276**, 12951–8 (2001).

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

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

105. Gressel, S. *et al.* CDK9-dependent RNA polymerase II pausing controls transcription initiation. *Elife* **6**, 1–24 (2017).

106. Ozer, A. *et al.* Quantitative assessment of RNA-protein interactions with high-throughput sequencing-RNA affinity profiling. *Nat. Protoc.* **10**, 1212–1233 (2015).

107. Kawai, J. *et al.* Functional annotation of a full-length mouse cDNA collection. *Nature* **409**, 685–689 (2001).

108. Lenhard, B., Sandelin, A. & Carninci, P. Metazoan promoters: Emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* **13**, 233–245 (2012).

109. Lai, W. K. M. & Pugh, B. F. Genome-wide uniformity of human 'open' pre-initiation complexes. *Genome Res.* **27**, 15–26 (2017).

110. Bird, J. G. *et al.* The mechanism of RNA 5′ capping with NAD+, NADH and desphospho-CoA. *Nature* **535**, 444–447 (2016).

111. Jiao, X. *et al.* 5′ End Nicotinamide Adenine Dinucleotide Cap in Human Cells Promotes RNA Decay through DXO-Mediated deNADding. *Cell* **168**, 1015-1027.e10 (2017).

112. Boswell, S. A. *et al.* Total RNA-seq to identify pharmacological effects on specific stages of mRNA synthesis. *Nat. Chem. Biol.* **13**, 501–507 (2017).

113. Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* **29**, 436–42 (2011).

114. Fraser, N. W., Sehgal, P. B. & Darnell, J. E. DRB-induced premature termination of late adenovirus transcription. *Nature* **272**, 590–3 (1978).

115. Yamashita, R. *et al.* Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome Res.* **21**, 775–789 (2011).

116. Steurer, B. *et al.* Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. *Proc. Natl. Acad. Sci.* 201717920 (2018). doi:10.1073/pnas.1717920115

117. Gao, W. wei *et al.* JMJD6 Licenses ERα-Dependent Enhancer and Coding Gene Activation by Modulating the Recruitment of the CARM1/MED12 Co-activator Complex. *Mol. Cell* 340–357 (2018). doi:10.1016/j.molcel.2018.03.006

118. Schaukowitch, K. *et al.* Enhancer RNA facilitates NELF release from immediate early genes. *Mol. Cell* **56**, 29–42 (2014).

119. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).

120. Sawado, T., Halow, J., Bender, M. A. & Groudine, M. The beta -globin locus control region (LCR) functions primarily by enhancing the transition from transcription initiation to elongation. *Genes Dev.* **17**, 1009–18 (2003).

121. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).

122. Ehrensberger, A. H., Kelly, G. P. & Svejstrup, J. Q. Mechanistic interpretation of promoter-proximal peaks and RNAPII density maps. *Cell* **154**, 713–715 (2013).

123. Krebs, A. R. *et al.* Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol. Cell* **67**, 411-422.e4 (2017).

124. Erickson, B., Sheridan, R. M., Cortazar, M. & Bentley, D. L. Dynamic turnover of paused Pol II complexes at human promoters. 1–11 (2018). doi:10.1101/gad.316810.118.

125. Tyner, C. *et al.* The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.* **45**, D626–D634 (2017).

126. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res.* **33**, 34–38 (2005).

127. Henriques, T. *et al.* Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev.* **32**, 26–41 (2018).

128. Petesch, S. J. & Lis, J. T. Rapid, Transcription-Independent Loss of Nucleosomes over a Large Chromatin Domain at Hsp70 Loci. *Cell* **134**, 74–84 (2008).

129. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–80 (2005).

130. Vernimmen, D. Uncovering Enhancer Functions Using the α-Globin Locus. *PLoS Genet.* **10**, e1004668 (2014).

131. Heidemann, M., Hintermair, C., Voß, K. & Eick, D. Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. *Biochim. Biophys. Acta* **1829**, 55–62 (2013).

132. Nilson, K. A. *et al.* THZ1 Reveals Roles for Cdk7 in Co-transcriptional Capping and Pausing. *Mol. Cell* **59**, 576–587 (2015).

133. Cisse, I. I. *et al.* Real-time dynamics of RNA polymerase II clustering in live human cells. *Science* **341**, 664–7 (2013).

134. Cho, W.-K. *et al.* Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361**, 412–415 (2018).

135. Soutourina, J. Transcription regulation by the Mediator complex. *Nat. Rev. Mol. Cell Biol.* **19**, 262–274 (2018).

136. Lu, H. *et al.* Phase-separation mechanism for C-terminal hyperphosphorylation of RNA polymerase II. *Nature* **558**, 318–323 (2018).

137. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13–23 (2017).